ỨNG DỤNG TRÍ TUỆ NHÂN TẠO VÀ HỌC MÁY VÀO MẠNG TRUYỀN THÔNG KHÔNG DÂY DI ĐỘNG 5G VÀ 6G

5-2025

Ứng dụng trí tuệ nhân tạo và học máy (AI/ML) vào mạng truyền thông không dây di động 5G và 6G là một xu thế tất yếu đã được 3GPP nghiên cứu nhiều năm. Tài liệu này cung cấp kên thức cơ bản về AI/ML và quá trình nghiên cứu của 3GPP để ứng dụng trí tuệ nhân tạo và học máy (AI/ML) vào mạng truyền thông không dây di động 5G và 6G. Tài liệu được chia làm ba phần:

1. Trình bầy tổng quan các kến thức cơ sở về AI/ML và ứng dụng AI/ML trong các hệ thống 5G và 6G
2. Khảo sát ứng dụng AI/ML cho các dịch vụ trong các mạng truyền thông di động 5G và 6G
3. Trình bày các nghiên cứu kết hợp AI/ML và các công nghệ không dây và mạng không dây di động 5G và 6G

PHẦN I

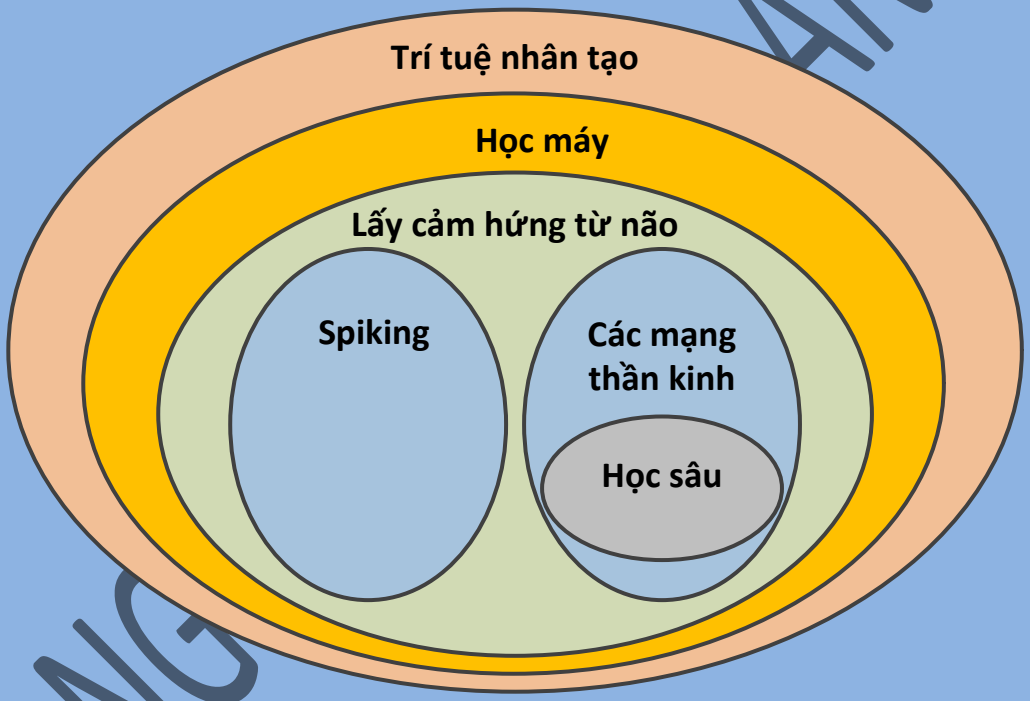
TỔNG QUAN AI/ML VÀ ỨNG DỤNG AI/ML TRONG CÁC HỆ THỐNG TRUYỀN THÔNG KHÔNG DÂY DI ĐỘNG 5G/6G

1. CÁC MÔ HÌNH AI/ML

Trí tuệ nhân tạo (AI)/Machine Learning (ML) đang được sử dụng trong một loạt các lĩnh vực ứng dụng trong các lĩnh vực công nghiệp, đạt được năng suất tăng đáng kể. Đặc biệt, trong các hệ thống truyền thông di động, các thiết bị di động (ví dụ: điện thoại thông minh, xe thông minh, UAV, robot di động) đang ngày càng thay thế các thuật toán thông thường (ví dụ: nhận dạng giọng nói, dịch máy, nhận dạng hình ảnh, xử lý video, dự đoán hành vi người dùng) bằng các mô hình AI/ML để cho phép các ứng dụng như chụp ảnh nâng cao, trợ lý cá nhân thông minh, VR/AR, trò chơi điện tử, phân tích video, đề xuất mua sắm được cá nhân hóa, lái xe tự động/điều hướng, thiết bị gia dụng thông minh, robot di động, y tế di động, cũng như tài chính di động.

1. Phân loại các cách tiếp cận AI

Trí tuệ nhân tạo (AI) là khoa học và kỹ thuật để chế tạo những cỗ máy thông minh có khả năng thực hiện các nhiệm vụ như con người, được xác định bởi John McCarthy vào năm 1956. Việc phân loại các phương pháp tiếp cận AI có thể được minh họa trong hình 1.1 [25].



Spiking: điện toán tăng vọt

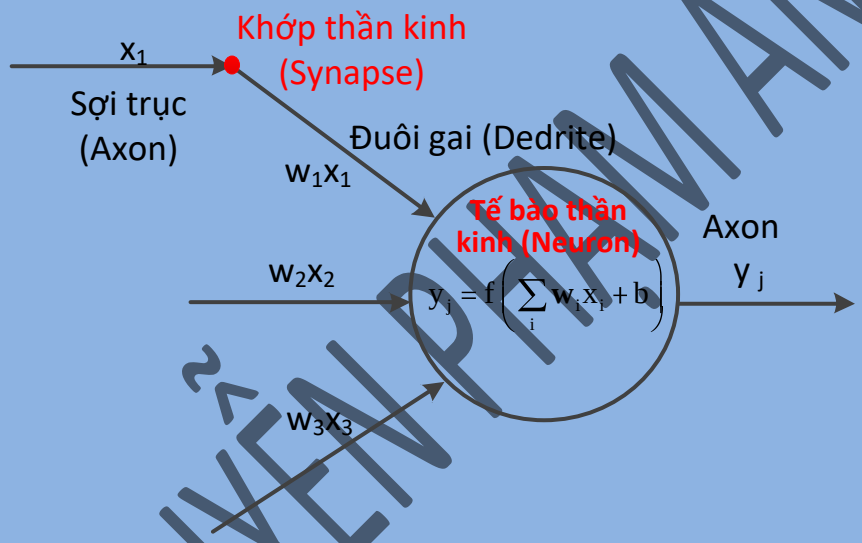
Hình 1.1. Phân loại các cách tiếp cận AI

Trong AI một lĩnh vực con lớn được gọi là học máy (ML), được Arthur Samuel xác định vào năm 1959 là lĩnh vực nghiên cứu cung cấp cho máy tính khả năng học mà không cần được lập

trình rõ ràng. Thay vì cách tiếp cận tốn nhiều công sức và may rủi là tạo ra một chương trình tùy chỉnh, riêng biệt để giải quyết từng vấn đề riêng lẻ trong một miền, một thuật toán ML duy nhất chỉ cần học, thông qua một quá trình gọi là đào tạo, để xử lý từng vấn đề mới [25]. Nhiều phương pháp ML được minh họa bởi cây quyết định (Decition Tree), phân cụm K-means (K-means Clusstering) và mạng Bayes đã được phát triển để đào tạo mô hình để thực hiện các phân loại và dự đoán, dựa trên dữ liệu thu được từ thế giới thực.

2. Các tiếp cận mạng thần kinh sâu (DNN)

Trong lĩnh vực ML, có một lĩnh vực thường được gọi là tính toán lấy cảm hứng từ não (brain-inspired computation), là một chương trình nhằm mô phỏng một số khía cạnh của cách chúng ta hiểu não hoạt động. Vì người ta tin rằng các yếu tố tính toán chính của bộ não con người là 86 tỷ tế bào thần kinh, hai phân khu vực của tính toán lấy cảm hứng từ não đều được lấy cảm hứng từ kiến trúc của tế bào thần kinh, như thể hiện trong hình 1.2.



Hình 1.2. Kết nối với tế bào thần kinh trong não. x_i, w_i, f(•) và b lần lượt là các kích hoạt, trọng số, hàm phi tuyến tính và thiên kiến

Bản thân các tế bào thần kinh (Neuron) được kết nối với nhau với một số nguyên tố đi vào chúng được gọi là đuôi gai (Dendrite) và một nguyên tố rời khỏi chúng được gọi là sợi trục như trong Hình 1.2. Tế bào thần kinh chấp nhận các tín hiệu đi vào nó thông qua các đuôi gai, thực hiện tính toán trên các tín hiệu đó và tạo ra tín hiệu trên sợi trục (Axon). Các tín hiệu đầu vào và đầu ra này được gọi là kích hoạt (Activation). Sợi trục của một tế bào thần kinh phân nhánh và được kết nối với các đuôi gai của nhiều tế bào thần kinh khác. Các kết nối giữa một nhánh của

sợi trục và một đuôi gai được gọi là khớp thần kinh (Synapse). Ước tính có 10^{14} đến 10^{15} khớp thần kinh trong não người trung bình.

Một đặc điểm chính của khớp thần kinh là nó có thể chia tỷ lệ tín hiệu (x_i) đi qua nó như thể hiện trong Hình 1.2. Hệ số tỷ lệ đó có thể được gọi là trọng số (w_i) và cách não được cho là học là thông qua những thay đổi đối với trọng số liên quan đến khớp thần kinh. Do đó, các trọng số khác nhau dẫn đến các phản hồi khác nhau đối với một đầu vào. Lưu ý rằng học tập là sự điều chỉnh các trọng số để đáp ứng với một kích thích học tập, trong khi tổ chức (những gì có thể được coi là chương trình) của não không thay đổi. Đặc điểm này làm cho bộ não trở thành nguồn cảm hứng tuyệt vời cho thuật toán kiểu học máy.

Trong mô hình máy tính lấy cảm hứng từ não bộ có một lĩnh vực phụ được gọi là điện toán spiking (Spike: gai hay tăng vọt). Trong tiểu khu vực này, cảm hứng được lấy từ thực tế là truyền thông trên các đuôi gai và sợi trục là các xung giống gai và thông tin được truyền tải không chỉ dựa trên biên độ của gai. Thay vào đó, nó cũng phụ thuộc vào thời gian xung đến và tính toán xảy ra trong tế bào thần kinh là một hàm không chỉ của một giá trị duy nhất mà còn là độ rộng của xung và mối quan hệ thời gian giữa các xung khác nhau. Một ví dụ về một dự án được lấy cảm hứng từ sự tăng đột biến của não là IBM TrueNorth . Trái ngược với điện toán gai, một lĩnh vực phụ khác của điện toán lấy cảm hứng từ não bộ được gọi là mạng nơ-ron được tập trung trình bầy trong các phần dưới đây.

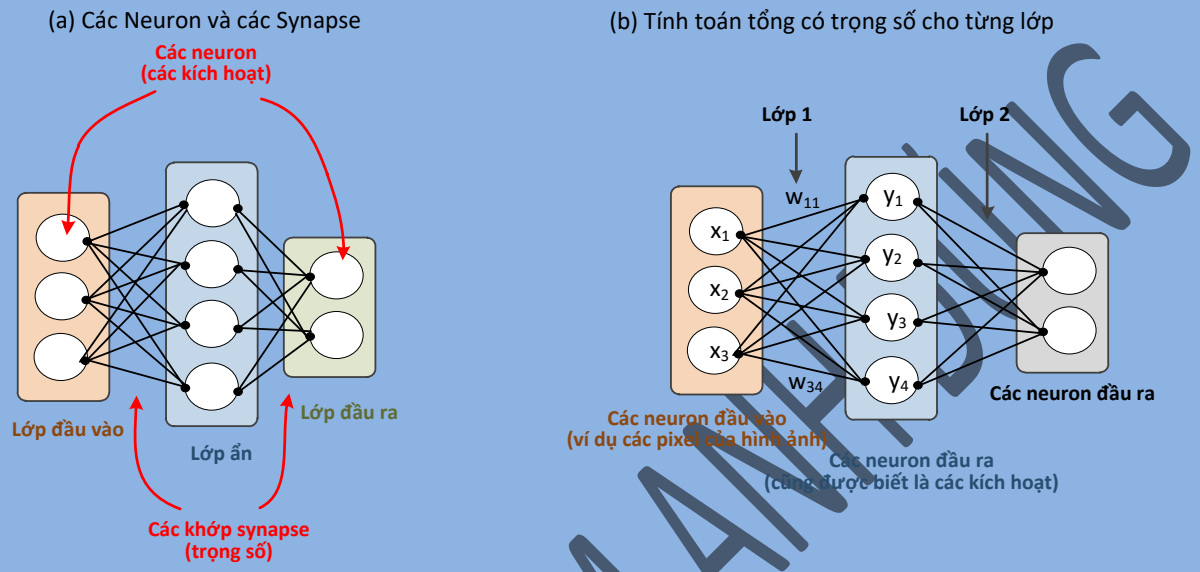
So với các phương pháp tiếp cận điện toán tăng vọt (Spike Approach), các phương pháp ML phổ biến hơn là sử dụng "mạng nơ-ron" làm mô hình. Dưới đây ta sẽ xét các mạng neuron.

1.3. Mạng nơ-ron và mạng nơ-ron sâu (DNN: Deep Learning Network)

Mạng nơ-ron lấy cảm hứng từ quan niệm rằng tính toán của tế bào thần kinh liên quan đến tổng có trọng số của các giá trị đầu vào. Các tổng có trọng số này tương ứng với tỷ lệ giá trị được thực hiện bởi các khớp thần kinh và sự kết hợp của các giá trị đó trong tế bào thần kinh. Hơn nữa, tế bào thần kinh không chỉ xuất ra tổng có trọng số đó, vì tính toán được liên kết với một tầng tế bào thần kinh sau đó sẽ là một phép toán đại số tuyến tính đơn giản. Thay vào đó, có một hoạt động chức năng trong tế bào thần kinh được thực hiện trên các đầu vào kết hợp. Phép toán này dường như là một hàm phi tuyến tính khiến tế bào thần kinh chỉ tạo ra đầu ra nếu các đầu vào vượt qua một số ngưỡng nào đó. Do đó, bằng cách tương tự, mạng nơ-ron áp dụng một hàm phi tuyến tính cho tổng trọng số của các giá trị đầu vào. Chúng ta xem xét một số hàm phi tuyến tính đó là gì trong các phần dưới đây.

Hình 1.3 (a) cho thấy một bức tranh sơ đồ của một mạng neuron tính toán. Các tế bào thần kinh trong lớp đầu vào (Input Layer) nhận một số giá trị và truyền chúng đến các tế bào thần kinh ở lớp giữa của mạng, thường được gọi là 'lớp ẩn' (Hidden Layer). Các tổng có trọng số (Weighted Sumes) từ một hoặc nhiều lớp ẩn cuối cùng được truyền đến lớp đầu ra (Output Layer) , lớp này

trình bày đầu ra cuối cùng của mạng cho người dùng. Để điều chỉnh thuật ngữ lấy cảm hứng từ não bộ với mạng neuron, đầu ra của các tế bào thần kinh thường được gọi là kích hoạt và các khớp thần kinh (Synapse) thường được gọi là trọng số như trong Hình 1.3 (a). Trong các phần dưới đây các các thuật ngữ kích hoạt/trọng số sẽ được sử dụng.



Neuron: tế bào thần kinh); Snapse: khớp thần kinh.

Hình 1.3. Thí dụ về mang neuron đơn giản và thuật ngữ

Hình 1.3 (b) cho thấy một ví dụ về tính toán tại mỗi lớp:

$$y_j = f\left(\sum_{i=1}^{3} w_{ij} \times x_i + b\right) \qquad (1.1)$$

trong đó W_{ij}, x_i và y_j là lần lượt là trọng số, kích hoạt đầu vào và kích hoạt đầu ra, và f (•) là một hàm phi tuyến tính được mô tả trong Phần 3. Để đơn giản, thuật ngữ thiên kiến (Bias) b được bỏ qua trong Hình 1.3 (b).

T có thể biểu diễn phương trình (1.1) vào dạng tích của các ma trận như sau:

$$\mathbf{y} = f\left(\mathbf{x}^T \mathbf{w} + b\right) \qquad (1.2)$$

Trong đó $\mathbf{y} = \left[y_1, \ldots, y_j, \ldots, y_N\right]^T$ là vectơ của N đầu vào của lớp sau và $\mathbf{x} = \left[x_i, \ldots, x_i, \ldots, x_K\right]^T$ là vectơ của K đầu ra của lớp trước, T ký hiệu cho đảo vị và ma trận trọng số:

$$\mathbf{w} = \begin{bmatrix} w_{11} & \cdots & w_{1j} & \cdots & w_{1N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{i1} & \cdots & w_{ij} & \cdots & w_{iN} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{K1} & \cdots & w_{kj} & \cdots & w_{KN} \end{bmatrix}$$

Trong lĩnh vực mạng nơ-ron, có một lĩnh vực được gọi là học sâu (Deep Laerning), trong đó mạng nơ-ron có nhiều hơn ba lớp, tức là nhiều hơn một lớp ẩn. Ngày nay, số lượng lớp mạng điển hình được sử dụng trong deep learning dao động từ năm đến hơn một nghìn. Trong phần này, ta sẽ sử dụng thuật ngữ mạng neuron sâu (DNN: Deep Learning Network) để chỉ các mạng neuron được sử dụng trong học sâu.

DNN có khả năng học các tính năng cấp cao với độ phức tạp và trừu tượng hơn so với mạng nơ-ron nông hơn. Một ví dụ minh họa điểm này là sử dụng DNN để xử lý dữ liệu thị giác. Trong các ứng dụng này, các pixel của hình ảnh được đưa vào lớp đầu tiên của DNN và đầu ra của lớp đó có thể được hiểu là đại diện cho sự hiện diện của các đặc điểm cấp thấp khác nhau trong hình ảnh, chẳng hạn như đường thẳng và cạnh. Ở các lớp tiếp theo, các đặc điểm này sau đó được kết hợp thành một thước đo về sự hiện diện có thể xảy ra của các đặc điểm cấp cao hơn, ví dụ: các đường thẳng được kết hợp thành các hình dạng, được kết hợp thêm thành các tập hợp các hình dạng. Và cuối cùng, với tất cả thông tin này, mạng cung cấp xác suất rằng các tính năng cấp cao này bao gồm một đối tượng hoặc cảnh cụ thể. Hệ thống phân cấp tính năng sâu này cho phép DNN đạt được hiệu năng vượt trội trong nhiều nhiệm vụ.

DNN đã tạo ra những bước đột phá đáng kinh ngạc kể từ những năm 2010 trong nhiều lĩnh vực ứng dụng thiết yếu vì chúng có thể đạt được độ chính xác ở cấp độ con người hoặc thậm chí vượt quá độ chính xác của con người. Các kỹ thuật học sâu sử dụng các chiến lược có giám sát và / hoặc không giám sát để tự động học các biểu diễn phân cấp trong kiến trúc sâu để phân loại. Với một số lượng lớn các lớp ẩn, hiệu suất vượt trội của DNN đến từ khả năng trích xuất các đặc điểm cấp cao từ dữ liệu cảm giác thô sau khi sử dụng học thống kê trên một lượng lớn dữ liệu để đạt được trình bày hiệu quả một không gian đầu vào. Trong những năm gần đây, nhờ dữ liệu lớn (Big Data) thu được từ thế giới thực, khả năng tính toán tăng nhanh và các thuật toán phát triển liên tục, các DNN đã trở thành các mô hình phổ biến bật nhất cho nhiều ứng dụng AI.

1.3. Đào tạo và suy luận

Vì DNN là một ví dụ của thuật toán học máy, chương trình cơ bản không thay đổi khi nó học cách thực hiện các nhiệm vụ nhất định của nó. Trong trường hợp cụ thể của DNN, việc học này liên quan đến việc xác định giá trị của trọng số (và độ thiên kiến: Bias) trong mạng và được gọi là đào tạo mạng. Sau khi được đào tạo, chương trình có thể thực hiện nhiệm vụ của mình bằng

cách tính toán đầu ra của mạng bằng cách sử dụng trọng số được xác định trong quá trình đào tạo. Chạy chương trình với các trọng số này được gọi là suy luận.

Khi thực hiện suy luận sử dụng DNN, một ảnh đầu vào được cho trước và đầu ra của DNN là một vectơ của các điểm số (Scores), mỗi điểm số cho một loại đối tượng. Loại có có điểm số cao nhất chỉ thị rằng loại này có khả năng cao nhất là đối tượng trong ảnh. Mục đích bao trùm đối với đào tạo là xác định trong số dẫn đến cực đại hóa điểm số của loại đúng trong ảnh. Khi đào tạo mạng, loại đúng thường đã được biết, nó được cho trước đối các ảnh được sử dụng cho đào tạo (ví dụ: tập đào tạo của mạng). Sự khác biệt giữa các điểm số đúng lý tưởng và các điểm số được tính toán bởi DNN dựa trên các trọng số của nó được gọi là tổn thất (L: Loss). Như vậy, mục đích của đào tạo các mạng DNN là tìm được tập các trọng số cho phép giảm thiểu tổn thất trung bình trên một tập đào tạo lớn.

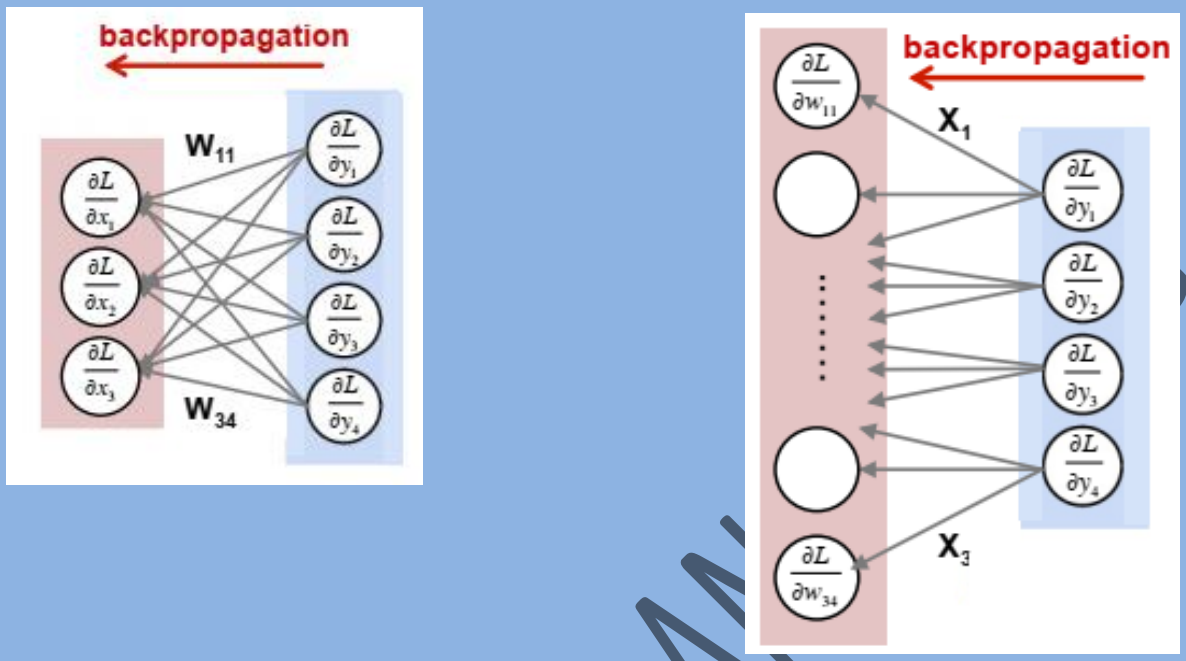
Đào tạo (Traning) là một quá trình trong đó mô hình AI/ML học cách thực hiện các nhiệm vụ nhất định của nó, cụ thể hơn là bằng cách tối ưu hóa giá trị của trọng số trong DNN. DNN được đào tạo bằng cách nhập một tập huấn luyện, thường là các mẫu đào tạo được dán nhãn chính xác. Ví dụ, phân loại hình ảnh, bộ đào tạo bao gồm các hình ảnh được phân loại chính xác. Khi đào tạo một mạng, các trọng số (w_{ij}) thường được cập nhật bằng cách sử dụng quy trình tối ưu hóa leo đồi (hill-climbing) được gọi là giảm độ dốc (gradient descent). Độ dốc cho biết trọng số sẽ thay đổi như thế nào để giảm tổn thất (khoảng cách giữa đầu ra chính xác và đầu ra do DNN tính toán dựa trên trọng số hiện tại của nó). Bội số của gradient của tổn thất so với từng trong số là đạo hàm riêng được sử dụng để cập nhật trong số (ví dụ: trọng số cập nhật $\mathbf{w}_{ij}^{t+1} = \mathbf{w}_{ij}^{t} - \alpha \frac{\partial L}{\partial \mathbf{w}_{ij}}$, trong đó α được gọi là tốc độ học) . Lưu ý rằng gradient chỉ thị rằng các trong số cần thay đổi như thế nào để giảm tổn thất. Quá trình đào tạo được lặp đi lặp lại để liên tục giảm tổn thất tổng thể. Cho đến khi tổn thất dưới ngưỡng xác định trước, DNN với độ chính xác cao sẽ thu được.

Cách tính toán các đạo hàm riêng của gradient hiệu quả là thông qua một quá trình được gọi là truyền lan ngược (Backpropagation). Truyền lan ngược, tính toán dựa được rút ra từ một quy tắc tính toán dây truyền (Chain Rule), hoạt động bằng cách chuyển ngược qua mạng các giá trị để tính toán tổn hao đã chịu ảnh hưởng của từng trọng số như thế nào.

Thực ra, tính toán truyền lan ngược có dạng rất giống với tính toán được sử dụng cho suy luận như thấy trên hình 1.4. Như vậy các kỹ thuật thực hiện hiệu quả suy luận có thể đôi khi hữu ích cho thực hiện hiệu quả đào tạo. Tuy nhiên cần lưu ý hai điểm. Thứ nhất, truyền lan ngược yêu cầu các đầu ra trung gian của mạng được dành trước cho tính toán ngược,vì thế đào tạo sẽ tăng các yêu cầu lưu trữ. Thứ hai, do sử dụng các gradient cho leo đồi (Hill-Climbing), yêu cầu độ chính xác cho đào tạo nói chung sẽ cao hơn suy luận.

(a) Tính toán gradient cho tổn thất liên quan đến các đầu vào đầu vào bộ lọc

(b) Tính toán gradient cho tổn thất liên quan đến các trọng số



Backpropagation: truyền lan ngược

Hình 1.4. Thí dụ về truyền lan ngược

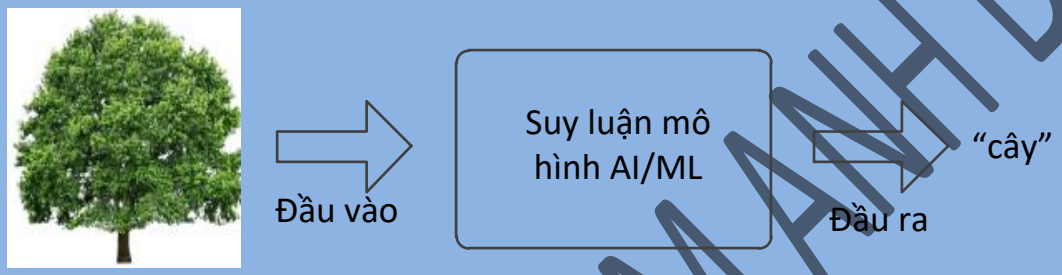
Nhiều loại kỹ thuật được sử dụng để cải thiện hiệu quả cũng như độ bền vững của đào tạo. Thí dụ, thường thì tổn thất từ nhiều tập dữ liệu đầu vào (ví dụ: một lô) được thu thập trước khi thực hiện chuyển một lần cập nhật trọng số. Điều này giúp tăng tốc và ổn định quá trình đào tạo.

Có nhiều cách để đào tạo mạng (đào tạo trọng số) cho các mục tiêu khác nhau. Giới thiệu ở trên là học có giám sát (supervised learning) sử dụng các mẫu đào tạo được gắn nhãn để tìm đầu ra chính xác cho một nhiệm vụ. Học không giám sát (unsupervised learning) sử dụng các mẫu đào tạo không được gắn nhãn để tìm cấu trúc hoặc cụm trong dữ liệu. Học tăng cường (Reinforcement learning) có thể được sử dụng để đưa ra hành động mà tác nhân nên thực hiện tiếp theo để tối đa hóa phần thưởng mong đợi. Học chuyển giao (transfer learning) là điều chỉnh các trọng số đã được đào tạo trước đó (ví dụ: trọng số trong mô hình toàn cầu) bằng cách sử dụng một tập huấn luyện mới, được sử dụng để đào tạo nhanh hơn hoặc chính xác hơn cho một mô hình được cá nhân hóa.

Mộ cách tiếp cận khác thường được sử dụng để xác định các trọng số là điều chỉnh tinh (Fine-Tuning), trong đó các trọng số được đào tạo trước đây có sẵn và được sử dụng như là một điểm bắt đầu và sau đó các trọng số này được điều chỉnh cho một tập dữ liệu mới (ví dụ: học chuyển

giao (Transfer Learning)). Điều này dẫn đến đào tạo nhanh hơn và đôi khi có thể dẫn đến độ chính xác tốt hơn.

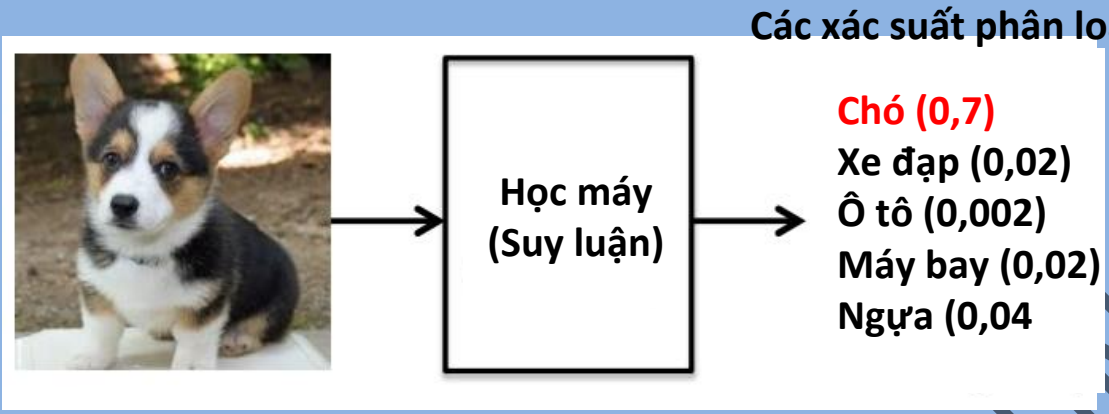
Sau khi DNN được đào tạo, nó có thể thực hiện nhiệm vụ của mình bằng cách tính toán đầu ra của mạng bằng cách sử dụng trọng số được xác định trong quá trình đào tạo, được gọi là suy luận (Inference). Trong quá trình suy luận mô hình, đầu vào từ thế giới thực được truyền qua DNN. Sau đó, dự đoán (Prediction) cho nhiệm vụ là đầu ra, như trong hình 1.3. Ví dụ: đầu vào có thể là pixel của hình ảnh, biên độ lấy mẫu của sóng âm thanh hoặc biểu diễn số của trạng thái của một số hệ thống hoặc trò chơi. Tương ứng, đầu ra của mạng có thể là xác suất mà một hình ảnh chứa một đối tượng cụ thể, xác suất mà một chuỗi âm thanh chứa một từ cụ thể hoặc một hộp giới hạn trong hình ảnh xung quanh một đối tượng hoặc hành động được đề xuất nên được thực hiện .



Hình 1.3. Thí dụ suy luận mô hình AI/ML

Hiệu năng của DNN đạt được với chi phí của độ phức tạp tính toán cao. Do đó, các công cụ tính toán hiệu quả hơn thường được sử dụng, ví dụ như đơn vị xử lý đồ họa (GPU: graphics processing unit) và bộ xử lý mạng (NPU: network processing unit). So với suy luận chỉ liên quan đến quá trình chuyển thẳng (feedforward process), việc đào tạo thường đòi hỏi nhiều tài nguyên tính toán và lưu trữ hơn vì nó cũng liên quan đến quá trình lan truyền ngược.

Trong phần này, để làm thí dụ ta sẽ sử dụng phân loại hình ảnh, như trong Hình 1.5, để làm thí dụ đào tạo và sử dụng DNN. Khi thực hiện suy luận bằng DNN, một hình ảnh được đưa và đầu vào của DNN và đầu ra của DNN là một vectơ điểm số (Score Vector), một cho mỗi lớp đối tượng; Lớp có điểm cao nhất cho biết lớp đối tượng có khả năng xảy ra nhất trong ảnh. Mục tiêu bao trùm để đào tạo DNN là xác định trọng số tối đa hóa điểm số của lớp chính xác và giảm thiểu điểm số của lớp không đúng. Khi đào tạo mạng, lớp chính xác thường được biết đến vì nó được đưa ra cho các hình ảnh được sử dụng để đào tạo (tức là tập huấn luyện của mạng). Khoảng cách giữa điểm số chính xác lý tưởng và điểm số do DNN tính toán dựa trên trọng số hiện tại của nó được gọi là tổn thất (L). Do đó, mục tiêu của việc đào tạo DNN là tìm ra một tập hợp các trọng số để giảm thiểu tổn thất trung bình trong một tập luyện lớn.



Hình 1.5. Ví dụ về nhiệm vụ phân loại hình ảnh. Nền tảng học máy lấy hình ảnh và cho ra điểm số độ tin cậy cho một tập hợp các lớp được xác định trước

2. CÁC ỨNG DỤNG CỦA DNN

Nhiều ứng dụng có thể được hưởng lợi từ các DNN khác nhau, từ đa phương tiện đến không gian y tế. Trong phần sẽ cung cấp các ví dụ về các lĩnh vực mà DNN hiện đang tác động và nêu bật các lĩnh vực mới nổi mà DNN hy vọng sẽ tạo ra tác động trong tương lai.

Hình ảnh và video. Video được cho là dữ liệu lớn nhất. Nó chiếm hơn 70% lưu lượng truy cập Internet ngày nay. Ví dụ, hơn 800 triệu giờ video được thu thập hàng ngày trên toàn thế giới để giám sát video. Thị giác máy tính (Computer Vision) là cần thiết để trích xuất thông tin có ý nghĩa từ video. DNN đã cải thiện đáng kể độ chính xác của nhiều tác vụ thị giác máy tính như phân loại hình ảnh, bản địa hóa và phát hiện đối tượng, phân đoạn hình ảnh và nhận dạng hành động.

Lời nói và ngôn ngữ. Các DNN đã cải thiện đáng kể độ chính xác của nhận dạng giọng nói [21] cũng như nhiều các tác vụ liên quan như dịch máy [2], xử lý ngôn ngữ tự nhiên [22] và tạo âm thanh.

Y tế. Các DNN đã đóng một vai trò quan trọng trong bộ gen để hiểu rõ hơn về di truyền của các bệnh như tự kỷ, ung thư và teo cơ cột sống. Chúng cũng đã được sử dụng trong hình ảnh y tế để phát hiện ung thư da, ung thư não và ung thư vú.

Chơi game. Gần đây, nhiều thử thách AI lớn liên quan đến chơi trò chơi đã được vượt qua bằng cách sử dụng DNN. Những thành công này cũng đòi hỏi những đổi mới trong kỹ thuật đào tạo và nhiều người dựa vào học tăng cường]. DNN đã vượt qua độ chính xác của con người khi chơi Atari cũng như Go], nơi việc tìm kiếm toàn diện tất cả các khả năng là không khả thi do số lượng nước đi có thể xảy ra khổng lồ không thể tưởng tượng được.

Robot. DNN đã thành công trong lĩnh vực các nhiệm vụ robot như nắm bằng cánh tay robot, lập kế hoạch chuyển động cho robot mặt đất, điều hướng trực quan, điều khiển để ổn định quadcopter (máy bay bốn cánh quạt) và chiến lược lái xe cho xe tự hành.

DNN đã được sử dụng rộng rãi trong các ứng dụng đa phương tiện ngày nay (ví dụ: thị giác máy tính, nhận dạng giọng nói). Trong tương lai, ta có thể hy vọng rằng DNN có thể sẽ đóng một vai trò ngày càng quan trọng trong lĩnh vực y tế và robot, như đã nói ở trên, cũng như tài chính (ví dụ: giao dịch, dự báo năng lượng và đánh giá rủi ro), cấu trúc khung (ví dụ: an toàn cấu trúc và kiểm soát giao thông), dự báo thời tiết và phát hiện sự kiện. Vô số lĩnh vực ứng dụng đặt ra những thách thức mới đối với việc xử lý hiệu quả DNN; khi này các giải pháp phải thích ứng và có thể mở rộng để xử lý các dạng DNN mới và đa dạng mà các ứng dụng này có thể sử dụng.

Nhúng so với đám mây

Các ứng dụng và khía cạnh khác nhau của xử lý DNN (tức là đào tạo so với suy luận) có nhu cầu tính toán khác nhau. Cụ thể, việc đào tạo thường yêu cầu một bộ dữ liệu lớn và tài nguyên tính toán đáng kể cho nhiều lần lặp lại cập nhật trọng số. Trong nhiều trường hợp, việc đào tạo mô hình DNN vẫn mất vài giờ đến nhiều ngày và do đó thường được thực hiện trên đám mây. Mặt khác, suy luận có thể xảy ra trên đám mây hoặc ở biên (ví dụ: IoT hoặc thiết bị di động).

Trong nhiều ứng dụng, thường mong muốn xử lý suy luận DNN gần cảm biến. Ví dụ, trong các ứng dụng thị giác máy tính, chẳng hạn như đo thời gian chờ đợi trong các cửa hàng hoặc dự đoán mô hình lưu lượng truy cập, nên trích xuất thông tin có ý nghĩa từ video ngay tại cảm biến hình ảnh thay vì trên đám mây để giảm chi phí truyền thông. Đối với các ứng dụng khác như xe tự hành, điều hướng máy bay không người lái và robot, xử lý tại chỗ là mong muốn vì độ trễ và rủi ro bảo mật của việc dựa vào đám mây là quá cao.

Tuy nhiên, video liên quan đến một lượng lớn dữ liệu đòi hỏi xử lý phức tạp về mặt tính toán; nên phần cứng chi phí thấp để phân tích video là một thách thức nhưng rất quan trọng để cho phép các ứng dụng này. Nhận dạng giọng nói cho phép chúng ta tương tác liền mạch với các thiết bị điện tử, chẳng hạn như điện thoại thông minh. Mặc dù hiện tại hầu hết quá trình xử lý cho các ứng dụng như Apple Siri và dịch vụ thoại Amazon Alexa là trên đám mây, nhưng vẫn nên thực hiện nhận dạng trên chính thiết bị để giảm độ trễ và phụ thuộc vào kết nối, đồng thời cải thiện quyền riêng tư và bảo mật.

Nhiều nền tảng nhúng thực hiện suy luận DNN có giới hạn nghiêm ngặt về mức tiêu thụ năng lượng, tính toán và chi phí bộ nhớ; do đó, việc xử lý hiệu quả các DNN đã trở nên quan trọng hàng đầu trong điều kiện những ràng buộc này. Các phần dưới đây sẽ tập trung vào các yêu cầu tính toán để suy luận hơn là đào tạo

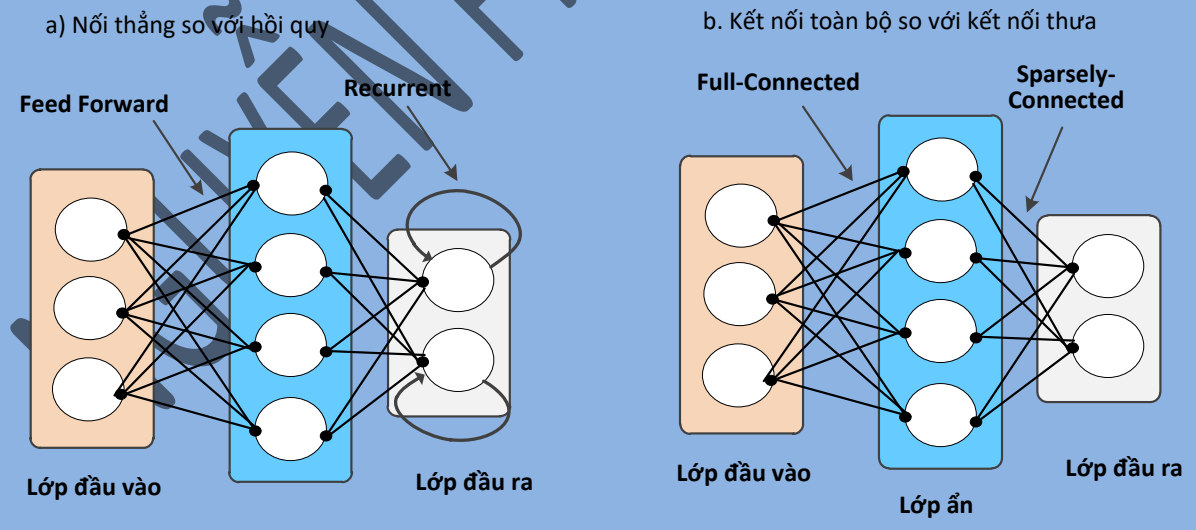
3.TỔNG QUAN KIẾN TRÚC DNN

DNN có nhiều hình dạng và kích thước khác nhau tùy thuộc vào ứng dụng. Các hình dạng và kích thước phổ biến cũng đang phát triển nhanh chóng để cải thiện độ chính xác và hiệu quả. Trong mọi trường hợp, đầu vào cho DNN là một tập hợp các giá trị đại diện cho thông tin sẽ được mạng phân tích. Ví dụ: các giá trị này có thể là pixel của hình ảnh, biên độ lấy mẫu của sóng âm thanh hoặc biểu diễn số của trạng thái của một số hệ thống hoặc trò chơi.

3.1. Các kiểu mạng thần kinh khác nhau

Các mạng xử lý đầu vào có hai dạng chính: nối thẳng (Feeed Forward) và hồi quy (Recurrent) như trên hình 3.1a). Trong các mạng nối thẳng (Feeed Forward Network), tất cả các tính toán được thực hiện như một chuỗi các hoạt động trên đầu ra của lớp trước đó. Tập hợp các hoạt động cuối cùng tạo ra đầu ra của mạng, ví dụ xác suất một hình ảnh chứa một đối tượng cụ thể, xác suất mà một chuỗi âm thanh chứa một từ cụ thể, một hộp giới hạn trong hình ảnh xung quanh một đối tượng hoặc hành động được đề xuất nên được thực hiện. Trong các DNN như vậy, mạng không có bộ nhớ và đầu ra cho một đầu vào luôn giống nhau bất kể chuỗi đầu vào trước đó được cung cấp trước đó cho mạng.

Ngược lại, mạng nơ-ron hồi quy (RNN: recurrent neural network), trong đó mạng bộ nhớ dài hạn ngắn (LSTM: Long Short-Term Memory network) là một biến thể phổ biến, có bộ nhớ trong để cho phép các phụ thuộc dài hạn ảnh hưởng đến đầu ra. Trong các mạng này, một số hoạt động trung gian tạo ra các giá trị được lưu trữ bên trong mạng và được sử dụng làm đầu vào cho các hoạt động khác kết hợp với việc xử lý đầu vào sau này.



Hình 3.1. Các kiểu mạng thần kinh khác nhau

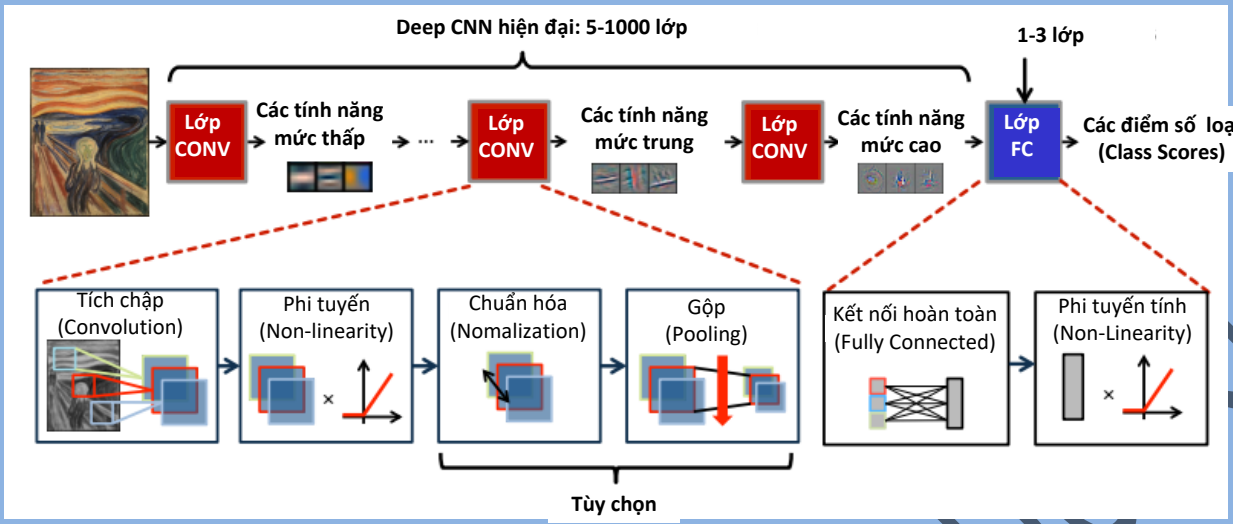
Các phần dưới đây sẽ tập trung vào các mạng nối thẳng (Feedforward Network) vì (1) tính toán chính trong RNN vẫn là tổng trọng số, được bao phủ bởi các mạng nôi thẳng và (2) cho đến nay ít chú ý dành cho tăng tốc phần cứng, đặc biệt cho RNN.

DNN có thể chỉ bao gồm các lớp được kết nối hoàn toàn (FC) (còn được gọi là bộ nhận thức nhiều lớp, hoặc MLP: multi-layer perceptron) như được biểu trong lớp ngoài cùng bên trái của Hình 3.1b). Trong lớp FC, tất cả các kích hoạt đầu ra bao gồm tổng trọng số của tất cả các kích hoạt đầu vào (tức là tất cả các đầu ra được kết nối với tất cả các đầu vào). Điều này đòi hỏi một lượng lưu trữ và tính toán đáng kể. Rất may, trong nhiều ứng dụng, chúng ta có thể loại bỏ một số kết nối giữa các kích hoạt bằng cách đặt trọng số bằng không mà không ảnh hưởng đến độ chính xác. Điều này dẫn đến một lớp được kết nối thưa thớt. Một lớp được kết nối thưa thớt được minh họa trong lớp ngoài cùng bên phải của hình 3.1b).

Ta cũng có thể làm cho việc tính toán hiệu quả hơn bằng cách giới hạn số lượng trọng số đóng góp vào đầu ra. Loại thưa thớt có cấu trúc này có thể phát sinh nếu mỗi đầu ra chỉ là một hàm của một cửa sổ đầu vào có kích thước cố định. Hiệu quả thậm chí có thể đạt được nếu cùng một bộ trọng lượng được sử dụng trong tính toán mọi đầu ra. Sử dụng lặp của cùng một giá trị trọng lượng này được gọi là chia sẻ trọng số (weight sharing) và có thể làm giảm đáng kể yêu cầu bảo quản trọng số.

3.2. Mạng thần kinh tích chập (CNN)

Một dạng phổ biến của DNN là Mạng thần kinh tích chập (CNN), bao gồm nhiều lớp CONV như trong hình 3.2. Trong các mạng như vậy, mỗi lớp tạo ra một sự trừu tượng hóa cấp độ cao hơn liên tiếp của dữ liệu đầu vào, được gọi là bản đồ tính năng (fmap: feature map) giữ lại thông tin thiết yếu nhưng duy nhất. CNN hiện đại có thể đạt được hiệu năng vượt trội bằng cách sử dụng hệ thống phân cấp các lớp rất sâu. CNN được sử dụng rộng rãi trong nhiều ứng dụng bao gồm hiểu hình ảnh nhận dạng giọng nói, chơi trò chơi, robot v.v. Phần này sẽ tập trung vào việc sử dụng nó trong xử lý hình ảnh, đặc biệt cho nhiệm vụ phân loại hình ảnh (image classification)

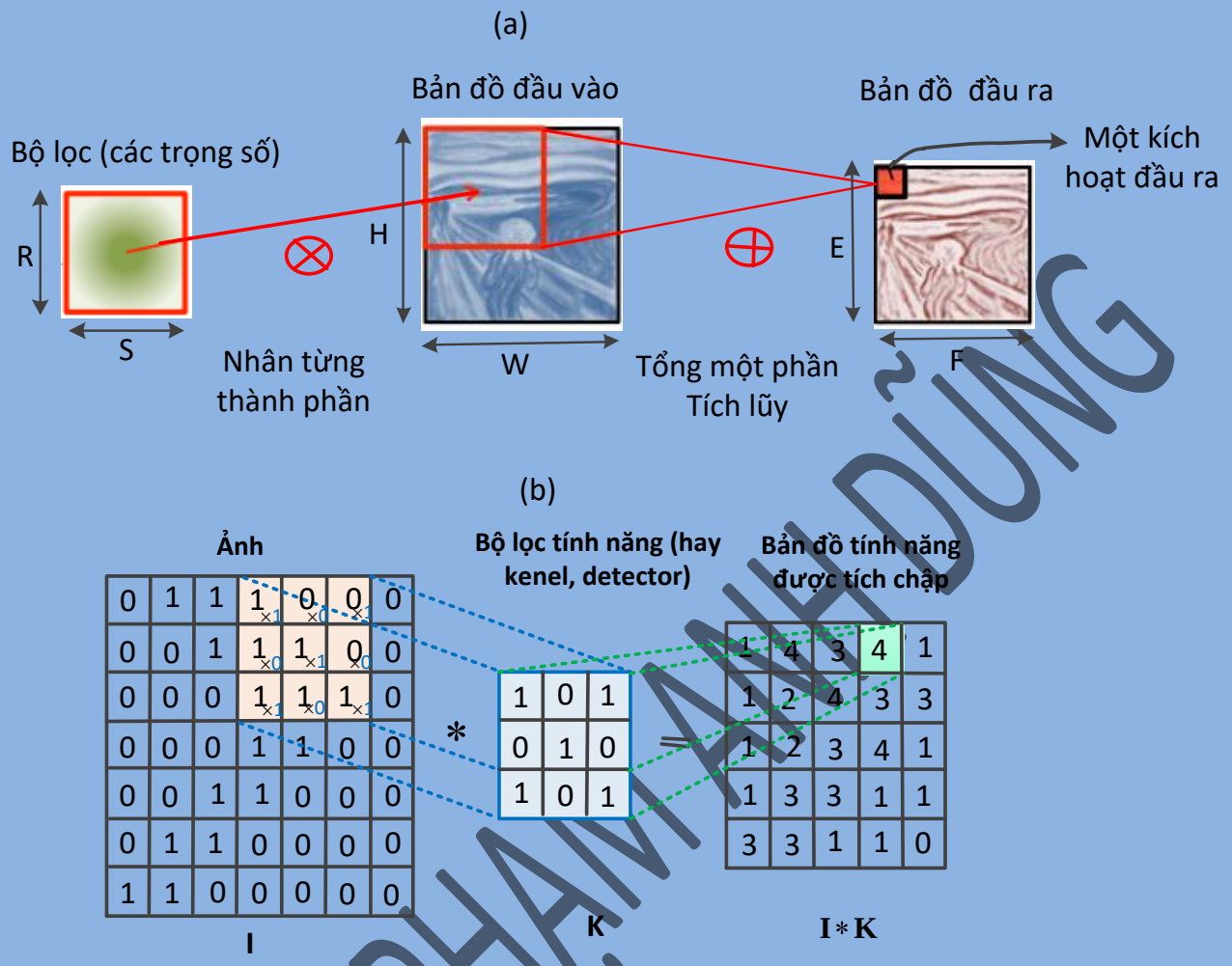


Hình 3.2. Mạng CNN

Hình 3.3 cho thấy tích chập 2D trong xử lý ảnh truyền thống. Hình 3.3(a), trong đó tổng trọng số cho mỗi kích hoạt đầu ra được tính toán chỉ bằng cách sử dụng một vùng lân cận nhỏ của các kích hoạt đầu vào (tức là tất cả các trọng số vượt ra ngoài vùng lân cận được đặt bằng không) và trong đó cùng một tập hợp trọng số được chia sẻ cho mọi đầu ra (tức là, bộ lọc là không gian bất biến).. Hình 3.3(b) minh họa tính toán tích chập cụ thể cho một bản đồ tính năng của hình ảnh đầu vào trong đó bộ lọc tính năng có kích thước 3×3 . Quá trình tích chập trên hình 3.3(b) được tính như sau:

$$1\times1++0\times0+0\times1+1\times1\ +0\times0++1\times1+0\times1\ +1\times0++0\times0+1\times1=4$$

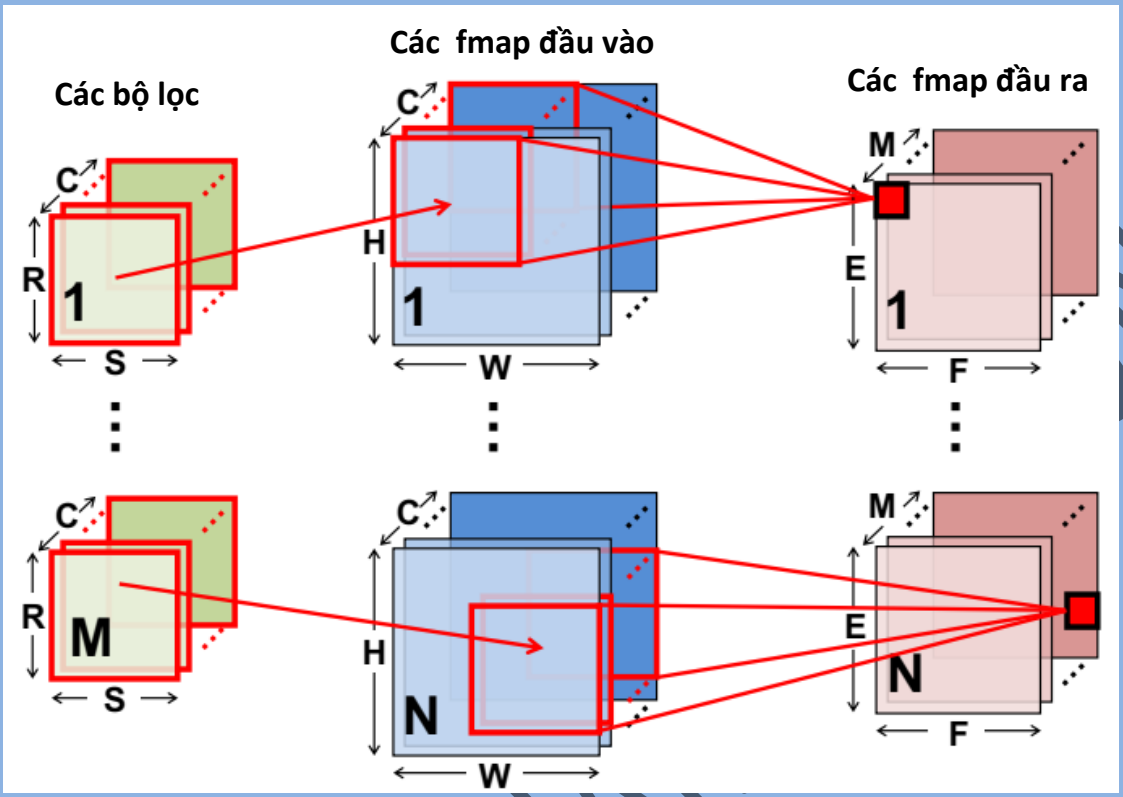
Chi tiết hơn, bộ lọc tính năng (kenel chập) sẽ dịch trên toàn bộ vị trí của bản đồ tính năng đầu vào, nhân phần tử và cộng để được tích chập đầu ra của vị trí tương ứng.



Hình 3.3. Tích chập 2D trong xử lý ảnh truyền thống

Mỗi lớp CONV trong CNN chủ yếu bao gồm các tích chập kích thước cao (High- Dimentional Convolution) như thể hiện trong hình 3.4. Trong tính toán này, các kích hoạt đầu vào của một lớp được cấu trúc thành một tập hợp các bản đồ tính năng đầu vào 2-D (ifmaps: input feature maps), mỗi bản đồ được gọi là một kênh. Mỗi kênh được ghép với một bộ lọc 2-D riêng biệt từ ngăn xếp bộ lọc, một bộ lọc cho mỗi kênh; ngăn xếp bộ lọc 2-D này thường được gọi là một bộ lọc 3-D duy nhất (single 3-D filter). Kết quả của tích chập tại mỗi điểm được tổng hợp (summed) trên tất cả các kênh. Ngoài ra, thiên kiến 1-D (1-D bias) có thể được thêm vào kết quả lọc, nhưng một số mạng gần đây loại bỏ việc sử dụng nó khỏi các phần của lớp. Kết quả của tính toán này là các kích hoạt đầu ra bao gồm một kênh của bản đồ tính năng đầu ra (ofmap: output feature map). Các bộ lọc 3-D bổ sung có thể được sử dụng trên cùng một đầu vào để tạo các kênh đầu ra bổ sung. Cuối cùng, nhiều bản đồ tính năng đầu vào (multiple input feature

maps) có thể được xử lý cùng nhau như một lô (batch) để có khả năng cải thiện việc tái sử dụng trọng số bộ lọc.



Hình 3.4. Các tích chập kích thước cao trong CNN

Với các tham số hình dạng trong Bảng 3.1, tính toán của một lớp CONV được định nghĩa như sau:

$$\mathbf{O}[z][x][y] = \mathbf{B}[u] + \sum_{k=0}^{C-1}\sum_{i=0}^{S-1}\sum_{j=0}^{R-1} \mathbf{I}[z][k][Ux+i][Uy+j] \times \mathbf{W}[u][k][i][j], \quad (3.1)$$

$$0 \le z < N;\ 0 \le u < M;\ 0 \le x < F;\ 0 \le y < E;$$

$$E = (H - R + U)/U;\ F = (W - S + U)/U.$$

Bảng 3.1. Các thông số hình dạng của một lớp CONV/FC

Thông số hình dạng (Shape Parameter)	**Mô tả**
N	Kích thước lô của các fmaps
M	Số của các bộ lọc 3-D/số của các kênh ofmap
C	Số của các kênh ifmap/lọc
H/W	Chiều cao/chiều rộng của mặt phẳng ifmap

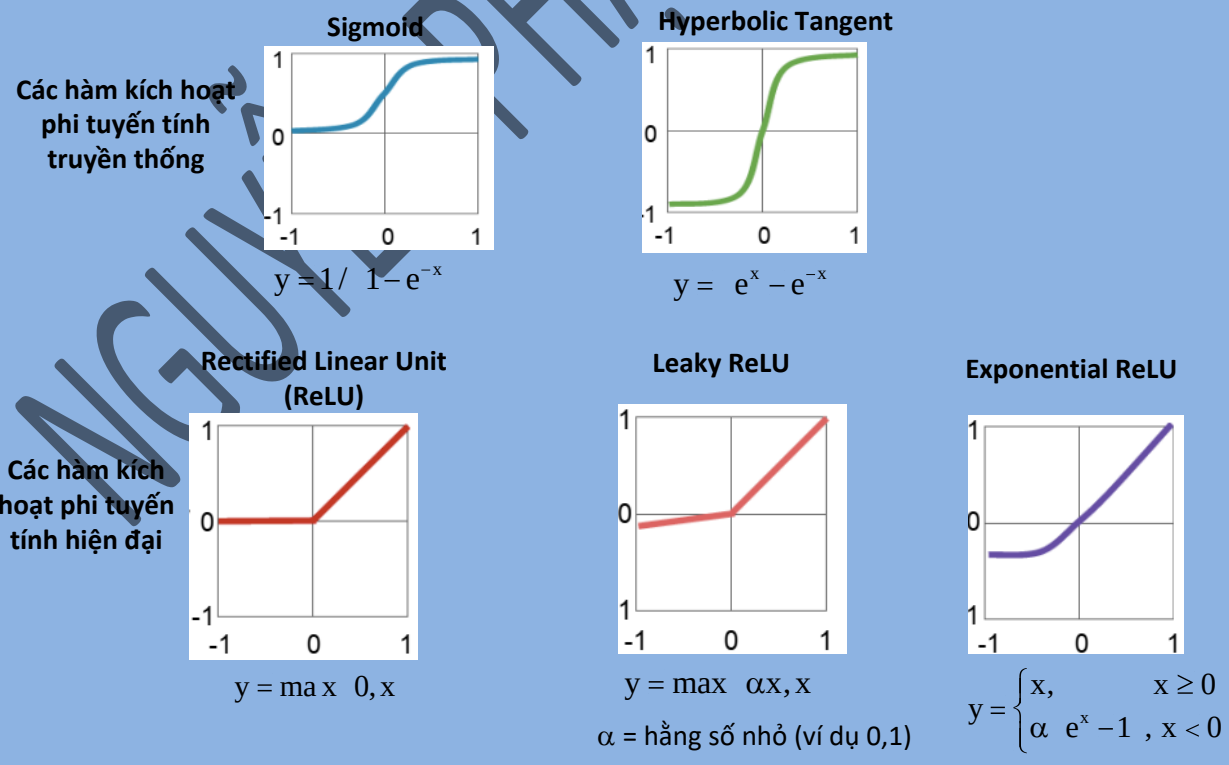
R/S	Chiều cao/chiều rộng của mặt phẳng bộ lọc (=H hoặc W trong FC)
E/F	Chiều cao/chiều rộng của mặt phẳng ofmap (=1 trong FC)

O, **I**, **W** và **B** lần lượt là ma trận của ofmaps, ifmaps, bộ lọc và bias (thiên kiến). **U** là một kích thước sải chân nhất định. Hình 3.4 cho thấy hình ảnh trực quan của tính toán này (bỏ qua các thiên kiến).

Để đồng bộ thuật ngữ của CNN với DNN chung,

- Các bộ lọc bao gồm các trọng số (tức là synapses: các khớp thần kinh)
- Bản đồ tính năng đầu vào và đầu ra (ifmaps, ofmaps) bao gồm các kích hoạt (tức là các tế bào thần kinh đầu vào và đầu ra
- Sải chân (Stride): Trong tích chập theo sải chân, thay vì dịch một dòng hay một cột bộ lọc mỗi lần, ta có thể dịch nó 2 hay 3 dòng hay cột mỗi lần.

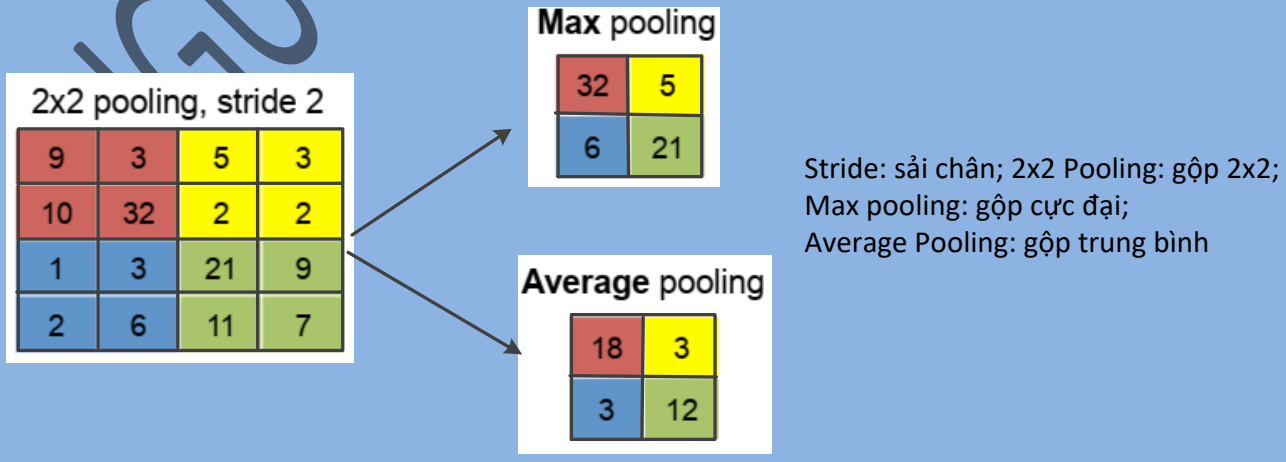
Từ năm đến hơn một nghìn lớp CONV thường được sử dụng trong các mô hình CNN gần đây. Một số lượng nhỏ, ví dụ: 1 đến 3, các lớp kết nối đầy đủ (FC) thường được áp dụng sau các lớp CONV cho mục đích phân loại. Một lớp FC cũng áp dụng các bộ lọc trên các ifmap như trong các lớp CONV, nhưng các bộ lọc có cùng kích thước với các ifmap. Do đó, nó không có thuộc tính chia sẻ trọng lượng của các lớp CONV. Phương trình (3.1) vẫn giữ nguyên cho việc tính toán các lớp FC với một vài ràng buộc bổ sung về các tham số hình dạng: H = R, F = S, E = F = 1 và U = 1.



Hình 3.5. Các dạng hàm kích hoạt phi tuyến tính khác nhau

Ngoài các lớp CONV và FC, các lớp tùy chọn khác nhau có thể được tìm thấy trong DNN như phi tuyến tính, gộp và chuẩn hóa. Chức năng và tính toán cho từng lớp này sẽ được trình bàu dưới đây.

1) *Phi tuyến tính*: Một hàm kích hoạt phi tuyến tính thường được áp dụng sau mỗi lớp CONV hoặc FC. Các hàm phi tuyến tính khác nhau được sử dụng để đưa tính phi tuyến tính vào DNN như trong Hình 3.5. Chúng bao gồm các hàm phi tuyến thông thường như sigmoid hoặc tiếp tuyến hyperbol cũng như đơn vị tuyến tính chỉnh lưu (ReLU: rectified linear unit) đã trở nên phổ biến trong những năm gần đây do tính đơn giản và khả năng cho phép đào tạo nhanh chóng. Các biến thể của ReLU, chẳng hạn như ReLU rò rỉ (LeakyRLU), ReLU tham số (Parametric RLU) và LU hàm mũ (exponential LU) cũng đã được khám phá để cải thiện độ chính xác. Cuối cùng, một phi tuyến tính được gọi là maxout, lấy giá trị tối đa của hai hàm tuyến tính giao nhau, đã được chứng minh là có hiệu quả trong các nhiệm vụ nhận dạng giọng nói.
2) *Gộp*: Lớp gộp thực hiện chiết suất các giá trị đặc biệt của một tập giá trị thường là giá trị cực đại hoặc giá trị trung bình của tất cả các giá trị. Điều này giảm kích thước của bản đồ tính năng đầu ra. Một loạt các phép tính làm giảm kích thước của bản đồ tính năng được gọi là gộp (Pooling). Pooling, được áp dụng riêng biệt cho từng kênh, cho phép mạng mạnh mẽ và bất biến trước những thay đổi và biến dạng nhỏ. Pooling kết hợp, hoặc gộp, một tập hợp các giá trị trong trường dễ tiếp nhận (receptive field) của nó thành một số giá trị nhỏ hơn. Nó có thể được cấu hình dựa trên kích thước của trường dễ tiếp nhận của nó (ví dụ: 2×2) và hoạt động gộp (ví dụ: tối đa hoặc trung bình), như thể hiện trong Hình 3.6. Thông thường, gộp lại xảy ra trên các khối không chồng lên nhau (tức là sải chân bằng kích thước của gộp). Thông thường một sải bước lớn hơn một được sử dụng sao cho có sự giảm kích thước của biểu diễn (tức là bản đồ tính năng).



Hình 3.6.Các dạng gộp khác nhau

3) *Chuẩn hóa*: Kiểm soát phân phối đầu vào trên các lớp có thể giúp tăng tốc độ đào tạo và cải thiện độ chính xác. Theo đó, sự phân bố của các kích hoạt đầu vào lớp (σ, μ) được chuẩn hóa sao cho nó có giá trị trung bình bằng không và độ lệch chuẩn đơn vị. Trong chuẩn hóa theo lô (BN: Batch Normalization), giá trị chuẩn hóa được chia tỷ lệ và dịch chuyển thêm, như thể hiện trong Phương trình (2), trong đó các tham số (γ, β) được học từ đào tạo [47]. ε là một hằng số nhỏ để tránh các vấn đề số (Numerical Problem). Trước đó, chuẩn hóa phản ứng cục bộ (LRN: Local Response Normalization) được sử dụng, được lấy cảm hứng từ sự ức chế bên trong sinh học thần kinh, nơi các tế bào thần kinh bị kích thích (tức là kích hoạt có giá trị cao) sẽ khuất phục các tế bào lân cận của nó (tức là gây ra các kích hoạt có giá trị thấp); tuy nhiên, BN hiện được coi là thông lệ tiêu chuẩn trong thiết kế CNN trong khi LRN hầu như không được dùng nữa. Lưu ý rằng trong khi LRN thường được thực hiện sau hàm phi tuyến tính, BN chủ yếu được thực hiện giữa lớp CONV hoặc FC và hàm phi tuyến tính.

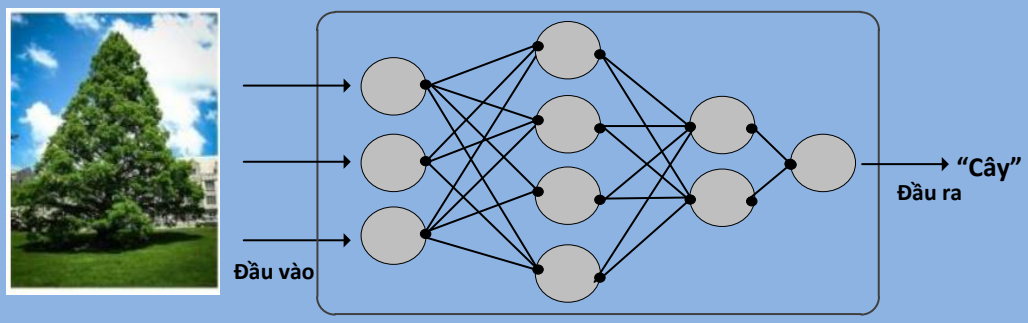
$$y = \frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}}\gamma + \beta \qquad (3.2)$$

4. CÁC MÔ HÌNH VÀ GIẢI THUẬT ĐƯỢC SỬ DỤNG RỘNG RÃI

Nhiều mô hình DNN đã được phát triển trong hai thập kỷ qua. Mỗi mô hình này có một "kiến trúc mạng" khác nhau về số lớp, loại lớp, hình dạng lớp (tức là kích thước bộ lọc, số kênh và bộ lọc) và kết nối giữa các lớp.

4.1. Mô hình Perceptron nhiều lớp (MLP)

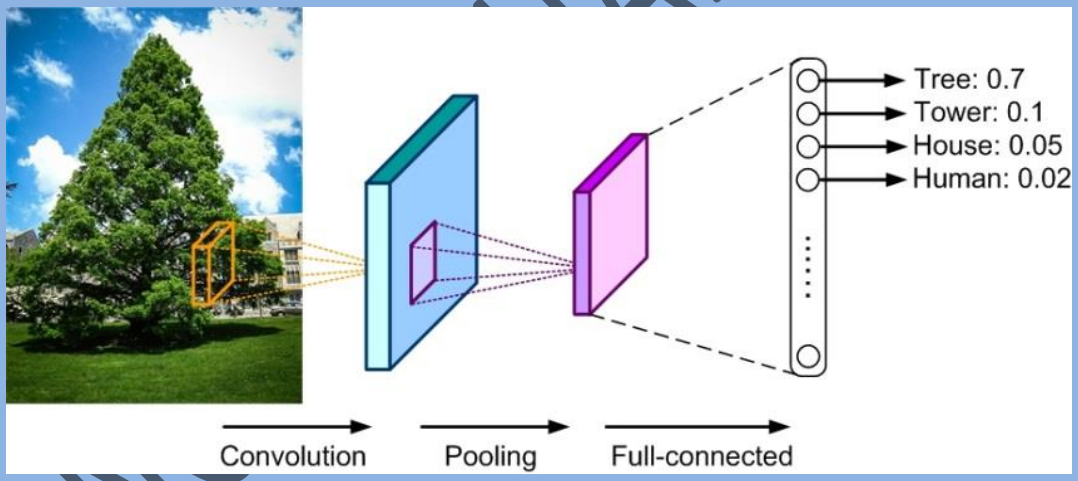
Hình 4 trình bày ba cấu trúc phổ biến của DNN: perceptron nhiều lớp (MLP: multilayer perceptron; perceptron là một neuron nhân tạo cố gắng xấp xỉ một neuron sinh học), mạng neuron tích chập (CNN: convolution neural network) và mạng nơ-ron hồi quy (RNN: recurrent neural network). Mô hình Perceptron nhiều lớp (MLP) là DNN cơ bản nhất, bao gồm một loạt các lớp được kết nối đầy đủ. Trong một lớp được kết nối hoàn toàn, tất cả các đầu ra được kết nối với tất cả các đầu vào, như thể hiện trong Hình 4. Do đó, MLP đòi hỏi một lượng lưu trữ và tính toán đáng kể.



Hình 4. Mô hình MLP DNN

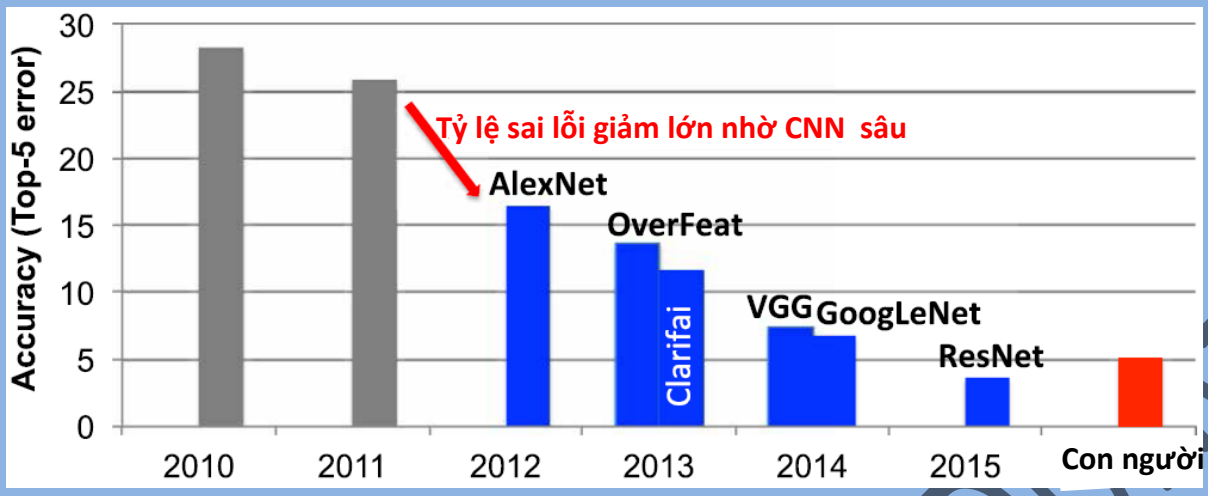
4.2. Mạng neuron tích chập (CNN)

Một cách tiếp cận để giới hạn số lượng trọng số đóng góp vào đầu ra là chỉ tính toán đầu ra bằng cách sử dụng một hàm của cửa sổ đầu vào có kích thước cố định. Một mô hình DNN dựa trên cửa sổ cực kỳ phổ biến sử dụng phép toán tích chập để cấu trúc tính toán, do đó nó được đặt tên là mạng neuron tích chập (CNN: convolution neural network). Một CNN bao gồm nhiều lớp tích chập, như thể hiện trong hình 4.1. Áp dụng các bộ lọc tích chập khác nhau, các mô hình CNN có thể nắm bắt biểu diễn cấp cao của dữ liệu đầu vào, làm cho nó trở nên phổ biến cho các nhiệm vụ phân loại hình ảnh đáng kể hiệu năng của các nhiệm vụ phân loại hình ảnh (ví dụ: AlexNet, VGG network, GoogleNet, ResNet , MobileNet) như thể hiện trong hình 4.2.



Convolution: tích chập; Pooling: gộp; Full-connected: kết nối toàn bộ; Tree: cây; Tower: tháp; Housse: nhà; Human: người

Hình 4.1. Mô hình CNN



Accuracy (Top-5 error): độ chính xác (sai lỗi top-5)

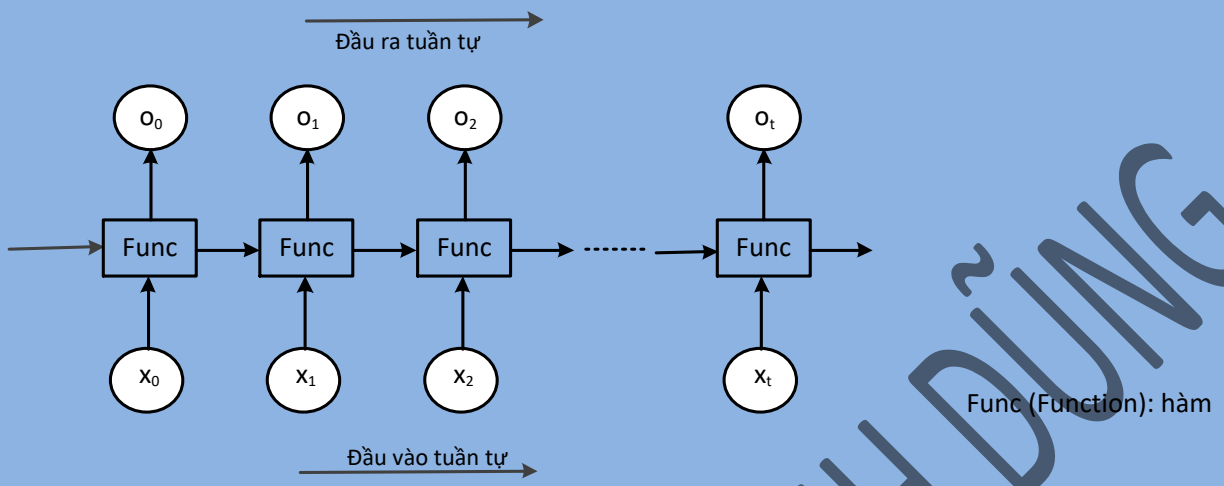
Hình 4.2. Cải thiện phân loại hình ảnh được thực hiện bởi các mô hình CNN

Hình 6 cho thấy hiệu năng của các người tham gia tốt nhất trong cuộc thi ImageNet trên nhiều năm. Ta có thể thấy rằng, độ chính xác của các thuật toán ban đầu có tỷ lệ lỗi từ 25% trở lên. Năm 2012, một nhóm từ Đại học Toronto đã sử dụng các đơn vị xử lý đồ họa (GPU) do chúng khả năng tính toán cao và cách tiếp cận mạng nơ-ron sâu, được đặt tên là AlexNet, và đã giảm tỷ lệ lỗi khoảng 10%. Thành tích của họ đã truyền cảm hứng cho một loạt các thuật toán phong cách học sâu dẫn đến một loạt các cải tiến đều đặn. Cùng với xu hướng tiếp cận học sâu cho Thử thách ImageNet, đã có sự gia tăng tương ứng về số lượng người tham gia sử dụng GPU. Từ năm 2012 khi chỉ có 4 người tham gia sử dụng GPU cho đến năm 2014 khi gần như tất cả (110) người tham gia sử dụng GPU. Điều này phản ánh sự chuyển đổi gần như hoàn toàn từ các phương pháp tiếp cận thị giác máy tính truyền thống sang các phương pháp tiếp cận dựa trên học sâu cho cuộc thi. Năm 2015, chiến thắng của ImageNet, ResNet, vượt quá độ chính xác ở cấp độ con người với tỷ lệ lỗi top 5 dưới 5%. Kể từ đó, tỷ lệ lỗi đã giảm xuống dưới 3% và hiện đang tập trung nhiều hơn vào các thành phần khó khăn hơn của cuộc thi, chẳng hạn như phát hiện và định vị đối tượng. Những thành công này rõ ràng là một yếu tố góp phần vào phạm vi rộng các ứng dụng trng đó các DNN được áp dụng.

4.3. Mô hình mạng nơ-ron hồi quy (RNN: Recurrent neural network)

Mô hình mạng nơ-ron hồi quy (RNN: Recurrent neural network) là một loại DNN khác, sử dụng nguồn cấp dữ liệu tuần tự. Đầu vào của RNN bao gồm đầu vào hiện tại và các mẫu trước đó. Mỗi tế bào thần kinh trong RNN sở hữu một bộ nhớ bên trong lưu giữ thông tin tính toán từ các mẫu trước đó. Như thể hiện trong hình 4.3, đơn vị cơ bản của RNN được gọi là tế bào, và mỗi tế bào bao gồm các lớp và một loạt các tế bào cho phép xử lý tuần tự các mô hình RNN. Các mô

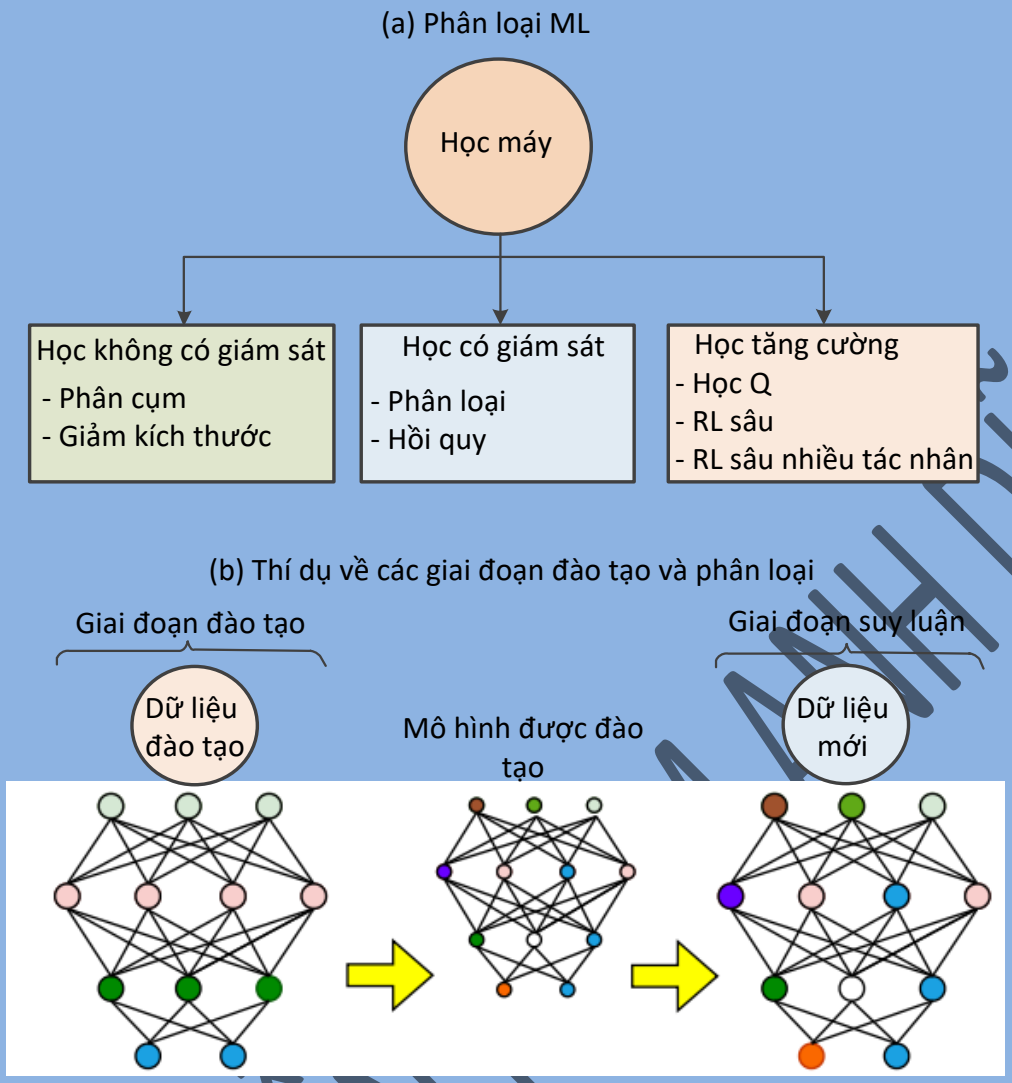
hình RNN đã được sử dụng rộng rãi trong nhiệm vụ xử lý ngôn ngữ tự nhiên trên thiết bị di động, ví dụ: mô hình ngôn ngữ, dịch máy, trả lời câu hỏi, nhúng từ và phân loại tài liệu.



hình 4.3. Mô hình RNN

4.4. Cơ sở của học máy (ML)

ML là một nhánh của AI, để dạy cho một hệ thống máy tính thực hiện dự báo dựa trên các dữ liệu. Nói chung, ML có thể được phân chia (Split) và ba cách tiếp cận, bao gồm học có giám sát (SL: Supervised Learning), học không có giám sát (UL: Unsupervised Learning) và học tăng cường (RL: Reinforcement Learning) như thấy trên hình 4.4(a). Dữ liệu đào tạo trong học có giám sát bao gồm cả hai: các đầu vào và các đầu ra được đánh nhãn và mục đích của nó là ước tính mô hình chưa được biết và lập bản đồ (ánh xạ) các đầu vào đã biết vào các nhãn. Hai loại chính của học có giám sát là phân loại (Classification) và hồi quy (Regression). Loại đầu học để dự đoán các nhãn lớp (Class Label) rời rạc (ví dụ: phân loại đối tượng), trong khi đó loại thứ hai học để dự đoán số lượng liên tục (ví dụ: dự báo đoán giá). Không giống như học có giám sát, trong học không có giám sát dữ liệu đào tạo không có nhãn và mục đích của học không có giám sát là học để để trình bày hiệu quả hơn một tập các đầu vào chưa được biết. Một thí dụ nổi tiếng của học không có giám sát là phân cụm k-means (k-means clustering), có thể áp dụng cho nhiều vấn đề phân cụm trong các hệ thống IoT, như phân cụm người dùng và tối ưu hóa sắp xếp. Trong học tăng cường, tác nhân không nói ra các hành động cần thực hiện mà tương tác liên tục với môi trường học và cố gắng tìm ra chính sách /hành động tốt để tạo ra phần thưởng tích lũy tốt nhất.

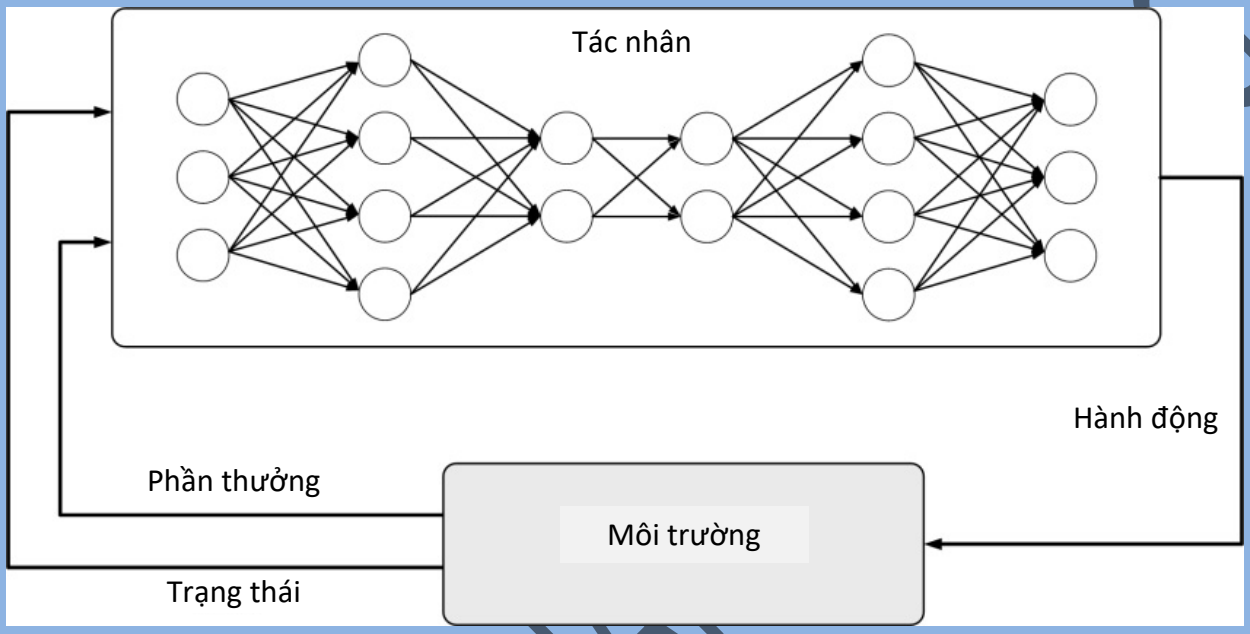


Hình 4.4. Học máy: (a) phân loại chung của ML và (b) Xử lý ML nói chung bao gồm giai đoạn đào tạo và giai đoạn suy luận

Như thấy trên hình 4.4(b), một vấn đề của ML có thể được chia thành hai giai đoạn: giai đoạn đào tạo và giai đoạn suy luận. Trong giai đoạn đào tạo, mô hình ML (dự báo hoặc phân loại) được xây dựng dựa trên dữ liệu đầu vào, các dữ liệu này có thể được đánh nhãn hoặc khôn đánh nhãn. Trong AI thông thường, dữ liệu được thu thập từ các người dùng cho đào tạo trung tâm trong đám mây tập trung có các tài nguyên lưu trữ và tính toán mạnh. Giai đoạn thứ hai của ML là suy luận, khi này dữ liệu trực tiếp (Live Data) được đưa vào mô hình được đào tạo để tạo ra các hành động đầu ra.

Học tăng cường sâu (DRL: Deep reinforcement learning)

Học tăng cường sâu (DRL: Deep reinforcement learning) không phải là một mô hình DNN khác. Nó bao gồm DNN và học tăng cường. Như minh họa trong hình 4.4, mục tiêu của DRL là tạo ra một tác nhân thông minh có thể thực hiện các chính sách hiệu quả để tối đa hóa phần thưởng của các nhiệm vụ dài hạn với các hành động có thể kiểm soát được. Ứng dụng điển hình của DRL là giải quyết các vấn đề lập lịch khác nhau, chẳng hạn như các vấn đề quyết định trong trò chơi, lựa chọn tốc độ truyền video, v.v.



Hình 4.4. Học tăng cường

5. CÁC MÔ HÌNH CNN PHỔ BIẾN

Nhiều mô hình CNN đã được phát triển trong hai thập kỷ qua. Mỗi mô hình này có một 'kiến trúc mạng' khác nhau về số lớp, loại lớp, hình dạng lớp (tức là kích thước bộ lọc, số kênh và bộ lọc) và kết nối giữa các lớp. Hiểu được những biến thể và xu hướng này là rất quan trọng để kết hợp tính linh hoạt phù hợp trong bất kỳ công cụ CNN hiệu quả nào.

Phần nay sẽ cung cấp một cái nhìn tổng quan về các CNN phổ biến khác nhau như LeNet cũng như những CNN đã cạnh tranh và / hoặc giành chiến thắng trong Thử thách ImageNet (ImageNet Chalange) như trong Hình 4.4, hầu hết các mô hình có trọng số được đào tạo trước đều có sẵn công khai để tải xuống; các mô hình CNN được tóm tắt trong Bảng 5.2. Hai kết quả cho kết quả lỗi top 5 (Top 5 Error) được báo cáo. Trong hàng đầu tiên, độ chính xác được tăng

cường bằng cách sử dụng nhiều crop (mùa vụ) từ hình ảnh và một tập hợp nhiều mô hình được đào tạo (tức là CNN cần được chạy nhiều lần); những kết quả này đã được sử dụng để cạnh tranh trong Thử thách ImageNet. Hàng thứ hai báo cáo độ chính xác nếu chỉ sử dụng một mùa vụ duy nhất (tức là DNN chỉ được chạy một lần), điều này phù hợp hơn với những gì có thể được triển khai trong các ứng dụng thời gian thực và/hoặc hạn chế năng lượng.

Bảng 5.2. Tổng kết các CNN phổ biến, độ chính xác được đo trên Top-5 Error

Metrics	LeNet 5	AlexNet	Overfeat fast	VGG 16	GoogLeNet v1	ResNet 50
Top-5 error†	n/a	16.4	14.2	7.4	6.7	5.3
Top-5 error (single crop)†	n/a	19.8	17.0	8.8	10.7	7.0
Input Size	28×28	227×227	231×231	224×224	224×224	224×224
# of CONV Layers	2	5	5	13	57	53
Depth in # of CONV Layers	2	5	5	13	21	49
Filter Sizes	5	3,5,11	3,5,11	3	1,3,5,7	1,3,7
# of Channels	1, 20	3-256	3-1024	3-512	3-832	3-2048
# of Filters	20, 50	96-384	96-1024	64-512	16-384	64-2048
Stride	1	1,4	1,4	1	1,2	1,2
Weights	2.6k	2.3M	16M	14.7M	6.0M	23.5M
MACs	283k	666M	2.67G	15.3G	1.43G	3.86G
# of FC Layers	2	3	3	3	1	1
Filter Sizes	1,4	1,6	1,6,12	1,7	1	1
# of Channels	50, 500	256-4096	1024-4096	512-4096	1024	2048
# of Filters	10, 500	1000-4096	1000-4096	1000-4096	1000	1000
Weights	58k	58.6M	130M	124M	1M	2M
MACs	58k	58.6M	130M	124M	1M	2M
Total Weights	60k	61M	146M	138M	7M	25.5M
Total MACs	341k	724M	2.8G	15.5G	1.43G	3.9G
Pretrained Model Website	[56]‡	[57, 58]	n/a	[57-59]	[57-59]	[57-59]

SUMMARY OF POPULAR DNNS [3, 15, 48, 50, 51]. †ACCURACY IS MEASURED BASED ON TOP-5 ERROR ON IMAGENET [14]. ‡THIS VERSION OF LENET-5 HAS 431K WEIGHTS FOR THE FILTERS AND REQUIRES 2.3M MACS PER IMAGE, AND USES RELU RATHER THAN SIGMOID.

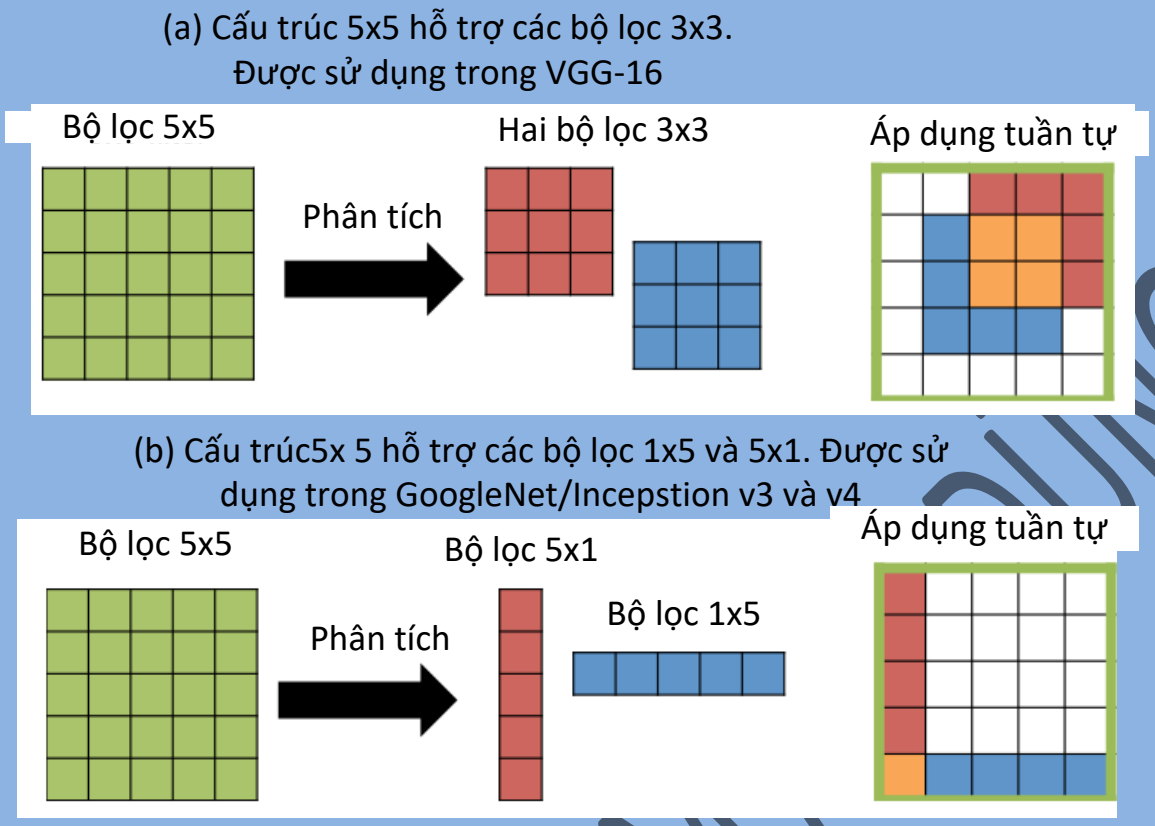
Phiên bản Lenet-5 này có 431K trọng số cho các bộ lọc và yêu cầu 2,3 triệu MAC (Multiply-and- Accumlate: nhân và tích lũy) cho mỗi hình ảnh và sử dụng ReLU thay vì Sigmoid.

LeNet là một trong những cách tiếp cận đầu tiên của CNN được giới thiệu vào năm 1989. Nó được thiết kế cho nhiệm vụ phân loại chữ số trong hình ảnh thang độ xám (Grayscale) có kích thước 28×28. Phiên bản nổi tiếng nhất, LeNet-5, chứa hai lớp CONV và hai lớp FC. Mỗi lớp CONV sử dụng các bộ lọc có kích thước 5×5 (1 kênh cho mỗi bộ lọc) với 6 bộ lọc ở lớp đầu tiên và 16 bộ lọc ở lớp thứ hai. Gộp trung bình 2×2 được sử dụng sau mỗi tích chập và một sigmoid được sử dụng cho tính phi tuyến tính. Tổng cộng, LeNet yêu cầu 60k trọng số và 341k nhân và tích lũy (MAC: Multiply-and-Accumlate) cho mỗi hình ảnh. LeNet đã dẫn đến thành công thương mại đầu tiên của CNN, vì nó được triển khai trong các máy ATM để nhận dạng các chữ số cho các khoản tiền gửi séc.

AlexNet [3] là CNN đầu tiên giành chiến thắng trong ImageNet Challenge vào năm 2012. Nó bao gồm năm lớp CONV và ba lớp FC. Trong mỗi lớp CONV, có 96 đến 384 bộ lọc và kích thước bộ lọc nằm trong khoảng từ 3×3 đến 11×11, với 3 đến 256 kênh mỗi kênh. Trong lớp đầu tiên, 3 kênh của bộ lọc tương ứng với các thành phần màu đỏ, xanh lá cây và xanh lam của hình ảnh đầu vào. Một phi tuyến tính ReLU được sử dụng trong mỗi lớp. Gộp tối đa 3×3 được áp dụng cho đầu ra của các lớp 1, 2 và 5. Để giảm tính toán, sải bước 4 được sử dụng ở lớp đầu tiên của mạng. AlexNet đã giới thiệu việc sử dụng LRN trong các lớp 1 và 2 trước khi gộp tối đa, mặc dù LRN không còn phổ biến trong các mô hình CNN sau này. Một yếu tố quan trọng phân biệt AlexNet với LeNet là số lượng trọng số lớn hơn nhiều và hình dạng khác nhau giữa các lớp. Để giảm số lượng trọng số và tính toán trong lớp CONV thứ hai, 96 kênh đầu ra của lớp đầu tiên được chia thành hai nhóm gồm 48 kênh đầu vào cho lớp thứ hai, sao cho các bộ lọc trong lớp thứ hai chỉ có 48 kênh. Tương tự, trọng số ở lớp thứ tư và thứ năm cũng được chia thành hai nhóm. Tổng cộng, AlexNet yêu cầu 61 triệu trọng số và 724 triệu MAC để xử lý một hình ảnh đầu vào 227×227.

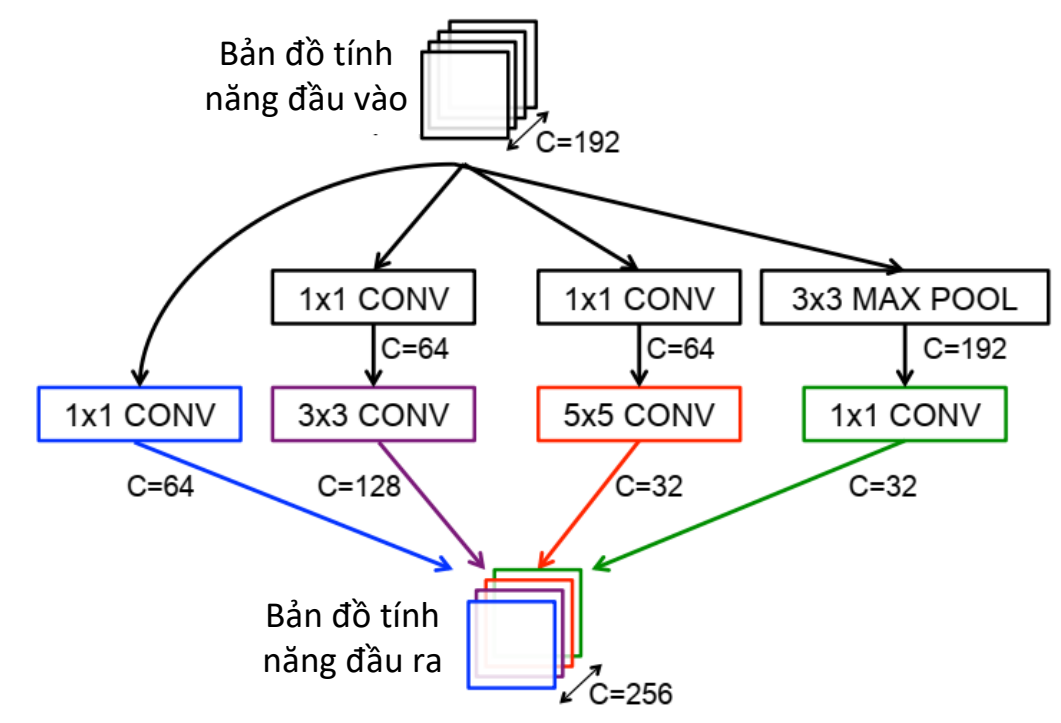
Overfeat có kiến trúc rất giống với AlexNet với năm lớp CONV và ba lớp FC. Sự khác biệt chính là số lượng bộ lọc được tăng lên cho các lớp 3 (384 đến 512), 4 (384 đến 1024) và 5 (256 đến 1024), lớp 2 không được chia thành hai nhóm, lớp đầu tiên được kết nối đầy đủ chỉ có 3072 kênh thay vì 4096 và kích thước đầu vào là 231×231 thay vì 227×227. Kết quả là, số lượng trọng số tăng lên 146 triệu và số lượng MAC tăng lên 2,8G cho mỗi hình ảnh. Overfeat có hai mô hình khác nhau: nhanh (được mô tả ở đây) và chính xác. Mô hình chính xác được sử dụng trong Thử thách ImageNet cho tỷ lệ lỗi top 5 thấp hơn 0,65% so với mô hình nhanh với chi phí 1,9× MAC nhiều hơn.

VGG-16 đi sâu hơn đến 16 lớp bao gồm 13 lớp CONV và 3 lớp FC. Để cân bằng chi phí đi sâu hơn, các bộ lọc lớn hơn (ví dụ: 5×5) được xây dựng từ nhiều bộ lọc nhỏ hơn (ví dụ: 3×3), có trọng số ít hơn, để đạt được các trường tiếp nhận (receptive fields) tương tự như trong Hình 5.1(a). Kết quả là, tất cả các lớp CONV có cùng kích thước bộ lọc là 3×3. Tổng cộng, VGG-16 yêu cầu 138M trọng số và 15,5G MAC để xử lý một hình ảnh đầu vào 224×224. VGG có hai mẫu khác nhau: VGG-16 (được mô tả ở đây) và VGG-19. VGG-19 cho tỷ lệ lỗi top 5 thấp hơn 0,1% so với VGG-16 với chi phí 1,27× MAC nhiều hơn.



Hình 5.1. Phân tích các bộ lọc lớn hơn thành các bộ lọc nhỏ hơn

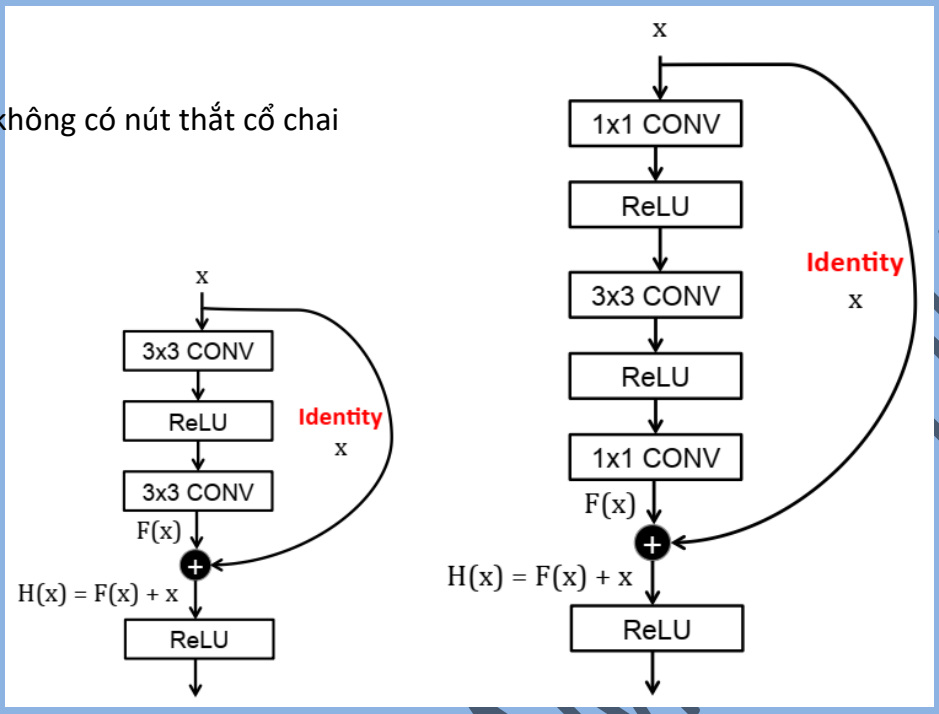
GoogLeNet thậm chí còn đi sâu hơn với 22 lớp. Nó giới thiệu một mô-đun khởi động (Inception Module), được hiển thị trên Hình 5.2, bao gồm các kết nối song song, trong khi trước đây chỉ có một kết nối nối tiếp duy nhất. Các bộ lọc có kích thước khác nhau (tức là 1×1, 3×3, 5×5), cùng với gộp tối đa 3×3, được sử dụng cho mỗi kết nối song song và các đầu ra của chúng được móc nối cho đầu ra mô-đun. Sử dụng nhiều kích thước bộ lọc có tác dụng xử lý đầu vào ở nhiều quy mô (scale). Để cải thiện tốc độ đào tạo, GoogLeNet được thiết kế sao cho trọng số và kích hoạt, được lưu trữ để lan truyền ngược trong quá trình đào tạo, tất cả đều có thể phù hợp với bộ nhớ GPU. Để giảm số lượng trọng số, bộ lọc 1×1 được áp dụng như một 'nút thắt cổ chai' để giảm số lượng kênh cho mỗi bộ lọc. 22 lớp bao gồm ba lớp CONV, tiếp theo là 9 lớp khởi đầu (Inception Layer), (mỗi lớp sâu hai lớp CONV) và một lớp FC. Kể từ khi được giới thiệu vào năm 2014, GoogleNet (còn được gọi là Inception) có nhiều phiên bản: v1 (mô tả ở đây), v3 và v4. Inception-v3 phân tách các tích chập bằng cách sử dụng các bộ lọc 1-D nhỏ hơn như trong Hình 5.1(b) để giảm số lượng MAC và trọng số để đi sâu hơn đến 42 lớp. Kết hợp với chuẩn hóa hàng loạt], v3 đạt được sai số top 5 thấp hơn 3% so với v1 với mức tăng 2,5× tính toán. Inception-v4 sử dụng các kết nối dư, được mô tả trong phần tiếp theo, để giảm 0,4% lỗi.



Hình 5.2. Mô-đun khởi đầu từ GoogleNet [51] với độ dài kênh ví dụ. Lưu ý rằng mỗi lớp CONV được theo sau bởi một ReLU (không vẽ)

ResNet, còn được gọi là Residual Net (Mạng dư), sử dụng các kết nối dư để đi sâu hơn nữa (34 lớp trở lên). Đây là DNN đầu tiên trong ImageNet Challenge vượt quá độ chính xác ở cấp độ con người với tỷ lệ lỗi top 5 dưới 5%. Một trong những thách thức với mạng sâu là gradient (độ dốc) biến mất trong quá trình đào tạo: khi lỗi lan truyền ngược qua mạng, gradient sẽ thu hẹp, điều này ảnh hưởng đến khả năng cập nhật trọng số trong các lớp trước đó cho các mạng rất sâu. Mạng dư giới thiệu một mô-đun 'phím tắt' hay kết nối bỏ qua (‘shortcut’ module) có chứa kết nối đồng dạng (Identity) sao cho các lớp trọng số (tức là các lớp CONV) có thể được bỏ qua như trong Hình 5.3. Thay vì học hàm F(x) cho các lớp trọng số, mô-đun phím tắt học ánh xạ dư (F(x) = H(x) - x). Ban đầu, F(x) bằng không và kết nối nhận dạng được thực hiện; sau đó dần dần trong quá trình tập luyện, kết nối thẳng (Forward) thực tế thông qua lớp trọng số được sử dụng. Điều này tương tự như các mạng LSTM được sử dụng cho dữ liệu tuần tự. ResNet cũng sử dụng cách tiếp cận 'nút thắt cổ chai' bằng cách sử dụng bộ lọc 1×1 để giảm số lượng thông số trọng số. Kết quả là, hai lớp trong mô-đun phím tắt được thay thế bằng ba lớp (1×1, 3×3, 1×1) trong đó 1×1 giảm và sau đó tăng (khôi phục) số lượng trọng số. ResNet-50 bao gồm một lớp CONV, tiếp theo là 16 lớp phím tắt (Shortcut Layer) (mỗi lớp sâu ba lớp CONV) và một lớp FC; nó yêu cầu 25,5 triệu trọng số và 3.9G MAC cho mỗi hình ảnh. Có nhiều phiên bản khác nhau của ResNet với nhiều độ sâu (ví dụ: không có nút cổ chai: 18, 34; với nút cổ chai: 50, 101, 152). ResNet với 152 lớp là người chiến thắng trong Thử thách ImageNet yêu cầu MAC 11.3G

và trọng lượng 60M. So với ResNet-50, nó giảm sai số top 5 khoảng 1% với chi phí là 2,9× MAC nhiều hơn và 2,5× trọng số hơn.



Hình 5.3. Mô-đun phím tắt từ ResNet. Lưu ý rằng ReLU theo sau lớp CONV cuối cùng trong đường tắt là sau khi cộng

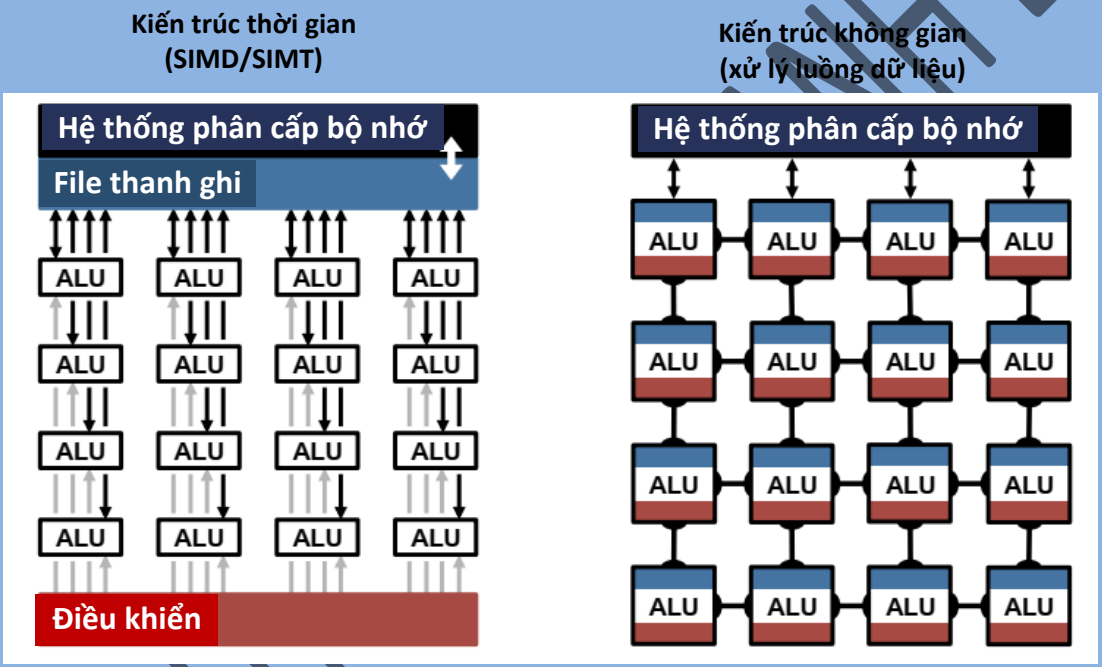
Một số xu hướng có thể được quan sát thấy trong các DNN phổ biến được biểu thị trong Bảng 5.2. Tăng độ sâu của mạng có xu hướng mang lại độ chính xác cao hơn. Bằng cách điều khiển số lượng trọng số, một mạng sâu hơn có thể hỗ trợ một phạm vi rộng hơn của các hàm phi tuyến tính tách biệt hơn và cũng cung cấp nhiều cấp độ phân cấp hơn trong biểu diễn đã học (Learned Representation). Số lượng hình dạng bộ lọc tiếp tục thay đổi giữa các lớp, do đó tính linh hoạt vẫn rất quan trọng. Hơn nữa, hầu hết các tính toán đã được đặt trên các lớp CONV thay vì các lớp FC. Ngoài ra, số lượng trọng số trong các lớp FC được giảm và trong hầu hết các mạng gần đây (kể từ GoogLeNet), các lớp CONV cũng chiếm ưu thế về trọng số. Do đó, trọng tâm của việc triển khai phần cứng nên là giải quyết hiệu quả của các lớp CONV, điều này trong nhiều lĩnh vực ngày càng quan trọng.

6. PHẦN CỨNG ĐỂ XỬ LÝ DNN

Do sự phổ biến của DNN, nhiều nền tảng phần cứng gần đây có các tính năng đặc biệt nhắm mục tiêu vào xử lý DNN. Chẳng hạn, CPU Intel Knights Landing có các hướng dẫn vectơ đặc

biệt để học sâu; GPU Nvidia PASCAL GP100 có hỗ trợ số học dấu phẩy động 16 bit (FP16: 16-bit floating point) để thực hiện hai phép toán FP16 trên một lõi chính xác duy nhất để tính toán học sâu nhanh hơn. Các hệ thống cũng đã được xây dựng đặc biệt cho xử lý DNN như Nvidia DGX-1 và máy chủ DNN tùy chỉnh Big Basin của Facebook. Suy luận DNN cũng đã được chứng minh trên các System-onChips (SoC) nhúng khác nhau như Nvidia Tegra và Samsung Exynos cũng như FPGA. Theo đó, điều quan trọng là phải hiểu rõ về cách xử lý đang được thực hiện trên các nền tảng này và cách thiết kế các bộ tăng tốc dành riêng cho DNN để cải thiện hơn nữa thông lượng và hiệu quả năng lượng.

Thành phần cơ bản của cả hai lớp CONV và FC là các phép toán nhân và tích lũy (MAC), có thực hiễn thể dễ dàng song song. Để đạt được hiệu suất cao, các mô hình tính toán song song cao được sử dụng rất phổ biến, bao gồm cả kiến trúc thời gian và không gian như trong Hình 6.1.



Hình 6.1. Các mô hình tính toán song song cao

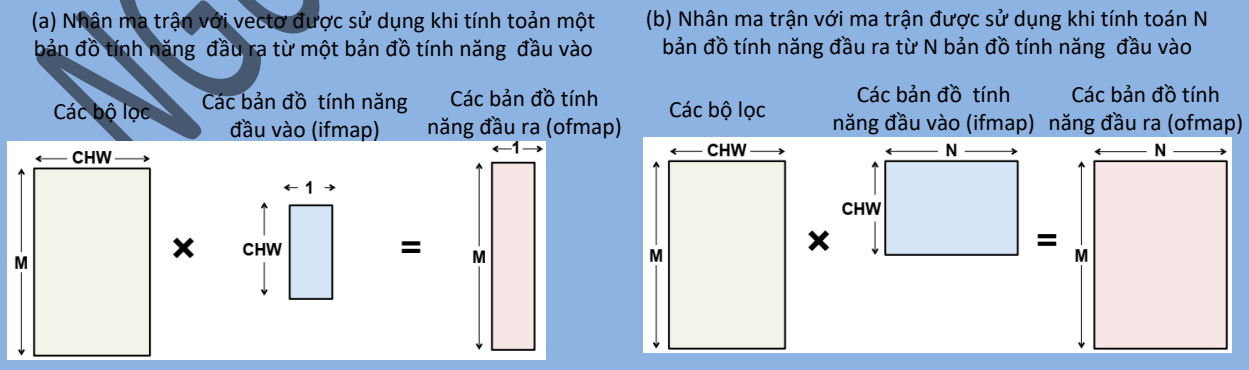
Các kiến trúc thời gian xuất hiện chủ yếu trong CPU hoặc GPU và sử dụng nhiều kỹ thuật khác nhau để cải thiện tính song song như vectơ (SIMD: Single Introduction, Multiple Data) hoặc luồng song song (SIMT: Single Introduction, Multiple Threads). Kiến trúc thời gian như vậy sử dụng điều khiển tập trung cho một số lượng lớn ALU (Arithmetric Logic Unit: Đơn vị logic số học). Các ALU này chỉ có thể tìm nạp dữ liệu từ hệ thống phân cấp bộ nhớ và không thể giao tiếp trực tiếp với nhau. Ngược lại, kiến trúc không gian sử dụng xử lý luồng dữ liệu, tức là ALU tạo thành một chuỗi xử lý để chúng có thể truyền dữ liệu trực tiếp từ người này sang người khác. Đôi khi mỗi ALU có thể có logic điều khiển và bộ nhớ cục bộ riêng, được gọi là scratchpad

hoặc file thanh ghi. ALU với bộ nhớ cục bộ riêng của nó được gọi là một công cụ xử lý (PE: processing engine). Kiến trúc không gian thường được sử dụng cho DNN trong các thiết kế dựa trên ASIC (Application Specific Integrated Circuit: Mạch tích hợp ứng dụng cụ thể) và FPGA (Field-Programmable Gate Array: Mảng cổng có thể lập trình tại hiện trường) . Trong phần này, chúng ta sẽ thảo luận về các chiến lược thiết kế khác nhau để xử lý hiệu quả trên các nền tảng khác nhau này mà không ảnh hưởng đến độ chính xác (tức là tất cả các cách tiếp cận trong phần này tạo ra kết quả giống hệt nhau theo bit); cụ thể là

- Đối với các kiến trúc thời gian như CPU và GPU, chúng ta sẽ thảo luận về cách các biến đổi tính toán trên hạt nhân (Kenel) có thể giảm số lần nhân để tăng thông lượng;
- Đối với các kiến trúc không gian được sử dụng trong bộ tăng tốc(Accelerator), chúng ta sẽ thảo luận về cách các luồng dữ liệu có thể tăng khả năng tái sử dụng dữ liệu từ các bộ nhớ chi phí thấp trong hệ thống phân cấp bộ nhớ để giảm tiêu thụ năng lượng.

Tăng tốc tính toán hạt nhân trên các nền tảng CPU và GPU

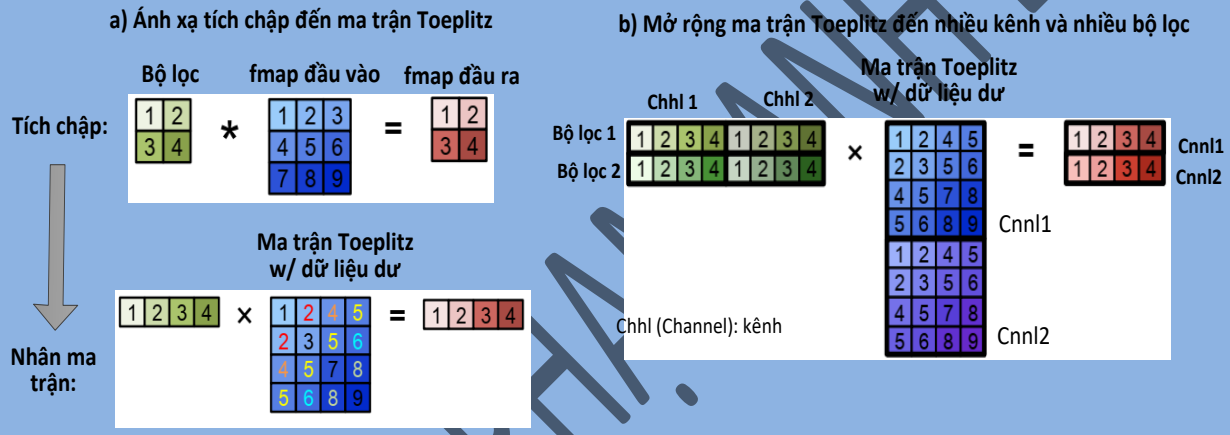
CPU và GPU sử dụng các kỹ thuật song song như SIMD (Single Introduction, Multiple Data) hoặc SIMT (Single Isocenter for Multiple Target) để thực hiện MAC song song. Tất cả các ALU chia sẻ cùng một điều khiển và bộ nhớ (tệp thanh ghi). Trên các nền tảng này, cả hai lớp FC và CONV đều được ánh xạ đến phép nhân ma trận (tức là tính toán hạt nhân). Hình 17 cho thấy cách phép nhân ma trận được sử dụng cho lớp FC. Chiều cao của ma trận bộ lọc là số lượng bộ lọc và chiều rộng là số trọng số trên mỗi bộ lọc (kênh đầu vào (C) × chiều rộng (W) × chiều cao (H), vì R = W và S = H trong lớp FC); chiều cao của ma trận bản đồ tính năng đầu vào là số lần kích hoạt trên mỗi bản đồ tính năng đầu vào (C × W × H) và chiều rộng là số lượng bản đồ tính năng đầu vào (một trong Hình 6.2 (a) và N trong Hình 6.6 (b)); cuối cùng, chiều cao của ma trận bản đồ tính năng đầu ra là số kênh trong bản đồ tính năng đầu ra (M) và chiều rộng là số lượng bản đồ tính năng đầu ra (N), trong đó mỗi bản đồ tính năng đầu ra của lớp FC có kích thước là 1×1×số kênh đầu ra (M).



Hình 6.2. Ánh xạ đến nhân ma trận cho các lớp kết nối hoàn toàn

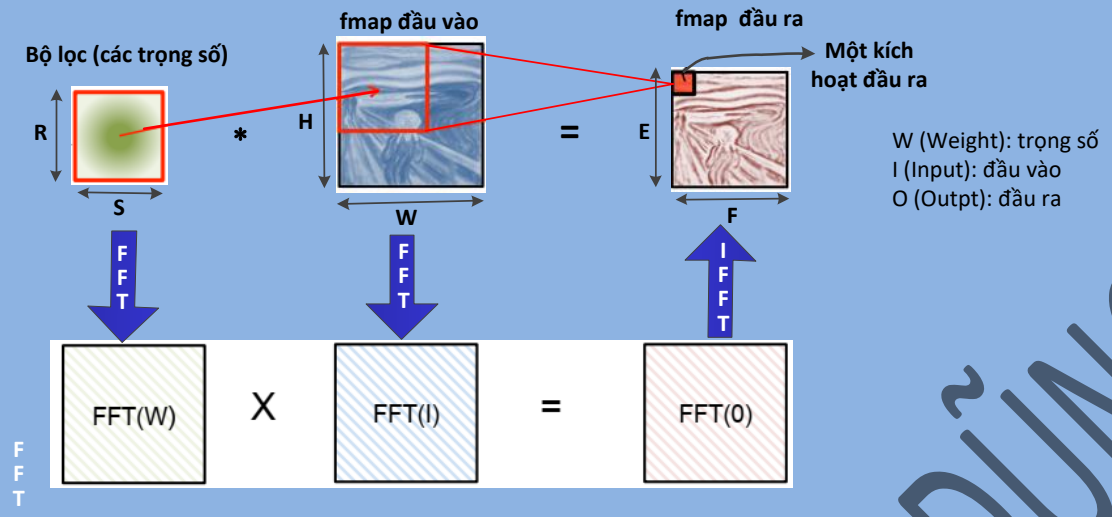
Lớp CONV trong DNN cũng có thể được ánh xạ đến phép nhân ma trận bằng cách sử dụng dạng thả long của ma trận Toeplitz như trong Hình 6.3. Nhược điểm của việc sử dụng phép nhân ma trận cho các lớp CONV là có dữ liệu dư thừa trong ma trận bản đồ tính năng đầu vào như được đánh dấu trong Hình 6.3(a). Điều này có thể dẫn đến sự kém hiệu quả trong lưu trữ hoặc mô hình truy cập bộ nhớ phức tạp.

Có các thư viện phần mềm được thiết kế cho CPU (ví dụ: OpenBLAS, Intel MKL, v.v.) và GPU (ví dụ: cuBLAS, cuDNN, v.v.) tối ưu hóa cho phép nhân ma trận. Phép nhân ma trận được xếp vào hệ thống phân cấp lưu trữ của các nền tảng này, theo thứ tự vài megabyte ở các cấp độ cao hơn. Các phép nhân ma trận trên các nền tảng này có thể được tăng tốc hơn nữa bằng cách áp dụng các phép biến đổi tính toán cho dữ liệu để giảm số lần nhân, trong khi vẫn cho cùng một kết quả theo từng bit. Thông thường, điều này có thể dẫn đến chi phí tăng số lượng bổ sung và mô hình truy cập dữ liệu không đều đặn hơn.



Hình 6.3. Ánh xạ nhân ma trận cho các lớp tích chập

Biến đổi Fourier nhanh (FFT) là một cách tiếp cận nổi tiếng, được thể hiện trong Hình 6.4 giúp giảm số lần nhân từ O ($N_o^2N_f^2$) xuống O ($N_o^2\log_2N_o$), trong đó kích thước đầu ra là $N_o \times N_o$ và kích thước bộ lọc là $N_f \times N_f$. Để thực hiện tích chập, ta lấy FFT của bộ lọc và bản đồ tính năng đầu vào, sau đó thực hiện phép nhân trong miền tần số; sau đó, ta áp dụng FFT nghịch đảo cho sản phẩm kết quả để khôi phục bản đồ tính năng đầu ra trong miền không gian. Tuy nhiên, có một số hạn chế khi sử dụng FFT: (1) lợi ích của FFT giảm theo kích thước bộ lọc; (2) kích thước của FFT được quyết định bởi kích thước bản đồ tính năng đầu ra thường lớn hơn nhiều so với bộ lọc; (3) các hệ số trong miền tần số rất phức tạp. Do đó, mặc dù FFT làm giảm tính toán, nhưng nó đòi hỏi dung lượng lưu trữ và băng thông lớn hơn.



Hình 6.4. FFT để tăng tốc DNN

Một số tối ưu hóa có thể được thực hiện trên FFT để làm cho nó hiệu quả hơn cho DNN. Để giảm số lượng hoạt động, FFT của bộ lọc có thể được tính toán trước và lưu trữ. Ngoài ra, FFT của bản đồ tính năng đầu vào có thể được tính toán một lần và được sử dụng để tạo nhiều kênh trong bản đồ tính năng đầu ra. Cuối cùng, vì một hình ảnh chỉ chứa các giá trị thực, nên biến đổi Fourier của nó là đối xứng và điều này có thể được khai thác để giảm chi phí lưu trữ và tính toán.

7. TỔNG QUAN NGHIÊN CỨU ỨNG DỤNG AI/ML VÀO HỆ THỐNG 5G/6G

Dự án Đối tác Thế hệ thứ 3 (3GPP) đã kết thúc thành công giai đoạn đầu của sự tiến bộ thế hệ thứ năm (5G) thông qua Bản phát hành 15–17. Bây giờ, nó bắt tay vào giai đoạn tiếp theo của quá trình tiến hóa 5G, được gọi là 5GAdvanced. Là bản phát hành đầu tiên của 5G-Advanced, Bản phát hành 18 bao gồm các dự án toàn diện không chỉ phục vụ nhu cầu thương mại trước mắt mà còn bao gồm những nỗ lực dài hạn đặt nền móng cho sự phát triển của truy cập không dây sang lĩnh vực của thế hệ thứ sáu (6G). Một trọng tâm đáng chú ý trong lĩnh vực này là tích hợp trí tuệ nhân tạo (AI)/máy học (ML) vào kết cấu của sự phát triển 5G-Advanced, được thiết lập để tạo điều kiện thuận lợi cho việc áp dụng rộng rãi AI/ML trong các hệ thống truyền thông không dây.

Trước sự ra đời của sự phát triển 5G-Advanced, 3GPP đã tham gia vào các sáng kiến AI / ML sơ bộ trong giai đoạn đầu tiên của quá trình phát triển 5G, trải dài trên nhiều lĩnh vực từ mạng lõi 5G (5GC), vận hành, quản trị và bảo trì (OAM) và mạng truy cập vô tuyến (RAN). Giao diện vô tuyến là một thành phần cơ bản trong bất kỳ hệ thống truyền thông không dây nào. Các ấn phẩm gần đây đã vẽ ra tầm nhìn về một giao diện không khí gốc AI (AI-native air interface). Tuy nhiên, điều đáng chú ý là những nỗ lực AI / ML sơ bộ do 3GPP thực hiện trước khi phát triển 5G-Advanced không bao gồm giao diện vô tuyến mới 5G (5G NR: 5G New Radio). 3GPP

giải quyết lỗ hổng này trong Bản phát hành 18 bằng cách khám phá tiềm năng của việc sử dụng các thuật toán dựa trên AI / ML để tăng cường giao diện vô tuyến NR [7]. Việc theo đuổi này đánh dấu một bước tiến đột phá vì đây là bước đầu tiên thuộc loại này trong sự phát triển của 3GPP về các tiêu chuẩn truyền thông không dây.

Kết hợp trí tuệ nhân tạo (AI: Artificial Intelligence) với các mạng truyền thông di động không dây 5G và 6G là đã trở thành một trong các xu hướng công nghệ hướng đến năm 2030.

Từ sau 2010, chúng ta đã chứng kiến làn sóng lớn trí tuệ nhân tạo AI, AI đã vượt qua khả năng của con người tromg nhận dạng hình ảnh và chơi các trò chơi phức tạp. Mặt khác khác khi truyền thông không dây 4G trở nên hiện thực, các công nghệ này bắt đầu hội nhập và tăng cường lẫn nhau. Một mặt các mạng không dây có thể hỗ trợ triển khai rộng rãi AI để đảm bảo kết nối không dây trong quá trình huấn luyện mô hình và suy luận mô hình cho các ứng dụng khác nhau như lái xe tự quản, nhận dạng hình ảnh v.v. Mặt khác AI được áp dụng để giải quyết vấn đề trong mạng không dây như quản lý búp sóng (BM: Beam Management) , ước tính nhiễu trong lớp vật lý và định tuyến trong lớp mạng cũng như ấn định tài nguyên đối với các nhà mạng di động.

Vào năm 2019, nhóm đặc tả kỹ thuật của 3GPP đã bắt đầu nghiên cứu ảnh hưởng của chuyển giao trí tuệ nhân tạo và mô hình máy học (AMMT: Artificial Intelligence and Machine Learning Model Trasfer) trong các hệ thống không dây. Các kỹ thuật AI đã được phát triển trong lớp ứng dụng chẳng hạn nhận dạng đối tượng dựa trên hình ảnh. 3GPP tiếp tục nghiên cứu các chức năng mới cần thiết cho: 1) hỗ trợ lưu lượng mới do sự phân bố các mô hình AI như phân bố theo thời gian các mô hình AI cho các nhiệm vụ nhận dạng hình ảnh khác nhau và cho các môi trường khác nhau, 2) để hỗ trợ phân chia mô hình AI để phát triển các lớp khác nhau của một mô hình AI, 3) để hỗ trợ thực hiện hiệu quả học liên kết (FL: Federated Learning), chẳng hạn sử dụng mạng lõi 5G (5GC) để cung cấp danh sách UE với khoảng cách giữa chúng lớn để hỗ trợ lựa chọn người sử dùng và giảm tương quan số liệu hướng dẫn. Hỗ trợ này được thực hiện bằng cách nâng cao các chức năng hiện có như quản lý chất lượng dịch vụ (QoS Management), để lộ hay để lộ/ lộ sáng (exposure) thông tin mạng cho các ứng dụng AI và ấn định tài nguyên cho các người sử dụng dựa trên một số tiêu chí chẳng hạn vị trí, trong các ứng dụng FL. Hỗ trợ mạnh hơn cho chọn lựa người dùng trong FL có thể được thực hiện bằng cách thay đổi kích thước cửa sổ lùi (backoff window size) trong cơ chế đa truy nhập cảm nhận sóng mang (CSMA: Carrier-Sense Multiple Access), chẳng hạn giảm cửa sổ lùi (Backoff Window) cho các người dùng được chọn để cung cấp cho họ có cơ hội cao hơn nhận được quyền truy nhập vào các mô hình AI được tải lên một server.

Mặt khác, các công nghệ "AI cho không dây" (AI for Wireless) được đề cập như là sử dụng AI để nâng cao các chức năng của mạng. Nhiều nghiên cứu khác nhau đã được tiến hành để áp dụng các công nghệ AI cho các lớp khác nhau của các giao thức. Công nghiệp bán dẫn cũng bắt

đầu nghiên cứu các chip truyền thông có các khả năng AI. Năm 2020, nhóm nghiên các đặc tả kỹ thuật mạng truy nhập vô tuyến (Radio Access Group Technical Specification Group) đã bắt đầu nghiên cứu ứng dụng AI trong các chức năng của RAN (Radio Access Network: mạng truy nhập vô tuyến). TSG (Technical Specification Group) nghiên cứu ba khía cạnh về đề tài này:

1) Bằng cách sử dụng các mô hình AI được triển khai cả ở phía trạm gốc (BS) và phía UE, thông tin trạng thái kênh (CSI: Channel State Information) có thể được nén, được phát và được dự đoán. Điều này hỗ trợ giảm chi phí truyền thông phản hồi CSI so với ước tính kênh và thủ tục báo cáo thông thường;
2) Bằng cách triển khai AI tại phía UE, phía BS hoặc phía mạng lõi đối với định vị, hiệu năng có thể được nâng cao trong các kịch bản không có tầm nhìn thẳng (NLoS);
3) Băng cách áp dụng AI trong thủ tục quản lý búp sóng (BM: Beam Management), mức độ phức tạp của hệ thống có thể được giảm so với thủ tục BM truyền thống.

Phần 2

ỨNG DỤNG AI/ML CHO CÁC DỊCH VỤ TRONG MẠNG TRUYỀN THÔNG DI ĐỘNG 5G VÀ 6G

1. TỔNG QUAN

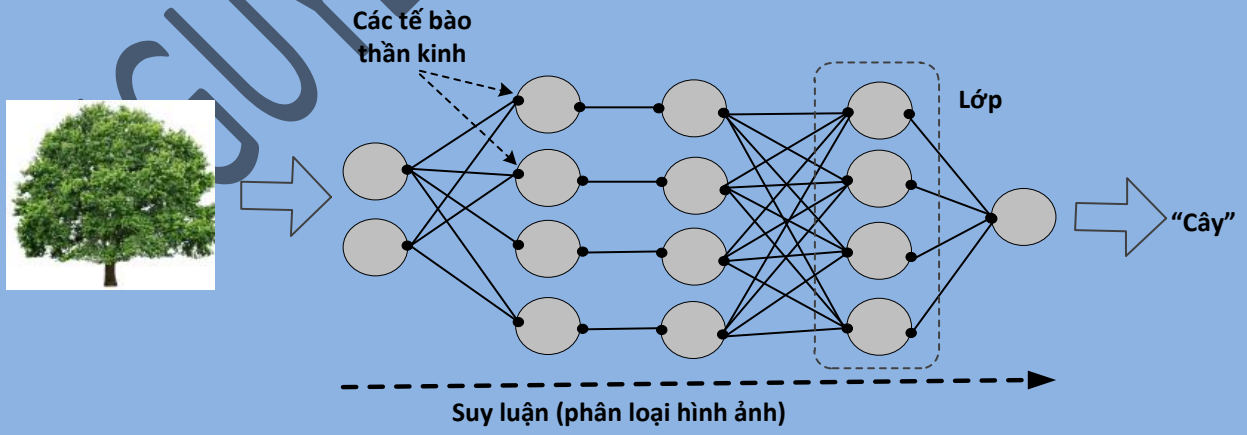
Trí tuệ nhân tạo (AI: Artificial Intellegence)/ học máy (ML: Machine Learng) đang được sử dụng trong nhiều lĩnh vức ứng dụng trong các bộ cảm biến công nghiệp. Trong các hệ thông truyền thông di động, các thiết bị di động (điện thoại thông minh, xe cô, robot) đang thay thế rất nhanh các giải thuật thông thường (nhận dạng tiêng nói, nhận dạng hình ảnh, xử lý video) bằng

các mô hình AM/ML để hỗ trợ các ứng dụng. Trong phần này ta sẽ khảo sát các ứng dụng của AI/ML cho các dịch vụ trong mạng truyền thông di động 5G và 6G.

1.1. Mạng nơron sâu (DNN: Deep Neural Network)

1.1.1. Tổng quan

Mạng nơron sâu (DNN: Deep Neural Network) là kỹ thuật lõi của học máy (ML). Các ứng dụng xử lý thị giác máy tính, giọng nói và ngôn ngữ tự nhiên tận dụng DNN làm thuật toán học máy cốt lõi của chúng. DNN được tổ chức trong một đồ thị có hướng (Directed Graph) trong đó mỗi nút là một phần tử xử lý (một tế bào thần kinh) áp dụng một hàm cho đầu vào của nó và tạo ra một đầu ra. Hình 1.1 mô tả mạng DNN 5 lớp để phân loại hình ảnh trong đó tính toán được thực hiện từ trái sang phải. Các rìa cạnh của biểu đồ là các kết nối giữa mỗi tế bào thần kinh xác định luồng dữ liệu. Nhiều tế bào thần kinh áp dụng cùng một chức năng cho các phần khác nhau của đầu vào xác định một lớp. Đối với một đường truyền thẳng (Feedforward) qua một DNN, đầu ra của một lớp là đầu vào cho lớp tiếp theo. Độ sâu của DNN được xác định bởi số lớp. Các ứng dụng Thị giác Máy tính (CV: Computer Vision) sử dụng DNN để trích xuất các tính năng từ hình ảnh đầu vào và phân loại hình ảnh thành một trong các loại được xác định trước. Các ứng dụng Nhận dạng Giọng nói Tự động (ASR: Automatic Speech Recognition) sử dụng DNN để tạo dự đoán cho các vectơ tính năng giọng nói, sau đó sẽ được xử lý hậu kỳ để tạo ra bản ghi văn bản có khả năng cao nhất. Các ứng dụng Xử lý ngôn ngữ tự nhiên (NLP: Natural Language Processing) sử dụng DNN để phân tích và trích xuất thông tin ngữ nghĩa và cú pháp từ các vectơ nhúng từ (Word Embedding-Vector) được tạo từ văn bản đầu vào.



Hình 1.1. Mạng thần kinh (DNN) 5 lớp phân loại ảnh đầu vào vào một trong các loại được định nghĩa

Tính toán DNN như sau có thể được xử lý:

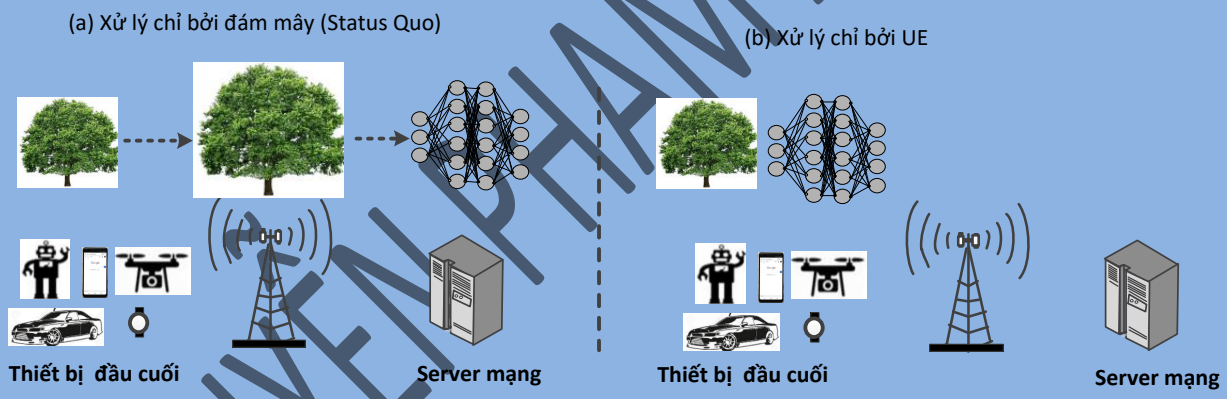
1.12. Xử lý DNN chỉ bởi đám mây: giữ nguyên hiện trạng (Status Quo)

Trong cách tiếp cận này, các yêu cầu xử lý DNN được gửi tử thiết bị đầu cuối đến đám may như trên hình 1.2(a). Tuy nhiên với các tiếp cận này một khối lượng dữ liệu lớn (hình ảnh, video và audio) được tải lên đến server qua mạng tô tuyến dẫn đến trễ và tiêu thụ năng lượng cao.

Hiện tại, cách tiếp cận giữ nguyên hiện trạng được các nhà cung cấp đám mây sử dụng cho các ứng dụng thông minh là thực hiện tất cả các xử lý DNN trên đám mây. Nên chi phí lớn của cách tiếp cận này là gửi dữ liệu qua mạng không dây.

1.1.3. Xử lý DNN chỉ bởi UE

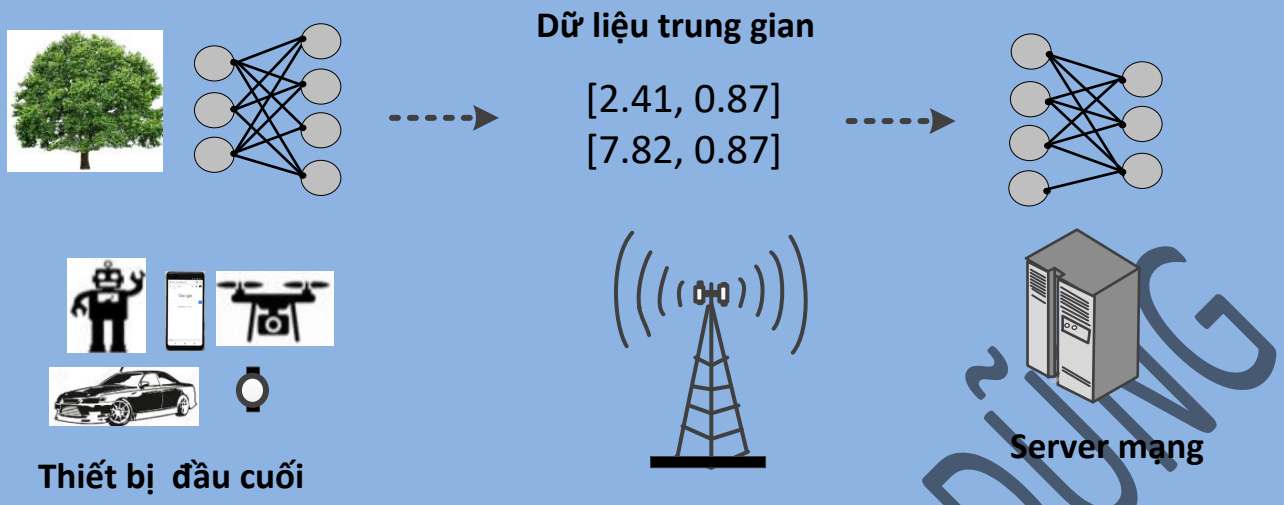
Trong cách tiếp cận này xử lý DNN được thực hiện ngay tại thiết bị đầu cuối như trên hình 1.2(b). Trong khi đó các thiết bị đầu cuối thường chỉ có khả năng tiêu thụ công suất giới hạn, khả năng tính toán và bộ nhớ giới hạn để thực hiện tính toán tại chỗ hoàn toàn ngoại tuyến.



Hình 1.2. (a) xử lý chỉ bởi đám mây;, (b) xử lý chỉ bởi UE

1.1.4.Phân chia tính toán giữa thiết bị đầu cuối và đam mây hay Neurosergion.

Trong cách tiếp cận Neurosergion tính toán được phân chia giữa thiết bị đầu cuối và đám mây (hình 1.3). Một phần tính toánn phù hợp với công suất, bộ nhớ và khả năng tính toán của thiết bị đầu cuối được thực hiện tại chỗ tại thiết bị đầu cuối. Dữ liệu trung gian được rút gọn sẽ được gửi đến đám mây, tại đây phần tính toán còn lại được thực hiện.



Hình 1.3. Phân chia tính toán giữa UE và đam mây hay Neurosergion

1.2. Các kiểu lớp trong DNN

Tồn tại các kiểu lớp (Layer Type) khác nhau trong DNN. Dưới đây ta sẽ liệt kê các kểu lớp này.

Lớp kết nối hoàn toàn (fc: Full-Connected Layer) - Tất cả các tế bào thần kinh trong một lớp được kết nối hoàn toàn với tất cả các tế bào thần kinh trong lớp trước. Lớp tính toán tổng trọng số của các đầu vào bằng cách sử dụng một tập hợp các trọng số đã học.

Tích chập và các lớp cục bộ (conv, local) – Tích chập và các lớp cục bộ (Convolution & Local Layer) tích chập hình ảnh với một tập hợp các bộ lọc được học để tao ra một tập các bản đồ tính năng (Feature Maps). Các lớp này chủ yếu khác nhau về kích thước bản đồ tính năng đầu vào, số lượng và kích thước các bộ lọc của chúng và sải chân hay sự dịch chuyển (Stride) mà các bộ lọc đang được áp dụng.

Lớp gộp (pool) - Các lớp gộp (Pooling Layer) áp dụng một hàm được xác định trước (ví dụ: tối đa hoặc trung bình) trên các vùng của bản đồ tính năng đầu vào để nhóm các tính năng lại với nhau. Các lớp này chủ yếu khác nhau về kích thước đầu vào của chúng, kích thước của vùng gộp và sải chân mà gộ được áp dụng.

Lớp kích hoạt (Activation Layer) - Các lớp kích hoạt áp dụng một hàm phi tuyến tính cho từng dữ liệu đầu vào của nó một cách riêng lẻ, tạo ra cùng một lượng dữ liệu như đầu ra. Các lớp kích hoạt có trong mạng nơ-ron được có thể bao gồm lớp sigmoid (sig), lớp tuyến tính chỉnh lưu (relu) và lớp Tanh cứng (htanh) v.v.

Lớp chuẩn hóa (norm) – Chuẩn hóa các tính năng trên các bản đồ tính năng được nhóm lại theo không gian.

Lớp softmax (softmax) - Tạo ra một phân bố xác suất trên một số loại có thể có đối với phân loại.

Lớp argmax (argmax) – Chọn loại có xác suất cao nhất.

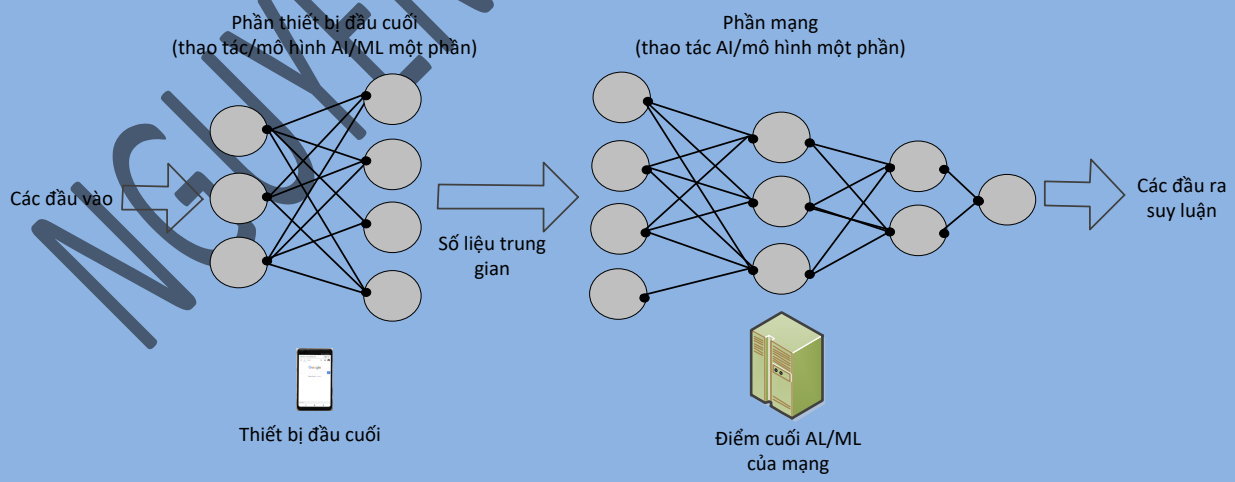
Dropout Layer (Dropout) - ngẫu nhiên bỏ qua các tế bào thần kinh trong quá trình đào tạo để tránh quá khớp mô hình và được chuyển qua trong quá trình dự đoán

1.3. Tổng quan AI/ML trong 5G và 6G

Hệ thống 5G và 6G có thể hỗ trợ ba kiểu hoạt động AI/ML:

- Phân chia hoạt động AI/ML giữa các điểm cuối;
- Phân bố mô hình/dữ liệu AI/ML và chia sẻ trên toàn bộ hệ thống 5G, 6G
- Học phân tán/học liên kết (Distributed/Federated Learning) trên hệ thống 5G, 6G.

Sơ đồ suy luận AI/ML được phân chia (Split AM/ML Inference) được mô tả trên hình 1.4. Hoạt động của AI/ML/mô hình được phân chia vào nhiều bộ phận tùy theo nhiệm vụ hiện thời của môi trường. Mục đích là giảm tải (Offloading) các phần đòi hỏi tính toán chuyên sâu và nhiều năng lượng vào các điểm cuối mạng trong khi để lại các phần nhạy cảm trễ và sự riêng tư tại các thiết bị đầu cuối. Thiết bị thực hiện thao tác/mô hình đến các bộ phận/lớp đặc thù sau đó gửi số liệu trung gian đến điểm cuối mạng. Điểm cuối mạng thực hiện các phần còn lại và cung cấp các kết quả suy luận ngược về thiết bị này.



Hình 1.4. Thí dụ về suy luận

Sơ đồ phân bố mô hình AI/ML được mô tả trên hình Hình 3.1 được xét trong phần 3. Các đầu cuối di động đa chức năng có thể cần chuyển mạch mô hình AI/ML để đáp ứng các thay đổi nhiệm vụ và môi trường. Điều kiện chọn mô hình đáp ứng là các mô hình này cần được chọn lựa phải khả dụng cho thiết bị di động. Tuy nhiên, do các mô hình AI/ML ngày càng đa dạng và tài nguyên được lưu giữ có hạn trong UE, nên không thể nạp trước tất cả các mô hình AI/ML ứng cử vào thiết bị. Phân bố mô hình trực tuyến (tải xuống mô hình mới) là cần thiết, trong đó một mô hình AI/ML có thể được phân bố từ một điểm cuối của mạng đến các thiết bị khi chúng cần thích ứng đến các nhiệm vụ và các môi trường bị thay đổi. Cho mục đích này, hiệu năng mô hình tai UE cần được thường xuyên giám sát.

Sơ đồ học liên kết (FL: Federated Learning) được mô tả trên Hình 4.1. trong phần 4. Server đám mây đào tạo một mô hình toàn cầu bằng cách kết hợp các mô hình địa phương được đào tạo một phần. Trong mỗi lần lặp đào tạo, UE thực hiện đào tạo dựa trên mô hình được tải xuống. Sau đó UE báo cáo các kết quả đào tạp tạm thời đến server đám mây thông qua các kênh đường lên của hệ thống 5G hoặc 6G. Server tổng hợp các kết quả đào tạo tạm thời từ các UE và cập nhật mô hình toàn cầu. Sau đó mô hình toàn cầu đã cập nhật được phân bố trở lại các UE và các UE thực hiện đào tạo cho lần lặp tiếp theo.

2. NGUYÊN LÝ HOẠT ĐỘNG AM/ML ĐƯỢC PHÂN CHIA GIỮA CÁC ĐIỂM CUỐI AI/ML (PRICIPE OF SPLIT AI/ML OPERATION BETWEEN AI/ML ENDPOINS)

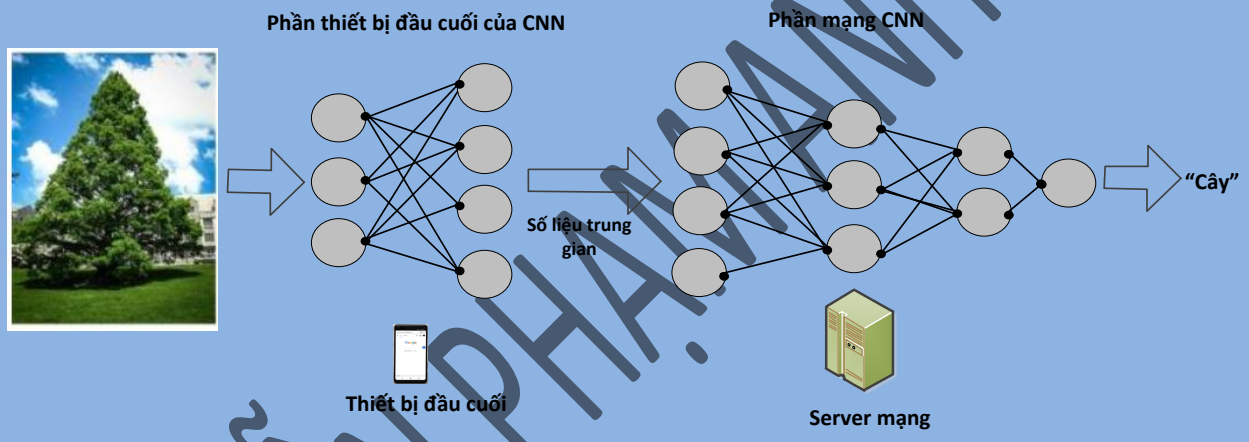
2.1. Tổng quan

Các ứng dụng AI/ML ngày càng đòi hỏi tính toán chuyên sâu hơn, bộ nhớ và thiêu thụ công suất lớn hơn. Trong khi đó các thiết bị đầu cuối thường chỉ có khả năng tiêu thụ công suất giới hạn, khả năng tính toán và bộ nhớ giới hạn để thực hiện suy luận AI/ML tại chỗ hoàn toàn ngoại tuyến . Rất nhiều các ứng dụng AI/ML, chẳng hạn nhận dạng hình ảnh, có xu thể giảm tải (Offloading) xử lý suy luận từ thiết bị di động đến trung tâm dữ liệu internet (IDC: Internet Data Center). Chẳng hạn các ảnh được chụp bởi điện thoại thông minh thường được xử lý tại đám mây AI/ML thước khi được hiện thị cho người dùng chụp chúng. Tuy nhiên các nhiệm vụ suy lận AI/ML dựa trên đám mây cũng cần phải xem xét đến áp lực tính toán tại các IDC, tốc độ số liệu/trễ yêu cầu và yêu cầu bảo vệ tính riêng tư.

Ảnh và video là các nguồn dữ liệu lớn nhất hiện nay trên internet.Các video chiếm trên 70% lưu lượng hàng ngày của internet. Các mô hình mạng thần kinh tích chập (CNN: Convolutional Neural Network) đang được sử dụng rộng rãi cho các nhiệm vụ nhận dạng hình ảnh/video trên

các thiết bị di động, chẳng hạn phân loại hình ảnh, phân đoạn hình ảnh, phát hiện và định vị đối tượng, xác thực mặt, nhận dạng hành động, tăng cường hình ảnh (Enhanced Photography) , VR/AR (Virtual Reality/ Augremented Reality: thực tế ảo/ thực tế tăng cường), các trò chơi video. Trong khi đó suy luận mô hình CNN đòi hỏi khả năng tính toán và nhớ lớn.

Rất nhiều nghiên cứu cho thấy rằng suy luận AI/ML cho xử lý hình ảnh với kết hợp thiết bị-mạng giảm áp lực tính toán, dấu chân bộ nhớ (Memory Foootprint), lưu giữ, công suất và tốc độ số liệu được yêu cầu trên thiết bị, giảm trễ đầu cuối đầu cuối và tiêu thụ năng lượng, cải thiện độ chính xác, hiệu năng và tính riêng tư đầu cuối đầu cuối đầu cuối so với các tiếp cận tại chỗ trên mỗi phía. Sơ đồ phân chia nhận dạng hình ảnh AI/ML được mô tả trên hình 2.1. CNN được phân chia thành hai phần dựa trên nhiệm vụ nhận dạng hình ảnh và môi trường hiện thời. Mục đích phân chia là chuyển tải các phần đòi hỏi tính toán và năng lượng cao đến server mạng. Serrver mạng thực hiện các lớp CNN còn lại. Trong khi mô hình được phát triển và được yêu cầu, hoạt động phân chia AI/ML vẫn dựa trên mô hình cũ.



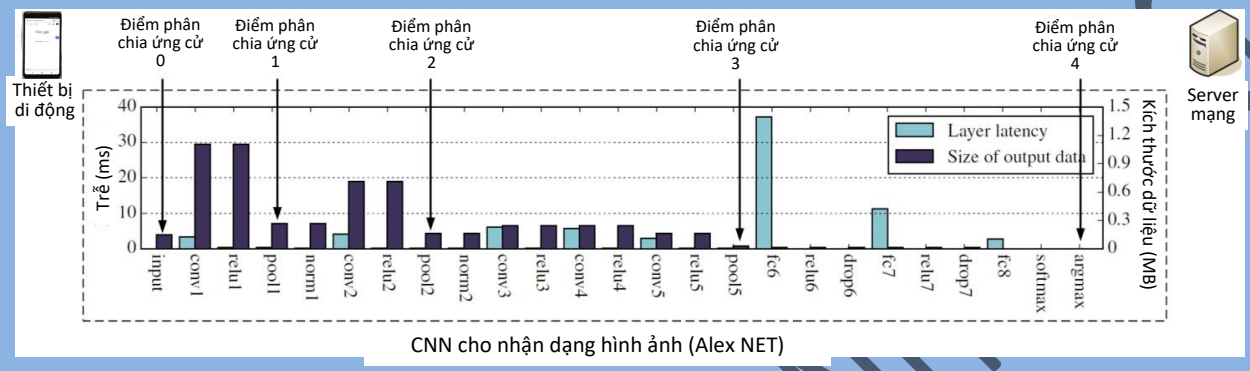
Hình 2.1. Thí dụ phân chia nhận dạng ảnh AI/ML

Do đặc điểm của một số thuật toán trong giai đoạn đào tạo mô hình, mỗi mô hình có một mức độ mạnh nhất định. Nên, nếu có lỗi trong quá trình truyền dữ liệu trung gian, mô hình có dung sai nhất định và vẫn có thể đảm bảo độ chính xác của kết quả suy luận. Vì kết quả suy luận cần được chuyển tiếp đến UE nên độ tin cậy của việc truyền kết quả suy luận cần được đảm bảo.

2.2. Lựa chọn điểm phân chia hoạt động của AI/ML

Các thuật toán nhận dạng hình ảnh AI/ML được phân chia có thể được phân tích dựa trên tính toán và đặc điểm dữ liệu của các lớp trong CNN. Như thể hiện trong hình 2.2 và 2.3 (dựa trên các số liệu được áp dụng từ nghiên cứu trong [Y. Kang et al., "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge", ACM SIGPLAN Notices, vol. 52, no. 4, pp. 615–629, 2017.]), kích thước dữ liệu trung gian được truyền từ lớp CNN này sang lớp tiếp theo phụ thuộc vào vị trí của điểm phân chia. Do đó, tốc độ dữ liệu UL (Uplink: đường lên) cần thiết

có liên quan đến điểm phân chia mô hình và tốc độ khung hình để nhận dạng hình ảnh như cũng được quan sát bởi nghiên cứu trên. Ví dụ: giả sử hình ảnh từ luồng video có tốc độ 30 khung hình / giây (FPS: Frame per Second) cần được phân loại, tốc độ dữ liệu UL cần thiết cho các điểm phân chia khác nhau nằm trong khoảng từ 4,8 đến 65 Mbit / s (được liệt kê trong Bảng 2.1). Kết quả dựa trên 227×227 hình ảnh đầu vào. Trong trường hợp hình ảnh có độ phân giải cao hơn, tốc độ dữ liệu cao hơn sẽ được yêu cầu.



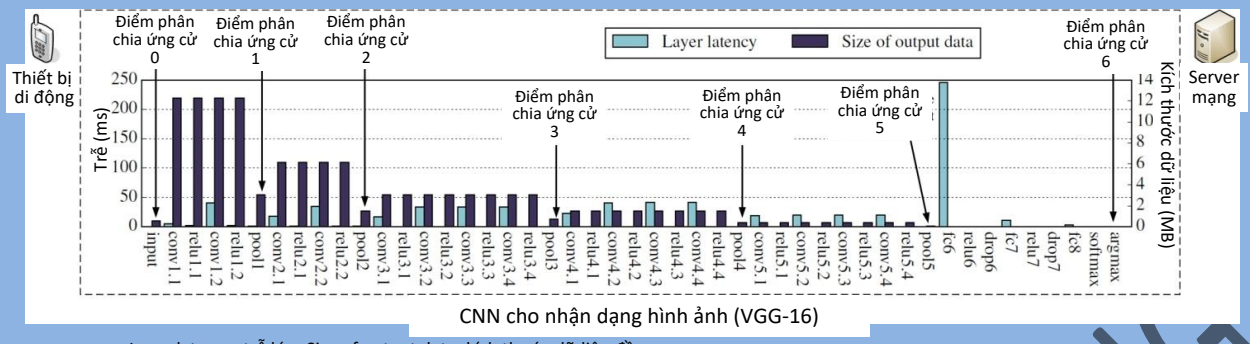
Layer latency: trễ lớp; Size of output data: kích thước dữ liệu đầu ra

Hình 2.2. Đánh giá tài nguyên tính toán / truyền thông mức lớp cho mô hình AlexNet

Bảng 2.1. Tốc độ dữ liệu UL yêu cầu cho các điểm phân chia khác nhau của mô hình nhận dạng hình ảnh @30FPS

Điểm phân chia	**Kích thước dữ liệu đầu ra gần đúng (Mbyte)**	**Tốc độ dữ liệu UL yêu cầu (Mbit/s)**
Điểm phân chia ứng cử 0 (suy luận dựa trên đám mây)	0,15	36
Điểm phân chia ứng cử 1 (sau lớp pool1)	0,27	65
Điểm phân chia ứng cử 2 (sau lớp pool2)	0,17	41
Điểm phân chia ứng cử 3 (sau lớp pool5)	0,02	4.8
Điểm phân chia ứng cử 4 (Suy luận dựa trên thiết bị)	NA	N/A

VGG-16 là một mô hình CNN được sử dụng rộng rãi khác của CNN. Vẫn giả sử hình ảnh từ luồng video có tốc độ 30 FPS cần được phân loại, tốc độ dữ liệu UL cần thiết cho các điểm phân chia khác nhau nằm trong khoảng từ 24 đến 720 Mbit / s (được liệt kê trong bảng 2).



Layer latency: trễ lớp; Size of output data: kích thước dữ liệu đầu ra

Hình 2.3. Đánh giá tài nguyên tính toán / truyền thông mứclớp cho mô hình VGG-16

Bảng 2.2. Tốc độ dữ liệu UL yêu cầu theo trải nghiệm người dùng cho các điểm phân chia khác nhau của mô hình VGG-16 @30FPS

Điểm phân chia	**Kích thước dữ liệu đầu ra gần đúng (Mbyte)**	**Tốc độ dữ liệu UL yêu cầu theo trải nghiệm người dùng (Mbit/s)**
Điểm phân chia ứng cử 0 (suy luận dựa trên đám mây)	0.6	145
Điểm phân chia ứng cử 1 (sau lớp pool1)	3	720
Điểm phân chia ứng cử 2 (sau lớp pool2)	1.5	360
Điểm phân chia ứng cử 3 (sau lớp pool3)	0.8	192
Điểm phân chia ứng cử 4 (sau lớp pool4)	0.5	120
Điểm phân chia ứng cử 5 (sau lớp pool5)	0.1	24
Điểm phân chia ứng cử 6 (Suy luận dựa trên thiết bị)	N/A	N/A

2.3. Các chế độ cho các hoạt động AI / ML phân chia giữa thiết bị và mạng

2.3.1. Kết nối giữa AI/ML và mạng 5G/6G qua đám mây biên

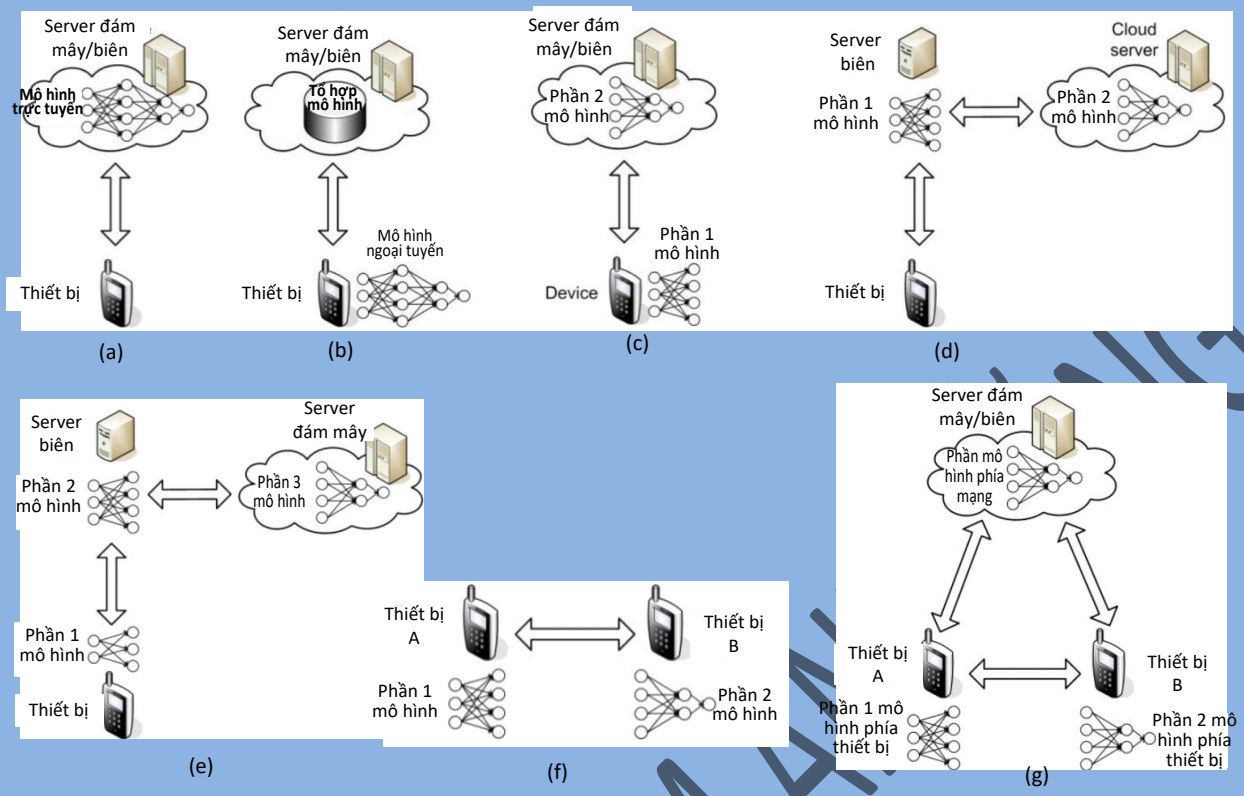
Các mạng 5G/6G tích hợp một dải rộng các nút được kết nối như: các thiết bị thông minh và các bộ cảm biến bằng cách cung cấp sô lượng lớn các ứng dụng. Để cho phép các ứng dụng trễ thấp hoạt động và đảm bảo an ninh cần chuyển dịch mô hình từ điện toán đám mây tập trung sang

đến điện toán đám mây biên gần thiết bị cuối hơn. Trong 5G điện toán đám mây biên được gọi là MEC (Mobile Edge Computing: điện toán di động biên). Nhiệm vụ chính của đám mây biên là: giảm tải quyết định, ấn định tài nguyên , triển khai server và chi phí truyền thông.

Một thách thức đối với mọi thiết bị người dùng là phải quyết định khi nào nó cần giam tải tính toán của mình lên server biên và khối lượng tính toán cần giảm tải. Quyết định vày rất quan trọng cả về tiết kiệm năng lượng lẫn hiệu quả tính toán. Quyết định này chịu ảnh hưởng của nhiều nhân tố như: an ninh và tính riêng tư của dữ liệu đối với người dùng, ưu tiên của người dùng thực hiện tính toán, chất lượng kết nối vô tuyến, mức độ khả dụng của đám mây biên (MEC đối với 5G) và đám mây trung tâm và khả năng thực hiện xử lý, v.v. Khi xét đến các nhân tố này, có ba chiến lược cho giảm tải đến server biên: thực hiện cục bộ (xảy ra trên thiết bị người dùng), giảm tải toàn bộ hoặc một phần đến server biên. Mục đích của ba giải thuật này là cân bằng chỉ số: tiết kiệm năng lượng trên thiết bị người dùng và thỏa mãn rang buộc trễ cần đảm bảo.

2.3.2. Các chế độ cho các hoạt động AI / ML phân chia giữa thiết bị và mạng

Các chế độ cho các hoạt động AI / ML phân chia giữa thiết bị và mạng được minh họa trong Hình 2.4. Các chế độ nói chung có thể áp dụng cho đào tạo AI / ML cũng như suy luận. Trong phần này, chúng ta tập trung vào xử lý suy luận. Chế độ a) và b) là các sơ đồ truyền thống vận hành suy luận AI / ML hoàn toàn trên một điểm cuối. Chế độ c) - g) cố gắng chia suy luận AI / ML hoặc thậm chí mô hình thành nhiều phần theo nhiệm vụ và môi trường hiện tại, để giảm bớt áp lực tính toán, bộ nhớ / lưu trữ, năng lượng và tốc độ dữ liệu cần thiết trên cả thiết bị và điểm cuối NW, cũng như để có được hiệu suất suy luận mô hình tốt hơn về độ trễ, độ chính xác và bảo vệ quyền riêng tư.



Hình 2.4. Các chế độ cho các hoạt động AI / ML phân chia giữa thiết bị và mạng

- Chế độ a): Suy luận dựa trên đám mây/biên (Cloud/edge)
  Ở chế độ này (như thể hiện trong Hình 2.4 (a)), suy luận mô hình AI/ML chỉ được thực hiện trong đám mây hoặc máy chủ biên. Thiết bị chỉ báo cáo dữ liệu cảm biến/nhận thức cho máy chủ và không cần hỗ trợ các hoạt động suy luận AI/ML. Máy chủ trả về kết quả suy luận cho thiết bị. Ưu điểm của chế độ này là hạn chế độ phức tạp của thiết bị. Một nhược điểm là hiệu suất suy luận phụ thuộc vào tốc độ dữ liệu truyền thông và độ trễ giữa thiết bị và máy chủ. Tải lên một số dữ liệu nhận thức theo thời gian thực (ví dụ: phát trực tuyến video có độ phân giải cao) yêu cầu tốc độ dữ liệu cao ổn định và một số dịch vụ AI/ML (ví dụ: robot điều khiển từ xa) yêu cầu độ trễ thấp ổn định, điều này rất khó đảm bảo trong hệ thống 5G do vùng phủ sóng mạng khác nhau. Và do việc tiết lộ dữ liệu nhạy cảm về quyền riêng tư cho mạng, các biện pháp bảo vệ quyền riêng tư tương ứng là cần thiết.
- Chế độ b): Suy luận dựa trên thiết bị
  Ở chế độ này (như thể hiện trong Hình 2.4 (b)), suy luận mô hình AI / ML được thực hiện tại chỗ trên thiết bị di động. Ưu điểm là, trong quá trình suy luận, thiết bị không cần giao tiếp với máy chủ đám mây/biên. Một động lực khác của chế độ này là bảo vệ quyền riêng tư tại nguồn dữ liệu, tức là thiết bị, mặc dù vấn đề bảo vệ quyền riêng tư cũng cần được xem xét ở phía thiết bị. Nhược điểm là có khả năng áp đặt tài nguyên tính toán / bộ nhớ / lưu trữ quá mức cho thiết bị. Và cũng không thể giả định rằng thiết bị luôn giữ tất cả các

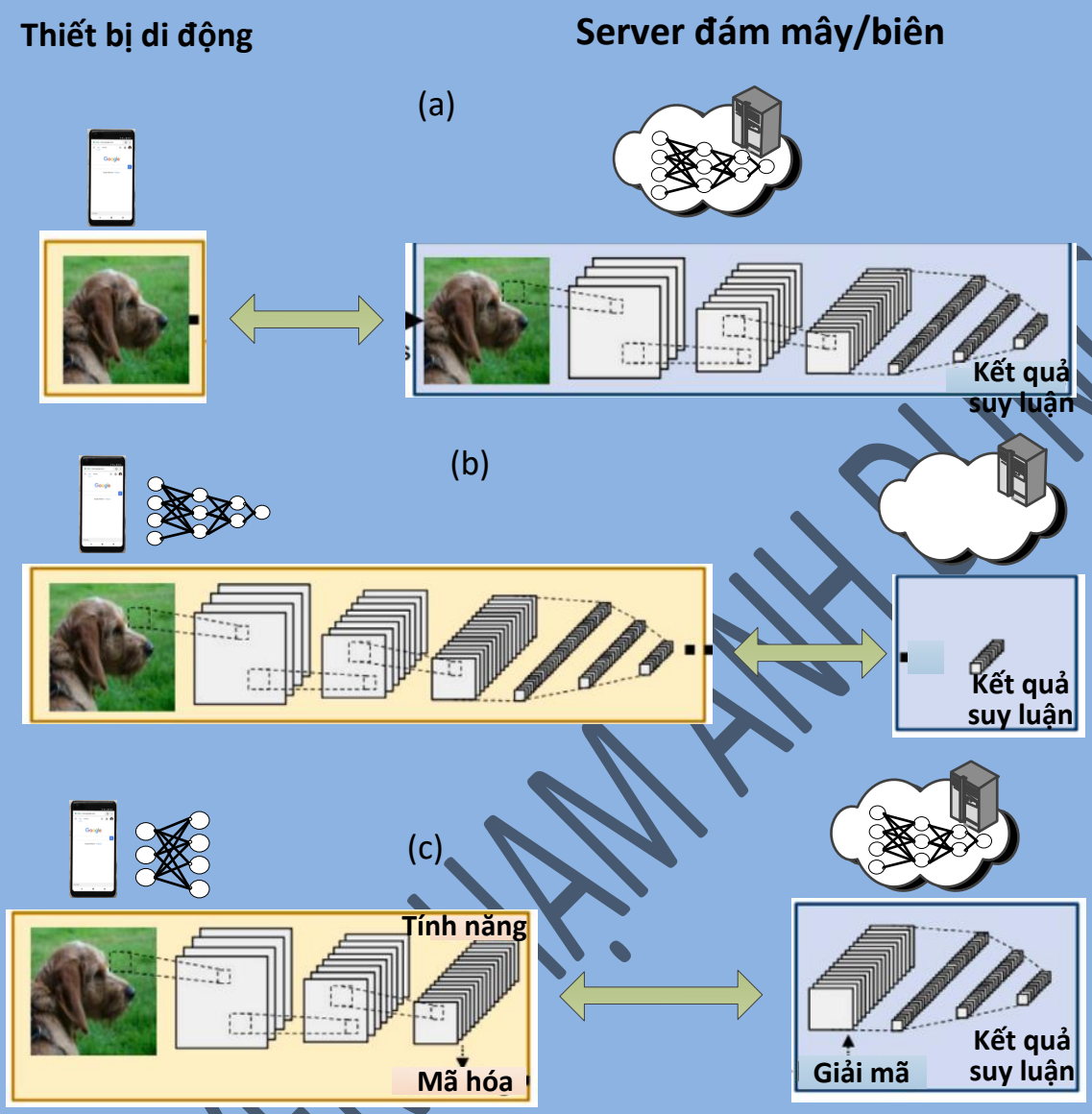
mô hình AI / ML tiềm năng cần thiết trên thiết bị. Trong một số trường hợp, thiết bị di động có thể cần lấy mô hình AI/ML từ đám mây/máy chủ biên, yêu cầu tốc độ tải xuống dữ liệu tương ứng từ hệ thống 5G và 6G.

- Chế độ c): Suy luận được phân chia thiết bị-đám mây/biên (Device-Cloud/edge)
  Ở chế độ này (như thể hiện trong Hình 2.4 (c)), một hoạt động hoặc mô hình suy luận AI / ML trước tiên được chia thành hai phần giữa thiết bị và máy chủ đám mây / biên theo các yếu tố môi trường hệ thống hiện tại như tốc độ dữ liệu truyền thông, tài nguyên thiết bị và khối lượng công việc máy chủ. Sau đó, thiết bị sẽ thực hiện suy luận AI/ML đến một phần cụ thể hoặc mô hình DNN lên đến một lớp cụ thể và gửi dữ liệu trung gian đến máy chủ đám mây/biên. Máy chủ sẽ thực hiện các phần/lớp còn lại và gửi kết quả suy luận đến thiết bị. So với Chế độ a) và b), chế độ này linh hoạt hơn và mạnh mẽ hơn đối với tài nguyên tính toán và điều kiện truyền thông khác nhau. Một liên kết quan trọng cho chế độ này là chọn đúng điểm phân chia tối ưu giữa phía thiết bị và phía mạng dựa trên các điều kiện.
- Chế độ d): Suy luận phân chia biên-đám mây (Edge-Cloud)
  Chế độ này (như thể hiện trong Hình 2.4 (d)) có thể được coi là phần mở rộng của Chế độ a). Sự khác biệt là mô hình DNN được thực thi thông qua sức mạnh tổng hợp biên-đám mây, thay vì chỉ được thực thi trên đám mây hoặc máy chủ biên. Phần nhạy cảm với độ trễ của hoạt động suy luận AI/ML hoặc các lớp của mô hình AI/ML có thể được thực hiện tại máy chủ biên. Các phần / lớp tính toán chuyên sâu mà máy chủ biên không thể thực hiện có thể được chuyển sang máy chủ đám mây. Thiết bị chỉ báo cáo dữ liệu cảm biến/nhận thức cho máy chủ và không cần hỗ trợ các hoạt động suy luận AI/ML. Dữ liệu trung gian được gửi từ máy chủ biên đến máy chủ đám mây. Cần chọn một điểm phân chia thích hợp để hợp tác hiệu quả giữa máy chủ biên và máy chủ đám mây.
- Chế độ e): Suy luận phân chia thiết bị-biên-đám mây (Device-Edge-Cloud)
  Chế độ này (như thể hiện trong Hình 2.4 (e)) là sự kết hợp của Chế độ c) và d). Hoạt động suy luận AI/ML hoặc mô hình AI/ML được phân chia trên thiết bị di động, máy chủ biên và máy chủ đám mây. Các phần/lớp tính toán chuyên sâu của hoạt động/mô hình AI/ML có thể được phân bố giữa đám mây và/hoặc máy chủ biên. Các phần/lớp nhạy cảm với độ trễ có thể được thực hiện trên thiết bị hoặc máy chủ biên. Dữ liệu nhạy cảm về quyền riêng tư có thể được để lại trên thiết bị. Thiết bị gửi kết quả dữ liệu trung gian từ tính toán của nó đến máy chủ biên. Và máy chủ biên gửi kết quả dữ liệu trung gian từ tính toán của nó đến máy chủ đám mây. Cần chọn hai điểm phân chia để hợp tác hiệu quả giữa thiết bị, máy chủ biên và máy chủ đám mây.
- Chế độ f): Suy luận phân chia thiết bị-thiết bị (Device-Device)
  Chế độ này (như thể hiện trong Hình 2.4 (f)) cung cấp một suy luận phân chia phi tập trung. Một hoạt động hoặc mô hình suy luận AI/ML có thể được phân chia trên các thiết bị di động khác nhau. Một nhóm thiết bị di động có thể thực hiện các phần khác nhau của

hoạt động AI/ML hoặc các lớp DNN khác nhau cho một nhiệm vụ suy luận và trao đổi dữ liệu trung gian với nhau. Tải tính toán có thể được phân bố trên các thiết bị trong khi mỗi thiết bị lưu giữ thông tin riêng tư cục bộ.

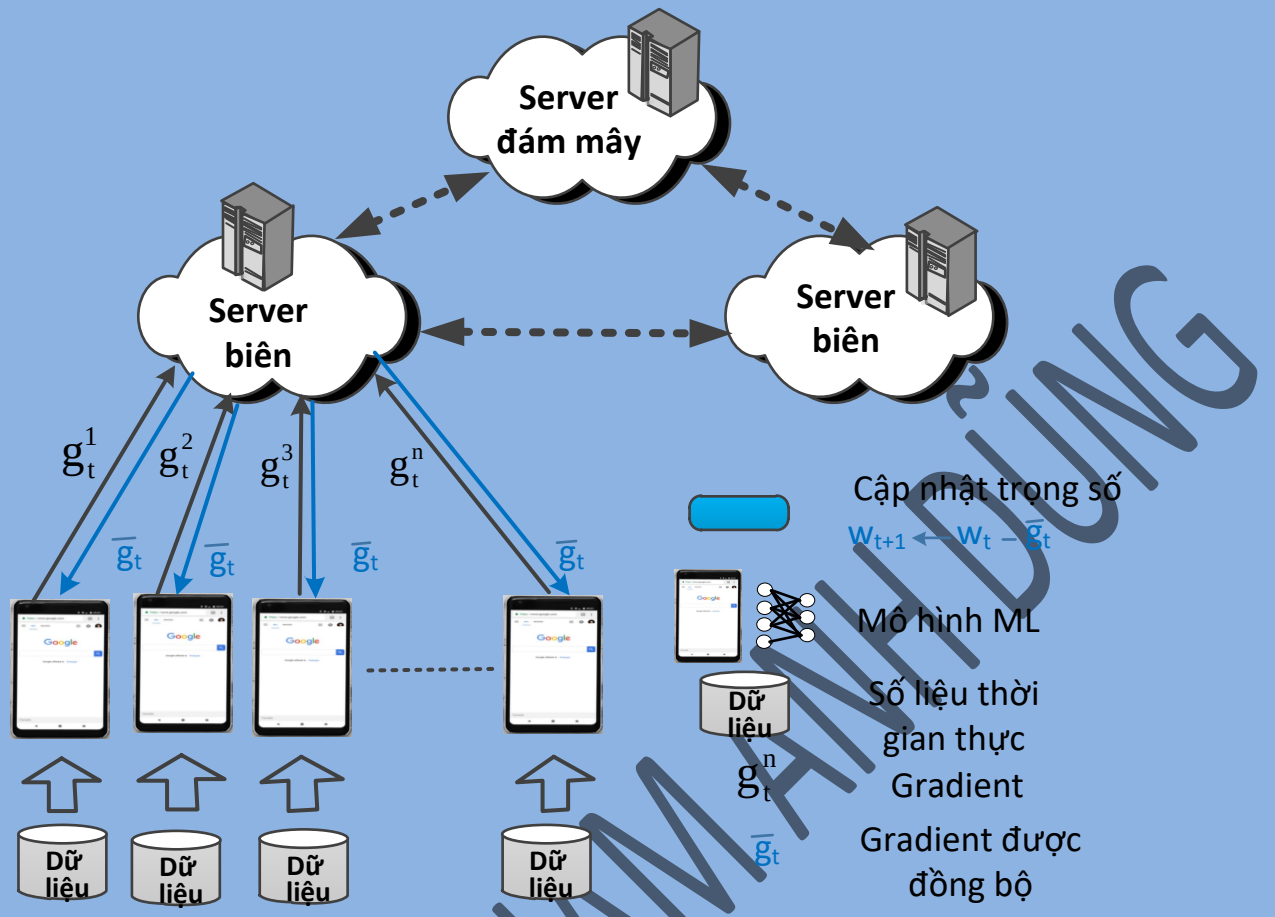
- Chế độ g): Suy luận phân chia thiết bị-thiết bị-đám mây/biên (Device-Edge/Cloud)
  Chế độ g) có thể được kết hợp thêm với chế độ c) hoặc e). Như thể hiện trong Hình 2.4 (g), một hoạt động hoặc mô hình suy luận AI / ML trước tiên được chia thành phần thiết bị và phần mạng. Sau đó, phần thiết bị có thể được thực thi theo cách phi tập trung, tức là phân chia thêm trên các thiết bị di động khác nhau. Dữ liệu trung gian có thể được gửi từ một thiết bị đến máy chủ đám mây / biên. Hoặc nhiều thiết bị có thể gửi dữ liệu trung gian đến máy chủ đám mây/biên.

Hình 2.5 cho thấy ba kịch bản xử lý AI/ML dựa trên CNN khác nhau: (a) Suy luận chỉ được thực hiện tại server đám mây/ biên, (b) suy luận chỉ được thực hiện tại UE và (c) suy luận được UE giảm tại đến server đám mây/biên.



Hình 2.5.Minh họa các kịch bản phân chia xử lý Ahoatj động AI/ML dựa trên mạng CNN

Trong kịch bản máy di động và server biên, để đạt được dự báo thời gian thực tại phía người dùng, server biên thường đóng vai trò một server thông số cho cả hai: thực hiện và đào tạo ML. Server biên thực hiện đồng bộ gradient bằng cách thu thập tất cả các gradient và lấy trung bình chúng cho các thông số cập nhật (hình 2.6). Các thông số được cập nhật được gửi trở lại các cho các máy di động biên (được gọi là đồng bộ thông số)

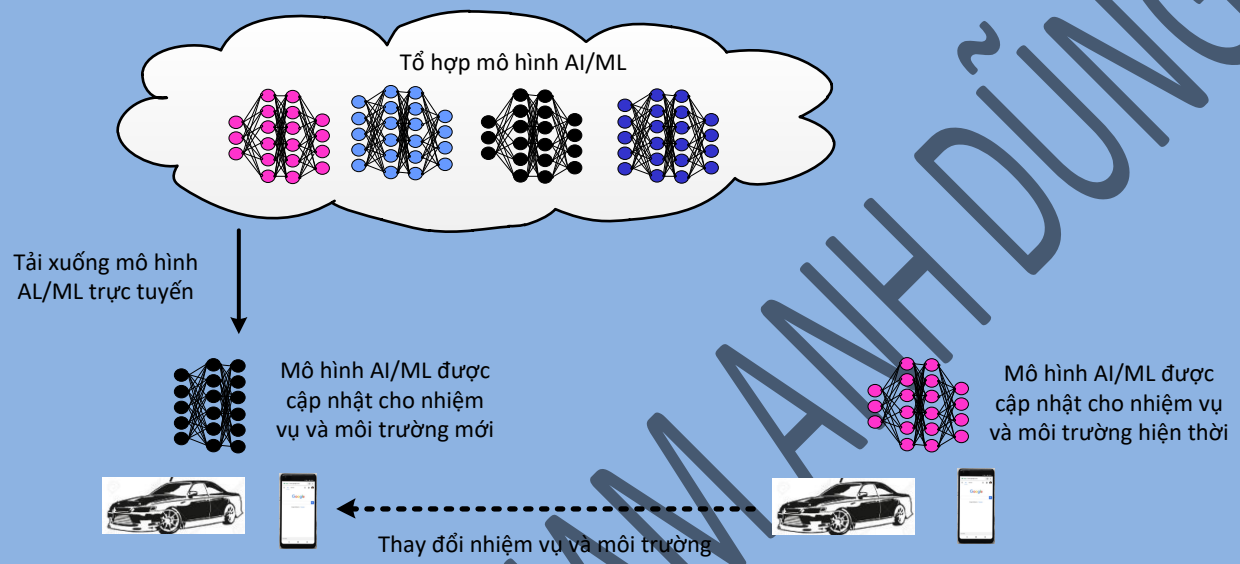


Hình 2.6. Thí dụ về đào tạo tại biên

3. NGUYÊN LÝ PHÂN BỐ VÀ CHIA SẺ SỐ LIỆU VÀ MÔ HÌNH AI/ML TRONG CÁC HỆ THỐNG 5G VÀ 6G

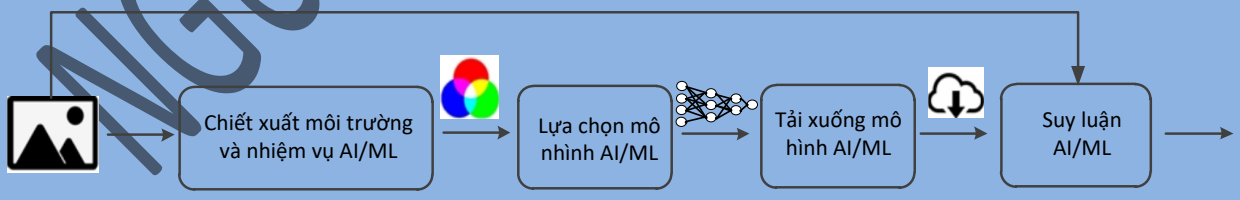
Đối với các nhiệm vụ suy luận yêu cầu độ trễ thấp và muốn lưu giữ dữ liệu nhạy cảm về quyền riêng tư ở phía UE, suy luận AI/ML ngoại tuyến là mong muốn, thay vì suy luận dựa trên đám mây. Tuy nhiên, mô hình AI/ML ngoại tuyến chạy trên thiết bị di động phải có độ phức tạp tính toán tương đối thấp và dung lượng lưu trữ nhỏ. Một cách tiếp cận để kích hoạt các mô hình DNN ngoại tuyến trên thiết bị di động là nén mô hình để giảm các yêu cầu về tài nguyên và tính toán của nó. Tuy nhiên, nén DNN sẽ dẫn đến mất độ chính xác của suy luận và khả năng thích ứng với các nhiệm vụ và môi trường khác nhau. Một giải pháp cho thách thức này là chọn mô hình thích ứng để suy luận từ một tập hợp các mô hình được đào tạo [10]. Việc lựa chọn mô hình được thúc đẩy bởi quan sát rằng mô hình tối ưu để suy luận phụ thuộc vào dữ liệu đầu vào và yêu cầu độ chính xác. Các thiết bị đầu cuối di động đa chức năng thường cần chuyển đổi mô hình AI/ML để đáp ứng các thay đổi nhiệm vụ và môi trường.

Điều kiện của việc lựa chọn mô hình thích ứng là các mô hình được chọn đã có sẵn cho thiết bị di động. Tuy nhiên, với thực tế là các mô hình DNN ngày càng trở nên đa dạng và với tài nguyên lưu trữ hạn chế trong UE, việc tải trước tất cả các mô hình AI/ML ứng cử viên trên bo mạch là không khả thi. Phân bố mô hình trực tuyến (tức là tải xuống mô hình mới) hoặc học chuyển giao trực tuyến (tức là cập nhật mô hình một phần) là cần thiết. Như minh họa trong Hình 3.1, mô hình AI/ML có thể được phân bố từ điểm cuối mạng (NW endpoint) đến các thiết bị khi chúng cần để thích ứng với các nhiệm vụ và môi trường AI / ML đã thay đổi.



Hình 3.1. Tải xuống mô hình AI/ML qua hệ thống 5G/6G

Mô hình được phân bố có thể được xác định theo hai cách: được yêu cầu bởi một thiết bị hoặc được điều khiển bởi một máy chủ mạng. Điều kiện của cơ chế đầu tiên là thiết bị có thể đưa ra quyết định lựa chọn / lựa chọn lại mô hình dựa trên sự hiểu biết về nhiệm vụ AI/ML sắp tới, môi trường và danh sách các mô hình có sẵn tại máy chủ mạng. Như thể hiện trong Hình 3.2, bộ chọn mô hình trên thiết bị được đào tạo để chọn DNN tốt nhất cho các dữ liệu đầu vào khác nhau.



Hình 3.2. Lựa chọn mô hình AI/ML và tải xuống

Tốc độ dữ liệu để tải xuống các mô hình cần thiết phụ thuộc vào các yếu tố sau:

- Kích thước của mô hình:

Điều này phụ thuộc vào các ứng dụng AI / ML khác nhau. Cùng với yêu cầu hiệu năng ngày càng tăng đối với các hoạt động AI/ML, kích thước của các mô hình cũng tiếp tục tăng lên, mặc dù các kỹ thuật nén mô hình đang được cải thiện.

- Độ trễ tải xuống yêu cầu
  Điều này phụ thuộc vào tốc độ mô hình cần sẵn sàng trên thiết bị. Nó bị ảnh hưởng bởi mức độ mà ứng dụng sắp tới có thể được dự đoán. Xem xét tính không thể đoán trước của hành vi người dùng và thời gian chờ đợi điển hình mà người dùng có thể chịu đựng, việc tải xuống mô hình AI/ML cần được hoàn thành trong vài giây hoặc thậm chí trong mili giây. Khác với video được phát trực tuyến có thể được phát khi một phần nhỏ được nhớ đệm, mô hình DNN chỉ có thể được sử dụng cho đến khi toàn bộ mô hình được tải xuống hoàn toàn.

Cần lưu ý rằng suy luận AI/ML dựa trên mạng và phân chia thường yêu cầu tốc độ dữ liệu đường lên cao và không đổi để liên tục giảm tải dữ liệu cảm biến/trung gian sang máy chủ đám mây/biên. Ngược lại, phân bố mô hình AI/ML chủ yếu yêu cầu tốc độ dữ liệu đường xuống cao trong một đợt bùng phát. Điều này làm cho việc phân bố mô hình phù hợp hơn với các hệ thống truyền thông di động chiếm ưu thế của đường xuống (ví dụ: sử dụng tỷ lệ DL-to-UL cao) hoặc các hệ thống có vùng phủ sóng không ổn định. Tất nhiên điều kiện là tài nguyên tính toán của thiết bị di động có thể đủ khả năng thực thi mô hình AI / ML trên bo mạch. Nếu tải tính toán vượt quá khả năng của thiết bị, suy luận dựa trên mạng hoặc phân chia phải được áp dụng.

4. NGUYÊN LÝ HỌC PHÂN TÁN/ LIÊN KẾT TRÊN CÁC HỆ THỐNG 5G VÀ 6G

Với khả năng liên tục cải tiến của camera và cảm biến trên thiết bị di động, dữ liệu đào tạo có giá trị, cần thiết cho việc đào tạo mô hình AI/ML, ngày càng được tạo ra trên các thiết bị. Đối với nhiều nhiệmvụ AI/ML, dữ liệu bị phân mảnh được thu thập bởi thiết bị di động là điều cần thiết để đào tạo một mô hình toàn cầu. Trong các phương pháp truyền thống, dữ liệu đào tạo được thu thập bởi thiết bị di động được tập trung vào trung tâm dữ liệu đám mây để đào tạo tập trung.

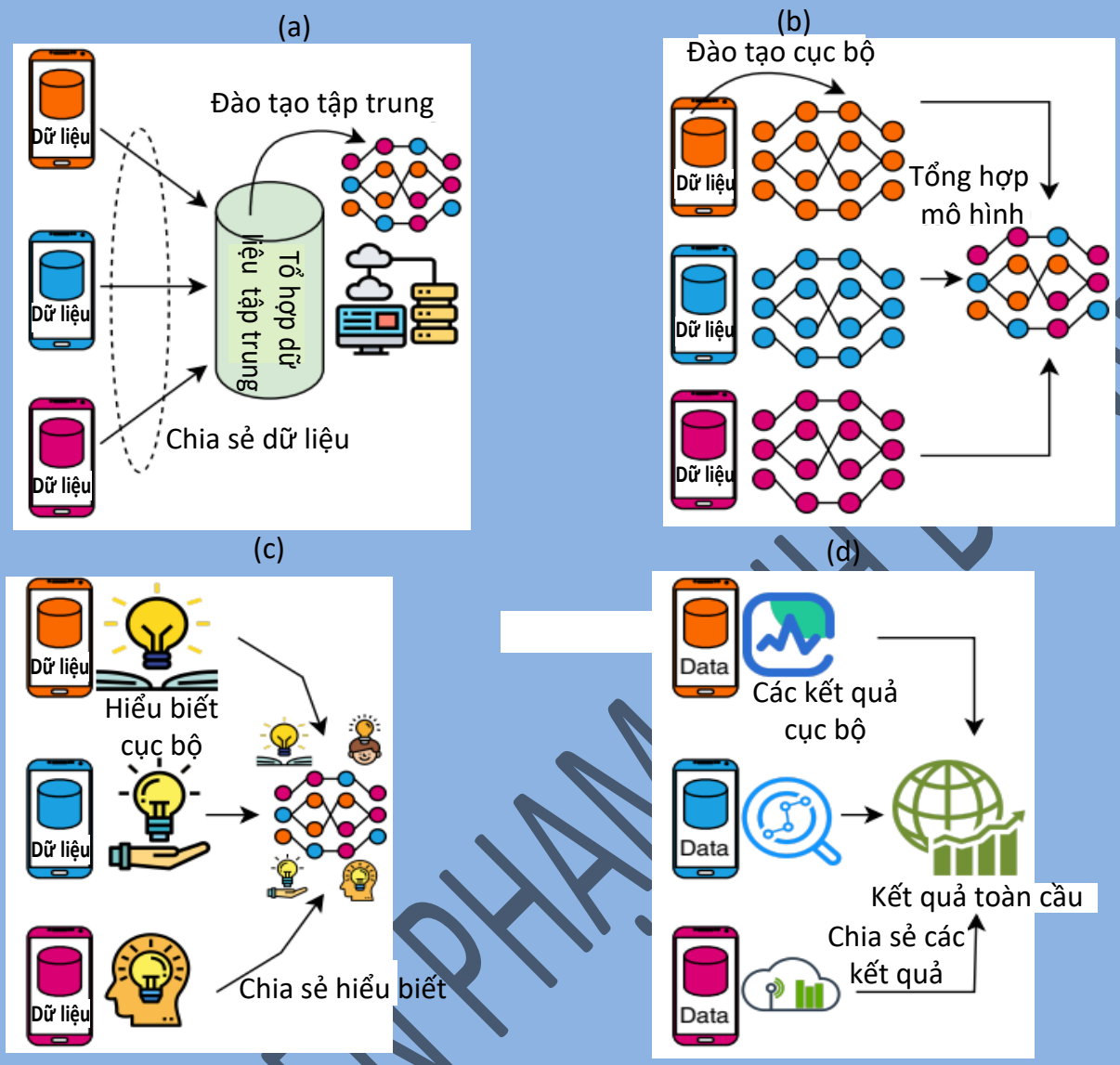
Tuy nhiên, việc đào tạo mô hình AI/ML thường yêu cầu một tập dữ liệu lớn và tài nguyên tính toán đáng kể cho nhiều lần lặp lại cập nhật trọng số. Ngày nay, hầu hết các nhiệm vụ đào tạo mô hình AI/ML được thực hiện trong các trung tâm dữ liệu đám mây lớn vì mức tiêu thụ tài nguyên của giai đoạn đào tạo vượt quá đáng kể giai đoạn suy luận. Trong nhiều trường hợp, việc đào tạo một mô hình DNN vẫn mất vài giờ đến nhiều ngày. Tuy nhiên, đào tạo dựa trên đám mây có nghĩa là lượng dữ liệu đào tạo khổng lồ phải được vận chuyển từ các thiết bị lên đám mây, gây ra chi phí giao tiếp quá lớn cũng như áp lực bảo mật dữ liệu ở phía mạng. Tương tự như suy luận AI/ML phân chia được giới thiệu trong mục 2, các nhiệm vụ đào tạo mô hình AI/ML cũng có thể hoạt động theo cách phối hợp thiết bị đám mây. Học tập phân tán và học tập liên kết là những ví dụ theo cách này.

4.1. Nguyên lý của học phân tán (Distributed Learning) trong hệ thống 5G/6G

4.1.1. Nguyên chung của kỹ thuật học phân tán

Khi các giải thuật ML cần được triển khai theo cách phân tán với nhiều server đám mây/biên và nhiều thiết bị (IoT hoặc di động), Giải pháp tiếp cận chung là chia sẻ dữ liệu, mô hình, sự hiểu biết hoặc kết quả giữa các client. Các cách tiếp cận này được đề cập đến như là cách tiếp cận chia sẻ dữ liệu, chia sẻ mô hình, chia sẻ sự hiểu biết và chia sẻ kết quả.

Hình 4.1 minh họa bốn cách tiếp cận học phân tán. Trong học phân tán, hệ thống thường bao gồm một server trung tâm và nhiều thiết bị phân tán. Trong cách tiếp cận chia sẻ số liệu (Data Sharing) như minh họa trên hình 4.1(a), server đào tạo mô hình học toàn cầu (Global Learning Model) dựa trên các dữ liệu được chia sẻ từ các thiết bị phân tán. Trong cách tiếp cận chia sẻ mô hình (Model Sharing) như minh họa trên hình 4.1(b), các thiết bị phân tán phân tán đào tạo các mô hình tương ứng của mình sử dụng các dữ liệu cục bộ của chúng và chỉ chia sẻ mô hình được đào tạo với server để tổng hợp mô hình. Học liên kết (FL) là thí dụ về cách tiếp cận này, trong đó mô hình được tổng hợp được phân phối từ server đến các thiết bị để khởi đầu một vòng mới. Khác với hai cách tiếp cận trên, trong đó dữ liệu và các mô hình được chia sẻ, cách tiếp cận thứ ba như minh họa trên hình 4.1.(c) chia sẻ hiểu biết (Sharing Knowledge) cho phép các thiết bị chiết suất hiểu biết của mình dựa trên dữ liệu cục bộ sau đó chia sẻ nó với server học. Trong cách tiếp cận cuối cùng như minh họa trên hình 4.1.(d), cách tiếp cận chia sẻ kết quả (Sharing Result) , các thiết bị phân tán chia sẻ đầu ra của các mô hình được đào tạo của chúng với server.



Hình 4.1. Minh họa các cách tiếp cận học phân tán chính

4.1.2. Học phân tán trong các hệ thống 5G/6G

Học phân tán trên biên được nghiên cứu rộng rãi, đây là một cách tiếp cận hiệu quả cho các mô hình học sâu. Nhân tố động lực chính là hứa hẹn một mô hình được cải thiện kết hợp được dữ liệu từ các nguồn khác nhau, trong khi có thể kiểm soát được an ninh, chi phí truyền thông và tính toán.

Học phân tán trên biên đang được nghiên cứu rộng rải, đây là một cách tiếp cận hiệu quả cho các mô hình học sâu. Nhân tố động lực chính là nó hứa hẹn một mô hình được cải thiện kết hợp được dữ liệu từ các nguồn khác nhau, trong khi có thể kiểm soát được an ninh, chi phía truyền thông và tính toán. Có hai cách tiếp cận học phân tán chính (xem A survey: Distributed

Machine Learning for 5G and beyond): học liên kết (FL) và học có giám sát (SL: Supervised Learning). Phần dưới đây ta sẽ xét học liên kết.

4.2. Nguyên lý của học liên kết (Federated Learning) trong hệ thống 5G/6G

4.2. 1. Kiến trúc tổng quát của học liên kết (Federated Learning)

Ở chế độ học liên kết (FL: Federated Learning), máy chủ đám mây đào tạo một mô hình toàn cầu bằng cách tổng hợp các mô hình cục bộ được đào tạo một phần bởi từng thiết bị cuối. Hình 4.2 minh họa kiến trúc tổng quát của học liên kết, trong đó một tập K người dùng (ví dụ các client hay cá thiết bị) đào tào các mô hình cục bộ bằng cách sử dụng các dữ liệu cục bộ của mình sau đó chia sẻ các cập nhật của mô hình cục bộ với server biên để tổng hợp mô hình (Model Aggregation). Mỗi người dùng có một tập dữ liệu gồm D_k mẫu rời rạc $D_k=\{1,\ldots, D_k\}$ và tổng số mẫu dữ liệu của tất cả các người dùng là: $D = \sum_{k=1}^{K} D_k$.

Như thấy trên hình 4.3, một vòng truyền thông của FL báo gồm ba bước sau:

- **Khởi đầu mô hình**. Server biên khởi đầu khởi đầu một trọng số w^0 của mô hình toàn cầu và phát quảng bá giá trị của nó đến các người dùng được chọn trong vòng truyền thông hiện thời. Lưu ý rằng do tính không đồng nhất của dữ liệu, các tài nguyên bị hạn chế và các điều kiện kênh, nên chỉ có một tập con các người dùng được chọn để đóng góp cho xử lý FL. Ngoài ra server biên phải đặc tả các thông số của các mô hình cục bộ (ví dụ: độ chính xác của mô hình cục bộ và tốc độ học) và các thông số của mô hình toàn cầu (ví dụ: độ chính xác của mô hình toàn cầu);
- **Đào tạo cục bộ**. Tại vòng truyền thông t, từng người dùng k đào tạo mô hình dựa trên tập dữ liệu thô của mình. Mục tiêu của đào tạo tại chỗ của người dùng k là cực tiểu hóa hàm tổn hao (Loss Function) $f_k(w)$ như sau:

$$w_k^t = \arg\min f_k \quad \mathbf{w}$$

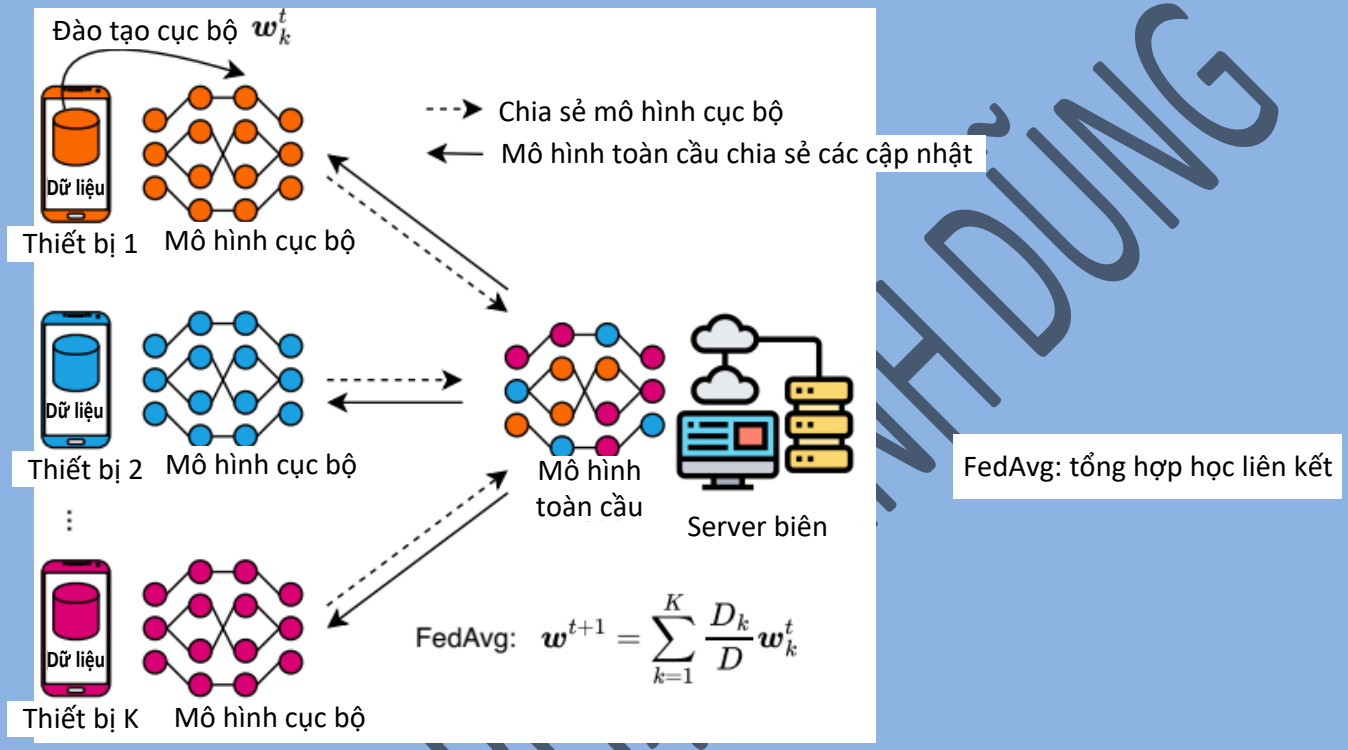
Khi này, các người dùng đã chia sẻ các thông số của mô hình cục bộ với server biên để thổng hợp mô hình (Model Aggregation). Các hàm tổn hao khác nhau có thể được sử dụng cho các nhiệm vụ học khác nhau. Chẳng hạn mô hình hồi quy tuyến tính trên tập dữ liệu (x,y) được nhận dạng bởi hàm tổn hao:

$$f(\mathbf{w}) = \frac{1}{N}(\mathbf{y} - \hat{\mathbf{y}})^2,$$

trong đó $\mathbf{y} = \mathbf{x}^T\mathbf{w}$ là vectơ giá trị thực tế theo phương trình (1.2) với bỏ qua thiên kiến b và $\hat{\mathbf{y}}$ là vectơ giá trị được dự báo.

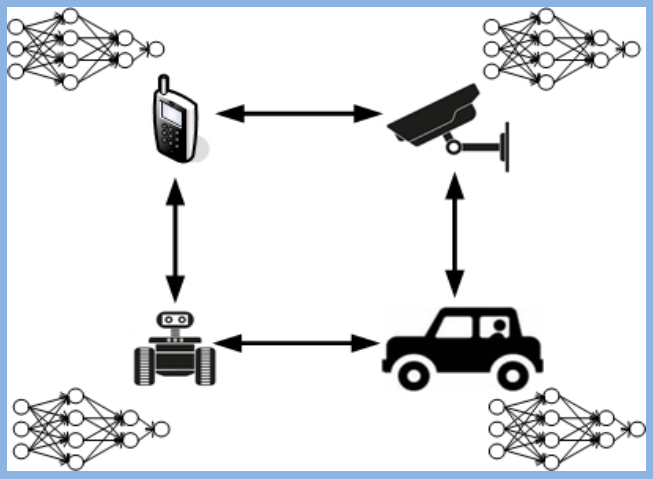
- **Tổng hợp mô hình**. Sau khi server biên nhận được các cập nhật từ các người dùng được lựa chọn, thông số mô hình toàn cầu có thể nhận được bằng cách cực tiểu hóa hàm tổn hao như sau:

$$\text{FedAvg} = f\ \mathbf{w}\ = \sum_{k=1}^{K} \frac{D_k}{D} f_k\ \mathbf{w}$$



Hình 4.2. Minh họa kiến trúc tổng quát của học liên kết

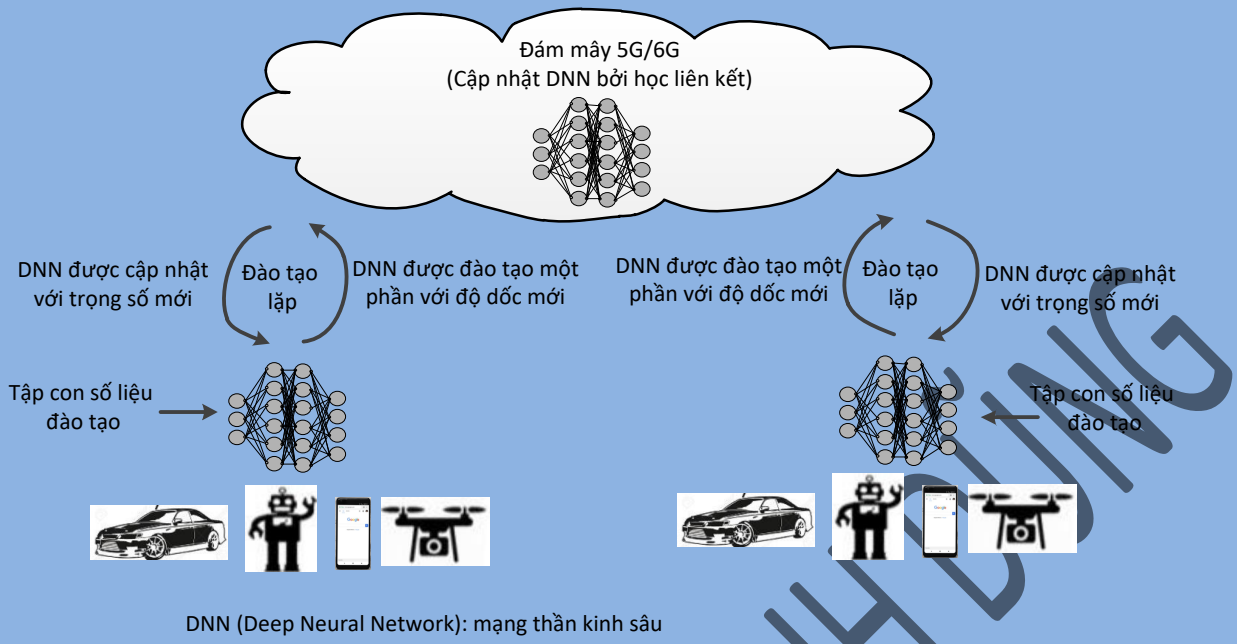
Trong chế độ học phân tán, như thể hiện trong Hình 4.3, mỗi nút tính toán đào tạo mô hình DNN của riêng mình cục bộ với dữ liệu cục bộ, bảo toàn thông tin cá nhân cục bộ. Để có được mô hình DNN toàn cầu bằng cách chia sẻ cải tiến đào tạo cục bộ, các nút trong mạng sẽ truyền thông với nhau để trao đổi các bản cập nhật mô hình cục bộ. Ở chế độ này, mô hình DNN toàn cầu có thể được đào tạo mà không cần sự can thiệp của trung tâm dữ liệu đám mây.



Hình 4.3. Học phân tán

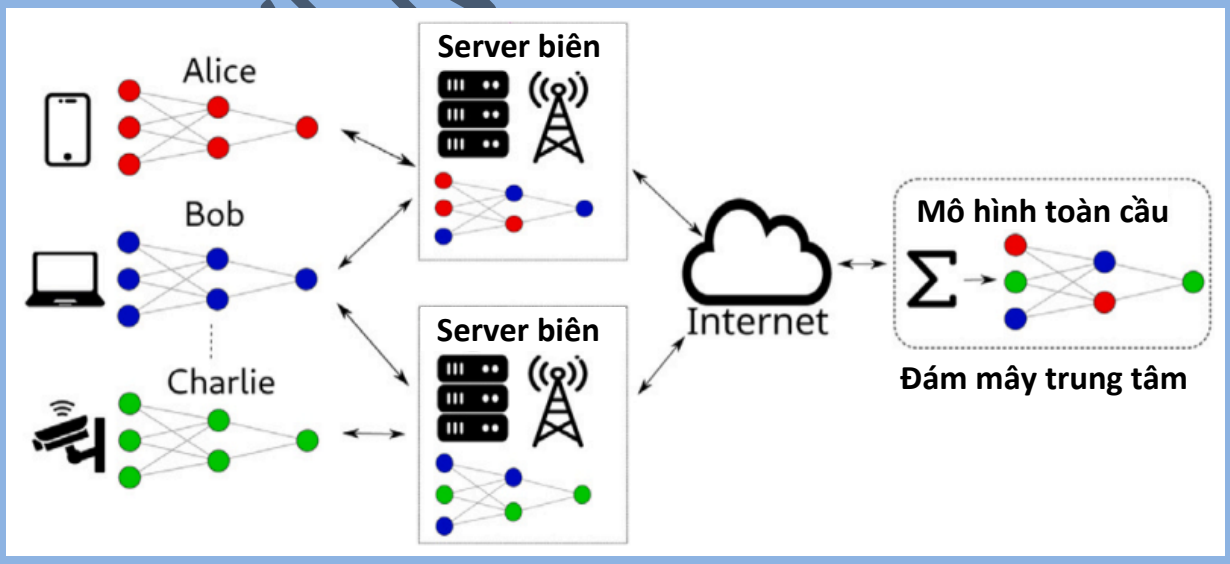
4.2.2. Học liên kết trong hệ thống 5G/6G

Thuật toán Federated Learning dễ chịu nhất cho đến nay dựa trên mô hình trung bình lặp đi lặp lại. Như được mô tả trong Hình 4.4, trong mỗi lần lặp lại đào tạo, UE thực hiện đào tạo dựa trên mô hình được tải xuống từ máy chủ AI bằng cách sử dụng dữ liệu đào tạo cục bộ. Sau đó, UE báo cáo kết quả đào tạo tạm thời (ví dụ: gradient hoặc trọng số cho DNN) cho máy chủ đám mây thông qua các kênh 5G/6G UL. Máy chủ tổng hợp các gradient/ trọng số từ UE và cập nhật mô hình toàn cầu. Tiếp theo, mô hình toàn cầu được cập nhật được phân bố cho các UE thông qua các kênh DL 5G/6G. Sau đó, các UE có thể thực hiện đào tạo cho lần lặp tiếp theo.



Hình 4.3. Học liên kết trên hệ thống 5G/6G

Hình 4.4 cho thấy kiến trúc FL với sự tham gia đồng thời của server biên và đám mây trung tâm. Các thiết bị cuối (di động hoặc IoT) chuyển các trọng số đến đám mây biên. Sau suy luận, đám mây biên chuyển các trọng số đến đám mây trung gian để tổng hợp mô hình toàn cầu. Cuối cùng mô hình mới này được phân phối ngược trở lại các thiết bị biên và từ Server biên nó được phân phối đến các thiết bị cuối.



Hình 4.4. Xử lý FL với sự tham gia đồng thời của server biên và đám mây trung tâm

4.3. Các yêu cầu về hiệu năng đối với học phân tán/ liên kết

Các yêu cầu về hiệu năng đối với học tập phân tán/liên kết được liệt kê dưới đây. Các yêu cầu đối với các liên kết truyền thông 5G/6G (ví dụ: tốc độ dữ liệu, độ trễ, độ tin cậy) có thể được rút ra từ các yêu cầu sau.

- Tổn thất đào tạo
  Tổn thất đào tạo là khoảng cách giữa đầu ra chính xác và đầu ra được tính toán bởi mô hình DNN, cho biết mô hình DNN được đào tạo phù hợp với dữ liệu đào tạo như thế nào. Mục đích của nhiệm vụ đào tạo là giảm thiểu tổn thất đào tạo. Tổn thất đào tạo chủ yếu bị ảnh hưởng bởi chất lượng của dữ liệu đào tạo và hiệu quả của các phương pháp đào tạo, tức là liệu ý nghĩa của dữ liệu đào tạo có thể được khám phá đầy đủ và đúng đắn hay không. Đối với học liên kết, chỉ khi dữ liệu đào tạo cục bộ có giá trị có thể được học đầy đủ trong thời gian lặp lại và các bản cập nhật đào tạo cục bộ có thể được báo cáo chính xác cho máy chủ đám mây trong khoảng thời gian mục tiêu, tổn thất đào tạo mới có thể được giảm thiểu.
  Điều này ngụ ý rằng các yêu cầu đối với các thiết bị tham gia vào quá trình đào tạo về tốc độ dữ liệu UL có thể đạt được, độ trễ và độ tin cậy để báo cáo các bản cập nhật được đào tạo và tốc độ dữ liệu UL có thể đạt được, độ trễ và độ tin cậy để phân bố mô hình để đào tạo trong lần lặp tiếp theo. Và để giảm thiểu tổn thất đào tạo với tính không đồng nhất của thiết bị (trong hiệu năng tính toán và truyền thông), cần lựa chọn thiết bị đào tạo và cấu hình đào tạo trước khi đào tạo được thực hiện trong một lần lặp lại (sẽ được giới thiệu ở phần sau trong phần này). QoS của các bản tin điều khiển liên quan, ví dụ như cho yêu cầu đào tạo, báo cáo tài nguyên đào tạo, lựa chọn thiết bị đào tạo, cấu hình đào tạo và phân bổ tài nguyên cho báo cáo cập nhật đào tạo, cũng cần được đảm bảo.
- Trễ đào tạo
  Độ trễ đào tạo là một trong những chỉ số hiệu năng cơ bản nhất của nhiệm vụ đào tạo mô hình AI/ML vì nó ảnh hưởng trực tiếp đến thời điểm mô hình được đào tạo có sẵn để sử dụng. Ngày nay, đào tạo dựa trên đám mây thường mất vài giờ đến nhiều ngày. Độ trễ của quá trình học phân tán/liên kết sẽ mất nhiều thời gian hơn nếu độ trễ tính toán hoặc độ trễ truyền thông không được giảm thiểu.
  Độ trễ của quá trình học tập phân tán/liên kết được xác định bởi tốc độ hội tụ (ví dụ: số lần lặp lại trước khi quá trình đào tạo hội tụ đến sự đồng thuận) và độ trễ của mỗi lần lặp bao gồm độ trễ tính toán và độ trễ truyền thông. Độ trễ tính toán phụ thuộc vào tài nguyên tính toán/bộ nhớ có sẵn trên thiết bị đào tạo. Độ trễ truyền thông phụ thuộc vào tốc độ dữ liệu DL khả dụng để phân bố mô hình và tốc độ dữ liệu UL khả dụng để cập nhật mô hình được đào tạo. Độ trễ của toàn bộ quá trình đào tạo được xác định bởi độ trễ

lớn hơn giữa độ trễ tính toán và độ trễ các liên kết truyền thông. Do đó, độ trễ của các liên kết tính toán và truyền thông cần được giảm thiểu một cách hợp tác. Nếu độ trễ giao tiếp không phù hợp với độ trễ tính toán, liên kết giao tiếp sẽ trở thành nút thắt cổ chai và kéo dài toàn bộ quá trình đào tạo.

Đối với học liên kết đồng bộ, trong mỗi lần lặp, độ trễ đào tạo được xác định bởi thiết bị cuối cùng báo cáo bản cập nhật đào tạo của nó vì tổng hợp liên kết có thể hoàn tất khi tất cả các bản cập nhật đào tạo cần thiết được thu thập chính xác. Điều đó có nghĩa là tính không đồng nhất của thiết bị (về hiệu suất tính toán và truyền thôngp) cũng sẽ ảnh hưởng lớn đến độ trễ đào tạo tổng thể. Thay vì yêu cầu độ trễ truyền UL của một thiết bị cụ thể, độ trễ tổng thể cần thiết cho tất cả các thiết bị đào tạo để tải lên các bản cập nhật đào tạo (độ trễ nhóm thiết bị) cần được xác định. Và QoS của các bản tin điều khiển để giảm thiểu độ trễ của nhóm thiết bị, ví dụ như cho yêu cầu đào tạo, báo cáo tài nguyên đào tạo, lựa chọn thiết bị đào tạo, cấu hình đào tạo và phân bổ tài nguyên cho báo cáo cập nhật đào tạo, cũng cần được đảm bảo.

- Hiệu quả năng lượng

  Đối với học phân tán/liên kết, cả quá trình tính toán và truyền thông đều tiêu tốn năng lượng đáng kể. Kiến trúc và giao thức học phân tán cũng nên xem xét các hạn chế về năng lượng trên các thiết bị đào tạo và hiệu quả năng lượng trên thiết bị cũng như phía mạng.

- Tính riêng tư

  Khi đào tạo mô hình DNN bằng cách sử dụng dữ liệu có nguồn gốc từ một lượng lớn các thiết bị đầu cuối, dữ liệu thô hoặc dữ liệu trung gian phải được chuyển ra khỏi các thiết bị đầu cuối. So với việc báo cáo cho máy chủ đám mây / biên, bảo vệ quyền riêng tư ở các thiết bị đầu cuối có thể giảm áp lực bảo vệ quyền riêng tư ở phía mạng. Ví dụ: học liên kết là một cách tiếp cận dễ chịu để tránh tải dữ liệu thô từ thiết bị lên mạng, như yêu cầu của đào tạo dựa trên đám mây.

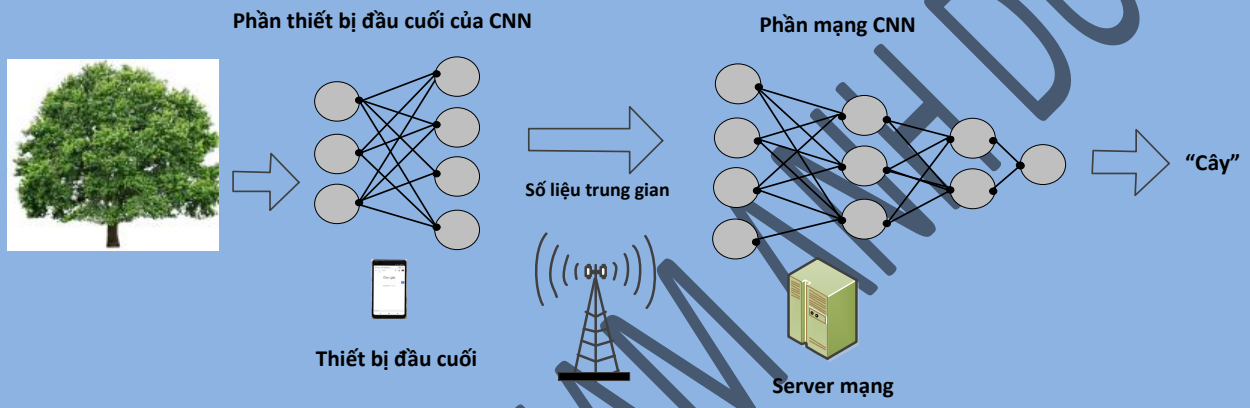
Các phần dưới đây sẽ xét các kịch bản sử dụng và các yêu cầu tiềm năng để các hệ thống 5G/6G hỗ trợ AI/ML, phân bố và chuyển giao mô hình (tải xuống, tải lên các cập nhật v.v.). Do hiện này chưa có 6G, nên các kịch bản sử dụng được 3GPP nghiên cứu trong các hệ thống 5G nhưng cũng hữu dụng cho khảo sát trong 6G. Các nghiên cứu trên các kịch bản này sử dụng này bao gồm các khía cạnh sau: phân chia hoạt động AI/ML giữa các điểm cuối, phân phối và chia sẻ mô hình/ dữ liệu AI/ML trren hệ thống 5G, học phân tán/liên kết trên hệ thống 5G.

5. CÁC KỊCH BẢN HOẠT ĐỘNG CỦA AI/ML ĐƯỢC PHÂN CHIA GIỮA CÁC ĐIỂM CUỐI TRONG CÁC HỆ THỐNG 5G/6G

5.1. Nhận dạng hình ảnh dựa trên AM/ML được phân chía

5.1.1. Kịch bản

Sơ đồ phân chia nhận dạng hình ảnh AI/ML được mô tả trên hình 5.1. CNN được phân chia thành hai phần dựa trên nhiệm vụ nhận dạng hình ảnh và môi trường hiện thời. Mục đích phân chia là giảm tải các phần đòi hỏi tính toán và năng lượng cao đến server mạng. Serrver mạng thực hiện các lớp CNN còn lại.



Hình 5.1. Thí dụ nhận dạng hình ảnh "cây" dựa trên AM/ML được phân chía

5.1.2. Quá trình diễn ra

Điều kiện ban đầu

Các điểm cuối AI/ML liên quan (ví dụ: UE, máy chủ đám mây/biên AI/ML) chạy các ứng dụng cung cấp khả năng suy luận mô hình AI/ML để nhận dạng hình ảnh và hỗ trợ hoạt động nhận dạng hình ảnh AI/ML phân chia.

Các luồng dịch vụ diễn ra như sau:

1) Ứng dụng nhận dạng hình ảnh dựa trên AI / ML được người dùng yêu cầu bắt đầu nhận dạng hình ảnh / video do UE chụp.
2) Trong điều kiện chế độ phân chia và điểm phân tách đã được xác định, ứng dụng nhận dạng hình ảnh dựa trên AI / ML trong một điểm cuối AI / ML liên quan thực hiện phần được ấn định của mô hình AI / ML và gửi dữ liệu trung gian đến điểm cuối tiếp theo trong đường ống AI / ML (AI/MLPipeline.)

3) Sau khi tất cả các điểm cuối AI/ML liên quan hoàn thành việc đồng suy luận, kết quả nhận dạng hình ảnh sẽ được cung cấp cho người dùng sử dụng kết quả.
4) Các ứng dụng nhận dạng hình ảnh dựa trên AI / ML trong các điểm cuối thực hiện nhận dạng hình ảnh phân chia cho đến khi nhiệm vụ nhận dạng hình ảnh kết thúc.

Làm lại Bước 3) và 4) để chuyển mạch/ chọn lại điểm/ chế độ phân chia nếu cần để thích ứng với các điều kiện thay đổi.

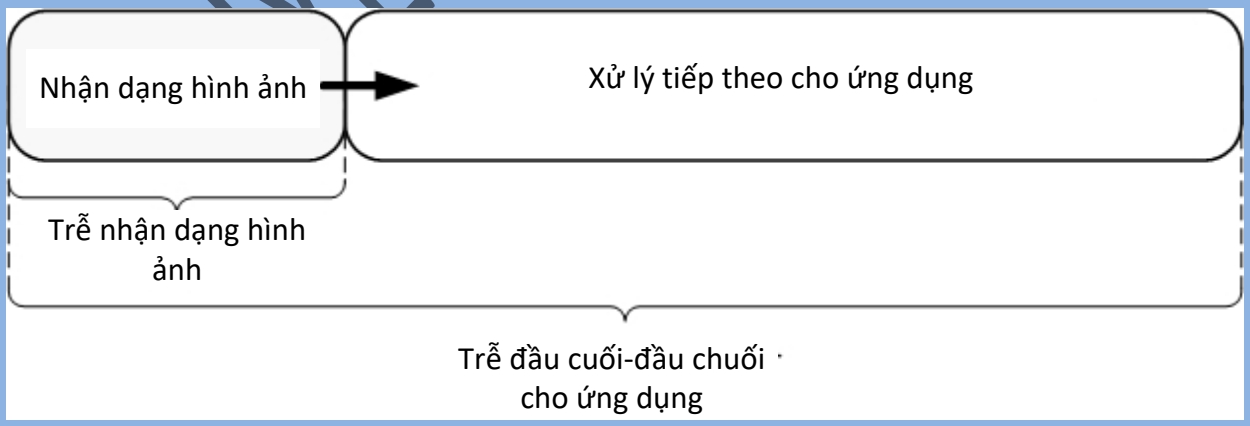
Điều kện kết thúc.Các đối tượng trong hình ảnh hoặc video đầu vào được nhận dạng, độ chính xác và độ trễ nhận dạng cần được đảm bảo.

Nhiệm vụ nhận dạng hình ảnh có thể được hoàn thành trong điều kiện tài nguyên năng lượng và tính toán sẵn có của UE. Và các các tài nguyên điện toán, truyền thông và năng lượng tiêu thụ trên các điểm cuối AI/ML được tối ưu hóa.

5.1. 3. Các yêu cầu mới tiềm năng cần thiết để hỗ trợ trường hợp sử dụng

"Độ trễ nhận dạng hình ảnh" có thể được định nghĩa là độ trễ từ hình ảnh được chụp đến kết quả nhận dạng hình ảnh được xuất ra ứng dụng người dùng. Theo nguyên tắc phân tích các yêu cầu về độ trễ và tốc độ dữ liệu của nhận dạng hình ảnh phân chia được giới thiệu trong mục 2.21, độ trễ nhận dạng hình ảnh có liên quan đến ứng dụng người dùng mà nhận dạng được sử dụng.

Thị giác máy tính và nhận dạng hình ảnh đã được sử dụng rộng rãi cho nhiều ứng dụng di động quan trọng như nhận dạng đối tượng, nâng cao ảnh, giám sát video thông minh, AR di động, ô tô điều khiển từ xa, điều khiển công nghiệp và robot. Nhận dạng hình ảnh thường là một bước của quy trình xử lý của ứng dụng. Và độ trễ nhận dạng là một phần của độ trễ đầu cuối, như được mô tả trong Hình 5.2.



Hình 5.2.Trễ nhận dạng hình ảnh là một phần của trễ đầu cuối đầu cuối

Ví dụ, nếu kết quả nhận dạng hình ảnh chỉ được sử dụng để nhận dạng đối tượng, ví dụ như nhận dạng đối tượng chưa được biết cho người dùng điện thoại thông minh hoặc tìm kiếm tội phạm trong cơ sở dữ liệu đối với an ninh thông minh (Intelligence Security), thì có thể chấp nhận được rằng việc nhận dạng hình ảnh kết thúc trong vài giây. Nếu kết quả nhận dạng hình ảnh được sử dụng làm đầu vào cho một ứng dụng nhạy cảm với thời gian khác, ví dụ như màn hình AR / chơi game, ô tô điều khiển từ xa, điều khiển công nghiệp và robot, độ trễ nghiêm ngặt hơn nhiều sẽ được yêu cầu. Dựa trên các yêu cầu về độ trễ đầu cuối của các ứng dụng, yêu cầu về độ trễ nhận dạng hình ảnh có thể được rút ra, như được liệt kê trong Bảng 5.1.

ng dụng ười dùng	**Độ trễ: maximum**			**Tốc độ dữ liệu UL trải kinh nghiệm của người dùng**	
	Trễ đầu cuối-đầu cuối	**Trễ nhận dạng hình ảnh**	**Trễ tải lên số liệu trung gian**	**AlexNet (H. 2.2, chú thích 4)**	**VGG-16 (H. 2.3, chú thích 4)**
Nhận dạng đối tượng một lần trên điện thoại thông minh	Vài giây	~1s	~100ms	1.6~21.6Mbit/s	8~240Mbit/s
Nhận dạng người trong hệ thống giám sát an ninh	Vài giây	~1s	~100ms	1.6~21.6Mbit/s	8~240Mbit/s
Nâng cấp ảnh trên điện thoại thông minh	Vài giây	~1s	~100ms	1.6~21.6Mbit/s	8~240Mbit/s
Nhận dạng video	Vài giây	33ms@30FPS	~10ms	16~216Mbit/s	80Mbit/s~2.4Gbit/s
Màn hình AR / chơi game	7~15ms (chú thích 1)	<5ms	2ms	80Mbit/s~1.08Gbit/s	0.4~12Gbit/s
Lái xe từ xa	10ms (chú thích 2)	<5ms	2ms	80Mbit/s~1.08Gbit/s	0.4~12Gbit/s
Robot điều khiển từ xa	10~100ms (chú thích 3)	<5ms	2ms	80Mbit/s~1.08Gbit/s	0.4~12Gbit/s

Chú thích 1: độ trễ chuyển động từ VR đến photon nằm trong khoảng 5-15ms.
Chú thích 2: độ trễ một chiều cần thiết để lái xe từ xa là 5ms. Độ trễ khứ hồi được giả định là 10ms.
Chú thích 3: độ trễ đầu cuối-đầu cuối cần thiết cho robot điều khiển từ xa hoạt động bằng video là 10 ~ 100ms.
Chú thích 4: Như được liệt kê trong bảng 2.1 và bảng 2.2, kích thước dữ liệu trung gian cho các điểm phân tách của AlexNet và VGG-16 lần lượt là 0,02 ~ 0,27MByte và 0,1 ~ 3MByte.

5.2. Nhận dạng phương tiện nâng cao: học sâu dựa trên các ứng dụng thị giác (Deep Learning Based On Vision Application)

5.2.1. Tổng quan

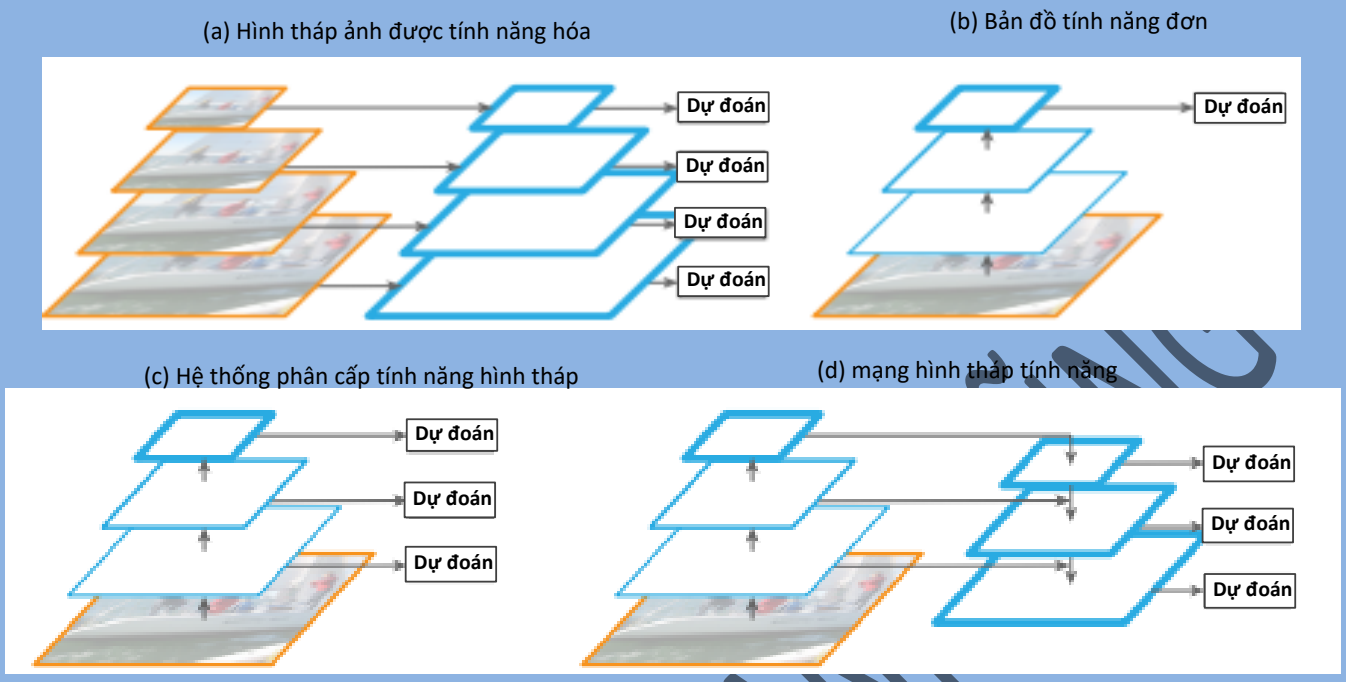
Nhận dạng đối tượng tại các tỷ lệ (kích thước) (tiếng Anh: Scale) rất khác nhau là một thách thức cơ bản trong thị giác máy tính (Computer Vision). Các kim tự tháp tính năng **(**Feature Pyramid**)** còn được gọi là kim tự tháp ảnh được tính năng hóa (Featerized Image Pyramid) tạo thành cơ sở của một giải pháp tiêu chuẩn như trên hình 5.3 (a). Các kim tự tháp là bất biến tỷ lệ/ kích thước (Scale Invariant) theo nghĩa là sự thay đổi tỷ lệ/ kích thước của đối tượng được bù

đắp bằng cách dịch mức của nó trong kim tự tháp. Thuộc tính này cho phép một mô hình phát hiện đối tượng trên một dải rộng các tỷ lệ bằng cách quét mô hình trên cả hai: các vị trí và các mức của hình tháp.

Kim tự tháp ảnh được tính năng hóa được sử dụng nhiều trong thời đại các tính năng được thiết kế thủ công. Đối với các nhiệm vụ nhận dạng, các tính năng được thiết kế thủ công phần lớn đã được thay đổi bằng các tính năng được tính toán bởi các mạng chập sâu (ConvNet). ConvNet cũng bền vững hơn đối với sự thay đổi tỷ lệ (kích thước) khi các tính năng được tính toán trên một đầu vào đơn (hình 5.3(b)).Tuy nhiên cho dù có tính bền vững, các kim tự tháp vẫn cần nhận được các kết quả chính xác nhất. Tất cả các mục hàng đầu (Top Entries) trong các thách thức phát hiện (Dection Challenge) của ImageNet và COCO sử dụng kiểm tra đa tỷ lệ (đa kích thước) trên các kim tự tháp hình ảnh được tính năng hóa. Ưu điểm nguyên tắc của tính năng hóa từng mức của một kim tự tháp hình ảnh là nó tạo ra một mô hình tính năng đa tỷ lệ (đa kích thước) trong đó tất cả các mức đều có ngữ nghĩa mạnh mẽ bao gồm cả các mức phân giải cao.

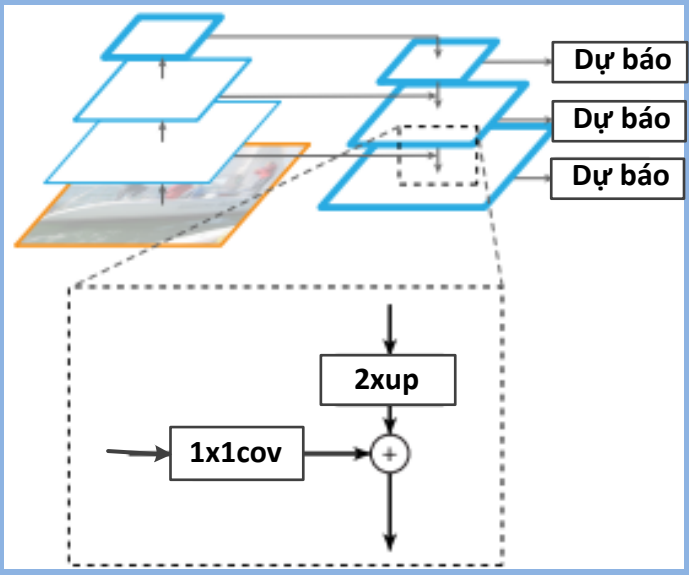
Tuy nhiên tính năng hóa từng mức của kim tự tháp hình ảnh có các hạn chế rõ ràng. Thời gian suy luận tăng đáng kể (bốn lần theo một nghiên cứu) làm cho cách tiếp cận này không thực tế đối với các ứng dụng thực tế. Hơn nữa mạng đào tạo sâu đầu cuối-đầu cuối trên kim tự tháp hình ảnh không khả thi liên quan đến bộ nhớ, vì thế các kim tự tháp hình ảnh chỉ được sử dụng trong thời gian thử nghiệm.

Tuy nhiên các kim tự tháp hình ảnh không phải là cách duy nhất để tính toán một mô hình tính năng đa kích thước (đa tỷ lệ). ConvNet sâu tính toán phân cấp tính năng theo lớp và với các lớp lấy mẫu phụ (Sub-Sampling Layers), phân cấp tính năng có hình dáng kim tự tháp đa kích thước cố hữu (hình 5.3(c)). Phân cấp tạo ra các khoảng trống ngữ nghĩa lớn gây ra nởi các độ sâu khác nhau. Các bản đồ độ phân giải cao có các tính năng mức thấp gây bất lợi cho khả năng tượng trưng đối với nhận biết đối tượng.



Hình 5.3.(a) Sử dụng một kim tự tháp hình ảnh (Image Pyramid) để xây dựng một kim tự tháp tính năng (Feature Pyramid). Các tính năng được tính toán dựa trên từng tỷ lệ (kích thước) của hình ảnh một cách độc lập, sơ đồ này phát hiện chậm. (b) Các hệ thống phát hiện gần đây sử dụng phân cấp các tính năng tỷ lệ (kích thước) đơn để phát hiện nhanh hơn. (c) Tái sử dụng phân cấp tính năng hình tháp được tính toán bởi Convnet như thể là kim tự tháp được tính năng hóa.(d) Mạng kim tự tháp tính năng (FPN: Feature Pyramid Network) nhanh giống như (b) và (c) nhưng chính xác hơn. Trong hình vẽ này, các bản đồ tính năng được biểu thị bằng các đường nét ngoài mầu xanh và các đường dầy hơn biểu thị các tính năng có ngữ nghĩa mạnh hơn.

Mạng kim tự tháp tính năng (FPN: Feature Pyramid Network) trên hình 5.3(d) cho phép lợi dụng hình dạng kim tự tháp của phân cấp tính năng của CinvNet trong khi tạo ra một kim tự tháp có ngữ nghĩa mạnh tại tất cả các tỷ lệ (kích thước). FPN dựa trên một kiến trúc cho phép kết hợp các tính năng ngữ nghĩa mạnh độ phân giải thấp với các tính năng ngữ nghĩa thấp độ phân giải cao thông qua đường đi từ đỉnh-xuống (Top-Down Pathway) và các kết nối bên cạnh (Laterral Conections). Kết quả cho ta một kim tự tháp tính năng giàu ngữ nghĩa và được xây dung nhanh chóng từ một tỷ lệ (kích thước) hình ảnh đầu vào đơn (hình 5.4).



up (upsampling): Lấy mẫu tăng (hay lấy mẫu lại)

Hình 5.4.Minh hoạ khối cơ sở của kết nối bên cạnh và đường từ đỉnh xuống (Top-Down Pathway) được hòa nhập bởi phép cộng

5.2.2. Kịch bản

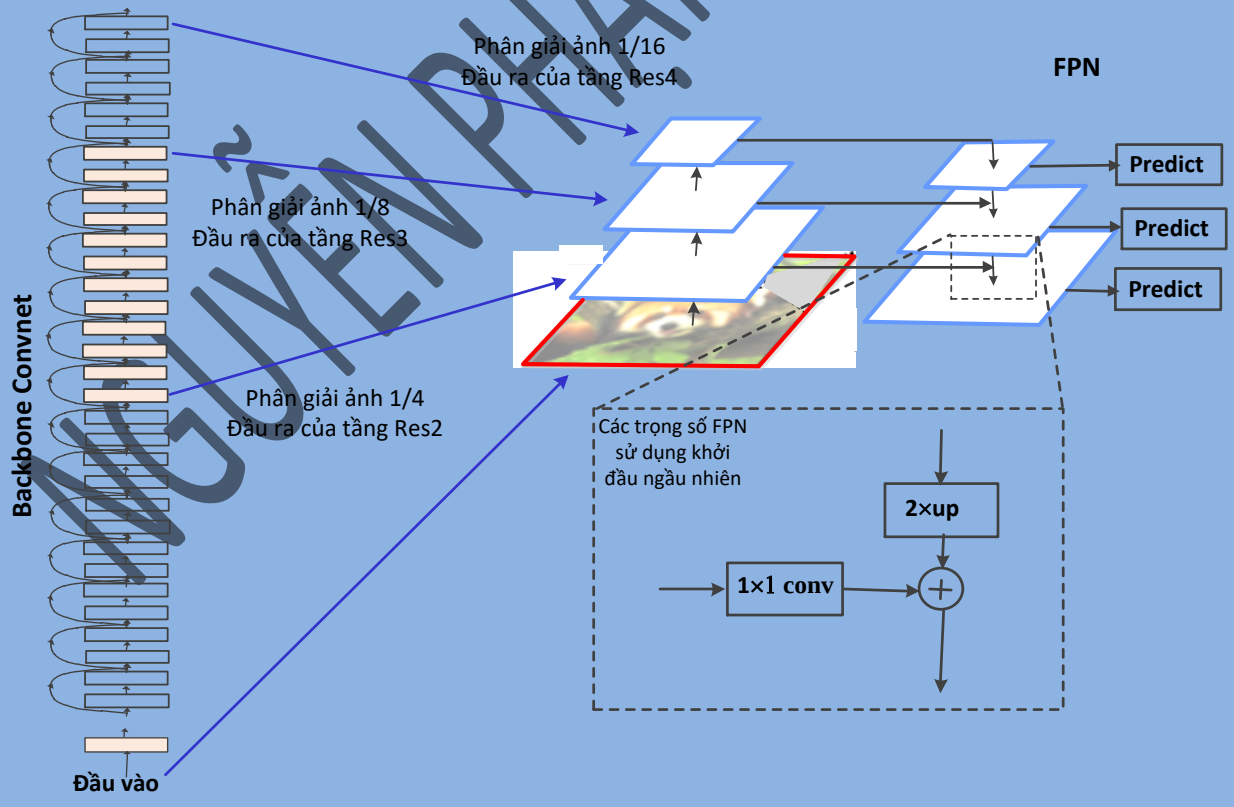
Ta xét một trường ứng dụng sau đây. Một khách du lịch đang lang thang quanh một thành phố và khám phá các điểm tham quan và danh lam thắng cảnh của thành phố. Người dùng nhìn thấy một đối tượng đẹp và cô ấy quyết định quay video về đối tượng đó. Ứng dụng sử dụng các thuật toán học sâu để xử lý video và xác định đối tượng quan tâm và cung cấp thông tin lịch sử về video đó cho người dùng. Hơn nữa, ứng dụng sử dụng deep learning để tái tạo mô hình 3D của đối tượng quan tâm bằng cách sử dụng video 2D đã chụp.

Để làm ví dụ, ta ngghiên cứu các phương pháp phát hiện đối tượng dựa trên Mạng kim tự tháp tính năng (FPN: Feature Paramid Network)). Các mạng này thường bao gồm một đường trục Blackbone), FPN và một đầu thực hiện suy luận theo nhiệm vụ cụ thể. FPN xử lý các hình ảnh đầu vào ở các tỷ lệ (kích thước) khác nhau để cho phép phát hiện các tính năng tỷ lệ nhỏ và tỷ lệ lớn. Ví dụ, đầu có thể phân đoạn các đối tượng, suy ra một hộp giới hạn cho các đối tượng hoặc phân loại các đối tượng.

Đường trục FPN tạo thành phần phức tạp nhất của mạng và rất thích hợp cho giảm tải tải đến biên / đám mây. Đường trục là một phần chung của một loạt các mạng có thể thực hiện các nhiệm vụ khác nhau. Các bản đồ tính năng được tạo ra sau đó có thể được gửi trở lại UE để suy luận theo nhiệm vụ cụ thể.

Phân tích kiến trúc mạng được thể hiện trong hình 5.5 dưới đây.

Như thể hiện trong hình 5.5, kiến trúc CNN cổ điển được sử dụng làm cốt lõi. FPN được sử dụng để trích xuất các đặc điểm (Feature) ở các tỷ lệ khác nhau của hình ảnh, làm cho nó bất biến theo tỷ lệ. Các nhiệm vụ dự đoán (Predict) tạo thành phần đầu mạng. Bằng cách cắm các đầu mạng khác nhau, các nhiệm vụ AI khác nhau có thể được thực hiện. Điều này làm cho mạng trở thành Mạng đa nhiệm (MTN: Multi-Task Network;). Ví dụ: Mạng đề xuất khu vực (RPN: Regional Proposal Network ; Mạng hoàn toàn tích chập cho phép dự đoán ranh giới đối tượng và điểm số đối tượng (Objectness Score) tại mọi vị trí) có thể được thêm vào để phát hiện và các đối tượng khung trong chuỗi đầu vào bằng cách xuất ra các hộp giới hạn. Các Đầu Nhiệm vụ cụ thể (Task Specific Head) khác có thể được thêm vào để phát hiện con người và tư thế, phân loại đối tượng, theo dõi đối tượng, v.v.



Hình 5.5. Thí dụ mạng đa nhiệm (Multi Task Network)

Từ hình 5.5 ta thấy, sau khi các tính năng được lấy ra bởi đường trục (Backbone) (hình kim tự tháp theo đường từ dưới lên đỉnh phía trái), mạng kim tự tháp tính năng (hình kim tự tháp theo đường từ đỉnh xuống dưới) cải thiện độ phân dải của các tính năng mức cao theo đường từ đỉnh xuống và hòa nhập chúng với các tính năng mức thấp bằng cách thực hiện lấy mẫu tăng (hay lấy mẫu lại) dựa trên phương pháp nội suy chẳng hạn và hòa nhập bằng phép cộng.

5.2.2. Quá trình diễn ra

Điều kiện ban đầu. Người dùng muốn nhận thông tin tức thời và cấu trúc lại để nâng cao trải nghiệm của họ. Thiết bị của người dùng hoạt động bằng pin.

Các luồng dịch vụ diễn ra như sau:

1. Người dùng mở ứng dụng máy ảnh của họ và bắt đầu quay video
2. Ứng dụng xử lý trước video để chuẩn bị cho việc suy luận
3. Ứng dụng truyền các tính năng (Feature) và / hoặc video được trích xuất đến biên/đám mây để xử lý.
4. Mạng thực hiện suy luận được phân chia (Split Infnerence) (ví dụ: chỉ chạy Brackbone) và truyền kết quả trở lại client
5. Ứng dụng chạy suy luận theo nhiệm vụ cụ thể để giải quyết nhiệm vụ cụ thể quan tâm (ví dụ: phát hiện đối tượng, theo dõi,...)
6. Ứng dụng sử dụng các nhãn (Label) và loại (Class) được suy ra để nâng cao chế độ xem của người dùng

Điều kện kết thúc. Người dùng nhận được thông tin nâng cao được trích xuất từ video về đối tượng quan tâm mà người dùng đang quay video.

5.2.3. Các yêu cầu mới tiềm năng cần thiết để hỗ trợ trường hợp sử dụng

Các yêu cầu KPI tiềm năng để hỗ trợ trường hợp sử dụng như được đưa ra trong Bảng 5.2, dựa trên hai ví dụ về mô hình/thuật toán phát hiện đối tượng:

- Độ trễ phát trực tuyến đường lên không cao hơn [100-200ms] và người dùng trải nghiệm tốc độ dữ liệu UL là [100-1500] kbit / s;
- Độ trễ phát trực tuyến đường xuống không cao hơn [100-500ms] và người dùng trải nghiệm tốc độ dữ liệu DL là [32-150] Mbit / s.

Bảng 5.2. Nhận dạng: phân tích độ trễ và tốc độ dữ liệu UL / DL mà người dùng đã trải nghiệm

Nhiệm vụ công nhận	Trễ : maximum (chú thích 4)	Tốc độ dữ liệu DL trải nghiệm của người dùng		Tốc độ dữ liệu DL trải nghiệm của người dùng	
		Faster R-CNN (chú thích 1)	YOLOv3 (chú thích 2)	Faster R-CNN	YOLOv3
Phát trực tuyến (Streaming) đường lên	100-200ms			100-1000 kbit/s	200-1500 kbit/s
Suy luận FPN chung	100-500ms	FPN: 4-10fps Sum(Pi)~1MB/khung 32-100Mbit/s không nén (chú thích 3) Hệ số nén 10~100	Nhiều thang đo (tương tự FPN): 1.5 MB bản đồ tính năng/khung 40-150 Mbit/s không nén Hệ số nén 10~100		
Phân loại đối tượng	20-50ms	Thực hiện trên UE	Thực hiện trên UE		
Phát hiện hộp giới hạn	20-50ms	Thực hiện trên UE	Thực hiện trên UE		
Theo dõi đối tượng	50-150ms	Thực hiện trên UE	Thực hiện trên UE		
Truy xuất thông tin nâng cao		vài kByte trên một yêu cầu		Vài kByte trên một yêu cầu	
Kết xuất (hiển thị) lớp phủ	10ms	Thực hiện trên UE		Thực hiện trên UE	

Chú thích 1: Faster R-CNN sử dụng kích thước hình ảnh đầuvào 3x224x224. Video được thu nhỏ trên UE xuống độ phân giải mục tiêu đó và sau đó được nén (ví dụ: sử dụng HEVC) và truyền trực tuyến đến biên để xử lý thêm.
Chú thích 2: YOLOv3 sử dụng kích thước hình ảnh đầu vào là 3x416x416. Video đã quay được giảm tỷ lệ trên UE xuống độ phân giải mục tiêu và nén trước khi phát trực tuyến đến biên (edge).
Chú thích 3: Faster R-CNN sử dụng FPN với ResNet 101 làm xương sống.; do đó dẫn đến các bản đồ tính năng {P2=(256x56x56), P3=(256x28x2), P4=(256x14x14), P5=(256x7x7)}.
Chú thích 4: các ước tính độ trễ giả định độ trễ tổng thể khoảng 1 giây từ khi người dùng hướng đến một đối tượng cho đến khi thông tin lớp phủ (Overlay Information) được hiển thị cho người dùng.

5.3. Nâng cao chất lượng phương tiện truyền thông: Nâng cấp chất lượng luồng video trực tuyến (Video Streaming Upgrade)

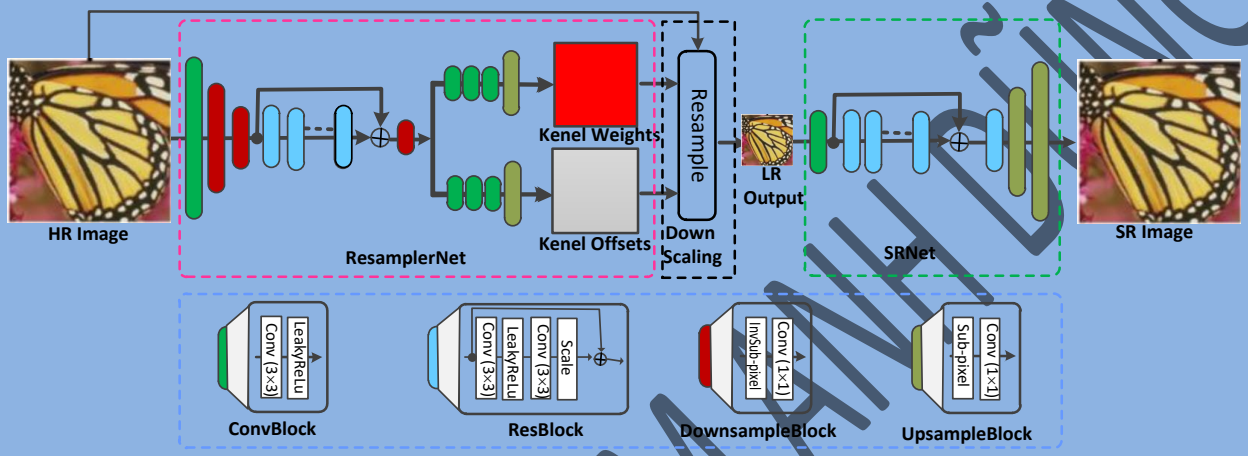
5.3.1 . Kịch bản

Ta xét một trường hợp cụ thể sau đây . Một người dùng chơi VR game với kính thực tế ảo (VR Headset). Game được chiết xuất (Render) trên đám mây và phát luồng ngược về UE. Người dùng muốn nhận được trải nghiệm chơi game mạnh đòi hỏi viedo chất lượng cao, chẳng 8K per eye at 120 frames per second (8K mỗi mắt tại tốc độ 120 khung hình trên giây) (8K có nghĩa là vào khoảng 8000 pixel theo chiều ngang của khung hình, chẳng hạn phân giải 7680×4320).

Game server chỉ có thể tạo ra dữ liệu video 4K (chẳng hạn phân giải 4096×2160 đối với 4K UHD) do hạn chế của phần cứng, tải và nối mạng.

AI được sử dụng để nâng cấp nội dung 4K thành nội dung 16K (chẳng hạn phân giải 15360×8640 đối với UHD) để có trải nghiệm người dùng tốt hơn.

Hình 5.6 dưới đây cho thấy một ví dụ về một mạng như vậy.



HR Image (High Resolution Image): ảnh phân giải cao; LR Image (Low Resolution Image): ảnh phân giải thấp; SR Image (Super Resolution Image): ảnh siêu phân giải; ResamplerNet: mạng lấy mẫu lại; SRNet: mạng siêu phân giải; Kenel Weight: trọng số hạt nhân, Kenel Offset: bù trừ hạt nhân; Down Scaling: giảm tỷ lệ; LR Output: đầu ra phân giải thấp; ConvBlock (Convolution Block): khối tích chập; ResBlock (Residual Block): khối dư; DownsmpleBlock: khối giảm mẫu; UpsampleBlock: khối tăng mẫu.

Các hàm trong các khối như sau:

ConvBlock bao gồm: Conv(3×3), LeakyReLu; ResBlock bao gồm: Conv(3×3), LeakyReLu, Conv(3×3), Scale;
DownBlock bao gồm: InvSub-pixel, Conv(1×1); UpsampleBlock bao gồm: Sub-pixel, Conv(1×1);

Hình 5.6. Ví dụ về một mạng nâng cấp phát luồng video

Video Độ phân giải thấp được truyền đến UE, UE sẽ xử lý video để suy ra phiên bản độ phân giải cao. Các phần lấy mẫu xuống (Down Sampling) và lấy mẫu lên (Up Sampling) của mạng được khớp để tạo ra kết quả tốt nhất. Bất kỳ bản cập nhật nào đối với phần lấy mẫu xuống của mạng sẽ yêu cầu cập nhật ở phía UE đến phần lấy mẫu lên của mạng. Ngoài phiên bản LR của video, trọng số mô hình và cập nhật cấu hình topo có thể cần được gửi đến UE.

Các mô hình được đào tạo trước được tối ưu hóa cho loại nội dung (ví dụ: thể thao, phim hoạt hình, phim, ...) và hệ số tỷ lệ (Scale Factor). Khi người dùng chuyển sang một phần nội dung khác, thiết bị sẽ kiểm tra xem trọng số mô hình được đào tạo trước tương ứng đã được tải xuống trong bộ nhớ đệm hay chưa. Nếu không có sẵn, UE sẽ tải xuống trọng số mô hình được đào tạo trước cho phần nội dung đã chọn và dựa trên hệ số nâng cấp mong muốn. Nội dung mới được hiển thị cho người dùng trong vòng chưa đầy 3 giây kể từ khi người dùng chọn nội dung đó.

5.3.2. Quá trình diễn ra

Toàn bộ quá trình diễn ra cho kịch bản này như sau.

Điều kiện ban đầu. Máy chủ chơi game từ xa tạo siêu dữ liệu luồng cùng với luồng được trích xuất bằng cách chạy bộ mã hóa tự động AI (AI Autocoder) trên nội dung được chụp ban đầu.

Các luồng dịch vụ diễn ra như sau:

1. Người dùng bắt đầu phiên chơi game VR trên đám mây trên HMD (Head Mounted Display: màn hình đeo trên đầu) của họ;
2. Trò chơi được khởi chạy trên máy chủ đám mây và trò chơi có thể bắt đầu;
3. Máy chủ đám mây hiển thị và ghi lại nội dung và giảm tỷ lệ (Downscale) xuống video 4K;
4. Máy chủ đám mây cũng chạy DNN để tạo ra luồng siêu dữ liệu (Metadata Stream) sẽ được sử dụng để nâng cấp quy mô/ tỷ lệ (Upscaling)
5. UE sử dụng mạng DNN ngược để tăng tỷ lệ (Upscale) luồng 4K đã nhận lên 16K. Đầu vào cho DNN là video 4K và luồng siêu dữ liệu.
6. UE hiển thị chế độ xem 16K chất lượng cao trên HMD.

Điều kện kết thúc. Người dùng tận hưởng trải nghiệm VR chất lượng cao.

5.3.3. Các yêu cầu mới tiềm năng cần thiết để hỗ trợ trường hợp sử dụng

Các yêu cầu KPI tiềm năng để hỗ trợ trường hợp sử dụng là:.
Hệ thống 5G/6G sẽ hỗ trợ cung cấp trọng số mô hình được đào tạo trước và cập nhật trọng lượng với độ trễ không quá 3 giây.
Bảng 5.3 dưới đây cung cấp ước tính về kích thước trọng số của mô hình.

Bảng 5.3. Kích thước trọng số của mô hình và tốc độ dữ liệu

Tên mô hình (chú thích 2)	Kích thước trọng số (chú thích 4)	Tóc độ số liệu (chú thích 3)
CAR 4x	Lấy mẫu giảm: ~40MB Lấy mẫu tăng: ~170MB	Tải xuống mô hình lấy mẫu tăng: 450Mbit/s (Tải xuống trong khoảng thời gian 3 giây)
SR-GAN 4x	Lấy mẫu giảm: N/A (bicubic) Lấy mẫu tăng: ~6MB	Tải xuống mô hình lấy mẫu tăng: 12Mbit/s (Tải xuống trong khoảng thời gian 3 giây)
SRResNet 4x	Lấy mẫu giảm: N/A (bicubic) Lấy mẫu tăng: ~6MB	Tải xuống mô hình lấy mẫu tăng: 12Mbit/s (Tải xuống trong khoảng thời gian 3 giây)
Chú thích 2: Kích thước của trọng số của các mô hình được đào tạo trước phụ thuộc vào bản chất của nội dung và hệ số tỷ lệ của video. Chú thích 3: Tải xuống trọng lượng mô hình được kích hoạt bởi sự thay đổi nội dung của người dùng và do đó dẫn đến lưu lượng truy cập bùng nổ về bản chất với thời gian nghỉ tương đối dài. Chú thích 4: tất cả các kích thước trọng sô mô hình là từ các mô hình PyTorch được đào tạo trước.		

5.4. Điều khiển được phân chia cho robot

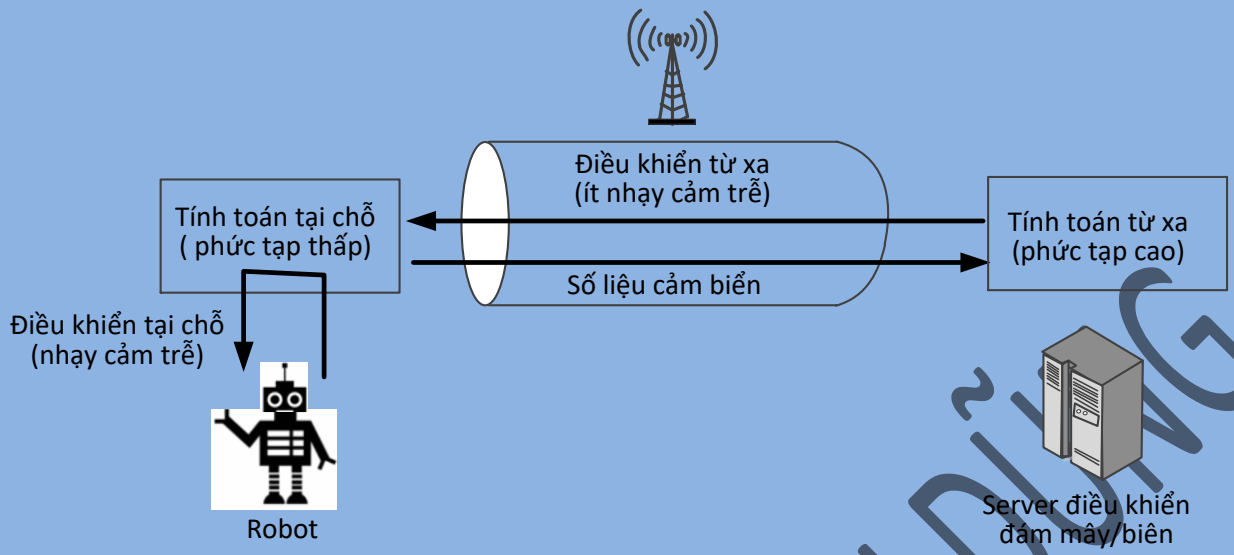
5.4.1. Mô tả

Robot di động đã và đang đóng một vai trò ngày càng quan trọng trong một số tình huống, ví dụ như nhà kho, cứu hộ thảm họa và nhà máy thông minh nhờ tính cơ động cao của chúng. Robot di động cần hoạt động trong một môi trường luôn thay đổi, do đó cần thực hiện cảm biến, lập kế hoạch và điều khiển nhanh chóng và đáng tin cậy. Nếu tính toán tương ứng được thực hiện trên bo mạch trong robot, nó sẽ yêu cầu các tính toán chuyên sâu dẫn đến tăng yêu cầu về khả năng tính toán và mức tiêu thụ điện năng. Tuy nhiên, yếu tố hình thức nhẹ luôn là yêu cầu đối với robot di động làm việc trong môi trường thực tế, điều này khiến robot không được trang bị một số lượng lớn CPU / GPU (Central Processing Unit/Graphic Processing Unit: đơn vị xử lý trung tâm/đơn vị xử lý đồ họa) và pin dung lượng lớn. Như ví dụ được đưa ra trong một nghiên cứu, một robot bốn chân tiên tiến có sẵn trên thị trường, mang 3kg pin có năng lượng khoảng 650Wh, trong khi GPU cao cấp tiêu thụ hơn 250W năng lượng, ảnh hưởng đáng kể đến tuổi thọ pin nếu sức mạnh tính toán như vậy được nhúng ...

Việc giảm tải các tính toán từ robot lên đám mây đã được nghiên cứu trong nhiều tài liệu. Trong khi đó, dựa vào dữ liệu hoặc code (đoạn chương trình) từ mạng để hỗ trợ hoạt động của robot, các nhà thiết kế robot di động tự động phải xem xét các tình huống mà robot phải bao gồm khả năng xử lý cục bộ cho các phản hồi có độ trễ thấp trong thời gian chất lượng truy cập mạng thay đổi kém hơn.

Hệ thống kết quả khác với hệ thống robot điều khiển hoàn toàn từ xa được mô tả trong [5], trong đó việc lập kế hoạch và điều khiển được thực hiện bằng điện toán đám mây và robot chỉ báo cáo dữ liệu cảm biến (bao gồm video) và nhận lệnh điều khiển. Vì điện toán đám mây hoàn chỉnh khó có thể đáp ứng yêu cầu độ trễ của vòng điều khiển phản hồi cấp ms của một số loại robot di động, ví dụ như robot có chân, nên việc điều khiển phân chia robot di động là một giải pháp dễ chịu trong trường hợp này.

[27] giới thiệu một robot điều khiển thăng bằng toàn bộ cơ thể được phân chia qua mạng 5G. Suy luận AI để điều khiển có thể được phân chia giữa robot và máy chủ đám mây: Như thể hiện trong Hình 5.5, phần phức tạp nhưng ít nhậy cảm trễ sẽ được chuyển tải đến tính toán từ xa trong máy chủ điều khiển đám mây hoặc biên. Phần độ phức tạp thấp chứa các điều khoản phản hồi lỗi và nhậy cảm trễcó thể được thực hiện hiệu quả bằng tính toán tại chỗ trong robot. Nếu rô bốt không nhận được điều khiển tối ưu từ "phần điều khiển từ xa" từ máy chủ điều khiển đám mây / biên (cloud/edge control server) do trễ truyền thông hoặc mất gói, nó có thể xấp xỉ hóa "phần điều khiển từ xa" bằng cách sử dụng ma trận phản hồi được tính toán trước đã nhận được trước đó. Và trong một khoảng thời gian nhất định, xấp xỉ hóa vẫn sẽ cho phép robot thực hiện kiểm soát phản hồi cho các nhiệm vụ bằng các xấp xỉ hóa và đảm bảo rằng robot vẫn có thể hoạt động.

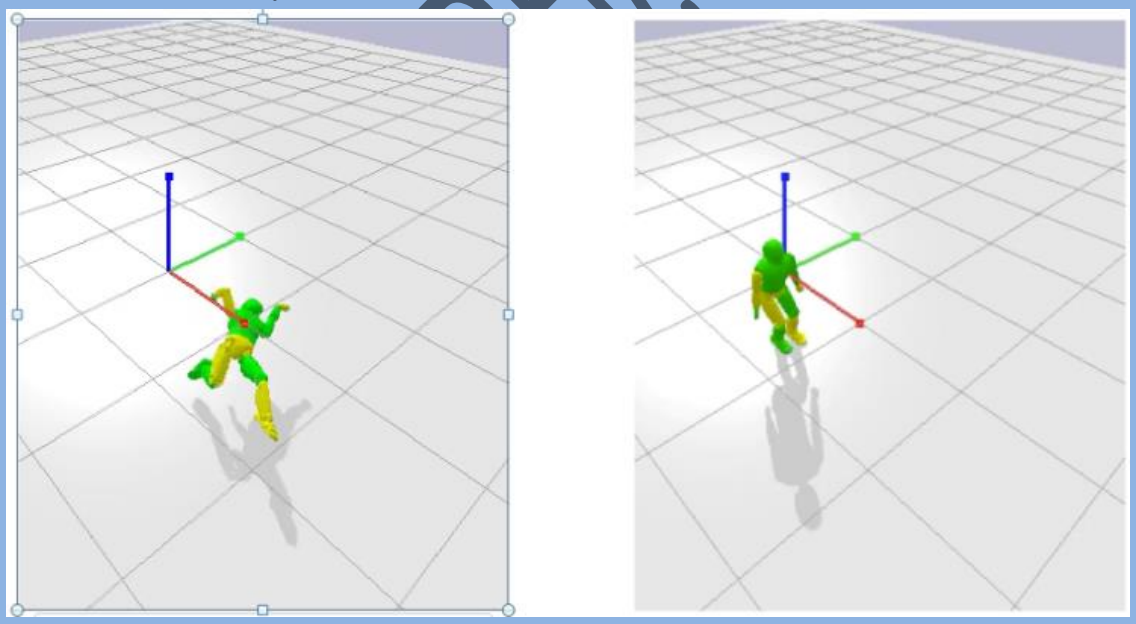


Hình 5.5. Điều khiển được phân chia cho robot có chân trên mạng 5G

Kết quả trong [27] cho thấy, trong trường hợp robot được điều khiển hoàn toàn bởi máy chủ đám mây, robot không thể hoàn thành nhiệm vụ đi bộ nếu độ trễ khứ hồi (Round Trip Latency) lớn hơn 3ms (từ gửi dữ liệu cảm biến đến nhận lệnh điều khiển, bao gồm cả xử lý ở đám mây / biên). Do các lệnh điều khiển bị trễ, robot sẽ bị đổ (như trong Hình 5.6(a). Tuy nhiên, nếu điều khiển phân chia được sử dụng, độ trễ khứ hồi 25ms trong trường hợp xấu hơn có thể được duy trì và robot vẫn có thể thực hiện nhiệm vụ đi bộ (như thể hiện trong Hình 5.6(b).

(a) Điều khiển từ xa 5G không có điều khiển tại chỗ

(b) Điều khiển từ xa 5G có điều khiển tại chỗ tại root



Hình 5.6. Hiệu suất mô phỏng của điều khiển thăng bằng toàn bộ cơ thể rô-bốt trên mạng 5G với độ trễ khứ hồi 25ms

5.4.2. Quá trình diễn ra

Quá trình trên hình 5.5 diễn ra như sau.

Điều kiện ban đầu. Các điểm cuối AI/ML liên quan (ví dụ: UE (robot), AI/ML đám mây/máy chủ biên) chạy các ứng dụng cung cấp khả năng suy luận mô hình AI/ML cho nhiệm vụ điều khiển robot và hỗ trợ hoạt động điều khiển robot phân chia.

Các luồng dịch vụ diễn ra như sau:

1. UE (robot) được kết nối với server điều khiển đám mây / biên thông qua mạng 5G;
2. Việc phân chia các hoạt động điều khiển ở phía robot và phía server điều khiển đám mây / biên cho nhiệm vụ điều khiển robot được xác định bởi robot hoặc mạng;
3. Trong điều kiện chế độ phân chia và điểm phân chia được xác định, robot thực hiện tính toán tại chỗ dựa trên dữ liệu cảm biến thu thập được và gửi các dữ liệu cảm biến này đến server điều khiển đám mây / biên. Server điều khiển đám mây / biên thực hiện tính toán từ xa và cung các đầu ra trở lại robot;
4. Robot điều khiển chuyển động của nó dựa trên kết hợp đầu ra của các tính toán điều khiển tại chỗ và từ xa;
5. Bắt đầu với bước 3) với nhiều thao tác điều khiển hơn, cho đến khi nhiệm vụ điều khiển robot kết thúc.

2. **Kết quả**. Robot nhận điều khiển tại chỗ và từ xa với độ chính xác và độ trễ cần thiết, để hoàn thành các nhiệm vụ di chuyển, có nghĩa là nhiệm vụ cân bằng và nhiệm vụ đi bộ.

5.4.3. Các yêu cầu mới tiềm năng cần thiết để hỗ trợ trường hợp sử dụng

Nếu mọi thứ được thực hiện trên phần biên (Edge), robot cần gửi dữ liệu cảm biến 592B trên mỗi chu kỳ điều khiển (mỗi mili giây) và nhận 200B trên mỗi chu kỳ điều khiển từ "bộ điều khiển từ xa", dẫn đến tốc độ dữ liệu UL là 4.7Mbit / s và tốc độ dữ liệu DL là 1.6Mbit / s. Tuy nhiên, độ trễ khứ hồi tối đa bị giới hạn ở 3ms và độ trễ một chiều để tải xuống "phần điều khiển từ xa" cần được giới hạn ở 1ms. Nếu tuân theo chiến lược phân chia, việc tải xuống "phần điều khiển từ xa" từ máy chủ điều khiển đám mây / biên yêu cầu tải xuống dữ liệu bùng nổ 40kB trên mỗi chu kỳ điều khiển (mỗi mili giây) vì cần thêm thông tin để đảm bảo bộ điều khiển cục bộ có thể tiếp quản trong trường hợp có độ trễ không mong muốn, dẫn đến tốc độ dữ liệu DL của người dùng là 320Mbit / s. Trong trường hợp đó, độ trễ khứ hồi tối đa là 25ms có thể được chấp nhận giữa mỗi chu kỳ điều khiển và độ trễ một chiều để tải xuống "phần điều khiển từ xa" có thể được nới lỏng xuống còn 12ms.

Điều này ngụ ý sự đánh đổi giữa tốc độ dữ liệu DL và độ trễ: So với toàn quyền kiểm soát ở biên, chế độ điều khiển phân chia yêu cầu tốc độ dữ liệu DL có trải nghiệm cao hơn của người dùng, nhưng nới lỏng yêu cầu độ trễ nghiêm ngặt.

Bảng 5.4 cung cấp các yêu cầu về tốc độ dữ liệu và độ trễ đối với điều khiển robot.

Bảng 5.4 Các yêu cầu về tốc độ dữ liệu và độ trễ đối với điều khiển rô-bốt.

Chế độ điều khiển	Tốc độ dữ liệu UL trải nghiệm của người dùng cho tải lên dữ liệu cảm biến	Kích thước của "Phần điều khiển từ xa": tối đa	Tốc độ dữ liệu DL trải nghiệm của người dùng để tải xuống "Phần điều khiển từ xa"	Độ trễ khi tải xuống "Phần điều khiển từ xa": tối đa
Điều khiển hoàn toàn tại biên	4,7Mbit/s	200Bytes	1,6Mbit/s	1ms
Điều khiển phân chia	4,7Mbit/s	40kBytes	320Mbit/s	12ms

5.5. Các hoạt động tính toán phân chia chuyển giao mô hình đặc thù phiên

5.5.1. Ngữ cảnh

Một UE, để đạt được kết quả cho người dùng, sử dụng tính toán phân chia (Tính toán phân chia là để giảm tải nhiệm vụ tính toán chuyên sâu giữa UE và mạng). Các nhiệm vụ tính toán chuyên sâu (học máy, tính toán phức tạp sử dụng dữ liệu đầu vào và mô hình, v.v.) có thể được giảm tải hoàn toàn hoặc một phần. Trường hợp sử dụng này xem xét một mục đích sử dụng cụ thể - hiển thị thực tế tăng cường trong kính thực tế ảo với tài nguyên tính toán khiêm tốn. Quyết định làm thế nào để phân chia nhiệm vụ tính toán giữa UE và các tài nguyên tính toán khác có thể phụ thuộc vào các điều kiện của mạng truyền thông và các tài nguyên tính toán có sẵn trong UE.

5.5.2. Quá trình diễn ra

Điều kiện ban đầu. Alice có kính Thực tế tăng cường, một UE với sức mạnh tính toán hạn chế. Cô rời khỏi một chiếc xe buýt và đứng ở trạm xe buýt, nơi đằng sau một màn hình quảng cáo lớn, một gNB được lắp đặt. Kính của Alice có thể truy cập thông qua điểm truy cập. Cô ấy tìm cách tăng cường tầm nhìn của mình về thành phố bằng các chỉ đường và chú thích (giờ mở cửa, lịch sử địa phương, mô tả doanh nghiệp, v.v.). Tăng cường cảnh trực quan của thành phố trong thời gian thực là một nhiệm vụ chuyên sâu về tính toán, được thực hiện bởi một mô hình được phát triển thông qua ML. Mô hình có hai điểm phân chia ứng cử, mỗi điểm phân chia ứng cử có khối lượng công việc và yêu cầu truyền thông khác nhau được (được thể hiện hiển thị trong phần dưới). Chiến lược phân chia tính toán này đã được cài đặt trong UE và server ứng dụng để điểm phân chia có thể được điều chỉnh động dựa trên sự thay đổi của hiệu suất truyền thông và / hoặc khả năng của UE. Server ứng dụng có thể nằm trong miền MNO (MNO: Mobile Network Operator: nhà mạng di động,tức là một ứng dụng đáng tin cậy) hoặc bên ngoài miền MNO (tức là một ứng dụng phía thứ ba được xác thực). Trường hợp sử dụng này không xem xét cách xác

định chiến lược phân chia, liệu điều này có thể được thực hiện một cách tự động hay chiến lược có hình thức nào.

Bảng 5.5 liệt kê khối lượng công việc và yêu cầu truyền thông cho các điểm phân chia.

Bảng 5.5. Khối lượng công việc và yêu cầu truyền thông cho các điểm phân chia

	Tốc độ dữ liệu đầu ra UL (Mbit/s)	**Tải tính toán trong UE**
Điểm phân chia ứng cử 1	120	Thấp
Điểm phân chia ứng cử 2	24	Cao

Kính có khả năng tính toán hạn chế và khả năng này thay đổi theo thời gian. Phương tiện để xác định trạng thái hiện tại (thông tin AI-ML) của UE có sẵn cho mạng (thông qua lớp ứng dụng). Điều kiện được xác định ban đầu là UE có khả năng hỗ trợ một trong hai điểm phân chia ứng cử 1 hoặc 2.

Tài nguyên truyền thông mạng rất lớn, nó được quyết định bởi dịch vụ thực tế tăng cường để áp dụng điểm phân chia ứng cử viên 1, nên tính toán được thực hiện chủ yếu trong mạng, nhận được lượng lớn dữ liệu do kính của cô ấy cung cấp và điều này giúp giảm tính toán trong UE. Lượng lớn dữ liệu được truyền qua luồng QoS với tốc độ dữ liệu đảm bảo (GBR: Garanteed Bit Rate) 200 Mbit / s.

Các luồng dữ liệu diễn ra như sau:

Trường hợp A (điểm phân chia được điều chỉnh dựa trên hiệu năng truyền thông):

Alice đi bộ ra khỏi trạm xe buýt và vùng lân cận của điểm nóng.

Vì Alice đứng cách điểm phát sóng điểm nóng gNB vài mét, Tài nguyên truyền thông không đủ dẫn đến gNB phục vụ không thể giữ luồng QoS với GBR 200 Mbit / s nữa (GBR: Guaranteed Bit Rate: tốc độ bit được đảm bảo). Do đó, điểm quyết định chính sách (có thể ở bất cứ đâu – ta bỏ qua những gì đưa ra quyết định và cách thức) xác định hạ cấp sẽ được thực hiện tại thời điểm A GBR từ 200 Mbit / s xuống 30 Mbit / s và ngay lập tức thông báo cho UE và server ứng dụng của việc hạ cấp này. Chiến lược phân chia tính toán cho ứng dụng AR bây giờ phải được điều chỉnh, tức là thay đổi sang điểm phân chia ứng cử viên 2, trong đó cần phải tính toán nhiều hơn tại chỗ nhưng tốc độ bit cần thiết cho truyền UL giảm xuống còn 24 Mbit / s.

Chiến lược và ràng buộc cho việc phân chia công việc nằm ngoài phạm vi của trường hợp sử dụng này. (Chúng có thể bao gồm, ví dụ: kết quả một phần có thể được gửi đến UE, có thể hoạt động dưới mức tối ưu với tài nguyên giảm, thông tin mô hình có thể được gửi ở dạng mất dữ

liệu / nén vẫn hữu ích, v.v.) Trong mọi trường hợp, một trong những đầu vào quan trọng cho quyết định phân chia công việc là tập hợp các tài nguyên truyền thông hiện có sẵn.

Mạng cung cấp thông tin tài nguyên mạng hiện tại liên quan đến hiệu năng truyền thông từ UE đến mạng mạng bao gồm các tham số QoS mới (GBR = 20Mbit / s), thông tin điều kiện (thời điểm-A) để cập nhật QoS mới, cũng như hiệu năng đầu cuối giữa UE và tài nguyên tính toán (ví dụ: trong Môi trường lưu trữ dịch vụ). Thông tin này được khả dụng (được để lộ) cho 'điểm quyết định chính sách' tính toán phân chia (có thể ở bất kỳ đâu - trong UE, tại biên, tại đám mây, v.v., điều này không liên quan đến trường hợp sử dụng.)

Trường hợp B (điểm phân chia được điều chỉnh dựa trên tong tin về AM/ML của UE):

Ban đầu, điểm phân chia ứng cử 2 được chọn khi khả năng của UE có thể hỗ trợ khối lượng công việc cao.

Thông tin tính toán của UE của Alice được giám sát tại lớp ứng dụng. Khi tài nguyên giao tiếp đủ để hỗ trợ điểm phân chia ứng cử 1, nếu các điều kiện của UE bị suy giảm đủ lớn (ví dụ: do pin cạn kiệt, thiếu bộ nhớ, giảm khả năng tính toán) thì đây sẽ là lý do để chọn điểm phân chia ứng cử 1.

Khi này, điểm quyết định tính toán phân chia điều chỉnh chiến lược tính toán phân chia:

Đối với trường hợp-A: để tránh gián đoạn dịch vụ, điểm quyết định tính toán phân chia chọn điểm phân chia-2 mới trước khi thời gian điểm-A đến. Cách điều này được truyền thông hoặc 'thực thi' nằm ngoài phạm vi của trường hợp sử dụng này và không đề xuất rằng điều này sẽ được tiêu chuẩn hóa.

Đối với trường hợp-B: để đảm bảo trải nghiệm người dùng, điểm quyết định tính toán phân chia chọn điểm phân chia-1 vì trạng thái của UE không đủ để hỗ trợ khối lượng công việc cao nữa.

Thông tin trạng thái UE được giả định là được thu thập thông qua lớp ứng dụng.

Điều kện kết thúc. Alice không nhận thức được sự thay đổi của điểm phân chia mô hình và tiếp tục tận hưởng hiệu năng có thể chấp nhận được khi cô ấy mạo hiểm vào thành phố, ngay cả khi có lẽ nó không tốt bằng khi cô ấy đứng ở trạm xe buýt. Lưu ý rằng trường hợp sử dụng này không kết thúc miễn là Alice tiếp tục sử dụng dịch vụ - vì hiệu suất giao tiếp UE với mạng có thể thay đổi bất cứ lúc nào.

6. PHÂN BỐ VÀ CHIA SẺ MÔ HÌNH /DỮ LIỆU AI/ML TRÊN HỆ THỐNG 5G

6.1. Phân bố mô hình/dữ liệu cho nhận dạng ảnh (AI/ML model distribution for image recognition)

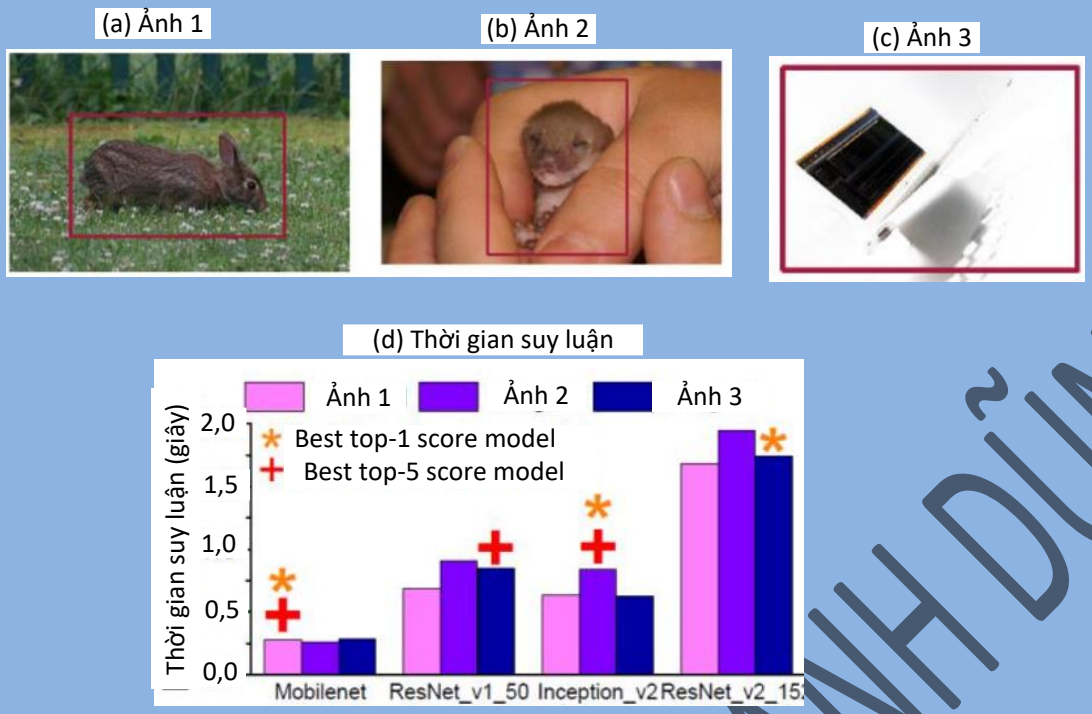
6.1.1. Mô tả

Nhận dạng hình ảnh là một lĩnh vực có sẵn một bộ mô hình AI/MLphong phú được đào tạo trước. Mô hình tối ưu phụ thuộc vào tính năng của hình ảnh / video đầu vào, môi trường và yêu cầu độ chính xác. Mô hình được sử dụng để xử lý thị giác tại thiết bị cần được cập nhật thích ứng cho các đối tượng thị giác, nền, điều kiện ánh sáng, mục đích khác nhau (ví dụ: khôi phục hình ảnh so với phân loại) và thậm chí cả tỷ lệ nén mục tiêu. Mặc dù mô hình tĩnh cũng có thể hoạt động như mặc định trong một số trường hợp, nhưng việc điều chỉnh mô hình cho phù hợp với các điều kiện làm việc khác nhau sẽ giúp cải thiện độ chính xác nhận dạng và trải nghiệm người dùng tốt hơn.

Một ví dụ đã được đưa ra trong [28] cho động lực lựa chọn mô hình tối ưu cho các nhiệm vụ nhận dạng hình ảnh và các môi trường khác nhau. Như thể hiện trong Hình 6.1, 4 mô hình CNN điển hình đã được đánh giá và so sánh cho các nhiệm vụ nhận dạng hình ảnh khác nhau, ví dụ MobileNet_v1_025, ResNet_v1_50 (ResNet với 50 lớp), Inception_v2 và ResNet_v2_152 (ResNet với 152 lớp). Ví dụ này cho thấy rằng mô hình tốt nhất phụ thuộc vào loại hình ảnh đầu vào và yêu cầu nhiệm vụ. Đối với một thiết bị di động cần nhận dạng các nhiều loại hình ảnh và đáp ứng các yêu cầu khác nhau cho các ứng dụng khác nhau, mô hình cần được chuyển đổi thích ứng.

Trong trường hợp mô hình cần được chọn chưa được tải trước trong thiết bị, thiết bị cần tải xuống từ mạng trước khi nhiệmvụ nhận dạng hình ảnh có thể bắt đầu. Một mô hình có thể được tái sử dụng nếu nó được giữ trong bộ nhớ sau lần sử dụng trước đó. Nhưng do tài nguyên lưu trữ hạn chế, thiết bị không thể giữ lại tất cả các mô hình để sử dụng tiềm năng trong lưu trữ. Tốc độ dữ liệu để tải xuống các mô hình cần thiết phụ thuộc vào kích thước của mô hình và độ trễ tải xuống cần thiết.

Cùng với các yêu cầu về hiệu năng ngày càng tăng đối với các hoạt động AI/ML, kích thước của các mô hình cũng tiếp tục tăng lên, mặc dù các kỹ thuật nén mô hình đang được cải thiện. Các kích thước điển hình của các mô hình DNN điển hình để nhận dạng hình ảnh được liệt kê trong Bảng 6..1. Tham số DNN có thể được biểu thị bằng 32 bit để có độ chính xác suy luận cao hơn. Kích thước mô hình và chi phí tải xuống có thể bị nén nếu kích thước của một tham số giảm xuống còn 8 bit, bằng cách có khả năng hy sinh độ chính xác nhận dạng hình ảnh.



Best top-1 score model: Mô hình đạt điểm số tốt nhất top 1
Best top-1 score model: Mô hình đạt điểm số tốt nhất top 5

Hình 6.1. Ví dụ về việc chọn mô hình tối ưu cho các nhiệm vụ/môi trường nhận dạng hình ảnh khác nhau (hình được phỏng từ [20])

Chú thích: Điểm số top-k (Top-k score) là một chỉ số đánh giá được sử dụng trong học máy để đo lường hiệu suất của một mô hình trong nhiệm vụ phân loại đa lớp (Multi Class Classification Task). Nó đánh giá tần suất mô hình dự đoán chính xác lớp thực (True Class) trong các lớp dự đoán top k.

Bảng 6.1. Kích thước của các mô hình nhận dạng hình ảnh điển hình và tốc độ dữ liệu DL cần thiết để tải xuống trong 1 giây

Mô hình DNN để nhận dạng ảnh	Số lượng thông số (Triệu)	32 bit trên mỗi thông số		8 bit trên mỗi thông số	
		Kích thước của mô hình (MByte)	Tốc độ dữ liệu DL yêu cầu (Mbit / s)	Kích thước của mô hình (MByte)	Tốc độ dữ liệu DL yêu cầu (Mbit / s)
AlexNet	60	240	1920	60	480
VGG16	138	552	4416	138	1104
ResNet-152	60	240	1920	60	480
ResNet-50	25	100	800	25	200
GoogleNet	6.8	27.2	217.6	6.8	54.4
Inception-V3	23	92	736	23	184
1.0 MobileNet-224	4.2	16.8	134.4	4.2	33.6

Độ trễ tải xuống mô hình yêu cầu phụ thuộc vào tốc độ mô hình cần sẵn sàng trên thiết bị. Nó bị ảnh hưởng bởi mức độ mà ứng dụng sắp tới có thể được dự đoán. Trong trường hợp sử dụng, chúng ta giả định thiết bị không thể dự đoán và tải xuống trước mô hình cần thiết. Độ trễ tải xuống mô hình yêu cầu phụ thuộc vào tốc độ mô hình cần sẵn sàng trên thiết bị. Nó bị ảnh hưởng bởi mức độ mà ứng dụng sắp tới có thể được dự đoán. Trong trường hợp sử dụng, chúng ta giả định thiết bị không thể dự đoán và tải xuống trước mô hình cần thiết. Khác với video được phát trực tuyến có thể được phát khi một phần nhỏ được nhớ đệm, mô hình DNN chỉ có thể được sử dụng cho đến khi toàn bộ mô hình được tải xuống hoàn toàn.

Ví dụ: nếu độ trễ tải xuống là 1 giây, tốc độ dữ liệu DL cần thiết nằm trong khoảng từ 134,4 Mbit/s đến 1,92Gbit/s trong trường hợp thông số 32 bit, như thể hiện trong Bảng 6.1.. Trong trường hợp thông số 8 bit, tốc độ dữ liệu DL yêu cầu có thể được giới hạn ở mức 33,6Mbit / s ~ 1,1Gbit / s.

Một mô hình bao gồm cấu hình topo của mô hình và các thừa số trọng số của mô hình. Cấu hình topo phản ánh cấu trúc của mạng thần kinh (ví dụ: tế bào thần kinh của mỗi lớp, kết nối của các tế bào thần kinh giữa hai lớp lân cận). Kích thước của các thông số mô hình được thể hiện trong Bảng 6.1. Kích thước của tệp cấu hình của một mô hình (tức là cấu hình topo mô hình), thường không vượt quá 1Mbits. Khi một ứng dụng của UE yêu cầu một mô hình, server của bên thứ ba gửi một cấu hình mô hình có thể bao gồm hai phần, tức là topo mô hình và các thừa số trọng số mô hình. Bởi vì topo mô hình và các thừa số số trọng số mô hình đến từ cùng một server, có thể chúng có cùng bộ IP (IP tuple). Sai số truyền của topo mô hình rất quan trọng so với các yếu tố trọng số của mô hình (UE khó có thể chạy mô hình nếu topo mô hình có sai số trong khi các hệ số trọng số có thể có khả năng chịu sai số truyền cao do độ bền của mô hình). Do đó, việc truyền dữ liệu của topo mô hình quan trọng hơn và đòi hỏi độ tin cậy cao hơn.

6.1.2. Quá trình diễn ra

Điều kiện ban đầu.

Server AI/ML quản lý nhóm mô hình AI/ML và có khả năng tải mô hình được yêu cầu xuống ứng dụng cung cấp nhận dạng hình ảnh dựa trên AI/ML.

Hệ thống 5G có khả năng cung cấp thông tin liên quan đến mạng 5G cho máy chủ AI/ML.

Các luồng dữ liệu diễn ra như sau:

1. Ứng dụng nhận dạng hình ảnh dựa trên AI / ML được người dùng yêu cầu bắt đầu nhận dạng hình ảnh / video do UE chụp.

2. Mô hình AI/ML được tải xuống từ máy chủ mô hình vào ứng dụng nhận dạng hình ảnh dựa trên AI/ML thông qua mạng 5G.

2. Mô hình AI/ML được tải xuống từ máy chủ mô hình đến ứng dụng nhận dạng hình ảnh dựa trên AI/ML thông qua mạng 5G.

3. Ứng dụng nhận dạng hình ảnh dựa trên AI / ML sử dụng mô hình AI / ML để suy luận cho đến khi hoàn thành nhiệm vụ nhận dạng hình ảnh.

4. Làm lại Bước 2) đến 3) để lựa chọn lại mô hình AI/ML và tải xuống nếu cần để thích ứng với các điều kiện thay đổi.

Điều kện kết thúc.

Các đối tượng trong hình ảnh hoặc video đầu vào được ứng dụng nhận dạng hình ảnh dựa trên AI / ML nhận dạng và độ chính xác và độ trễ suy luận cần được đảm bảo.

Nhiệm vụ nhận dạng hình ảnh có thể được hoàn thành theo nguồn năng lượng và tính toán sẵn có của UE.

6.1.3. Các yêu cầu mới tiềm năng cần thiết để hỗ trợ trường hợp sử dụng

Các yêu cầu KPI tiềm năng cần thiết để hỗ trợ trường hợp sử dụng bao gồm:

- Hệ thống 5G sẽ hỗ trợ tải xuống mô hình AI/ML để nhận dạng hình ảnh với độ trễ tối đa như trong Bảng 6.2.
- Hệ thống 5G sẽ hỗ trợ tải xuống mô hình AI / ML để nhận dạng hình ảnh với tốc độ dữ liệu DL trải nghiệm của người dùng như trong Bảng 6.2.
- Hệ thống 5G sẽ hỗ trợ tải xuống mô hình AI/ML để nhận dạng hình ảnh với tính khả dụng của dịch vụ truyền thông không thấp hơn 99.999%.

Bảng 6.2. Phân tích độ trễ tải xuống mô hình nhận dạng hình ảnh tải xuống ứng dụng ví dụ (thông số 8 bit cho DNN)

ng dụng của người dùng	Độ trễ: tối đa		Tốc độ dữ liệu DL trải nghiệm người dùng đối với tải xuống mô hình
	Độ trễ nhận dạng hình ảnh	Độ trễ tải xuống mô hình	
Nhận dạng đối tượng một lần trên điện thoại thông minh	~1s	1s (chú thích 1)	33.6Mbit/s~1.1Gbit/s
Nhận dạng người trong hệ thống giám sát an ninh	~1s	1s (chú thích 2)	33.6Mbit/s~1.1Gbit/s
Nâng cấp ảnh trên điện thoại thông minh	~1s	1s (chú thích 1)	33.6Mbit/s~1.1Gbit/s
NHận dạng video	33ms@30FPS	1s (chú thích 2)	33.6Mbit/s~1.1Gbit/s
Hiển thị AR / chơi game	<5ms (chú thích 3)	1s (chú thích 2)	33.6Mbit/s~1.1Gbit/s
Lái xe từ xa	<5ms (chú thich 4)	1s (chú thích 2)	33.6Mbit/s~1.1Gbit/s
Robot điều khiển từ xa	<5ms (chú thích 5)	1s (chú thích 2)	33.6Mbit/s~1.1Gbit/s

chú thích 1: Một mô hình dưới tối ưu được cài đặt sẵn có thể được sử dụng tạm thời trong khi tải xuống mô hình tối ưu. Mô hình tối ưu nên được tải xuống khoảng 1 giây.

chú thích 2: Đối với các ứng dụng có camera luôn bật, thiết bị có thể dự đoán mô hình cần thiết và bắt đầu tải xuống trước. Tải xuống mô hình trong vòng 1 giây là chấp nhận được.

chú thích 3: độ trễ chuyển động từ VR đến photon nằm trong khoảng 7-15ms. Có thể giả định rằng màn hình AR hoặc chơi game yêu cầu độ trễ đầu cuối tương tự. Xem xét thời gian cần thiết để kết xuất AR (e.g. 3D kết xuất các đối tượng ảo, kết xuất tăng cường trong lớp phủ), việc nhận dạng video nền sẽ được hoàn thành trong vòng 5ms.

chú thích 4: độ trễ một chiều cần thiết để lái xe từ xa là 5ms. Độ trễ khứ hồi được giả định là 10ms. Xem xét thời gian cần thiết của suy luận lái xe dựa trên thị giác tại máy chủ, quỹ độ trễ để nhận dạng đối tượng giao thông có thể được ước tính là 5ms.

chú thích 5: độ trễ đầu cuối cần thiết cho robot điều khiển từ xa hoạt động bằng video là 10 ~ 100ms. Xem xét thời gian cần để suy luận điều khiển robot tại máy chủ, nhận dạng thị giác robot cần được hoàn thành trong vòng 5ms.

6.2. Biên tập phương tiện theo thời gian thực với suy luận AI tại chỗ (Real time media editing with on-board AI inference)

6.2.1. Mô tả kịch bản

Điện thoại thông minh là thiết bị mà mọi người mang theo và sử dụng ở khắp mọi nơi để ghi âm và quay video. Nó cũng trở thành thiết bị đầu tiên trao đổi nội dung truyền thông với bạn bè và gia đình, để xuất bản trên phương tiện truyền thông xã hội. Điện thoại thông minh cao cấp ngày càng tích hợp CPU và GPU mạnh mẽ hơn và thậm chí cả bộ tăng tốc phần cứng AI chuyên dụng. Khi chất lượng máy ảnh và hình ảnh / video trở thành một yếu tố khác biệt giữa các điện thoại thông minh cao cấp, các mô hình AI / ML để nâng cao khả năng chụp ảnh xuất hiện tại chỗ trên các thiết bị cao cấp này. Các bộ tăng tốc AI dự kiến sẽ cho phép thực hiện các mô hình AI/ML phức tạp trực tiếp trên các thiết bị được kết nối của người dùng cuối; không chỉ nâng cấp ảnh mà cả nâng cao cao âm thanh chất lượng cao và phân tích nội dung video dự kiến sẽ được thực hiện tại chỗ trên điện thoại thông minh. Do đó, điện thoại thông minh sẽ trở thành một thiết bị để chỉnh sửa nội dung phương tiện trước khi chia sẻ qua mạng. Với sự ra đời của 5G/6G, các dịch vụ mới dựa trên tải xuống theo yêu cầu của các mô hình AI/ML lớn được thực thi (gần) thời gian thực trên thiết bị người dùng cuối sẽ xuất hiện; tùy thuộc vào dịch vụ, môi trường, sở thích của người dùng, đặc điểm của thiết bị, v.v., các mô hình DNN này sẽ cần được điều chỉnh hoặc cập nhật theo các ràng buộc về độ trễ nghiêm ngặt để không phải lưu trữ trước tất cả chúng tại chỗ.

Phân tích nội dung phương tiện kết hợp các nhiệm vụ như phát hiện đối tượng, phân đoạn, nhận dạng khuôn mặt, đếm người, theo dõi hoạt động của con người. Trong bảng 6.3, các ví dụ về các mô hình DNN thường được sử dụng để phát hiện đối tượng (với kích thước tương ứng của chúng) được liệt kê.

Bảng 6.3. Kích thước của các mô hình phát hiện đối tượng điển hình

Mô hình để phát hiện đối tượng	Số lượng thông số (Triệu)	Kích thước mô hình (MByte) các thông số 32 bit	Kích thước mô hình (MByte) các thông số 8 bit
MobileNet	3,2	12,8	3,2
DarkNet	20	80	20
SE ResNet	26	104	26
Inception v4	41	164	41
YOLONet	64	256	64
VGGNet	134	536	134

Biên tập nội dung phương tiện kết hợp các nhiệm vụ như cải thiện chất lượng âm thanh và video, dịch ngôn ngữ, ẩn danh khuôn mặt. Trong bảng 6.3 và 6.4, các ví dụ về các mô hình DNN thường được sử dụng cho siêu phân giải hình ảnh và siêu phân giải video được liệt kê.

Bảng 6.3. Kích thước của các mô hình siêu phân giải hình ảnh điển hình

Mô hình cho siêu phân giải ảnh	Số lượng thông số (Triệu)	Kích thước mô hình (MByte) các thông số 32 bit	Kích thước mô hình (MByte) các thông số 8 bit
RCAN	15,44	61,78	15.44
SAN	15,71	62,82	15.71
RDN	22,12	88,48	22.12
EDSR	40,73	16,92	40.73
OISR-RK3	41,91	167,64	41.91

Bảng 6.4. Kích thước của các mô hình siêu phân giải hình ảnh điển hình

Mô hình cho siêu phân giải ảnh	Số lượng thông số (Triệu)	Kích thước mô hình (MByte) các thông số 32 bit	Kích thước mô hình (MByte) các thông số 8 bit
RBPN/4-PF	12,7	50,8	12,7
RBPN/6-PF	12,7	50,8	12,7
VSR-DUF	6,8	27,2	6.8
DRDVSR	0,7	2,8	0,7

Hai cài đặt được xem xét cho trường hợp sử dụng:

a) Người dùng độc lập: một người quay video hoặc bắt đầu cuộc gọi video trên UE của mình trong môi trường ồn ào, có điều kiện ánh sáng khó khăn và gắn thẻ cảnh tự động và các đối tượng được nhúng vào video.

a) Đám đông: Trong một sự kiện lớn, như một buổi hòa nhạc trực tiếp, vài nghìn người sử dụng UE của họ để quay phim hoặc chụp ảnh ban nhạc cùng một lúc và yêu cầu thêm thông tin về buổi hòa nhạc như danh sách đĩa nhạc, lời bài hát, nhận dạng khuôn mặt nghệ sĩ, thương hiệu nhạc cụ/thiết bị. Trong bối cảnh này, các UE yêu cầu tải xuống các mô hình DNN để cải thiện việc ghi lại và ghi lại buổi hòa nhạc, đồng thời cung cấp thông tin theo yêu cầu của những người tham dự buổi hòa nhạc. Mỗi UE có thể yêu cầu một số mô hình DNN để thực hiện các nhiệmvụ sau: chụp ảnh và tối ưu hóa video, nhận dạng khuôn mặt nghệ sĩ, nhận dạng nhạc cụ, cải thiện âm thanh và tạo lời bài hát. Với nhóm UE không đồng nhất, hàng nghìn mô hình DNN theo yêu cầu của ứng dụng / dịch vụ - có thể được yêu cầu tải xuống. Các mô hình DNN này được điều chỉnh hoặc cập nhật theo loại và phiên bản hệ điều hành UE, đặc điểm phần cứng và môi trường.

6.2.2. Quá trình diễn ra

Điều kiện ban đầu

Cài đặt cho trường hợp sử dụng này như sau. Alice đang tham dự một buổi hòa nhạc trực tiếp đông đúc. Cô ấy rất háo hức để có được các clip phim và hình ảnh như một món quà lưu niệm tuyệt vời của buổi hòa nhạc, nhưng cô ấy lo lắng về điều kiện khó khăn để có được món quà lưu niệm tuyệt vời này vì điều kiện về ánh sáng và âm thanh rất thay đổi, với ánh sáng hạn chế cho khán giả và quá nhiều ánh sáng trên hiện trường. Âm thanh nổi có thể thay đổi và không cân bằng tốt tùy thuộc vào vị trí của cô ấy trong số khán giả và có nền rất ồn ào.

Alice muốn lưu trữ các clip phim và hình ảnh chất lượng tốt trên tài khoản cá nhân của mình trên internet cho tương lai, đồng thời đăng ảnh và video được gắn thẻ tên nghệ sĩ và các thông tin liên quan khác trong buổi hòa nhạc. Vì Alice cũng là một nhạc sĩ nghiệp dư, cô ấy cũng muốn có được thông tin chi tiết theo thời gian thực về cấu trúc của bài hát, lời bài hát và nhạc cụ.

Các điều kiện ban đầu là:

a) Alice đang tham dự một buổi hòa nhạc trực tiếp đông đúc.

b) UE của cô đã được đăng ký vào mạng 5G.

c) Các ứng dụng của điện thoại thông minh của Alice có thể dựa vào các mô hình máy học được tinh chỉnh có sẵn thông qua mạng bao phủ phòng hòa nhạc. Các mô hình này là:

   1) Một mô hình ML cải thiện khả năng chụp ảnh cho phòng hòa nhạc này (một mô hình được tinh chỉnh đặc biệt cho địa điểm hòa nhạc này).

   2) Mô hình ML cải thiện khả năng thu âm cho phòng hòa nhạc này (một mô hình được tinh chỉnh đặc biệt cho địa điểm hòa nhạc này).

   3) Một mô hình ML chuyên về danh sách đĩa nhạc và lời bài hát của ban nhạc.

   4) Một mô hình ML chuyên về nhận dạng khuôn mặt của nghệ sĩ.

   5) Một mô hình ML chuyên về nhận dạng nhạc cụ.

Các mô hình được liệt kê ở trên là ví dụ cho trường hợp sử dụng này và không phải là danh sách đầy đủ.

Các luồng dịch vụ.

1) Ngay sau khi bắt đầu buổi hòa nhạc, Alice, cũng như hầu hết người hâm mộ, khởi chạy ứng dụng máy ảnh trên điện thoại di động của mình để quay cảnh và nhận thêm thông tin về ban nhạc, từng nghệ sĩ hoặc các bài hát hoặc nhạc cụ.

2) Cô ấy hướng máy ảnh của thiết bị của mình về phía hiện trường.

3) Môi trường rất tối với các đốm sáng mạnh. Để có đầy đủ chức năng và mang lại trải nghiệm người dùng tốt nhất, ứng dụng máy ảnh sẽ tải xuống các mô hình ML đặc biệt.

4) Các mô hình ML được đề xuất rất hiệu quả trong môi trường này nhưng cũng rất nặng về kích thước.

5) Các mô hình ML có thể được sử dụng cho toàn bộ quá trình chụp, đặc biệt nếu môi trường vẫn ổn định. Nếu môi trường thay đổi đáng kể hoặc có sẵn các mô hình ML tốt hơn, các mô hình ML có thể được cập nhật dần dần cho phù hợp với một số quy tắc vận hành (như số lượng cập nhật ML tối đa mỗi giây). Trong mọi trường hợp, ứng dụng máy ảnh tiếp tục hoạt động liền mạch.

6) Các luồng âm thanh và video được ghi lại, cải thiện chất lượng, xử lý để trích xuất và hiển thị thông tin bổ sung, đồng thời được lưu trữ trong thời gian thực trên chính điện thoại di động hoặc được cung cấp dưới dạng luồng trực tiếp.

Điều kiện kết thúc

Alice có thể thấy rằng ngay cả trong điều kiện ánh sáng khắc nghiệt và với nền ồn ào, ảnh và video rất tuyệt, thông tin bổ sung được cung cấp và tất cả đều được gắn thẻ chính xác theo yêu cầu.

Các điều kiện kết thúc là:

1) Ảnh và video được lưu trữ trên điện thoại di động với chất lượng cao được cải thiện, sẵn sàng để tải lên và chia sẻ trên phương tiện truyền thông xã hội hoặc chia sẻ trực tiếp trên phương tiện truyền thông xã hội.

2) Ghi âm có chất lượng cao với khả năng giảm tiếng ồn xung quanh, cải thiện cân bằng âm thanh nổi.

3) Thông tin bổ sung về ban nhạc, bài hát / lời bài hát, nhạc cụ, v.v. được hiển thị trên điện thoại di động và được lưu trữ trong siêu dữ liệu của bản ghi phương tiện.

4) Alice có thể hình dung thông tin bổ sung và tải ảnh và video lên (các) mạng xã hội của mình với các thẻ và thông tin liên quan do mô hình cung cấp, đồng thời lưu trữ những thông tin trên máy chủ truyền thông cá nhân của cô ấy.

6.2.3. Các yêu cầu mới tiềm năng cần thiết để hỗ trợ trường hợp sử dụng

Các yêu cầu mới tiềm năng cần thiết để hỗ trợ kết quả trường hợp sử dụng của các thông số sau:

a) Kích thước mô hình AI/ML.

b) Độ chính xác của mô hình.

c) Hạn chế độ trễ của ứng dụng hoặc dịch vụ.

d) Số lượt tải xuống đồng thời, tức là số lượng UE yêu cầu tải xuống mô hình AI/ML trong cùng một khoảng thời gian.

Số lượng tải xuống đồng thời phụ thuộc vào mật độ UE trong khu vực được bao phủ và kích thước khu vực được bao phủ.

Các bảng 6.5, 6.6 và 6.7 chứa KPI cho các khía cạnh khác nhau của trường hợp sử dụng biên tập phương tiện theo thời gian thực.

Bảng 6.5. Kích thước điển hình của mô hình AI/ML cho trường hợp sử dụng

Mô hình AI/ML	Số lượng các thông số (trệu)	Kích thước mô hình AI (MByte)	Chú thích
MobileNet	3,2	3,2	Các thông số 8 bit
MobileNet	3,2	12,8	Các thông số 32 bit
RCAN	15,44	15,44	Các thông số 8 bit
DarkNet	20	20	Các thông số 8 bit
Inception v4	41	41	Các thông số 8 bit
RCAN	15,44	61,78	Các thông số 32 bit
YOLONet	64	64	Các thông số 8 bit
DarkNet	20	80	Các thông số 32 bit
VGGNet	134	134	Các thông số 8 bit
Inception v4	41	164	Các thông số 32 bit
YOLONet	64	256	Các thông số 32 bit
VGGNet	134	536	Các thông số 32 bit

Từ bảng 6.5, các mô hình AI/ML hiện có sẵn để xây dựng trường hợp sử dụng có kích thước thay đổi từ 3,2 MB đến 536 MB.

Như đã chỉ ra ở trên, kích thước của các mô hình AI/ML có thể được giảm bớt trước khi truyền bằng các kỹ thuật nén mô hình chuyên dụng. Ngược lại, các mô hình AI/ML với nhiều lớp mạng nơ-ron hơn và kiến trúc phức tạp hơn phát sinh để giải quyết các tác vụ phức tạp hơn và cải thiện độ chính xác. Xu hướng này dự kiến sẽ tiếp tục trong những năm tới. Kích thước mô hình điển hình trong khoảng từ 3 MB đến 500 MB dường như là một sự thỏa hiệp hợp lý để xem xét cho trường hợp sử dụng này.

Trong phần sau, hai danh mục được xem xét cho kích thước mô hình AI/ML:

a) Kích thước mô hình AI / ML dưới 64 MB, có thể được liên kết với các mô hình được tối ưu hóa để truyền nhanh,

b) Kích thước mô hình AI/ML dưới 500 MB, có thể được liên kết với các mô hình được tối ưu hóa để có độ chính xác cao hơn.

Độ trễ tối đa trong chức năng của ứng dụng hoặc dịch vụ:

a) Dịch vụ videocall: độ trễ đầu cuối-đầu cuối dưới 200 ms,

b) Ứng dụng quay video, phát trực tuyến video và nhận dạng đối tượng: độ trễ dưới 1 giây.

Các kết quả tốc độ dữ liệu DL trải nghiệm của người dùng từ kích thước và giá trị độ trễ tối đa của mô hình AI/ML ở trên được tóm tắt trong bảng 6.6.

Bảng 6.6. Tải xuống mô hình AI/ cho trường hợp sử dụng- UE đơn – các KPI

Tải xuống mô hình trường hợp sử dụng	Kích thước mô hình AI/ML model	Độ trễ: tối đa	Tốc độ dữ liệu DL trải nghiệm bởi người dùng
Mô hình đơn / UE đơn	[3 MB – 64 MB]	<1 s	[24 Mb/s ~ 512 Mb/s]
Mô hình đơn / UE đơn	[3 MB – 64 MB]	<200 ms	[120 Mb/s ~ 2.56 Gb/s]

Mô hình đơn / UE đơn	[64 MB – 500 MB]	<1 s	[512 Mb/s ~ 4 Gb/s]
Mô hình đơn / UE đơn	[64 MB – 500 MB]	<200 ms	[2.56 Gb/s ~ 20 Gb/s]

Như đã chỉ ra ở trên, số lượt tải xuống đồng thời là thông số thứ ba để xác định các yêu cầu mới tiềm năng. Điều này tương ứng với số lượng UE tối đa yêu cầu tải xuống mô hình AI/ML trong cùng một cửa sổ thời gian và cùng một khu vực/ô được bao phủ.

Trường hợp của một phòng hòa nhạc là một minh họa cho kịch bản, "Truy nhập băng thông rộng trong đám đông (từ TS 22.261). Kịch bản này giả định mật độ người dùng tổng thể là 500 000 UE / km2 (tức là 0,5 UE / m2) và hệ số hoạt động là 30%.

Số lượng UE điển hình trong một phòng hòa nhạc thay đổi từ ~ 1000 chỗ ngồi đến ~ 5000 chỗ ngồi.

Dựa trên các số liệu này và giả định hoạt động UE, số lượng tải xuống đồng thời được ước tính như trong bảng 6.7.

Bảng 6.7. Số lượt tải xuống đồng thời ước tính

Số lượng UE	Diện tích ước tính (Mật độ 0.5 UE / m^2)	Hệ số hoạt động	Số lượt tải xuống đồng thời trong cùng một ô
1000	2000 m^2	1 %	10
5000	10000 m^2	1 %	50

Từ bảng 6.6 và bảng 6.7, các yêu cầu về khu vực được bao phủ được ước tính như sau:

Bảng 6.8. Yêu cầu về tốc độ dữ liệu DL khu vực được bao phủ ước tính

Số lượng UE	Hệ số hoạt động	Số lượt tải xuống đồng thời	Kích thước mô hình AI/ML model	Độ trễ: cực đại	Tốc độ dữ liệu DL trải nghiệm người dùng	Tốc lệ dữ liệu DL trải nghiệm người dùng tổng hợp cho khu vực được bảo hiểm
1000	1 %	10	[3 MB – 64 MB]	<1 s	[24 Mb/s ~ 512 Mb/s]	[240 Mb/s ~ 5,12 Gb/s]
			[3 MB – 64 MB]	<200 ms	[120 Mb/s ~ 2,56 Gb/s]	[1.2 Gb/s ~ 25,6 Gb/s]
			[64 MB – 500 MB]	<1 s	[512 Mb/s ~ 4 Gb/s]	[5,12 Gb/s ~ 40 Gb/s]
5000	1 %	50	[3 MB – 64 MB]	<1 s	[24 Mb/s ~ 512 Mb/s]	[1,2 Gb/s ~ 25,6 Gb/s]
			[3 MB – 64 MB]	<200 ms	[120 Mb/s ~ 2,56 Gb/s]	[6 Gb/s ~ 128 Gb/s]
			[64 MB – 500 MB]	<1 s	[512 Mb/s ~ 4 Gb/s]	[25,6 Gb/s ~ 200 Gb/s]

6.3. Phân phối mô hình AI/ML để nhận dạng giọng nói

6.3.1. Mô tả

Xử lý giọng nói dựa trên AI/ML đã được sử dụng rộng rãi trong các ứng dụng trên thiết bị di động (ví dụ: điện thoại thông minh, trợ lý cá nhân, dịch ngôn ngữ), bao gồm nhận dạng giọng nói tự động (ASR: Automatic Speech recognition), dịch giọng nói, tổng hợp giọng nói. Nhận dạng giọng nói để đọc chính tả, tìm kiếm và ra lệnh bằng giọng nói đã trở thành một tính năng tiêu chuẩn trên điện thoại thông minh và thiết bị đeo.

Các yêu cầu dịch vụ đối với ASR đã được giải quyết trong [29]. Các hệ thống ASR truyền thống dựa trên mô hình Markov ẩn (HMM: hidden Markov model) và mô hình hỗn hợp Gaussian (GMM: Gaussian mixture model). Tuy nhiên, các hệ thống HMM-GMM chịu WER (Word Error Rate :Tỷ lệ lỗi từ) tương đối cao với sự hiện diện của tiếng ồn môi trường. Mặc dù một số cải tiến đã được phát triển bao gồm "tăng cường tính năng" (cố gắng loại bỏ tiếng ồn làm hỏng khỏi các quan sát trước khi nhận dạng) và "thích ứng mô hình" (giữ nguyên các quan sát và thay vào đó cập nhật các tham số mô hình của bộ nhận dạng để đại diện hơn cho giọng nói được quan sát), các mô hình truyền thống khó có thể đáp ứng các yêu cầu của các ứng dụng thương mại. Các mô hình âm thanh dựa trên mạng nơ-ron sâu (DNN) có độ bền tiếng ồn đáng chú ý và đã được sử dụng rộng rãi trong các ứng dụng ASR trong các thiết bị di động.

Ngày nay, hầu hết các ứng dụng ASR trên điện thoại thông minh đều hoạt động trong các máy chủ đám mây. Thiết bị đầu cuối tải giọng nói lên máy chủ đám mây, sau đó tải kết quả được giải mã trở lại thiết bị. Tuy nhiên, nhận dạng giọng nói dựa trên đám mây có khả năng gây ra độ trễ cao hơn (không chỉ do độ trễ mạng 4G / 5G mà còn do độ trễ internet) và vấn đề bảo mật và kết nối mạng độ tin cậy cần được xem xét.

Hệ thống nhận dạng giọng nói nhúng chạy trên thiết bị di động đáng tin cậy hơn và có thể có độ trễ thấp hơn. Hiện tại, một số ứng dụng ASR sẽ chuyển từ suy luận mô hình dựa trên đám mây sang suy luận mô hình ngoại tuyến khi vùng phủ sóng đường lên của người dùng di động trở nên yếu, ví dụ như khi vào tầng hầm hoặc thang máy. Tuy nhiên, các mô hình ASR cho máy chủ đám mây quá phức tạp đối với tài nguyên tính toán và lưu trữ trên thiết bị di động. Kích thước của mô hình ASR dựa trên ML chạy trên máy chủ đám mây đã tăng nhanh chóng trong năm gần đây, từ ~ 1GByte lên ~ 10GByte, không thể chạy trên thiết bị di động. Do hạn chế, chỉ các ứng dụng ASR đơn giản, ví dụ như phát hiện từ đánh thức (wakeword), mới có thể được triển khai trên điện thoại thông minh. Thực hiện các ứng dụng ASR phức tạp hơn, ví dụ như nhận dạng giọng nói liên tục từ vựng lớn (LVCSR: large vocabulary continuous speech recognition) vẫn là một lĩnh vực thách thức đối với một trình nhận dạng giọng nói ngoại tuyến.

Vào năm 2019, một trình nhận dạng LVCSR ngoại tuyến hiện đại dành cho thiết bị di động Android đã được công bố. Bộ nhận dạng đầu cuối phát trực tuyến dựa trên mô hình đầu bộ chuyển đổi mạng neuron hồi quy (RNN-T: recurrent neural network transducer). Người ta nói rằng, bằng cách sử dụng tất cả các loại cải tiến và tối ưu hóa, dung lượng bộ nhớ có thể được giảm đáng kể và tính toán có thể được tăng tốc. Mô hình có thể được nén thành 80MB. Trong khi đó, mô hình ASR được nén để phù hợp với việc sử dụng trên thiết bị di động, độ bền đối với các loại tiếng ồn xung quanh khác nhau đã bị hy sinh. Khi môi trường tiếng ồn thay đổi, mô hình cần được chọn lại và trong trường hợp mô hình không được giữ trong thiết bị, mô hình cần được tải xuống từ máy chủ đám mây / biên của chủ sở hữu mô hình AI / ML thông qua mạng 5G.

6.3.2. Quá trình diễn ra

Điều kiện ban đầu

UE chạy một ứng dụng cung cấp khả năng suy luận mô hình AI / ML để nhận dạng giọng nói.

Máy chủ AI/ML quản lý nhóm mô hình AI/ML (AI/ML Model Pool) và có khả năng tải xuống mô hình được yêu cầu đến ứng dụng cung cấp nhận dạng giọng nói dựa trên AI/ML.

Hệ thống 5G có khả năng cung cấp thông tin liên quan đến mạng 5G cho máy chủ AI/ML.

Các luồng dịch vụ

1. Ứng dụng nhận dạng giọng nói dựa trên AI / ML được người dùng yêu cầu bắt đầu nhận dạng giọng nói được ghi lại.

2. Mô hình AI/ML được tải xuống từ máy chủ mô hình về ứng dụng nhận dạng giọng nói dựa trên AI/ML thông qua mạng 5G.

3. Ứng dụng nhận dạng giọng nói dựa trên AI / ML sử dụng mô hình AI / ML để suy luận cho đến khi nhiệm vụ nhận dạng giọng nói kết thúc.

4. Làm lại Bước 2) đến 3) để lựa chọn lại mô hình AI/ML và tải xuống lại nếu cần để thích ứng với các điều kiện thay đổi.

Điều kiện kết thúc

Nội dung trong giọng nói đầu vào được ứng dụng nhận dạng giọng nói dựa trên AI / ML nhận dạng và độ chính xác và độ trễ suy luận cần được đảm bảo.

Nhiệm vụ nhận dạng giọng nói có thể được hoàn thành theo nguồn năng lượng và tính toán sẵn có của UE.

6.3.4. Các yêu cầu mới tiềm năng cần thiết để hỗ trợ trường hợp sử dụng

Mô hình ASR chỉ có thể được sử dụng cho đến khi toàn bộ mô hình được tải xuống hoàn toàn. Và thiết bị cần áp dụng mô hình ASR mạnh mẽ về tiếng ồn thích ứng với môi trường tiếng ồn đang thay đổi. Nói chung, micrô của thiết bị chỉ được bật khi ứng dụng nhận dạng giọng nói được kích hoạt. Thiết bị cần xác định môi trường tiếng ồn và tải xuống mô hình ASR tương ứng với độ trễ thấp (mức 1 giây). Kích thước của một số mô hình ASR điển hình được liệt kê trong Bảng 6.9, được sử dụng để lấy tốc độ dữ liệu cần thiết.

Bảng 6.9. Kích thước của các mô hình nhận dạng giọng nói điển hình và tốc độ dữ liệu DL trải nghiệm của người dùng để tải xuống trong 1 giây (8 bit cho mỗi tham số)

Mô hình DNN để nhận dạng giọng nói	Số lượng thông số (Triệu)	Kích thước của mô hình (MByte)	Tốc độ dữ liệu DL trải nghiệm của người dùng (Mbit / s)
RNN-CTC	26,5	26,5	212
ResCNN-LAS	6,6	6,6	52,8
QuartzNet-15×5	19	19	152
Bộ nhận dạng giọng nói Gboard (Gboard speech recognizer)	N.A.	80	640

6.4. Quản lý mô hình AI như một dịch vụ (AI model management as a Service

6.4.1. Mô tả

Các mô hình AI/ML có thể được phân loại và tạo ra từ các góc độ khác nhau, chẳng hạn như dịch vụ, người dùng, thời gian, vị trí, v.v. Vì vậy, có thể có một số lượng lớn các mô hình AI/ML (bao gồm dữ liệu cho các yếu tố cấu trúc và trọng số, độ dốc (gradient)) được sử dụng để suy luận và đào tạo. Do hạn chế về dung lượng lưu trữ, UE không phải lúc nào cũng tải trước tất cả các mô hình AI/ML cho các tác phẩm khác nhau. Đây có thể là một cơ hội thương mại mới khi các nhà khai thác cung cấp dịch vụ giúp quản lý và phân phối các mô hình AI/ML để UE có thể có được một mô hình phù hợp ngay lập tức.

Trong tương lai, dự kiến các công ty bên thứ 3 sẽ sử dụng các mô hình AI để hỗ trợ các loại dịch vụ khác nhau như hướng dẫn du lịch toàn cảnh sử dụng thực tế tăng cường (AR) trong khu nghỉ dưỡng. Tuy nhiên, hầu hết bên thứ 3 không có tài nguyên (máy chủ) ở những nơi phân tán. Xem xét kích thước lớn của các mô hình AI, thời gian tải xuống nghiêm ngặt (được minh họa trong trường hợp sử dụng 6.1, 6.2) và dung lượng lưu trữ UE hạn chế, các công ty bên thứ 3 ủy quyền cho các công ty chuyên nghiệp quản lý các mô hình AI của họ thay vì tự làm điều tương tự cục bộ.

Vì tài nguyên đám mây của nhà điều hành (hay nhà mạng) có lợi thế để quản lý các mô hình AI/ML khác nhau một cách tập trung (ví dụ: máy chủ đám mây) và cục bộ (ví dụ: máy chủ MEC), các dịch vụ đó cũng có thể được để lộ cho người dùng bên thứ 3. Đặc biệt,

• Nhà điều hành có nhu cầu riêng để duy trì một nhóm mô hình AI / ML cho hoạt động kinh doanh của riêng mình như tối ưu hóa mạng;

• Bên thứ 3 muốn sử dụng một số mô hình AI/ML phổ biến đã được lưu trữ trong đám mây của nhà khai thác;

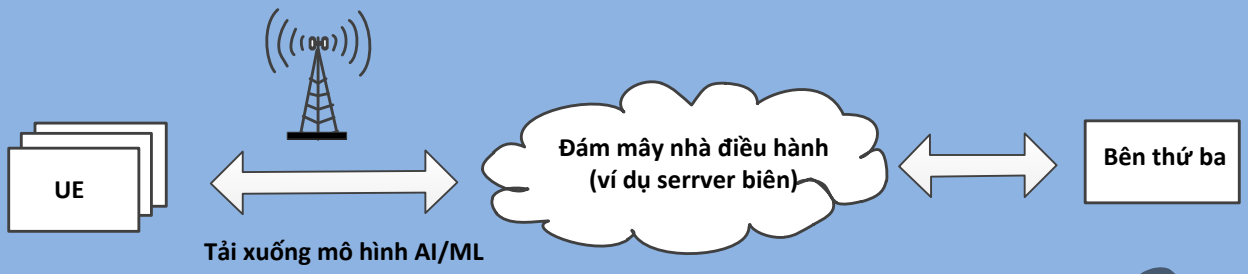
• Do hạn chế tài nguyên của bên thứ 3, họ muốn thuê tài nguyên đám mây của nhà điều hành để quản lý các mô hình AI của họ.

Do đó, như thể hiện trong Hình 6.2, Quản lý mô hình AI bao gồm:

• Cho phép bên thứ 3 gọi các khả năng do hệ thống 5G/6G tiết lộ để tải lên / tải xuống / cập nhật / xóa / lưu trữ / giám sát các mô hình AI.

• Truyền mô hình AI cho người dùng một cách hiệu quả theo tình huống (ví dụ: đi vào một khu vực nhất định).

• Vì hệ thống 5G có thể thu thập dữ liệu truyền thông, trải nghiệm người dùng, v.v., nó có thể thực hiện phân tích dữ liệu cho trải nghiệm người dùng, có thể tạo ra kết quả phân tích để giúp đào tạo đám mây của nhà khai thác và cải thiện các mô hình AI cho bên thứ 3.



Hình 6.2. Đám mây nhà điều hành để quản lý mô hình

6.4.2. Quá trình diễn ra

Điều kiện ban đầu

Đám mây của nhà điều hành (ví dụ: máy chủ biên) lưu trữ nhiều mô hình AI/ML theo yêu cầu của nhà khai thác hoặc bên thứ 3.

Đám mây của nhà điều hành (ví dụ: máy chủ biên) có khả năng phân phối mô hình được lưu trữ trong đám mây của nhà điều hành cho các thiết bị.

Các luồng dịch vụ

1. Công ty A cung cấp hướng dẫn viên du lịch toàn cảnh (panorama tourist) sử dụng công nghệ thực tế tăng cường (AR). Họ cung cấp dịch vụ hướng dẫn trong khu thương mại và khu du lịch. UE cần tải xuống mô hình A và B (cả hai đều là VGGnet 32 bit, 536MByte) tương ứng;

2. Để đạt được SLA (Service Level Agreement: thỏa thuận cấp độ dịch vụ) , công ty A cho nhà điều hành biết rằng UE yêu cầu mô hình A ở khu vực A và mô hình B ở khu vực B;

3. Khi UE di chuyển đến một nơi, máy chủ biên địa phương sẽ kích hoạt để thiết lập khả năng tăng tốc QoS và UE tải xuống mô hình tương ứng (A hoặc B) được lưu trữ trong máy chủ biên địa phương kịp thời để người dùng có thể tận hưởng thế giới AR liên tục mà không bị gián đoạn rõ ràng khi mô hình thay đổi.

Điều kiện kết thúc

UE sử dụng Mô hình hướng dẫn du lịch toàn cảnh.

6.4.3. Các yêu cầu mới tiềm năng cần thiết để hỗ trợ trường hợp sử dụng

Tùy thuộc vào sự đồng ý của người dùng, chính sách của nhà khai thác và các yêu cầu quy định của khu vực hoặc quốc gia, hệ thống 5G/6G sẽ có thể cung cấp khả năng để lộ thông tin (ví dụ: tốc độ dữ liệu đo được, độ trễ, kết quả phân tích mạng) cho ứng dụng của bên thứ ba được ủy quyền để hỗ trợ đào tạo và giám sát các mô hình AI/ML.

Hệ thống 5G/6G sẽ có thể hỗ trợ ứng dụng của bên thứ ba được ủy quyền để phân phối mô hình AI/ML từ 3,2~536MB đến ứng dụng của bên thứ ba chạy trên thiết bị trong vòng chưa đầy 1 giây với mật độ người dùng lên đến 5000~10000/km2 trong khu vực đô thị.

6.5. Hệ thống nối mạng ô tô dựa trên AI / ML (AI/ML based Automotive Networked Systems)

6.5.1. Mô tả kịch bản

Trường hợp sử dụng này là minh họa cho một ví dụ cụ thể nhưng có thể dễ dàng mở rộng cho các tình huống nâng cao hơn liên quan đến các kịch bản thảm họa, hoạt động cứu hộ hoặc thậm chí là hoạt động của xe tự lái. Trọng tâm của kịch bản này là các hệ thống suy luận và học chuyển giao AI/ML giữa máy với máy được kết nối với nhau bằng mạng 5G/6G để cung cấp các ứng dụng và dịch vụ ô tô thông minh.

Có hai loại mô hình ML chính và xử lý mô hình được xem xét trong trường hợp sử dụng này:

1) Cập nhật các mô hình ML lớn bằng cách sử dụng đào tạo không thời gian thực – Các mô hình ML này được đào tạo và tối ưu hóa với hàng triệu hoặc hàng tỷ tham số bằng cách sử dụng tài nguyên điện toán mở rộng để đạt được độ chính xác cao nhất có thể dựa trên các bộ dữ liệu đào tạo đầu vào cụ thể. Đào tạo này được thực hiện trong một khoảng thời gian dài và mô hình ML được đào tạo đầy đủ là mô hình cơ sở được cài đặt khi sản xuất ban đầu cho các thiết bị trong các trường hợp sử dụng được mô tả bên dưới. Các mô hình được đào tạo đầy đủ này có thể được cập nhật với dữ liệu mô hình AI-ML được cung cấp bên ngoài và cũng có thể tự cải thiện dựa trên dữ liệu cảm biến bên ngoài.

2) Chuyển giao và trao đổi một phần dữ liệu mô hình AI-ML – Trường hợp sử dụng này mô tả các ứng dụng khác nhau trong đó các loại hệ thống ML khác nhau được nối mạng để trao đổi các phần của dữ liệu mô hình AI-ML nhằm cải thiện độ chính xác dự đoán. Trong một số hệ thống, các mô hình ML cục bộ có thể liên tục cải thiện mô hình cơ sở của chúng bằng cách sử dụng dữ liệu thu thập được từ môi trường hiện có và đầu vào cảm biến khác. Các hệ thống này tải dữ liệu mô hình được cải thiện với tốc độ tương đối chậm lên mạng dựa trên đám mây lớn hơn, nơi xử lý thêm diễn ra để tinh chỉnh hơn nữa mô hình ML đầy đủ. Phương pháp cụ thể được sử dụng để cải tiến liên tục nằm ngoài phạm vi của trường hợp sử dụng này.

Các kịch bản ví dụ khác nhau được mô tả dưới đây.

Hệ thống AI/ML ứng phó khẩn cấp đối với phương tiện bị khuyết tật

Hình 6.3 minh họa một chiếc xe bị hỏng bị che khuất bởi một khúc cua mù trên đường. Phân tích này được coi là mức độ nghiêm trọng cao cho mục đích minh họa trường hợp sử dụng này.

Hình 6.3. Sự cố xe bị che khuất bởi một khúc cua mù nguy hiểm

Các cảm biến trên xe phát hiện những điều sau: hỏng hóc cơ học, giảm tốc độ, tọa độ GPS và nhiều thông số kỹ thuật khác nhau được sử dụng bởi các hệ thống suy luận ML trên xe để chẩn đoán ngay sự cố và thông báo cho các mạng xung quanh về mức độ nghiêm trọng của sự cố.

Các mô hình AI không thời gian thực khác nhau có thể sử dụng dữ liệu đào tạo lỗi thu thập được để tinh chỉnh các bản cập nhật mô hình dự đoán và phòng ngừa.

Camera giám sát giao thông và hệ thống an toàn giao thông MEC thành phố.

Camera giám sát sử dụng mô hình ML để phân tích dữ liệu cảm biến hình ảnh nhằm phát hiện và phân loại loại kiểu tai nạn bằng cách sử dụng suy luận ML với cải tiến mô hình AI-ML liên tục. Thông tin tai nạn đã xử lý này được tải lên cùng với các bản cập nhật mô hình lên MEC gần biên địa phương để dự đoán mô hình hệ thống AI / ML và xử lý thuật toán để đề xuất quá trình hành động tiếp theo ngay lập tức. Các hệ thống AI/ML này đề xuất một chương trình phát quảng bá cảnh báo tự động. Các hệ thống AI/ML tự động dựa trên các mô hình ML được đào tạo để dự đoán vị trí của phương tiện đặc biệt nguy hiểm do đường cong mù ẩn và thông báo cho Đơn vị Ứng phó Khẩn cấp địa phương để gửi trợ giúp thiết lập đường tránh giao thông sớm hơn trên đường. Gửi trở lại dữ liệu mô hình này trở lại camera giám sát để cải thiện thêm độ chính xác suy luận.

Việc tải lên dữ liệu, có thể được xử lý trước bởi mô hình AI/ML cục bộ, từ camera giám sát đến MEC có thể xảy ra với độ trễ thấp tùy thuộc vào mức độ nghiêm trọng của tai nạn trong khi dữ liệu được trả lại để cập nhật mô hình cục bộ có thể xảy ra trong khung thời gian chậm hơn nhiều.

Phản ứng đến gần của xe thông minh

Hình 6.4 minh họa một xe thông minh nhanh chóng đến gần một xe car được minh họa trong Hình 6.3. Các cảm biến của xe này không thể phát hiện mối nguy hiểm và phải dựa vào các hệ thống AI / ML khác để cảnh báo.

Hình 6.4. Xe tự hành đang tiến gần

Xe tự hành tiến gần nhận được nhiều thông báo cảnh báo và hệ thống AI/ML trong phương tiện này sử dụng các mô hình được đào tạo dự đoán và đề xuất các thao tác tự động thích hợp để di chuyển phương tiện đến làn đường xa để tránh bất kỳ nguy hiểm nào. Chia sẻ dữ liệu với các hệ thống khác cho phép liên tục tinh chỉnh các mô hình của họ; Việc chia sẻ dữ liệu này không có yêu cầu phi thực.

6.5.2. Quá trình diễn ra

Điều kiện ban đầu

Các kịch bản được mô tả ở trên giả định các điều kiện tiên quyết sau:

- Tất cả các hệ thống AI/ML hiện có đều có các mô hình AI-ML mặc định được đào tạo sẵn sẽ thực hiện dự đoán ở độ chính xác cơ bản và có khả năng cập nhật/tinh chỉnh lại các mô hình AI-ML để liên tục cải thiện độ chính xác cho mục đích cụ thể.

- Mộ số điều kiện bổ sung được cho tyrong bảng 6.10.

Bảng 6.10. Điều kiện tiên quyết trong trường hợp sử dụng

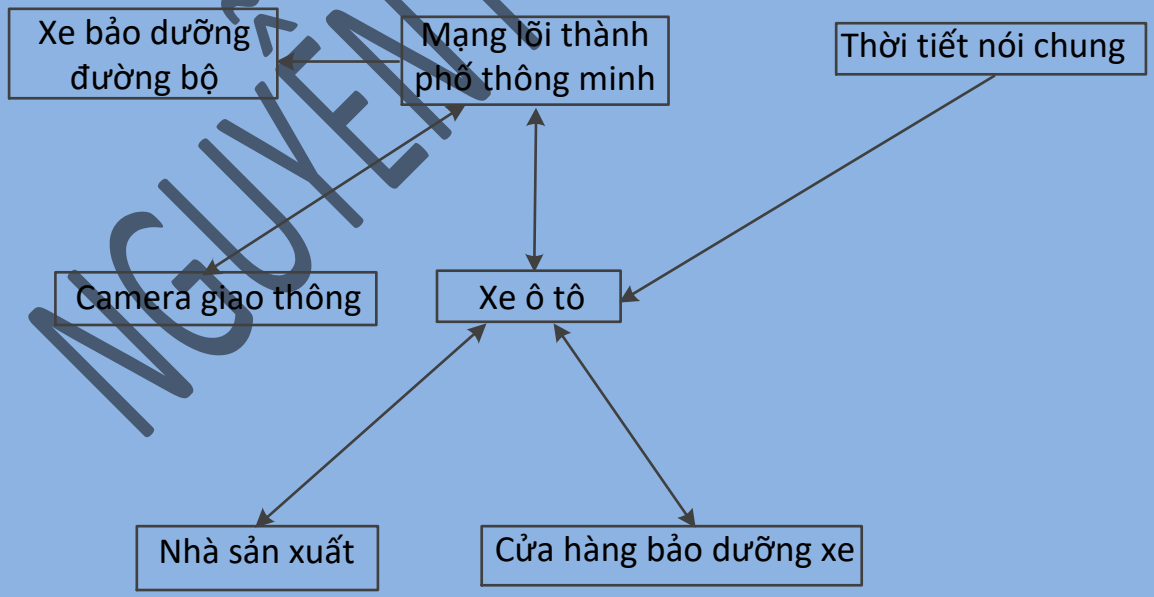
Xe tự hành bị sự cố	Xe tự hành đang tiên gần	Camera giám sát đường bộ	Điều khiển giao thông thành phố thông minh	Mạng lưới dịch vụ khẩn cấp thành phố thông minh
Mô hình xe AI / ML đã phát hiện ra lỗi. Tất cả các hệ thống truyền thông đều có sẵn và đang hoạt động.	Các mô hình Xe AI / ML đều hoạt động bình thường. Tất cả các hệ thống truyền thông đã sẵn sàng và hoạt động.	Camera được cấp nguồn, hoạt động và kết nối mạng.	Phần mềm và các hệ thống truyền thông đang chạy bình thường.	Phần mềm, Truyền thông và tài sản khẩn cấp đang chạy bình thường.

Các luồng dịch vụ

Các hệ thống AI/ML thích hợp sẽ cung cấp bảo mật và bảo vệ dữ liệu theo quy định của pháp luật và các chính sách thích hợp khác.

Các luồng dịch vụ được mô tả trong Hình 6.4 là tất cả các tương tác hệ thống thông minh AI/ML giữa máy với máy được thiết kế để chia sẻ thông tin với một loạt các nhu cầu từ độ trễ thấp nhất đối với các tình huống quan trọng (ví dụ: thông báo cảnh báo) và độ trễ nới lỏng cho các vấn đề ít nghiêm trọng hơn. Thông tin được chia sẻ (có thể bao gồm một bộ dữ liệu mở rộng) cũng có thể được trao đổi với độ trễ nới lỏng để hỗ trợ đào tạo và cập nhật mô hình AI/ML. Mỗi cấp độ dịch vụ có thể được sở hữu và điều hành bởi một tổ chức khác nhau.

Trường hợp sử dụng này giả định rằng không cần chia sẻ hoặc cập nhật quá [10%] tổng kích thước mô hình ML trong một lần truyền độ trễ thấp.



Hình 6.4. Ví dụ về tương tác AI-ML

Phương tiện đến các hệ thống AI khác

Ví dụ về trao đổi dữ liệu mô hình AI-ML là:

• Giám sát tình trạng xe và kiểm tra các mô hình ML - Các mô hình ML cục bộ cảm nhận và phản ứng với thông tin cảm biến va chạm, hỏa hoạn, nhiệt độ và điện. Các mô hình này liên tục cải tiến và tối ưu hóa nhưng không nhất thiết phải chia sẻ dữ liệu nếu hệ thống là danh nghĩa. Tuy nhiên, một khi tình huống khẩn cấp đã xảy ra, thông tin khẩn cấp có thể được gửi đến các dịch vụ ứng phó khẩn cấp thích hợp trong mạng lõi thành phố thông minh, camera cục bộ và cảm biến đường bộ. Thông tin mở rộng cũng có thể được chia sẻ với nhu cầu độ trễ thoải mái để hỗ trợ cập nhật/đào tạo mô hình.

• Cảm biến phương tiện cho động lực học của phương tiện dưới cảm biến và điều khiển mô hình ML. Các mô hình ML cục bộ cảm nhận và phản ứng với trượt, trượt, tăng tốc và gãy. Các mô hình này liên tục cải tiến và tối ưu hóa nhưng không nhất thiết phải chia sẻ dữ liệu nếu hệ thống là danh nghĩa. Tuy nhiên, một khi tình huống khẩn cấp đã xảy ra, thông tin khẩn cấp ngay lập tức được phát đến các dịch vụ ứng phó khẩn cấp thích hợp và hệ thống khác áp dụng các mô hình ML liên quan để thông báo cho chúng. Thông tin mở rộng cũng có thể được chia sẻ với nhu cầu độ trễ nới lỏng để hỗ trợ cải thiện độ chính xác dự đoán của hệ thống.

Camera giám sát giao thông nối đến mạng các dịch vụ khẩn cấp thành phố

Camera cảm nhận tai nạn bằng cách sử dụng các mô hình AI/ML cục bộ và gửi thông báo đến kiểm soát giao thông. Kiểm soát giao thông có thể sử dụng nhiều camera để đưa tin về cùng một sự cố và do đó, cải thiện độ chính xác suy luận của nó. Thông tin mở rộng được trao đổi với nhu cầu độ trễ thấp có thể được sử dụng để đào tạo lại và phân phối lại các lớp mô hình AI tham chiếu cho phù hợp.

Ví dụ về trao đổi dữ liệu Mô hình AI-ML là:

• Cảm biến máy ảnh, đường và thời tiết sử dụng các mô hình ML. Các cảm biến mô hình ML liên tục cải tiến và tối ưu hóa kết quả dự đoán. Khi tình huống khẩn cấp được phát hiện, thông tin khẩn cấp từ các hệ thống này có thể được chia sẻ với các xe cấp cứu, ví dụ: cải thiện thời gian định tuyến của họ qua giao thông dựa trên tình huống.

Điều khiển giao thông thành phố nối đến mạng các dịch vụ khẩn cấp thành phố

Các mô hình AI/ML sử dụng tự động hóa và dự đoán để gửi các xe EMS (Emerrgency Medical Service: dịch vụ cấp cứu y tế) thích hợp đến hiện trường vụ tai nạn.

Ví dụ về trao đổi dữ liệu Mô hình AI-ML là:

• Camera , các bộ cảm biến đường xa và thời tiết sử dụng các mô hình ML. Các cảm biến mô hình ML liên tục cải thiện và tối ưu hóa kết quả dự đoán. Khi tình huống khẩn cấp được phát hiện, thông tin khẩn cấp từ các hệ thống này có thể được chia sẻ với các xe cấp cứu để, ví dụ: cải thiện thời gian định tuyến của họ qua giao thông dựa trên tình huống.

Xe bị chết máy nối đến nhà sản xuất

Xe bị chết máy gửi dữ liệu (có thể đã được xử lý trước, ví dụ: dữ liệu mô hình AI/ML) cho nhà sản xuất ô tô để giúp chẩn đoán sự cố. Nhà sản xuất sử dụng thông tin để cải thiện chất lượng, độ tin cậy và hiệu suất của sản phẩm. Các kỹ thuật học liên kết và/hoặc học tập phân tán có thể được sử dụng để cải thiện mô hình AI/ML trên xe.

Xe bị chết máy nối đến điều khiển giao thông thành phố (MTC: Metro Traffic Control)

Xe chết máy gửi thông tin khẩn cấp đến MTC để giúp cảnh báo những người khác. Ví dụ: vị trí đường cong mù cụ thể cùng với dữ liệu khu vực khác được sử dụng để cải thiện tất cả các phản ứng của MTC. Thông tin mở rộng cũng có thể được chia sẻ với nhu cầu độ trễ nới lỏng để cải thiện các mô hình phản hồi MTC.

Xe bị chết máy nối đến dịch vụ sửa chữa

Xe bị chết máy gửi dữ liệu (có thể đã được xử lý trước, ví dụ: dữ liệu mô hình AI/ML) đến dịch vụ sửa chữa để đảm bảo các xe phản ứng thích hợp cung cấp các công cụ và thiết bị phù hợp. Phản hồi dịch vụ sửa chữa sử dụng các kỹ thuật học liên kết và / hoặc học tập phân tán để cải thiện mô hình liên tục.

Xe đang đến gần truyền thông với điều khiển giao thông thành phố (MTC)

MTC cảnh báo phương tiện đang đến gần về mối nguy hiểm sắp xảy ra và tùy thuộc vào mức độ SAE (Society Automotive Engineers: hiệp hội kỹ sử ô tô) của tự động hóa lái xe, xe phản ứng sẽ phản ứng thích hợp. Thông tin mở rộng cũng có thể được chia sẻ với nhu cầu độ trễ nới lỏng để hỗ trợ cải thiện độ chính xác dự đoán của hệ thống.

Điều kiện kết thúc

Tất cả các hệ thống được xác định trong kịch bản này sử dụng các thuật toán học liên kết hoặc học phân tán AI/ML độc lập, được nối mạng để tổng hợp thông tin từ nhiều nguồn nhằm cải thiện lớp của mô hình AI/ML hệ thống. Các hệ thống đào tạo AI/ML độc lập đảm bảo rằng các mô hình cải tiến được phân phối trở lại từng hệ thống cuối để cải thiện độ an toàn tổng thể và độ mạnh mẽ của phản ứng tiếp theo.

6.4.3. Các yêu cầu tiềm năng mới cần thiết để hỗ trợ trường hợp sử dụng

Bảng 6.11 cung cấp các mô hình ML ví dụ (kích thước và tốc độ dữ liệu DL) được xem xét cho các trường hợp sử dụng ở trên và Bảng 6.12 phác thảo các yêu cầu về độ trễ. Giả định rằng tối đa 10% kích thước mô hình (đầy đủ) được trao đổi giữa các hệ thống tham gia.

Bảng 6.11. Các mạng / các mô hình ML và tốc độ dữ liệu DL tiềm năng

Mô hình DNN	32 bits per parameter		
	Mô hình đầy đủ (MBytes)	Dữ liệu được trao đổi (Mbytes) [10% mô hình đầy đủl]	Tốc độ dữ liệu người dùng cực đại (Mbit/s)
1.0 MobileNet-224	16,8	1,68	13,4
SSD-ResNet34	81	8,1	64,8
SSD-MobileNet-v1	27,3	2,37	21,8
MASK R-CNN	245	24,5	~100
DLRM	400	40,0	10

Bảng 6.12. Yêu cầu về độ trễ tiềm năng

Ứng dụng người dùng	Yêu cầu về độ trễ tiềm năng Dữ liệu trao đổi - Độ trễ tải xuống
Xe phát hiện lỗi và tấp vào lề để tránh tai nạn	~500ms – 1sec
Camera bên đường phát hiện nguy hiểm đường và cảnh báo mạng thành phố thông minh	~500ms – 1sec
Giao thông sắp tới phát hiện nguy hiểm và tránh tai nạn	~500ms – 1sec
Nhận dạng video	~500ms – 1sec
Thành phố thông minh phát hiện tai nạn và đưa ra cảnh báo cục bộ	~ lên đến vài giây
Nhà sản xuất ô tô/Bảo hiểm được cảnh báo về lỗi	Lên đến vài phút

Yêu cầu:

Hệ thống 5G/6G có thể hỗ trợ tải dữ liệu với dung lượng tối đa ~2-40 MB để cập nhật mô hình AI/ML cục bộ với độ trễ lên đến 500ms – 1s.

Hệ thống 5G/6G có thể hỗ trợ tải dữ liệu để cập nhật mô hình AI/ML cục bộ với tính khả dụng dịch vụ truyền thông lên đến 99.999%.

6.6. Giám sát mô hình AI/ML được chia sẻ (Shared AI/ML model monitoring)

6.6.1. Mô tả

Các mô hình AI/ML được đào tạo trên một tập dữ liệu đào tạo để hoàn thành một nhiệm vụ đã thiết lập. Các nhiệm vụ có thể thay đổi từ nhận dạng hình ảnh hoặc giọng nói đến dự báo để tối ưu hóa, ví dụ: hiệu suất chuyển giao (xem 3GPP TR 28.809) hoặc điều chỉnh các thông số được hỗ trợ của Mạng lõi (xem phần 5.4.6.2, TS 23.501).

Trong mỗi nhiệm vụ này, nhà cung cấp mô hình AI/ML được chia sẻ có thể được hưởng lợi từ việc chia sẻ mô hình AI/ML được đào tạo với (các) người tiêu dùng của mô hình AI/ML được chia sẻ hoặc có thể được hưởng lợi từ việc đào tạo mô hình AI/ML phân tán/liên kết hoặc đào tạo mô hình AI/ML phân chia qua hệ thống 5G/6G. Hơn nữa, giám sát mô hình AI/ML là một yêu cầu để cho phép học trực tuyến trong mạng (ví dụ: thông qua Học tăng cường), một tập hợp các kỹ thuật phù hợp hơn để phản ứng kịp thời với sự suy giảm dịch vụ.

Các thuật ngữ sau được sử dụng trong trường hợp sử dụng này:

Mô hình AI/ML được chia sẻ (Shared AI/ML model): Mô hình AI/ML được chia sẻ giữa các ứng dụng khác nhau, ví dụ: mô hình AI/ML được đào tạo trước và cung cấp cho các người tiêu dùng khác nhau hoặc mô hình AI/ML được đào tạo bằng cách sử dụng phương pháp học tập phân tán/liên kết hoặc bằng cách chia giai đoạn đào tạo mô hình thành các phần khác nhau được thực hiện ở các vị trí mạng khác nhau.

Nhà cung cấp mô hình AI/ML được chia sẻ (Shared AI/ML model provider): máy chủ ứng dụng đang cung cấp hoặc quản lý "mô hình AI/ML được chia sẻ".

Người tiêu dùng mô hình AI/ML được chia sẻ (Shared AI/ML model consumer): ứng dụng, ví dụ: chạy trên UE, đang sử dụng/tiêu thụ "mô hình AI/ML được chia sẻ".

Do những thay đổi trong kịch bản (tức là trong bối cảnh thu thập dữ liệu đào tạo), mô hình AI/ML có thể mang lại hiệu suất kém so với hiệu suất của mô hình AI/ML được đo trong giai đoạn thử nghiệm mô hình. Điều này có thể xảy ra khi theo thời gian, sự phân phối của dữ liệu đầu vào để suy luận khác với sự phân phối của dữ liệu đào tạo hoặc nếu mô hình AI/ML được sử dụng trong một bối cảnh khác. Trong trường hợp này, nhà cung cấp mô hình AI/ML được chia sẻ sẽ có thể kịp thời phát hiện sự suy giảm hiệu suất và phản ứng để tránh sự suy giảm hoặc gián đoạn dịch vụ. Thông thường, việc cập nhật mô hình AI/ML được chia sẻ không chỉ phụ thuộc vào kết quả suy luận của một người tiêu dùng mô hình AI/ML được chia sẻ vì sự suy giảm hiệu suất có thể là do một số nguồn lỗi khác, ví dụ: lỗi đo lường đầu vào của mô hình. Để phát hiện mô hình AI/ML được chia sẻ lỗi thời, nhà cung cấp mô hình AI/ML được chia sẻ có thể sử dụng kết quả suy luận từ nhiều người tiêu dùng mô hình AI/ML được chia sẻ và .

Do đó, nhà cung cấp mô hình AI/ML được chia sẻ, sau khi chia sẻ mô hình AI/ML qua hệ thống 5/6G với người tiêu dùng mô hình AI/ML được chia sẻ, cần theo dõi hiệu suất của mô hình để phát hiện sự suy giảm hiệu suất có thể xảy ra của mô hình AI/ML được chia sẻ (ví dụ: dựa trên phản hồi suy luận từ người tiêu dùng mô hình AI/ML như mức độ tin cậy thấp hơn).

Ngoài ra, nhà cung cấp mô hình AI/ML được chia sẻ có thể phân chia đào tạo mô hình AI/ML với người tiêu dùng mô hình AI/ML được chia sẻ để liên tục cải thiện hiệu suất của mô hình AI/ML được chia sẻ dựa trên phản hồi đào tạo và/hoặc suy luận cục bộ từ người tiêu dùng mô hình AI/ML được chia sẻ. Phần cục bộ của mô hình AI/ML được chia sẻ sẽ được đào tạo/tinh chỉnh theo hướng dẫn của nhà cung cấp mô hình AI/ML được chia sẻ. Đầu vào cho khóa đào tạo này sẽ là dữ liệu có sẵn tại người tiêu dùng mô hình AI/ML được chia sẻ. Đầu ra của đào tạo mô hình cục bộ hoặc suy luận tại người tiêu dùng mô hình AI/ML được chia sẻ có thể được cung cấp cho nhà cung cấp mô hình AI/ML được chia sẻ và được nhà cung cấp mô hình AI/ML dùng chung để cung cấp thêm thông tin cho hệ thống 5G/6G nhằm cải thiện hoạt động của nó.

6.6.2. Quá trình diễn ra

Điều kiện ban đầu

Nhà cung cấp mô hình AI/ML được chia sẻ lưu trữ nhiều mô hình AI/ML cùng với hiệu năng đã được đo của chúng trong giai đoạn thử nghiệm.

Nhà cung cấp mô hình AI/ML được chia sẻ có khả năng chia sẻ mô hình AI/ML với người tiêu dùng mô hình AI/ML được chia sẻ để tận dụng trên hệ thống 5G/6G.

Người tiêu dùng mô hình AI/ML được chia sẻ có thể chạy các ứng dụng yêu cầu sử dụng mô hình AI/ML và tải xuống từ nhà cung cấp mô hình AI/ML thông qua hệ thống 5G/6G.

6.6.3. Các luồng dịch vụ

1. Nhà cung cấp mô hình AI/ML được chia sẻ muốn tối ưu hóa hiệu suất của một số quy trình bằng cách sử dụng hiệu năng của mô hình AI/ML được chia sẻ.

2. Nhà cung cấp mô hình AI/ML được chia sẻ gửi mô hình AI/ML được chia sẻ đã được đào tạo đến người tiêu dùng mô hình AI/ML được chia sẻ tại UE để tận dụng trên hệ thống 5G/6G.

3. UE nhận mô hình và sử dụng mô hình để thực hiện đào tạo và suy luận cục bộ bằng cách sử dụng dữ liệu có sẵn trên UE.

4. Nhà cung cấp mô hình AI/ML được chia sẻ giám sát kịch bản ngữ cảnh (ví dụ: dữ liệu UE có sẵn cho ứng dụng) trong đó UE đang chạy mô hình AI/ML được chia sẻ và hiệu năngcủa mô hình.

5. Nếu phát hiện thấy sự thay đổi trong kịch bản ngữ cảnh hoặc hiệu năng mô hình, ví dụ: dựa trên phản hồi suy luận từ người tiêu dùng mô hình AI/ML được chia sẻ, nhà cung cấp mô hình AI/ML được chia sẻ, để tránh suy giảm hiệu suất của mô hình, chia sẻ với người tiêu dùng mô hình AI/ML phiên bản cập nhật của mô hình AI/ML được đào tạo lại để nắm bắt ngữ cảnh mới với hiệu suất tốt hơn dự kiến.

6. UE tiếp tục chạy mô hình cập nhật mà không bị giảm hiệu suất

Điều kiện kết thúc

Theo ví dụ trong trường hợp sử dụng, nhà cung cấp mô hình AI/ML được chia sẻ nhận được dự báo chính xác về hiệu năng mô hình AI/ML và người tiêu dùng mô hình AI/ML đang sử dụng mô hình AI/ML có hiệu suất cao.

6.6.3. Các yêu cầu mới tiềm năng cần thiết để hỗ trợ trường hợp sử dụng

Hệ thống 5G/6G sẽ có thể chuyển giao mô hình AI/ML cập nhật từ nhà cung cấp mô hình AI/ML được chia sẻ sang người tiêu dùng mô hình AI/ML được chia sẻ trong khoảng thời gian [1 giây-1 phút] đối với các model AI/ML có kích thước tối đa là [100-500] MB.

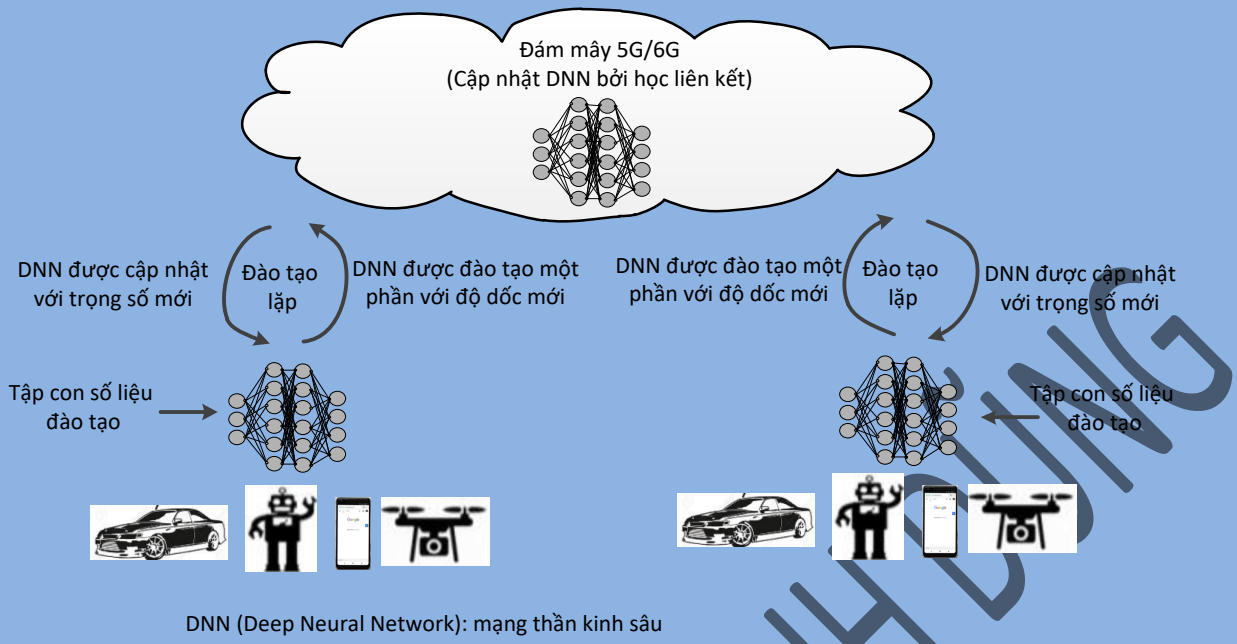
7. HỌC PHÂN TÁN/ LIÊN KẾT TRÊN HỆ THỐNG 5G (DISTRIBUTED/FEDERATED LEARNING OVER 5G SYSTEM

7.1. Học liên kết không nén để nhận dạng hình ảnh (Uncompressed Federated Learning for image recognition)

7.1.1. Mô tả

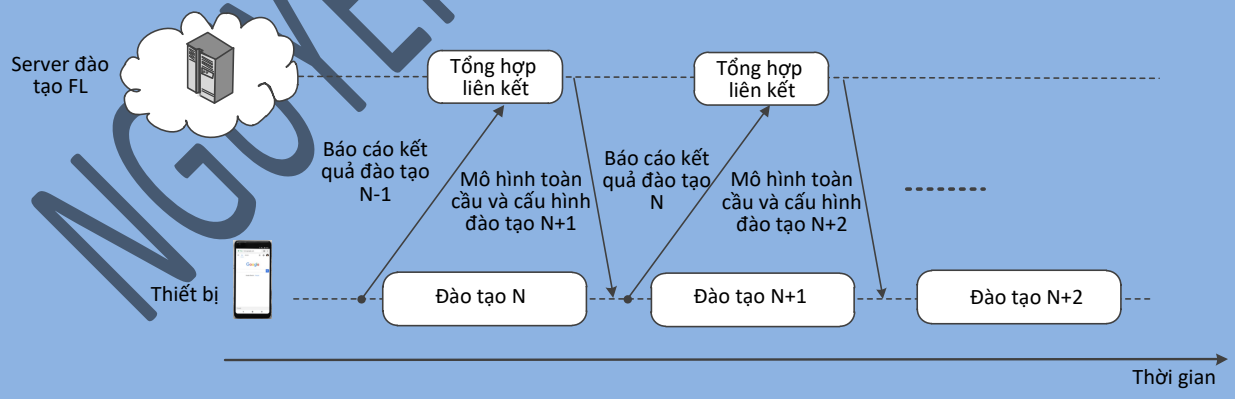
Ngày nay, máy ảnh trên điện thoại thông minh đã trở thành công cụ phổ biến nhất để chụp ảnh và quay video, chứa nhiều dữ liệu thị giác có giá trị để đào tạo mô hình nhận dạng hình ảnh. Đối với nhiều nhiệm vụ nhận dạng hình ảnh, hình ảnh/video được thu thập bởi thiết bị di động là điều cần thiết để đào tạo một mô hình toàn cầu. Federated Learning (FL) là một cách tiếp cận ngày càng được sử dụng rộng rãi để đào tạo thị giác máy tính và các mô hình nhận dạng hình ảnh.

Như đã trình bầy trong phần 4, trong chế độ Federated Learning, máy chủ đám mây đào tạo một mô hình toàn cầu bằng cách tổng hợp các mô hình cục bộ được đào tạo một phần bởi mỗi thiết bị đầu dựa trên mô hình trung bình lặp đi lặp lại Như được mô tả trong Hình 7.1, trong mỗi lần lặp lại đào tạo, một thiết bị thực hiện đào tạo dựa trên mô hình được tải xuống từ máy chủ AI bằng cách sử dụng dữ liệu đào tạo cục bộ. Sau đó, thiết bị báo cáo kết quả đào tạo tạm thời (ví dụ: độ dốc cho DNN) cho máy chủ đám mây thông qua các kênh 5G/6G UL. Máy chủ tổng hợp các gradient từ các thiết bị và cập nhật mô hình toàn cầu. Tiếp theo, mô hình toàn cầu được cập nhật được phân phối đến các thiết bị thông qua các kênh DL 5G/6G, các thiết bị có thể thực hiện đào tạo cho lần lặp tiếp theo.



Hình 7.1. Học liên kết trên hệ thống 5G

Một quy trình Federated Learning lặp đi lặp lại được minh họa trong Hình 7.2. Trong lần lặp lại đào tạo thứ N, thiết bị thực hiện đào tạo dựa trên mô hình được tải xuống từ máy chủ đào tạo FL bằng cách sử dụng hình ảnh / video được thu thập cục bộ. Sau đó, thiết bị báo cáo kết quả đào tạo tạm thời lặp lại thứ N (ví dụ: độ dốc (gradient) cho DNN) cho máy chủ thông qua các kênh UL 5G/6G. Trong khi đó, mô hình toàn cầu và cấu hình đào tạo cho lần lặp thứ (N + 1) được gửi đến thiết bị. Khi máy chủ tổng hợp các gradient từ các thiết bị cho lần lặp thứ N, thiết bị sẽ thực hiện đào tạo cho lần lặp thứ (N + 1). Đầu ra tổng hợp liên kết được sử dụng để cập nhật mô hình toàn cầu, mô hình này sẽ được phân phối cho các thiết bị, cùng với cấu hình đào tạo được cập nhật.



Hình 7.2. Dòng thời gian học liên kết để nhận dạng hình ảnh

Để sử dụng đầy đủ tài nguyên đào tạo tại thiết bị và giảm thiểu độ trễ đào tạo, quy trình đào tạo được hiển thị trong Hình 7.2 yêu cầu báo cáo kết quả đào tạo cho lần lặp thứ (N-1) và phân phối mô hình / cấu hình đào tạo toàn cầu cho lần lặp thứ (N + 1) được hoàn thành trong quá trình đào tạo của thiết bị cho lần lặp thứ N. Phân tích trong Mục 7.1.2 sẽ được phát triển dựa trên thời gian xử lý. Trong thực tế, dòng thời gian FL thư giãn hơn cũng có thể được xem xét với việc hy sinh tốc độ hội tụ đào tạo.

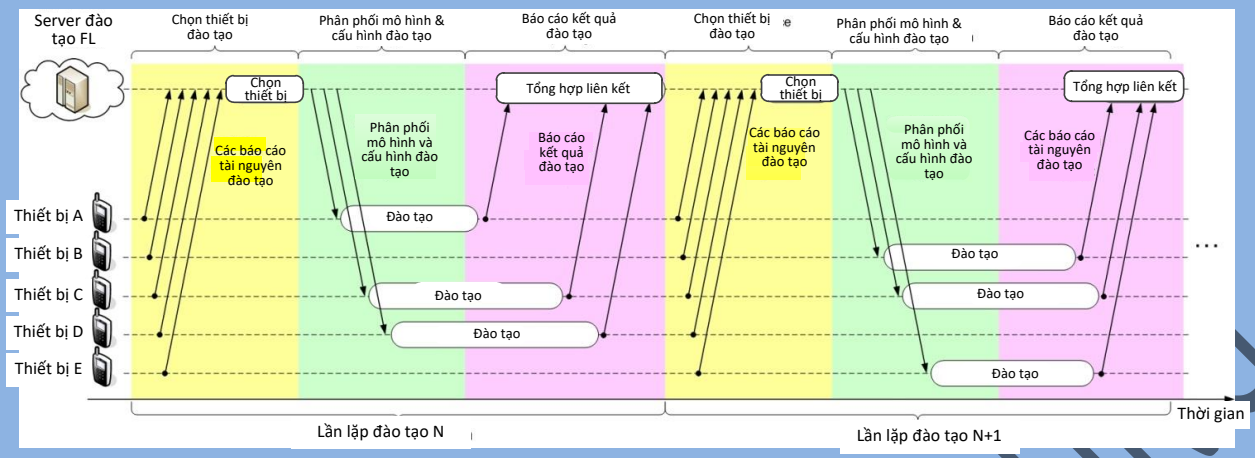
Thời gian đào tạo nên được giảm thiểu vì các thiết bị di động chỉ có thể ở trong môi trường trong một khoảng thời gian ngắn. Hơn nữa, xem xét dung lượng lưu trữ hạn chế tại thiết bị, có thể không thực tế khi yêu cầu thiết bị đào tạo lưu trữ một lượng lớn dữ liệu đào tạo trong bộ nhớ để đào tạo sau khi nó di chuyển ra ngoài môi trường.

Khác với đào tạo phi tập trung được vận hành trong các trung tâm dữ liệu đám mây, Federated Learning qua các hệ thống truyền thông không dây cần được sửa đổi để thích ứng với các điều kiện kênh không dây thay đổi, tài nguyên đào tạo không ổn định trên thiết bị di động và tính không đồng nhất của thiết bị . Giao thức Học liên kết cho truyền thông không dây có thể được mô tả trong Hình 7.3.

Đối với mỗi lần lặp, trước tiên các thiết bị đào tạo có thể được chọn. Các thiết bị đào tạo của ứng viên báo cáo tài nguyên tính toán của họ có sẵn cho nhiệm vụ đào tạo cho máy chủ FL. Máy chủ FL thực hiện lựa chọn thiết bị đào tạo dựa trên các báo cáo từ thiết bị và các điều kiện khác, ví dụ: điều kiện kênh không dây của thiết bị, vị trí địa lý, v.v.

Vì vậy, bên cạnh việc thực hiện nhiệm vụ học liên kết, các thiết bị đào tạo trong hệ thống truyền thông có dữ liệu khác để truyền ở đường lên (ví dụ: đối với các giao dịch dịch vụ đang diễn ra), có thể có mức độ ưu tiên cao và không chịu được độ trễ và việc truyền nó có thể ảnh hưởng đến khả năng tải lên mô hình được đào tạo cục bộ của thiết bị. Do đó, việc lựa chọn thiết bị phải tính đến sự đánh đổi để tải lên kết quả đào tạo so với việc tải lên dữ liệu đường lên khác. Hơn nữa, việc bỏ qua một thiết bị khỏi tổng hợp mô hình học tập liên kết cho một hoặc nhiều lần lặp lại ảnh hưởng đến sự hội tụ của mô hình học liên kết. Do đó, việc lựa chọn thiết bị đào tạo ứng viên qua các liên kết không dây phức tạp hơn so với học liên kết trong các trung tâm dữ liệu.

Sau khi các thiết bị đào tạo được chọn, máy chủ FL sẽ gửi các cấu hình đào tạo đến các thiết bị đào tạo đã chọn, cùng với mô hình toàn cầu để đào tạo. Một thiết bị đào tạo bắt đầu đào tạo dựa trên mô hình toàn cầu và cấu hình đào tạo nhận được. Khi kết thúc đào tạo cục bộ, một thiết bị báo cáo kết quả đào tạo tạm thời của nó (ví dụ: độ dốc cho DNN) cho máy chủ FL. Trong Hình 7.3, việc lựa chọn thiết bị đào tạo được thực hiện và các cấu hình đào tạo được gửi đến các thiết bị đào tạo khi bắt đầu mỗi lần lặp. Nếu các điều kiện (ví dụ: tài nguyên tính toán của thiết bị, điều kiện kênh không dây, các giao dịch dịch vụ khác của thiết bị đào tạo) không được thay đổi, việc lựa chọn lại thiết bị đào tạo và cấu hình lại đào tạo có thể không cần thiết cho mỗi lần lặp, tức là cùng một nhóm thiết bị đào tạo có thể tham gia đào tạo với cùng một cấu hình cho nhiều lần lặp. Tuy nhiên, việc lựa chọn các thiết bị đào tạo nên được xen kẽ theo thời gian để đạt được một lấy mẫu độc lập và phân phối giống hệt nhau từ tất cả các thiết bị, tức là tạo cơ hội công bằng cho tất cả các thiết bị đóng góp vào mô hình tổng hợp.



Hình 7.3. Giao thức Federated Learning điển hình trên hệ thống truyền thông không dây

7.1.2. Các yêu cầu tiềm năng mới cần thiết để hỗ trợ trường hợp sử dụng

Như đã xét ở trên, để giảm thiểu độ trễ đào tạo đói với học liên kết cho nhận dạng hình ảnh, tài nguyên tính toán cho nhiệm cụ đào tạo phải được sự dụng hết, nghĩa là cần duy trì ống đào tạo trên hình 7.1.

Nếu xét đào tạo mô hình CNN VGG16-BN 7 bit sử dụng ảnh 224×224×3 làm dữ liệu đào tao, thì bảng 7.1 phải chỉ ra rằng tổng trễ độ trễ tải lên gradient, độ trễ tổng hợp liên kết và trễ tải xuống mô hình toàn cầu không lớn hơn thời gian tính toán của GPU tại thiết bị cho một lần lặp. Đối với các kích thước lô (Batch) khác nhau, tải lên gradient và tải xuống mô hình toàn cầu cho mỗi lần lặp cần phải kết thúc trong thời gian [52~162 ms], tương ứng.

Ngay cả khi nhiều thiết bị đào tạo có mặt trong một ô, tất cả các thiết bị đào tạo phải kết thúc tải lên trong thời gian trễ như trong bảng 7.1.

Kích thước của mô hình VGG_BN 8 bit là 32 MByte cho cả các gradient được đào tạo và mô hình toàn cầu. Vì thế, để kết thúc tải lên gradient và tải xuống mô hinh toàn cầu, tốc độ dữ liệu đường lên và đường xuống được chỉ ra trong bảng 6.13 là [6,5 Gbit/s đến 20,3 Gbit/s], tương ứng. Cũng cần lưu ý rằng 132 Mbyte là kích thước không nén. Kích thước này có thể giảm nếu các kỹ thuật nén mô hình/ dữ liệu được tiếp nhận.

Trong các yêu cầu cũ đối với hệ thống 5G, vùng phủ đầy đủ luôn được mong muốn đối với các UE. Tuy nhiên, nhiệm vụ đào tạo mô hình AI/ML có thể giảm nhẹ đến một mức độ nhất định các yêu cầu về vùng phủ mạng liên tục. Khi một server FL chọn các thiết bị đào tạo cho một nhiệm vụ học liên kết, nó có thể tìm cách chọn các UE trong một vùng phủ thỏa mãn, nếu chúng có thể thu thập dữ liệu đào tạo cần thiết. Điều này có nghĩa là, thậm chí trong một vùng phủ sóng không liên tục của sóng mm 5G, nhiệm vụ học liên kết vẫn có thể được thực hiện tốt. Điều này cung cấp cho các nhà mạng 5G một dịch vụ tốt hơn khi khảo sát sử dụng tài nguyên phổ tần FR2 (dải tần số 2 của 5G).

Các yêu cầu KPI cần thiết để hỗ trợ trường hợp sử dụng:

- Hệ thống 5G sẽ phải hỗ trợ tốc độ dữ liệu trải nghiệm DL và UL được cho trong bảng 7.1để cho phép đào tạo liên kết không nén.

Bảng 7.1. Độ trễ và tốc độ dữ liệu trải nghiệm DL/UL cho học liên kết để nhận dạng ảnh

Kích thước lô mini (các ảnh)	Thời gian tính toan của GPU (ms)	Độ trễ cực đại cho tải lên gradient được đào tạo và phân bố mô hình toàn cầu (xem chú thích 1)	Tốc độ dữ liệu DL/UL trải nghiệm người dùng cho tải lên gradient được đào tạo và phân bố mô hình toàn cầu (xem chú thích 2)
64	325	3,24s	325Mbit/s
32	191	1,9s	55Mbit/s
16	131	1,3s	810Mbit/s
8	111	1,1s	960Mbit/s
4	105	1,04s	1,0Gbit/s

Chú thíc 1: Độ trễ trong bảng được cho là 20 lần thời gian tính toán của GPU của thiết bị cho kích thước lô mini được cho trước.

Chú thích: Giá trị được cung cấp trong bảng là các yêu cầu có tính toán đối với mô hình VGG 16 BN 8bit với kích thước 132 Mbyte, khi cho trước kích thước lô min và khoảng thời gian của một lần lặp tính theo giây. Có thể giảm tốc độ dữ liệu DL/UL trải nghiệm người dùng, ví dụ: thiết lập tời gian một lần lặp lớn hơn, áp dụng FL nén hoặc sử dụng một mô nhình AI/ML khác.

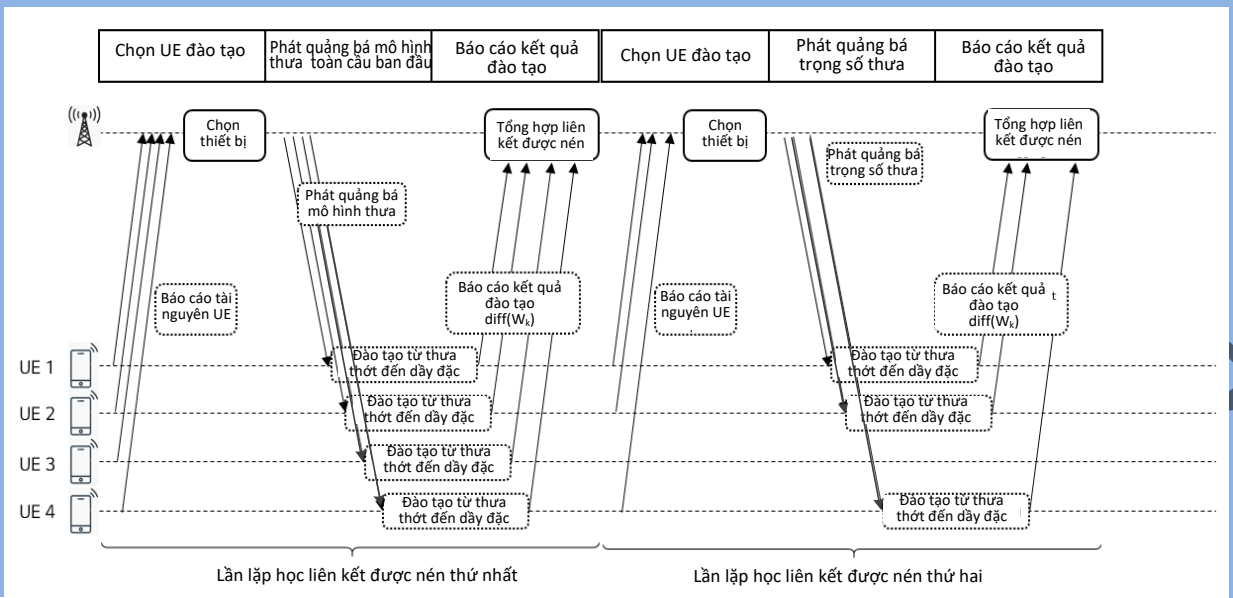
7.2. Học liên kết nén để xử lý hình ảnh/ video

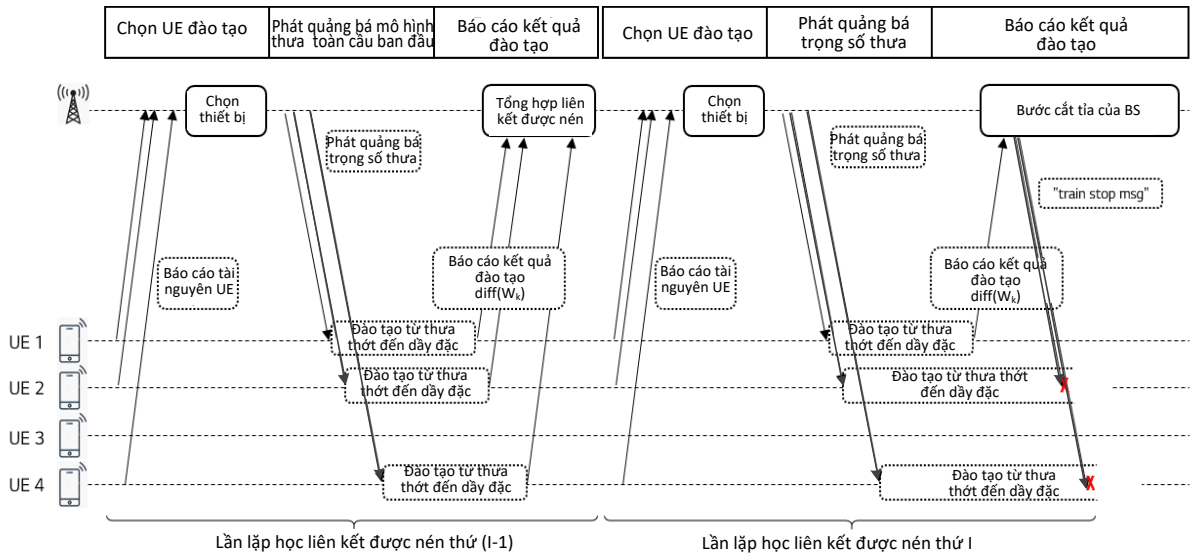
7.2.1. Mô tả

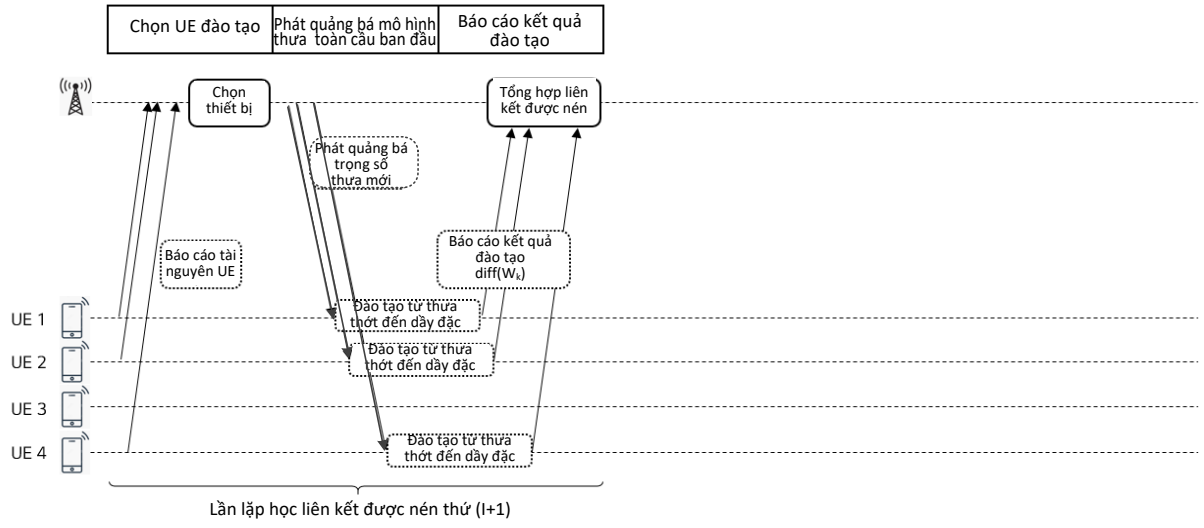
Học liên kết có thể được sử dụng để đào tạo các mô hình AI/ML dựa trên số lượng hình ảnh và video được tạo bởi camera trong thiết bị di động bằng cách trao đổi lặp đi lặp lại gradient của các mô hình cập nhật thay vì hình ảnh và video trực tiếp của người dùng. Bởi vì phương pháp này có thể sử dụng hình ảnh và video từ nhiều người dùng, hiệu năng của mô hình AI/ML được đào tạo có thể cao hơn đáng kể so với một trường hợp độc lập. Tuy nhiên, các phương pháp học liên kết cơ bản có thể có nhược điểm bởi lưu lượng truy cập đường lên lớn và chi phí tính toán cao trên thiết bị di động. Do đó, sẽ có lợi khi xem xét phương pháp học tập liên kết nén (CFL: Compressed Federated Learning), cho phép chuyển các mô hình nén (không đầy đủ) trong thời gian học tập.

Hình 7.4 cho thấy quy trình thiết yếu của CFL. CFL lặp đi lặp lại thực hiện một tập hợp ba giai đoạn hoạt động. Để mô tả các lần lặp lại trong CFL, ta xet một chu kỳ I: mỗi chu kỳ bắt đầu với lần lặp thứ 1 và kết thúc bằng lần lặp thứ I, ngay sau đó là lần lặp thứ 1 của chu kỳ tiếp theo (ví dụ: (I + 1) lần lặp thứ nhất). Đối với mỗi lần lặp, ba giai đoạn hoạt động bao gồm lựa chọn UE đào tạo, phân phối trọng số thưa thớt và các giai đoạn báo cáo kết quả đào tạo. Các hoạt động của ba giai đoạn này trong lần lặp đầu tiên và hoạt động trong lần lặp cuối cùng khác với các hoạt động trong lần lặp lại khác.

Mỗi lần lặp lại trong CFL bắt đầu với giai đoạn lựa chọn UE đào tạo, tại đó máy chủ CFL chọn một tập hợp người dùng có sẵn từ những người dùng ứng cử viên để liên kết với mô hình AI/ML cùng mục đích. Đối với những người dùng được chọn, sẵn sàng tham gia vào quá trình học tập vì đang ở trạng thái sẵn sàng, máy chủ CFL truyền thông tin cấu hình đào tạo. Ở giai đoạn tiếp theo, máy chủ CFL gửi mô hình toàn cầu thưa thớt, có thể là phiên bản ban đầu của mô hình AI/ML trong lần lặp đầu tiên. Nếu không, mô hình toàn cầu thưa thớt là một phiên bản tổng hợp dựa trên thông tin báo cáo của người dùng. Sau đó, mỗi UE đào tạo một mô hình nhận được sau khi mở rộng mô hình không gian và báo cáo kết quả đào tạo trung gian cho máy chủ CFL, trong đó kết quả đào tạo được nén để chỉ có các gradient trọng số giá trị đáng kể để áp dụng nén mô hình. Bằng cách đó, đường lên trong suốt yêu cầu có thể được giảm đáng kể so với phương pháp học liên kết cơ bản mà không cần nén. Trong lần lặp lại cuối cùng, lần lặp thứ I, Máy chủ CFL gửi 'thông báo dừng đào tạo' đến UE để UE có thể ngừng gửi bản cập nhật của nó lâu hơn và Máy chủ CFL thực hiện tinh chỉnh bằng cách cắt tỉa các nút không cần thiết. Trong suốt nhiều lần lặp lại này của một chu kỳ (tức là từ lần lặp lần 1 đến lần lặp lần thứ I) như trong hình, mô hình AI/ML sẽ được nâng cao dần dần dựa trên dữ liệu người dùng trong mạng di động với các yêu cầu giảm về thông lượng đường lên và đường xuống.


Chọn UE đào tạo
Phát quảng bá mô hình thưa toàn cầu ban đầu
Báo cáo kết quả đào tạo
Chọn UE đào tạo
Phát quảng bá trọng số thưa
Báo cáo kết quả đào tạo
Chọn thiết bị
Tổng hợp liên kết được nén
Phát quảng bá mô hình thưa
Báo cáo tài nguyên UE
Báo cáo kết quả đào tạo diff(W_k)
Đào tạo từ thưa thớt đến dầy đặc
UE 1
UE 2
UE 3
UE 4
Lần lặp học liên kết được nén thứ nhất
Lần lặp học liên kết được nén thứ hai


Chọn UE đào tạo
Phát quảng bá mô hình thưa toàn cầu ban đầu
Báo cáo kết quả đào tạo
Chọn UE đào tạo
Phát quảng bá trọng số thưa
Báo cáo kết quả đào tạo
Chọn thiết bị
Tổng hợp liên kết được nén
Phát quảng bá trọng số thưa
Bước cắt tỉa của BS
"train stop msg"
Báo cáo tài nguyên UE
Báo cáo kết quả đào tạo diff(W_k)
Đào tạo từ thưa thớt đến dầy đặc
UE 1
UE 2
UE 3
UE 4
Lần lặp học liên kết được nén thứ (l-1)
Lần lặp học liên kết được nén thứ l


Chọn UE đào tạo
Phát quảng bá mô hình thưa toàn cầu ban đầu
Báo cáo kết quả đào tạo
Chọn thiết bị
Tổng hợp liên kết được nén
Phát quảng bá trọng số thưa mới
Báo cáo tài nguyên UE
Báo cáo kết quả đào tạo diff(W_k)
Đào tạo từ thưa thớt đến dầy đặc
UE 1
UE 2
UE 3
UE 4
Lần lặp học liên kết được nén thứ (l+1)

Hình 7.4. Dòng thời gian học liên kết nén để nhận dạng hình ảnh

7.2.2.Quá trình diễn ra

Điều kiện ban đầu

UE có thể có khả năng phần cứng và thuật toán tính toán để đào tạo mô hình AI/ML chẳng hạn như nhận thức hình ảnh và video.

UE có thể gửi kết quả đào tạo trung gian đến máy chủ CFL.

Máy chủ CFL có thể chọn thiết bị đào tạo và xác định cấu hình đào tạo.

Máy chủ CFL có thể tổng hợp các kết quả đào tạo trung gian và tạo ra một mô hình toàn cầu thưa thớt cho lần lặp lại học tập tiếp theo.

Máy chủ CFL có thể phân phối chế độ AI/ML toàn cầu cho một nhóm người dùng đã chọn.

Các luồng dịch vụ

Bước 1: Máy chủ CFL chọn người dùng đào tạo từ người dùng ứng viên.

Bước 2: Máy chủ CFL gửi thông tin cấu hình đến người dùng đã chọn.

Bước 3: Máy chủ CFL phân phối mô hình toàn cầu thưa thớt ban đầu (hoặc, tổng hợp) cho những người dùng được chọn thông qua mạng 5G.

Bước 4: Mỗi UE mở rộng mô hình toàn cầu thưa thớt và đào tạo mô hình mở rộng bằng cách sử dụng dữ liệu cục bộ của nó. Sau đó, mỗi UE chỉ gửi gradient trọng số có giá trị đáng kể đến máy chủ CFL.

Bước 5: Máy chủ CFL tổng hợp các kết quả đào tạo nhận được từ các UE đào tạo và cập nhật mô hình toàn cầu bằng cách sử dụng kết quả tổng hợp.

Bước 6: Cho đến khi mô hình AI/ML đạt được mức tăng cường hiệu suất bão hòa, quá trình sẽ chạy lặp đi lặp lại từ bước 1.

Bước 7: Nếu không, máy chủ CFL thực hiện tinh chỉnh để nén mô hình toàn cầu cho mô hình toàn cầu. Quá trình này có thể được áp dụng thường xuyên để cải thiện băng thông và hiệu quả tài nguyên tính toán trước khi hoàn thiện đào tạo.

Cuối cùng, máy chủ CFL phân phối mô hình toàn cầu thưa thớt mới cho tất cả người dùng cần cùng một mô hình AI/ML.

Điều kiện kết thúc

Đối với một UE tiềm năng, CFL có thể giảm yêu cầu thông lượng đường lên và đường xuống cho quá trình học liên kết. Ngoài ra, độ phức tạp tính toán trong UE có thể giảm đáng kể do cho phép mô hình nén.

7.2.3. Các yêu cầu mới tiềm năng cần thiết để hỗ trợ trường hợp sử dụng

Hệ thống 5G/6G sẽ hỗ trợ tải lên một gradient được đào tạo cho mỗi lần lặp lại của học liên kết nén với độ trễ tối đa là 1,05 ~ 3,25 giây.

Hệ thống 5G/6G sẽ hỗ trợ tải xuống mô hình toàn cầu cho mỗi lần lặp lại của học liên kết nén với độ trễ tối đa là 1,05 ~ 3,25 giây.

Hệ thống 5G sẽ hỗ trợ truyền đơn hướng UL với tốc độ dữ liệu UL trải nghiệm 26.13 ~ 80.88Mbit / s và tính khả dụng của dịch vụ truyền thông không thấp hơn [99.9%] để báo cáo các gradient được đào tạo cho Học liên kết nén.

Hệ thống 5G sẽ hỗ trợ truyền đa hướng DL với tốc độ dữ liệu DL trải nghiệm của người dùng 26,13 ~ 80,88Mbit/s và tính khả dụng của dịch vụ truyền thông không thấp hơn [99,9%] để phân phối mô hình toàn cầu cho học liên kết nén.

PHẦN 3

KẾT HỢP AI VÀ CÁC CÔNG NGHỆ KHÔNG DÂY VÀO CÁC MẠNG KHÔNG DÂY DI ĐỘNG 5G VÀ 6G

Kết hợp AI/ML và các công nghệ truyền thông không dây đã trở thành một trong các xu thế công nghệ chính hướng đến 2030. Sự kết hợp này bao gồm sử dụng AI/ML để cải thiện hiệu năng truyền dẫn không dây và hỗ trợ triển khai AI/ML với các mạng không dây. Phần này sẽ trình bày các nghiên cứu của 3GPP để kết hợp AI/ML với các công nghệ truyền thông cho mạng 5G và 6G. Do mạng 6G chưa được chuẩn hóa, nên các trình bày ở đây chỉ đề cập đến các nghiên cứu cho mạng 5G. Nhưng các nghiên cứu này cũng tạo tiền đề cho các nghiên cứu cho mạng 6G hướng đến 2030.

1. BỘ KHUNG 3GPP AI/ML CHUNG

Để thiết lập một bộ khung AI/ML 3GPP chung cho giao diện không vô tuyến, những nỗ lực đáng kể từ các thành viên 3GPP đã được dành để xây dựng các thuật ngữ chung liên quan đến các chức năng, thủ tục và giao diện AI/ML. Hình 1 phác thảo bộ khung chức năng 3GPP AI/ML cho giao diện vô tuyến NR. Bộ khung mô tả một tập hợp các chức năng cốt lõi, bao gồm thu thập dữ liệu, đào tạo mô hình, quản lý, suy luận và lưu trữ mô hình.

Hình 1. Phác thảo bộ khung chức năng 3GPP AI/ML cho giao diện vô tuyến NR

Mô hình AI/ML cần được phát triển, triển khai và quản lý trong toàn bộ vòng đời — một quy trình được gọi là quản lý vòng đời mô hình AI/ML (AI/ML Model LCM: AI/ML Model Life Cycle Managment). 3GPP đã nghiên cứu hai phương pháp riêng biệt để quản lý vòng đời của mô hình AI/ML tại thiết bị người dùng (UE). Phương pháp đầu tiên được phân loại là LCM dựa trên chức năng (Function Based LCM) . Chức năng đề cập đến một tính năng được AI/ML hỗ trợ hoặc nhóm tính năng được hỗ trợ bởi một cấu hình. Nhận dạng chức năng AI/ML (AI/ML Function Identification) thúc đẩy sự hiểu biết lẫn nhau giữa mạng và UE về chức năng AI/ML. Quá trình nhận dạng chức năng có thể được tích hợp trong bộ khung báo hiệu theo khả năng UE

hiện có. Về cơ bản, cấu hình được điều chỉnh phù hợp với các điều kiện được chỉ ra bởi khả năng UE. Sau đó, sau khi nhận dạng các chức năng, UE có thể báo cáo các bản cập nhật liên quan đến các chức năng có thể áp dụng trong số những chức năng được định cấu hình hoặc được nhận dạng. Trong LCM dựa trên chức năng, mạng chỉ ra lựa chọn, kích hoạt, hủy kích hoạt, chuyển đổi và quay lại (Fallback) chức năng AI / ML thông qua báo hiệu 3GPP. Đáng chú ý, (các) mô hình AI/ML chính xác hỗ trợ cho một chức năng nhất định có thể không được xác định trên mạng.

Phương pháp thứ hai được phân loại là LCM dựa trên nhận dạng mô hình (Model Identity based LCM) . ID mô hình (Model ID) đóng vai trò là mã định danh đặc biệt cho mô hình AI/ML, trong đó mô hình có thể là logic và ánh xạ của nó đến mộ mô hình vật lý khi triển khai. Việc xác định mô hình AI/ML đảm bảo sự hiểu biết lẫn nhau giữa mạng và UE liên quan đến mô hình AI/ML được đề cập. Cụ thể, mô hình AI/ML được nhận dạng bằng ID mô hình được chỉ định tại mạng và UE cho biết mô hình AI/ML được hỗ trợ của nó cho mạng. Bên cạnh ID mô hình, mô hình có thể có các điều kiện đi kèm như một phần của định nghĩa khả năng UE cũng như các điều kiện bổ sung (ví dụ: kịch bản, trang web và bộ dữ liệu) xác định khả năng ứng dụng của mô hình. Trong LCM dựa trên ID mô hình, cả mạng và UE đều có thể thực hiện lựa chọn, kích hoạt, hủy kích hoạt, chuyển đổi và quay lại mô hình AI/ML bằng cách sử dụng ID mô hình tương ứng.

Việc triển khai thương mại tính năng hỗ trợ AI/ML phụ thuộc vào khả năng mang lại hiệu suất đáng tin cậy trên nhiều kịch bản, cấu hình và điều kiện cụ thể của địa điểm trong hệ thống truyền thông di động. Để đạt được mục tiêu này, 3GPP đã nghiên cứu ba cách tiếp cận: tổng quát hóa mô hình (model generalization), chuyển đổi mô hình (model switching) và cập nhật mô hình (model update). Tổng quát hóa mô hình nhằm mục đích phát triển một mô hình có thể tổng quát hóa cho các tình huống, cấu hình hoặc địa điểm khác nhau. Ngoài ra, một tập hợp các mô hình cụ thể có thể được phát triển — từ kịch bản cụ thể đến cấu hình hoặc địa điểm cụ thể.

Trong phạm vi mô hình này, kỹ thuật chuyển đổi mô hình (Model Switching Technique) được khai thác để giải quyết hiệu quả các tình huống, cấu hình hoặc địa điểm khác nhau. Quá trình cập nhật mô hình, thường liên quan đến việc tinh chỉnh, đòi hỏi sự thích ứng linh hoạt của cấu trúc mô hình hoặc các tham số của nó để đáp ứng với những thay đổi trong kịch bản, cấu hình hoặc các site. Một nguyên tắc then chốt làm nền tảng cho các phương pháp tiếp cận này là đảm bảo rằng hiệu suất của các tính năng hỗ trợ AI/ML vẫn ở mức bằng hoặc tốt hơn so với các hoạt động không dựa trên AI/ML cũ. Do đó, giám sát hiệu suất là điều bắt buộc đối với các tính năng hỗ trợ AI/ML, yêu cầu các chức năng như tính toán các chỉ số hiệu suất được giám sát, báo cáo kết quả giám sát và cơ chế báo hiệu điều khiển để nhanh chóng phục hồi sau lỗi.

Các trường hợp sử dụng khác nhau yêu cầu mức độ cộng tác khác nhau giữa mạng và UE cho các hoạt động AI/ML tương ứng. 3GPP đã xác định ba cấp độ hợp tác mạng và UE đặc biệt:

- **Cấp độ x–không hợp tác**: Ở cấp độ này, các hoạt động AI/ML dựa trên việc triển khai độc quyền mà không có bất kỳ cải tiến tiêu chuẩn cụ thể nào được điều chỉnh cho các chức năng của AI/ML.
- **Cấp độ y – cộng tác dựa trên báo hiệu mà không cần chuyển giao mô hình**: Ở cấp độ này, các hoạt động AI/ML tích hợp các cải tiến tiêu chuẩn dành riêng để tạo điều kiện thuận lợi cho quá trình mà không liên quan đến việc chuyển mô hình. Ở đây, 'chuyển mô hình' đề cập đến việc phân phối mô hình AI/ML qua giao diện vô tuyến từ thực thể này sang thực thể khác, được thực hiện theo cách không trong suốt (not transparent) đối với các cơ chế tín hiệu 3GPP.
- **Cộng tác dựa trên tín hiệu cấp z với chuyển giao mô hình**: Trong cấp này, các hoạt động AI/ML không chỉ bao gồm tích hợp báo hiệu mới mà còn tận dụng các khả năng chuyển giao mô hình nâng cao.

Về bản chất, các cấp độ cộng tác này bao gồm một phạm vi từ sự tham gia tối thiểu đến tích hợp sâu, biểu thị tính linh hoạt và khả năng thích ứng của các hoạt động AI/ML trong các bối cảnh khác nhau.

2. AI TRONG NGHIÊN CỨU CÁC KHÍA CẠNH HỆ THỐNG CỦA 3GPP (3GPP SA: 3GPP SYSTEM ASPECT)

2.1. Công nghệ cho phép tự động hóa mạng (eNA: enabler for Network Automation)

Tự động hóa mạng được đề cập như là sử dụng phần mềm (thường được tăng cường bởi các chức năng tiên tiến của các công nghệ AI) để thực hiện nhiều nhiệm vụ. Các nhiệm vụ này bao gồm: quy hoạch, triển khai, lập cấu hình, sắp xếp và điều phối (Orchestration), phân tích và đảm bảo các mạng và các dịch vụ. Ngay từ phát hành 15, 3GPP đã đưa ra thực thể Chức năng phân tích số liệu mạng (NWDAF: Network Data Analytics Function). Chức năng của thực thể này như là một hộp đen AI để đảm bảo phân tích hiệu năng và dự báo liên quan mạng được rút ra từ chọn lựa dữ liệu đặc thù từ một hay nhiều nguồn dữ liệu, chẳng hạn các chức năng mạng (Network Functions), quản trị và bảo trì vận hành (OAM: Operation Administration and Maintenance) và UE.

Các phát hành tiếp theo đã mở rộng khả năng của nó để bao hàm cả thu thập dữ liệu và các tính năng để lộ các phân tích mạng (Network Anlalytics Exprosure).Các nâng cao này là then chốt trong việc tăng chức năng và sự hữu ích của tính năng này trong kiến trúc 5GC. Đặc biệt là hiện nay các kiểu phân tích khác nhau được đặc tả và phụ thuộc và nhận dạng (hay các nhận dạng) của phân tích (Analytics ID) được yêu cầu bởi một phần tử mạng người tiêu dùng (Consumer-Network Function: chức năng người tiêu dùng- mạng). Có lẽ là, thực hiện trên học liên kết (FL) sớm nhất trong 3GPP là tại eNA, tại đây các NWDAF khác nhau có thể liên kết hướng dẫn một

mô hình để phân tích mạng. Tiêu chuẩn này cũng đặc tả các giao thức để các NWDAF hình thành quá trình FL hoặc như là server hoặc client.

2.2. AMMT của Phát hành 18

Năm 2019, cuộc họp TSG SA#86 chấp thuận mục làm việc (Work Item) tập trung lên nghiên cứu các đặc trưng lưu lượng và các yêu cầu hiệu năng liên quan đến AMMT (Artificial Intelligence and Machine Learning Model Trasfer: chuyển giao mô hình trí tuệ nhân tạo và máy học) trong hệ thống 5G (5G System). Mục tiêu là nói rõ các tiêu chí hiệu năng đặc thù bao gồm trễ đầu cuối đầu cuối, tốc độ số liệu trải nghiệm và tính khả dụng của dịch vụ truyền thông. Ngoài ra các yêu cầu dịch vụ như quản lý QoS AI/ML, phân bố và chuyển giao mô hình/dữ liệu AI/ML, cũng như hiệu năng mạng và giám sát/dự báo sử dụng tài nguyên cũng được định nghĩa. Các nỗ lực này nhằm cho phép 5GS hỗ trợ một loạt các hoạt động AI/ML cho các ứng dụng đa dạng khác nhau bao gồm (nhưng không phải tất cả) nhận biết hình ảnh/tiếng nói, biên tập/các nâng cao truyền thông (Media Editing/Enhancements), điều khiển robot và các ứng dụng ô tô.

- Phân chia hoạt động AM/ML giữa các điểm cuối AM/ML;
- Phân bố mô hình/dữ liệu AM/ML và chia sẻ trên 5GS;
- Phân bố/FL trên 5GS.

Nhóm công tác về các khía cạnh hệ thống 3GPP SA1 (AS: System Aspect) đã hoàn thiện nghên cứu gai đoạn 1 về AMMT bằng việc giải quyết các trường hợp sử dụng và các yêu cầu hiệu năng tiềm năng để hỗ trợ AI lớp ứng dụng, phân bố và chuyển giao mô hình ML trong 5GS. Nghiên cứu này cũng bao gồm cả nhận dạng các đặc trưng lưu lượng đi kèm với chuyển giao, phân bố AI/ML và đào tạo các ứng dụng khác nhau. Bắt đầu của nghiên cứu các thủ tục này là khám phá các khả năng của nền tảng 5GS để hỗ trợ các hoạt động AI/ML của lớp ứng dụng được bắt đầu bởi 3GPP SA2 vào cuối năm 2021.

SA 2 đã giải quyết thành công một số vấn đề quan trọng không chỉ giới hạn đến giám sát sử dụng tài nguyên mạng mà còn bao gồm cả bảo vệ thông tin 5GC không bị lộ đến UE và phía thứ ba, cải thiện QoS và chính sách, đặc biệt là miền các hoạt động FL phức tạp. Các giải quyết này đóng góp vào sự bền vững và hiệu quả tổng thể của hệ thống, đảm bảo tính riêng tư được nâng cao, hiệu quả và các khả năng phát triển các tiêu chuẩn viễn thông.

Trong phát hành 18, việc đưa FL vào 3GPP đánh dấu một cột mốc quan trọng. Các thực thể trước hết liên quan đến đóng góp vào FL là NWDAF và AF (Application Function: chức năng

ưng dụng). NWDAF đóng vai trò trung tâm trong hướng dẫn đào tạo mô hình và các cập nhật thông qua một FL server NWDAF thực hiện điều phối nhiều FL Client NWDAF. Sự cộng tác này hỗ trợ việc kết hợp dữ liệu từ nhiều dạng nguồn khác nhau trong khi vẫn đảm bảo an ninh và tính riêng tư. Mặt khác, AF chịu trách nhiệu lập lịch cho các UE tham gia và phơi sáng (Explosure) các yêu cầu thông qua NEF (Network Explosure Function: chức năng để lộ hay phơi sáng mạng) của 5GC. Quá trình này có một tập các nguyên tắc để đảm bảo sự tham gia hiệu quả cả các UE trong quá trình FL trong khi vẫn duy trì tính toàn vẹn và hiệu năng mạng.

2.3. FL được hỗ trợ bởi truyền thông thiết bị đến thiết bị (D2D: Device to Device)

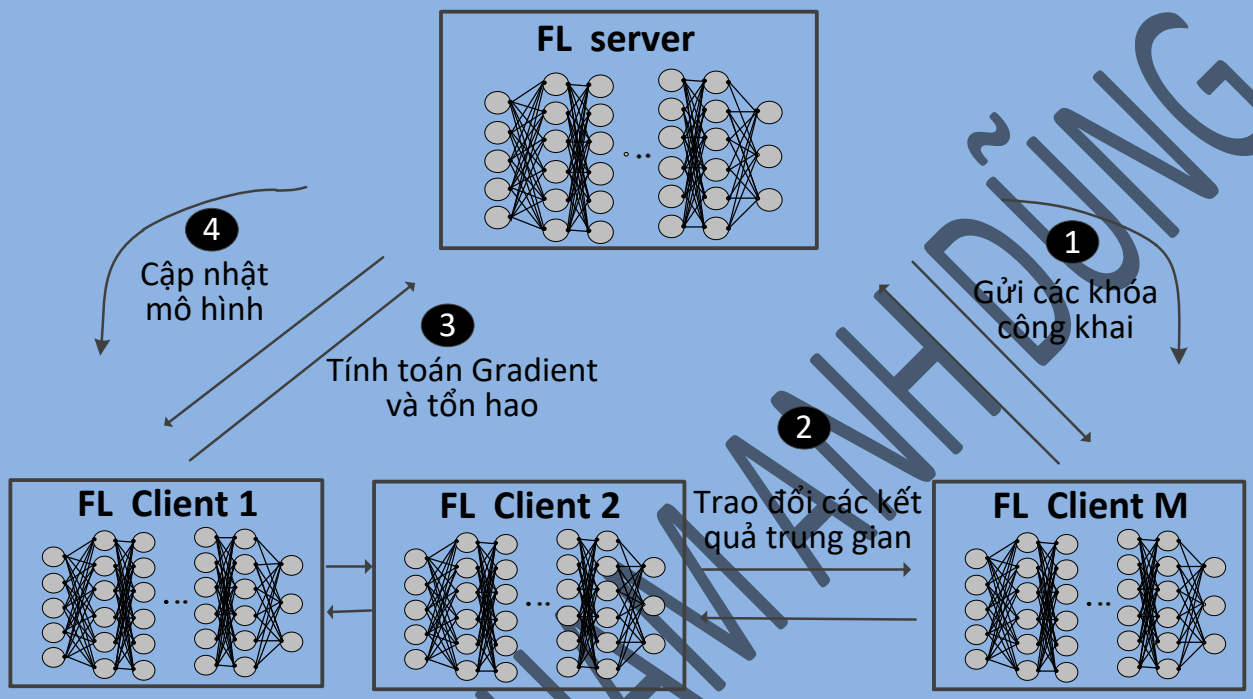
Để giảm tải (Offloading) các nhiệm vụ từ một UE này sang một UE khác trong vùng gần, nhiều nhà cung cấp dịch vụ đầu cuối đã đề xuất FL dựa trên (hay được hỗ trợ) bởi D2D.Giải pháp này nhằm giảm bớt các công sức tính toán và bảo tồn công suất bằng cách phân phát các nhiệm vụ giữa các tiết bị gần nhau. Trong trường hợp này, các UE chuyển tiếp (Relay UE) cần thủ tục khám phá khi chọn các FL client, đôi khi với sự hỗ trợ của AF. Vì thế cần nghiên cứu chính sách và điều khiển QoS bao gồm các ngưỡng QoS được tổng hợp, để hỗ trợ FL trên D2D hiệu quả.

Đáng tiếc là còn quá sớm để quyết định quy mô mà chức năng AI/ML dựa trên D2D có thể đóng góp cho hệ thống khi chưa có sự triển khai trước tiên một bộ khung AI/ML đầy đủ trong cả RAN và 5GC. Tuy nhiên có thể kỳ vọng rằng sử dụng FL được D2D hỗ trợ có thể cung cấp tiềm năng các lợi ích đang kể liên quan đến hiệu suất chuyển tải nhiệm vụ và tối ưu hóa tài nguyên một khi một bộ khung AL/ML bền vững được thiết lập trên toàn mạng.

2.4. AMMT của Phát hành 19

Đối với các mục tiêu đặc thù để nâng cao hỗ trợ cho AI/ML trong 5GC của phát hành 19, ưu tiên được dành cho dịch vụ định vị (LCS: Location Service) liên quan đến các nâng cao như là một tính năng trụ cột trên toàn miền của AM/ML. Ưu tiên này nhấn mạnh tầm quan trọng đặc biệt của tích hợp xuôn xẻ các khả năng AI/ML vào các dịch vụ định vị trong bộ khung của 5GC. Trung tâm của nỗ lực này là nghiên cứu một hô hình AI/ML cho định vị dựa trên LMF (Location Management Function: chức năng quản lý vị trí). Nghiên cứu này bao hàm nhiều khía cạnh khác nhau như: huấn luyện mô hình như thế nào, thu thập số liệu như thế nào và suy luận được như thế nào. Đồng thời các khía cạnh quan trọng như: thu thập số liệu của UE, chuyển giao/chuyển phát mô hình đến UE, đồng bộ nhận dạng và quản lý cũng được lập lịch cho nghiên cứu và thảo luận tiếp theo. Các khía cạnh này được giải quyết cùng với các các quyết định và các kết luận được đưa ra trong thời gian hội nghị toàn thể vào tháng 9 năm 2024 để đảm bảo xem xét và đồng bộ toàn diện với việc phát triển các tiêu chuẩn và và các yêu cầu từ góc nhìn RAN. Hai mục tiêu được ưu tiên khác là hỗ trợ học liên kết dọc chiều dọc (VFL: Vertical

Federated Learning), điều khiển chính sách được NWDAF hỗ trợ (NWDAF Assisted Policy) và giảm thiểu hành vi không bình thường của mạng (Network Abnormal Behavior Mitigation). Về tính năng VFL, nó đánh dấu một cột mốc quan trọng, vì nhóm nghiên cứu xem xét kỹ lưỡng sự bắt nguồn của dữ liệu từ cùng một không gian nhưng lại được đặc trưng bởi các không gian tính năng khác nhau. Hình 2 cho thấy thí dụ về một kiến trúc điển hình liên quan đến VFL.



Hình 2. Kiến trúc đối với hệ thống VFL

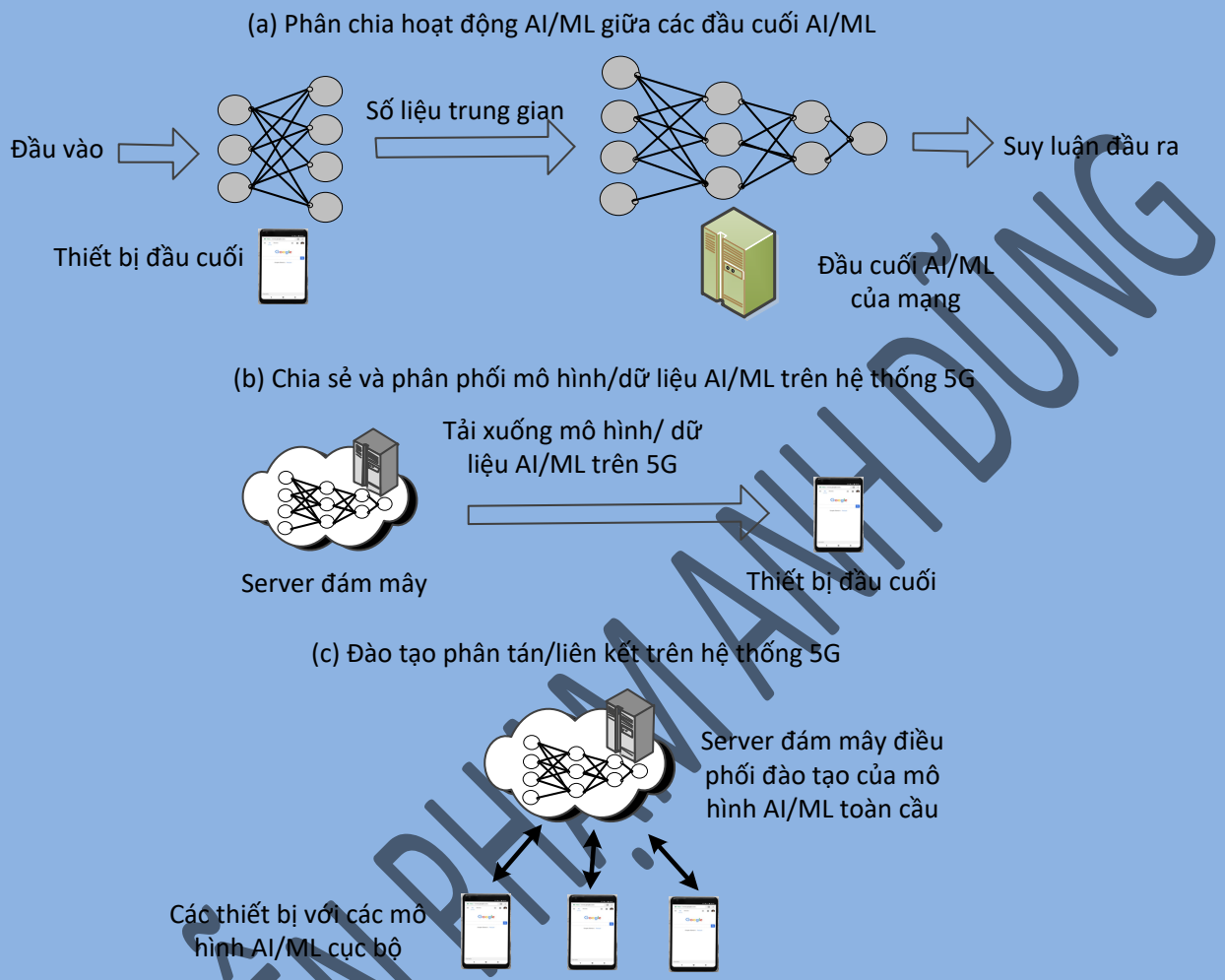
3. HOẠT ĐỘNG CỦA AI/ML LỚP ỨNG DỤNG TRONG HỆ THỐNG 5G

3.1. Chuyển giao mô hình AI/ML (AI/ML Model Transfer)

Các mô hình AI/ML tìm thấy ứng dụng đầu tiên trong các thiết bị di động trong hệ thống 5G, ví dụ nhận dạng ảnh, nhận dạng giọng nói và xử lý video. Do đa dạng ứng dụng và lưu trữ có hạn, nên không thể tải trước tất cả các mô hình xuống UE. Vì thế, tải xuống các mô hình AI/ML là cần thiết. Đối với một số ứng dụng, UE có thể không có đủ tài nguyên tính toán để hực hiện suy luận, trong trường hợp này có thể phải giảm tải suy luận từ UE đến đám mây hoặc biên 5G. Ngoài ra cần chia sẻ dữ liệu đào tạo trên các thực thể khác nhau để liên kết đào tạo một mô hình AI/ML toàn cầu trong hệ thống 5G.

Khuynh hướng chuyển giao các mô hình AI/ML và dữ liệu dẫn đến các kiểu lưu lượng mới cần phải phục vụ trong hệ thống 5G. 3GPP SAI chịu trách nhiệm nhận dạng dịch vụ và các yêu cầu

hiệu năng mới trong các hệ thống 3GPP. Các nghiên cứu đề cập đến ba trường hợp sử dụng như mô tả trên hình 3.



Hình 3. Các trường hợp sử dụng của các hoạt động AI/ML trong hệ thống 5G được nhận dạng trong phát hành 18 của 3GPP

Trong trường hợp sử dụng thứ nhất (hình 3(a)), hoạt động AI/ML được phân chia giữa các điểm cuối. Mục đích của phân chia này là để giữ các phần dịch vụ nhậy cảm tính riêng tư – hay độ trễ của hoạt động này tại UE, nhưng giảm tải các phần nhậy cảm tính toán – năng lượng đến các điểm cuối mạng. Trong trường hợp sử dụng thứ hai (hình 3(b) chia sẻ và phân bố mô hình/ dữ liệu AI/ML (Model/ Data Distribution and Sharing) được thực hiện trên hệ thống 5G để có thể tải xuống từ mạng đến thiết bị mô hình thích ứng khi cần thiết. Trong trường hợp sử dụng thứ ba (hình 3(c)), học phân tán/liên kết (Distributed/ Federated Learning) được thực hiện trên hệ thống 5G, tại đây UE thực hiện đào tạo một phần dựa trên dữ liệu cục bộ và thực thể trung tâm đào tạo mô hình toàn cầu bằng cách tổng hợp các kết quả cục bộ từ các UE.

Nghiên cứu đã nhận dạng các các yêu cầu dịch vụ mới và các chỉ số hiệu năng quan trọng (KPI: Key Performance Indicator) cho đào tạo, suy luận, tải xuống, giám sát, dự đoán và quản lý mô hình AI/ML trên hệ thống 5G. Sau khi hoàn thành nghiên cứu, 3GPP SA1 tiếp tục với mục công tác tiếp theo cho hệ thống 5G để hỗ trợ ba trường hợp sử dụng nói trên. Các yêu cầu này được đưa ra trong đặc tả kỹ thuật (TS 22.261).

Từ góc độ các yêu cầu dịch vụ, TS 22.261 đặc tả rằng đối với chuyển giao mô hình AI/ML, hệ thống 5G sẽ có thể hỗ trợ chức năng dịch vụ mói liên quan đến, ví dụ: giám sát sử dụng tài nguyên, giám sát hiệu năng mạng, dự báo và thay đổi thông báo và quản lý chất lượng dịch vụ (QoS). Việc áp dụng các yêu cầu này tùy thuộc và chính sách nhà mạng, sự chấp thuận của người dùng và các yêu cầu pháp lý.

Từ góc độ yêu cầu hiệu năng, TS 22.261 đặc tả các KPI cho chuyển giao mô hình AI/ML trong hệ thống 5G gồm độ trễ đầu cuối đầu cuối, tốc độ dữ liệu trải nghiệm, mức độ khả dụng của dịch vụ truyền thông, trong số các KPI khác. Chẳng hạn, tốc độ dữ liệu đường xuống trải nghiệm yêu cầu là 1,1 Gbit/s cho nhận dạng ảnh liên quan đến phân phối mô hinhd AI/ML. Một thí dụ khác là nhân dạng ảnh AI/ML phân chia giữa UE và server mạng, trễ tối thiểu được phép đường lên đầu cuối đầu cuối là 2 ms cà tốc độ dữ liệu đường lên trải nghiệm yêu cầu là 1,08 Gbit/s.

3.2. Hỗ trợ hệ thống 5G

3GPP SA1 tiếp tục cải thiện chuyển giao AI/ML trong dịch vụ được nhận dạng bởi hệ thống 5G và các yêu cầu hiệu năng, nhiều trong số chúng liên quan đến hỗ trợ hoạt động AI/ML của lớp ứng dụng trong UE. Mặc dù hệ thống 5G có NWDAF trong 5GC để cho phép tự động hóa mạng, vẫn chưa có đặc tả các giải pháp truyền tải được tối ưu cho hoạt động AI/ML lớp ứng dụng trong UE. Vì thế vẫn cần phải phát triển hệ thống 5G đế hỗ trợ các dịch vụ dựa trên AI/ML. Đây chính là động lực cho mục nghiên cứu hỗ trợ của hệ thống 5G cho các dịch vụ dựa trên AI/ML đang được tiến hành bởi 3GPP SA2, một nhóm công tác chịu trách nhiệm cho phát triển kiến trúc hệ thống 3GPP tổng thể.

Từ góc nhìn các yêu cầu kiến trúc, một chức năng ứng dụng (AF Application Function) điều khiển logic hoạt động AI/ML lớp ứng dụng. AF là một chức năng của mặt phẳng điều khiển tương tác với 5GC để cung cấp các dịch vụ hỗ trợ , chẳng hạn ảnh hưởng của ứng dụng lên định tuyến lưu lượng. Các yêu cầu AF cho sự trợ giúp của 5GC để hỗ trợ hoạt động AI/ML lớp ứng dụng phải được 5GC cấp quyền.

3GPP SA2 khám phá một số mở rộng kiến trúc và chức năng để hõ trợ hoạt động AI/ML lớp ứng dụng. Một vấn đề quan trọng là cách thức giám sát sử dụng tài nguyên mạng liên quan đến hiệu năng của UE như tốc độ dữ liệu và độ trễ. Chức năng để lộ mạng (NEF: Network Exposure Function) trong 5GC có thể hỗ trợ giám sát khả năng cho phép cấu hình, phát hiện và báo cáo

các sự kện giám sát đến phía ngoài được cấp quyền. Các sự kiện giám sát mới có thể được đưa ra để đo và dự đoán sử dụng tài nguyên mạng để hỗ trợ hoạt động AI/ML lớp mạng. Ngoài ra vệc mở rộng để lộ thông tin 5GC cho UE hoặc cho phía ngoài được cấp quyên có thể có ích để hỗ trợ hoạt động AI/ML lớp ứng dụng. Thông tin trợ giúp cho để lộ có thể gồm các tình trạng của UE hoặc mạng, và các dự báo hiệu năng theo, ví dụ: vị trí UE, tải và QOS. Một vấn đề quan trọng nữa là tăng cường khả năng cung cấp, cho phép một bên ngoài cung cấp thông tin cho 5GC tạo điều kiện hỗ trợ hoạt động AI/ML lớp ứng dụng. Một thí dụ về cung cấp thông tin thông số từ bên ngoài là các hành vi của UE được kỳ vọng, như tốc độ UE kỳ vọng và các đặc tính UE kỳ vọng. Cũng cần nghiên cứu các tăng cường 5GC khác như cơ chế định tuyến lưu lượng để hỗ trợ lưu lượng AI/ML.

Hệ thống 5G có các cơ chế đầy đủ cho định nghĩa QoS, thực hiện, điều khiển và giám sát để hỗ trợ các ứng dụng khác nhau. Hoạt động AI/ML lớp ứng dụng dựa trên QoS được cung cấp bởi hệ thống 5GC phía dưới. Cần nghiên cứu xem có cần tăng cường bộ khung chính sách và mô hình QoS hiện thời cho lưu lượng ứng dụng AI/ML, để lộ giám sát và thông tin trạng thái về sử dụng tài nguyên mạng cho phía thứ ba được cấp quyền và thông báo hoạt động AI/ML lớp ứng dụng về các dự báo về sự các sự thay đổi tình trạng mạng. Một mục tiêu khác của nghiên cứu 3GPP SA2 là việc nghiên cứu sự trợ giúp của 5GC để tạo điều kiện thuận lợi cho hoạt động học liên kết của lớp ứng dụng trên hệ thống 5G, bao gồm chọn lựa và quản lý thành viên liên kết, giám sát và để lộ, và ấn định tài nguyên mạng.

3.3. Các khía cạnh về tính riêng tư và an ninh

3GPP SA2 tiếp tục nghiên cứu hỗ trợ hệ thống 5G cho hoạt động AI/ML được mô tả trong phần trước nhận thấy rằng các dữ liệu khác nhau cần được truyền giữa 5GC và AF. Một số dữ liệu (thí dụ,: phân tích QoS, thông tin phân bố địa lý) có thể nhạy cảm tính riêng tư của người dùng và không được để lộ đến AF không được cấp quyền. Một số dữ liệu (ví dụ: phân tích tải mạng) phản ánh trạng thái mạng phải được xử lý một các an ninh để phòng ngừa dữ liệu này bị sử dụng bởi các kẻ tấn công tiềm năng. Vì thế nghiên cứu các khía cạnh tính riêng tư và an ninh của các hoạt động AI/ML lớp ứng dụng là rất quan trọng để tránh rủi do và đe dọa tiềm năng đối với hệ thống 5G và các người dùng.

3GPP SA3, một nhóm công tác chịu trách nhiệm cho an ninh và tính riêng tư trong các hệ thống 3GPP, đang tiến hành nghiên cứu an ninh và tính riêng tư của các dịch vụ dựa trên AI/ML và các ứng dụng trong 5G. Mục đích của nghiên cứu này là để nhận dạng các vấn đề then chốt, các đe dọa tiềm năng, các yêu cầu và các giải pháp. Một vấn đề then chốt được nhận dạng là tính riêng tư và cấp quyên cho để lộ thông tin hỗ trợ của 5GC, bao gồm xác định thông tin hỗ trợ nhậy cảm tính riêng tư và các phương pháp để 5GC bảo vệ và cấp quyền AF truy nhập thông tin

này. Các cơ chế bảo vệ tính riêng tư và cấp quyền thích hợp cần phải có để tôn trọng tính riêng tư và đảm bảo an ninh mạng. Một giải pháp là sử dụng lại các cơ chế hiện thời (ví dụ: cấp quyền mở (OAuth: Open Authorization)) cho để lộ thông tin hỗ trợ của 5GC đến AF.

Một giải pháp khác đang được nghiên cứu trong 3GPP SA3 là cấp quyền để lộ thông tin hỗ trợ 5GC dựa trên hồ sơ (Profile Based). Một hồ sơ của UE sẽ xác định liệu một AF đặc thù có thể yêu cầu hoặc thay đổi thông tin nhất định của một UE liên kết. Nó có thể gồm nhận dạng UE, nhận dạng AF, định danh dịch vụ kỳ vọng (Expected Service Identifier), thông tin hỗ trợ 5GC đích, thời gian hết hạn và các chính sách bảo vệ và cấp quyền. Hồ sơ có thể được lưu trữ trong quản lý dữ liệu thống nhất (UDM: Uniffied Data Management) / một kho số liệu thống nhất (UDR: Uniffied Data Repository). Một AF gửi yêu cầu truy nhập thông tin hỗ trợ 5GC đến NEF/NWDAF (Network Exposure Function/ Network Data Analytics Function: chức năng để lộ mạng/ chức năng phân tích dữ liệu mạng). Nhận được yêu cầu này, NEF/ NWDAF trực tiếp xác định hồ sơ UE nếu có trong NEF/NWDAF; nếu không NEF/NWDAF nhận hồ sơ UE từ UDM/UDR. Dựa trên hồ sơ UE, NEF/NWDAF gửi đi thông tin hỗ trợ 5GC cùng với cơ chế bảo vệ an ninh tương ứng đến AF yêu cầu.

4. AI/ML CHO CÁC PHƯƠNG TIỆN TRUYỀN THÔNG (MEDIA) VÀ QUẢN LÝ AI/ML TRONG 5G

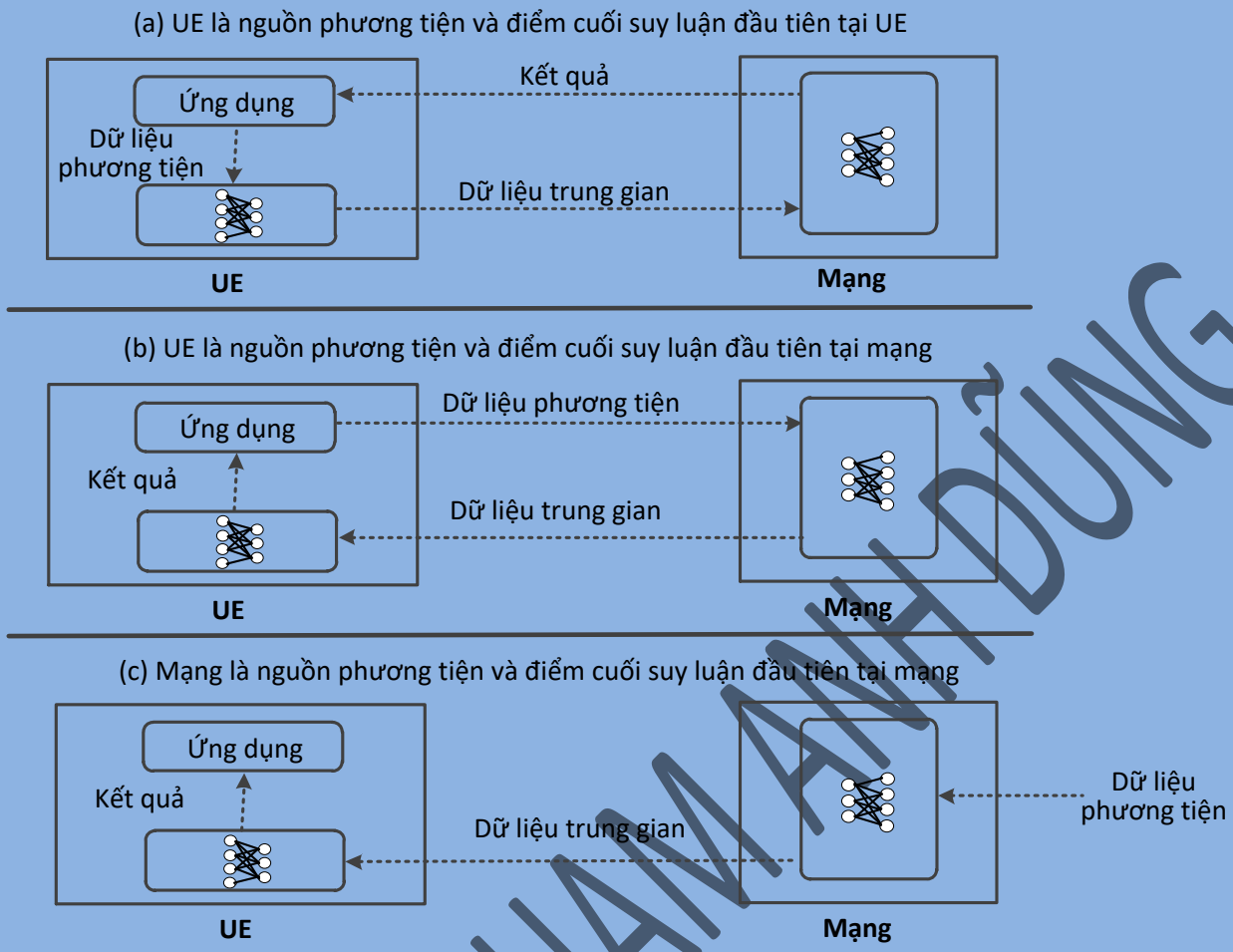
4.1. AI/ML cho các phương tiện truyền thông

Rất nhiều các ứng dụng AI/ML là các phương tiện truyền thông (gọi tắt là phương tiện) như phân loại ảnh, nhận dạng giọng nói và tăng cường chất hượng video. Các ứng dụng này đòi hỏi xử lý tính toán ngay càng cao và chuyên sâu hơn để xử lý khối lượng dữ liệu AI/ML và mức độ mức tạp của mô hình ngày càng tăng. Trong khi đó, các dịch vụ phương tiện mới nổi như thực tế ảo (VR), thực tế tăng cường (AF) và trò chơi đám mây cũng đang rất thành công. Các ứng dụng di động như vậy phụ thuộc rất nhiều vào các chức năng AI/ML. Chẳng hạn, các dịch vụ phương tiện AR dựa trên kính đeo đòi hỏi ngữ cảnh của môi trường xung quanh, có thể được cung cấp bằng cách sử dụng phân đoạn, nhận dạng và phân loại đối tượng với trợ giúp của AI/ML. AI/ML cũng tìm được ứng dụng trong các giải thuật nén phương tiện để làm cho các bộ mã hóa thích ứng hơn. Ngoài ra các mô hình AI/ML cho các ứng dụng phương tiện đòi hỏi nén phù hợp để tạo điều kiện dễ dàng hơn cho truyền dẫn hiệu suất. Tất nhiên nhóm cácchuyên gia hình ảnh động (MPEG: Moving Picture Experts Group) cũng đang nghiên cứu mã hóa video mạng thần kinh sâu để tối ưu hóa nén phương tiện và nghiên cứu đại diện mạng thần kinh cho các mạng thần kinh nén cho các ứng dụng phương tiện.

Các xu hướng mới nhất nói trên trong tác động lẫn nhau giữa AI/ML và các ứng dụng phương tiện đòi hỏi nghiên cứu sâu để hiểu biết được ảnh hưởng của chúng lên hệ thống 5G. Vi thế, 3GPP SA4 chịu trách nhiệm cho phát triển các đặc tả codec phương tiện, các khía cạnh hệ thống và các khía cạnh chuyển phát nội dung phương tiện, đạng nghiên cứu AI/ML trong các dịch vụ phương tiện 5G. Mục tiêu chính của nghiên cứu này là nhận dạng các yêu cầu tương tác liên quan và các ràng buộc thực hiện của AI/ML trong các dịch vụ phương tiện 5G.

Lúc đầu, công việc của 3GPP SA4 có mục đích là mô tả các trường hợp sử dụng dựa trên AI/ML dựa trên kết quả nghiên cứu về chuyển giao mô hình AI/ML để phân loại ba hoạt động then chốt bao gồm phân chia hoạt động AI/ML, phân phối mô hình và học phân tán/liên kết. Tập các trường hợp sử dụng thứ nhất là nhận dạng đối tượng trong hình ảnh và video được thực hiện bởi chỉ suy luận tại UE, chỉ suy luận tại mạng và phân chia các suy luận. Tập thứ hai của các trường hợp sử dụng là tăng cường chất lượng video trong phát luồng trực tiếp. Một trong các trường hợp sử dụng này là mã hóa video đầu cuối đầu cuối dựa trên mạng thần kinh, trong đó phía gửi xử lý một video chất lượng cao bằng cách sử dụng một mạng thần kinh để tạo ra một luồng số liệu trung gian và phát nó cùng với mã hóa video với độ phân giải thấp hơn và phía thu xử lý luồng dữ liệu trung gian thu được và luồng video để tái tạo lại video chất lượng cao cho kết suất (trình bày). Ta cũng có thể áp dụng mạng thần kinh để hậu xử lý một video đã được giải mã để nâng cao chất lượng video. Tập các trường hợp sử dụng thứ ba là thu tập phương tiện truyền thông từ một nhóm đông (Crowd-Sourcing Media Capture), trong đó mỗi UE sử dụng một mạng thần kinh chia sẻ để xử lý video/audio thu được của mình hay mạng thực hiện suy luận trên dự liệu phương tiện từ nhiều UE. Tập các trường hợp sử dụng cuối cùng là xử lý ngôn ngữ tự nhiên trên lời nói (Speech) chẳng nhận dạng lời nói và dịch tiếng nói (Voice).

Mục tiêu chính của nghiên cứu của 3GPP SA4 về AI/ML cho phương tiện truyền thông là mô tả kiến trúc dịch vụ của phương tiện cho AI/ML và các luồng dịch vụ liên quan. Cuối cùng, một nghiên cứu quan trọng là phân chia suy luận AI/ML giữa mạng và UE. Điểm phân chia phụ thuộc vào một số nhân tố bao gồm các khả năng của UE (ví dụ: bộ nhớ, tính toán, tiêu thụ năng lượng và độ trễ suy luận). Hình 4 minh họa thứ tự khác nhau của các hoạt động và các luồng dịch vụ tương ứng cho phân chia suy luận giữa mạng và UE.



Hình 4. Các cấu hình topo suy luận AI/ML được phân chia cho phương tiện

4.2. Quản lý AI/ML trong hệ thống 5G

AI/ML đang được sử dụng trong hệ thống 5G bao gồm quản lý và điều phối (Ochestration), mạng lõi và RAN. Để cho phép và tạo điều kiện thuận lợi cho hoạt động AI/ML trong hệ thống 5G, các mô hình AI/ML cần được tạo ra, được đào tạo, được kiểm tra, được phát triển và được quản lý trong suốt vòng đời. Trong phát hành 17, 3GPP tiên hành nghiên cứu chuẩn hóa quản lý đào tạo AI/ML, nhưng không để cập đến các khía cạnh khác như triển khai và suy luận AI/ML. Ngoài ra các khả năng quản lý AI/ML có thể không cần phối hợp với các khả năng của AI/ML trong mạng lõi và hỗ trợ các khả năng AI/ML trong RAN. 3GPP SA5, chịu trách nhiệm cho quản lý, điều phối và tính cước cho các hệ thống 3GPP, đang nghiên cứu quản lý để phối hợp các chức năng AI/ML trên các hệ thống 5G.

Quy trình làm việc của hoạt động AI/ML gồm ba giai đoạn: giai đoạn đào tạo (bao gồm đào tạo mô hình và kiểm tra), giai đoạn triển khai và giai đoạn suy luận. Quản lý AI/ML cho giai đoạn

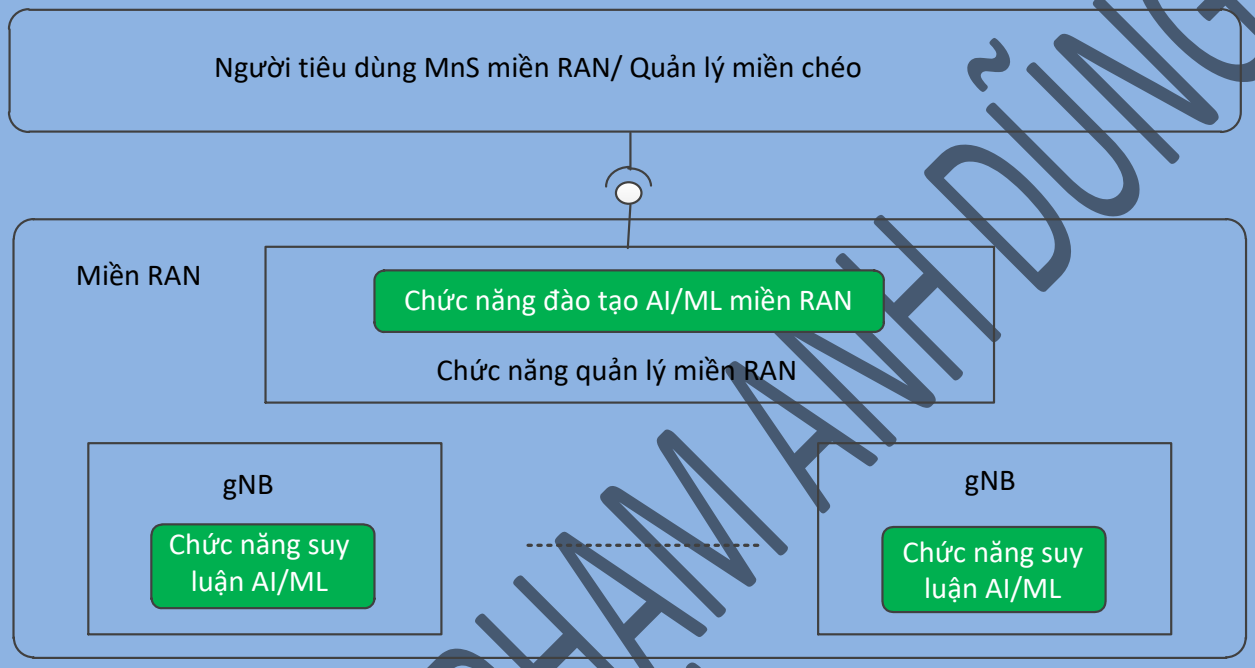
đào tạo cần hỗ trợ quản lý dữ liệu đào tạo, quản lý đào tạo, quản lý kiểm tra và xác nhận. AI/ML cho giai đoạn triển khai cần hỗ trợ điều khiển và giám sát triển khai, cho phép mô hình đã được đào tạo và đã được kiểm tra triển khai đến chức năng suy luận đích cũng như giám sát quá trình triển khai. Quản lý AI/ML cho giai đoạn suy luận cần hỗ trợ kích hoạt và hủy kích hoạt, điều kiển chức năng suy luận, quản lý hiệu năng suy luận, quản lý sự tin cậy (Trustworthiness) và điều phối suy luận.

3GPP SA5 đã nghiên cứu một mục đầy đủ các trường hợp sử dụng cho quản lý AI/ML và các yêu cầu tương ứng của chúng cũng như các giải pháp có thể có. Các khả năng của các trường hợp sử dụng cho quản lý cho giai đoạn đào tạo bao gồm dữ liệu sự kiện cho đào tạo, xác nhận thực thể, kiểm tra thực thể, đào tạo lại thực thể, đào tạo liên kết các thực thể, phân tích và báo cáo hiệu quả số liệu đào tạo, lập bản đồ và khai phá khả năng, quản lý cập nhật, đáng giá hiệu năng đào tạo, quản lý cấu hình đào tạo, học chuyển giao v.v. Các khả năng của các trường hợp sử dụng cho quản lý cho giai đoạn suy luận bao gồm theo dõi lịch sử suy luận, điều phối suy luận, điều khối khả năng, tải thực thể, mô phỏng suy luận, đánh giá hiệu năng suy luận, quản lý cấu hình suy luận, điều khiển cập nhật v.v.

Cần lưu rằng sự tin cậy được nhận dạng như là một khả năng quản lý cho cả giai đoạn đào tạo và giai đoạn suy luận. Mục tiêu của quản lý sự tin cậy là để đảm bảo mô hình AI/ML là bền vững, có thể giải thích được và tốt. Các yêu cầu của sự tin cậy có thể thay đổi tùy theo mức độ rủi do của trường hợp sử dụng yêu cầu cơ chế sự tin cậy. Từ bước đàu tiên, các chỉ số về sự tin cậy của AI/ML cần được định nghĩa. Phụ thuộc vào trường hợp sử dụng, người tiêu dùng dịch vụ dịch vụ quản lý AI/ML (AI/ML MnS: AI/ML Management Service) có thể lựa chọn một tập phù hợp các chỉ số sự tin cậy và yêu cầu một nhà sản xuất AI/ML MnS giám sát và và đánh giá các chỉ số được chọn. Tiền xử lý dữ liệu đào tạo/ kiểm tra/ suy luận có thể cần thiết tùy thuộc và phạm vi sự tin cậy mong muốn của mô hình AI/ML tương ứng. AI/ML phải trang bị cho người tiêu dùng khả năng cung cấp đến nhà sản suất các yêu cầu sự tin cậy của xử lý dữ liệu cũng như cho phép nhà sản xuất để lộ các khả năng xử lý dữ liệu liên quan đến sự tin cậy được hỗ trợ cho người tiêu dùng. Tương tự, người tiêu dùng AI/ML MnS phải có thể truy vấn nhà sản xuất đào tạo AI/ML, nhà sản xuất suy luận, (hoặc) và nhà sản xuất đánh giá và các khả năng sự tin cậy được hỗ trợ và yêu cầu cấu hình, đo và báo cáo tập các đặc tính sự tin cậy được chọn.

3GPP SA5 cũng đã nghiên cứu các kịch bản triển khai cho chức năng đào tạo, chức năng kiểm tra, chức năng suy luận và các khả năng quản lý tương ứng. Chức năng đào tao/ suy luận có thể được đặt trong hệ thống quản lý miền chéo (Cross-Domain) hay trong hệ thống quản lý miền, ví dụ: phân tích dữ liệu trong MDAF (Management Data Analytics Function: chức năng phân tích dữ liệu quản lý) hay trong mạng lõi, ví dụ: trong NWDAF (Network Data Analytics Function: chức năng phân tích số liệu mạng) và RAN, ví dụ: trí tuệ RAN (RAN Intelligence) trong gNB. Các khả năng quản lý do nhà sản xuất MnS được đặt trong chức năng quản lý tương ứng MnF

(Management Function: chức năng quản lý). Các chức năng khác nhau có thể được đặt cùng một chỗ hoặc được triển khai trong các thực thể khác nhau. Hình 5 minh họa một kịch bản triển khai cho trí tuệ RAN trong đó chức năng đào tạo và suy luận của AI/ML được đặt trong thực thể quản lý RAN và gNB và thực thể quản lý RAN cung cấp khả năng quản lý cho cả hai chức năng đào tạo và chức năng suy luận.



Hình 5. Thí dụ về kịch bản triển khai cho trí tuệ RAN

5. AI/ML CHO MẠNG TRUY NHẬP VÔ TUYẾN 5G NR-RAN

Có thể nói RAN là phần tử phức tạp nhất của các mạng tổ ong. Trong phát hành 18, 3GPP tham gia vào công việc chuẩn hóa cũng như nghiên cứu sử dụng AI/ML để tăng cường giao diện vô tuyến 5G NR, để cải thiện hiệu năng hệ thống và trải nghiệm người dùng. Phần này sẽ trình bày sự kết hợp AI vào mạng truy nhập vô tuyến của 5G NR-RAN (5G NR-RAN: New Radio –Radio Access Network: mang truy nhập vô tuyến – vô tuyến mới của 5G)

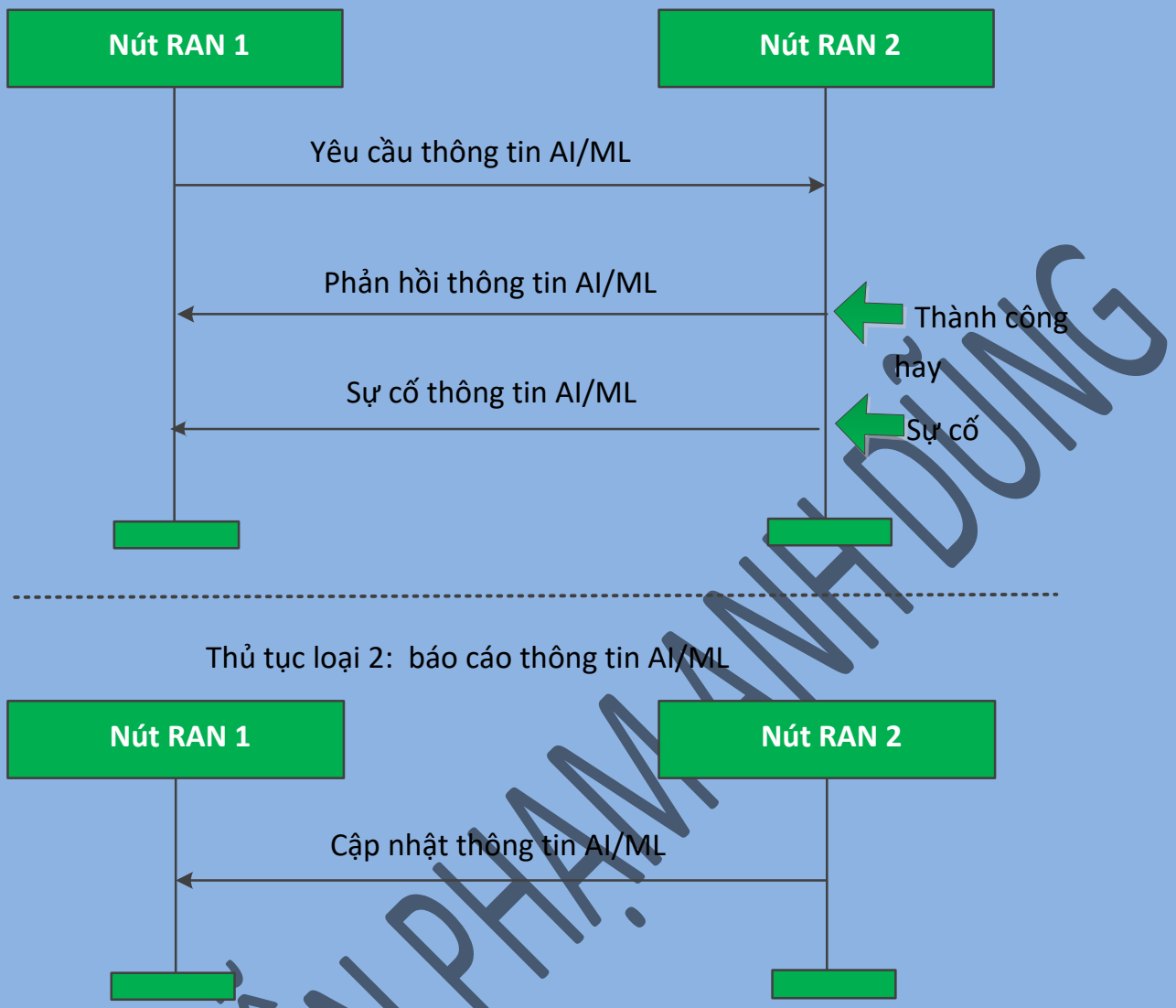
5.1. NG-RAN được AI/ML hỗ trợ

Tiếp sau việc hoàn thiện của nghiên cứu trong phát hành 17, 3GPP RAN3, chịu trách nhiệm cho kiến trúc RAN tổng thể và đặc tả các giao thức cho các giao diện mạng liên quan, tiến hành

công việc chuẩn hoá trong phát hành 18 cho trí tuệ RAN được hỗ trợ bởi AI/ML. Mục đích của công việc chuẩn hoá là đặc tả các tăng cường thu thập dữ liệu đặc thù và hỗ trợ báo hiệu để tiết kiệm năng lượng mạng dựa trên AI/ML, cân bằng tải và tối ưu hóa di động. Các tăng cường được đưa ra trong các giao diện và kiên trúc NR-RAN.Cả hai kiến trúc phân chia (thí dụ: gNB nguyên khối) và kiên trúc phân chía (ví dụ: gNB phân chia vào đơn vị trong tâm (CU) và đơn vị phân tán (DU) là phạm vi của công việc này. Đào tạo mô hình AT/ML và suy luận AI/ML có thể được đặt trong OAM (Operation and Maitenance: khai thác và bảo dưỡng) và gNB (hay gNB-CU cho kiến trúc phân chia) hay cả hai có thể được đặt trong gNB (hay gNB-CU cho kiến trúc phân chia).

So với các xử lý dựa trên đo truyền thống trong các đặc tả của 3GPP, trí tuệ RAN dựa trên AI/ML dựa rất nhiều vào sử dụng các dự đoán có thể được trao đổi giữa các gNB trên giao diện Xn. Các thí dụ về thông tin dự đoán bao gồm quỹ đạo UE chi tiết đến ô được dự báo và trạng thái tài nguyên được dự báo (ví dụ: số lượng UE hoạt động được dự báo, số lượng kết nối RRC (Radio Resource Control: điều khiển tài nguyên vô tuyến) được dự báo và tài nguyên vô tuyến được dự báo). Ngoài ra thông tin phản hồi cho biết các hoạt động dựa trên AI/ML ảnh hưởng lên hiệu năng RAN như thế nào có thể được trao đổi giữa các gNB. Các thí dụ về thông tin phản hồi bao gồm hiệu năng UE (ví dụ: thông lượng UE đường lên/ đường xuống, trễ gói, tỷ lệ lỗi gói) và số liệu đánh giá hiệu quả năng lượng.

Để hỗ trợ trao đổi thông tin giữa các gNB cho trí tuệ RAN được hỗ trợ bởi AI/ML, các thủ tục báo hiệu mới được đưa ra cho giao diện Xn. Có hai loại thủ tục cơ bản cho giao thức Xn là: loại 1: các thủ tục cơ bản với phản hồi (thành công hay thất bại) và loại 2: các thủ tục cơ bản không có phản hồi. Thủ tục loại 1 được sử dụng cho khởi đầu báo cáo thông tin AI/ML trong đó một nút của NG-RAN yêu cầu báo cáo thông tin liên quan đến AI/ML cho một nút NR-RAN khác. Thủ tục loại 2 được sử dụng để báo cáo thông tin AI/ML trong đó một nút NR-RAN báo cáo thông tin AI/ML sau khi thủ tục khởi đầu báo cáo thông tin AI/ML thành công. Hai loại thủ tục này được minh họa trên Hình 6.



Hình 6. Thủ tục trao đổi thông tin AI/ML trên Xn cho NR-RAN

5.2. AI/ML cho giao diện vô tuyến gốc 5G (5G Native Interface)

AI/ML cho giao diện vô tuyến gốc được dự đoán sẽ là giải pháp quan trọng trong 6G. Vì thế, điều quan trọng là ngay từ bây giờ phải nghiên cứu sử dụng AI/ML trong giao diện vô tuyến 5G để chuẩn bị cho chuẩn hóa 6G, quá trình sẽ sẽ bắt đầu vào khoảng đầu năm 2025. Mục nghiên cứu của phát hành 18 của 3GPP về AI/ML cho giao diện vô tuyến NR đánh dấu một mốc quan trọng trong phát triển các mạng tổ ong vì đây là lần đầu tiên cách tiếp cận như vậy được tiếp nhận. Mục đích chính của mục nghiên cứu này thiết lập một bộ khung tổng quát để tăng cường giao diện vô tuyến bằng cách sử dụng AI/ML, cùng với nhiều đề tài khác nhau được nghiên cứu. Các đề tài này bao gồm định nghĩa các giai đoạn của các giải thuật AI/ML và các mức độ cộng

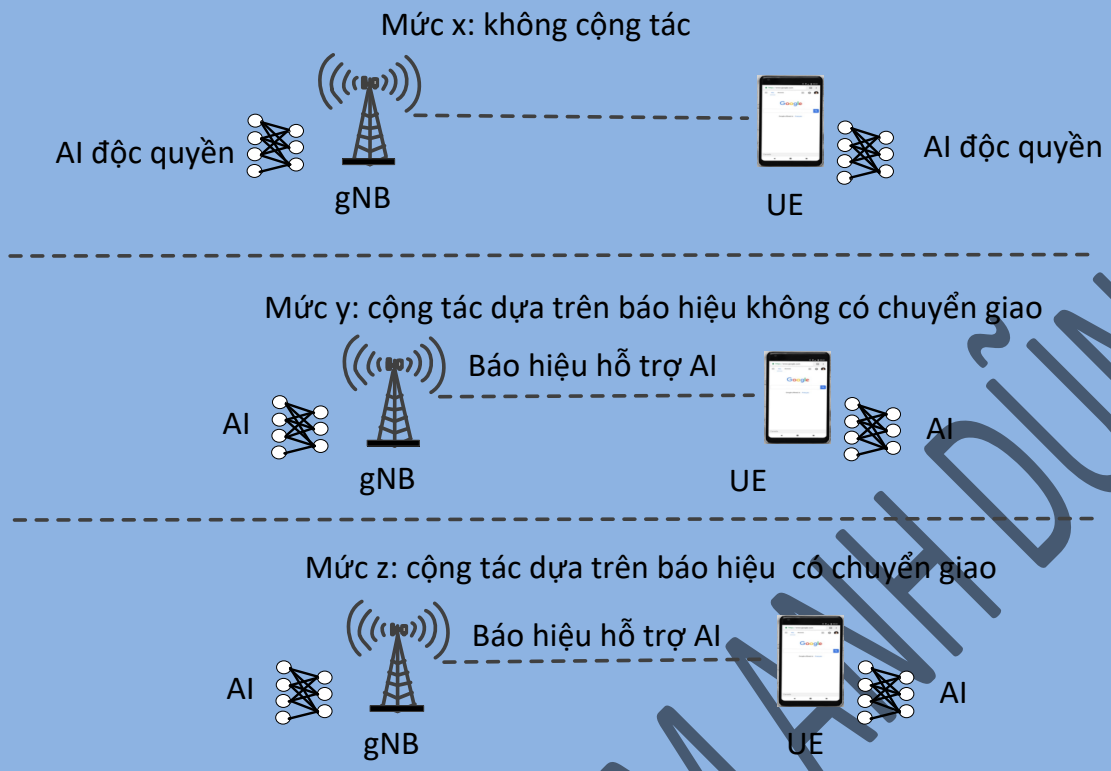
tác giữa gNB và UE, các tập dữ liệu được yêu cầu cho đào tạo mô hình AI/ML, xác nhận và kiểm tra, quản lý vòng đời của các mô hình AI/ML. .

3GPP đã xác định ba mức độ hợp tác giữa gNB và UE:

- Cấp độ x–không cộng tác.
- Cấp độ y – hợp dựa trên báo hiệu mà có chuyển giao mô hình.
- Cộng tác dựa trên tín hiệu cấp z với chuyển giao mô hình.

Hình 7 minh họa các mức độ hợp tác gNB-UE x, y và z. Tại mức x, không có cộng tác giữa gNB và UE. Việc sử dụng các kỹ thuật AI/ML trong trường hợp này chỉ đơn thuần dựa trên các thực hiện và các giải pháp bản quyền. Cụ thể là, không có tăng cường đặc thù AI/ML dành riêng (ví dụ,: báo hiệu liên quan đến quản lý vòng đời) cho các hoạt động AI/ML tại mức x. Tại mức y, có cộng tác dựa trên báo hiêu không có chuyển giao. So với mức x, sự khác biệt ở chỗ đối với mức y ta có thể thay đổi giao diện vô tuyến và đưa ra báo hiệu mới để tạo điều kiện dễ dàng cho các tính năng dựa trên AI/ML hiệu quả khác, chẳng hạn đưa ra các phép đo và báo cáo mới. Mức z được định nghĩa là hợp tác dựa trên báo hiệu có chuyển giao mô hình, trong đó chuyển giao mô hình đề cập đến chuyển phát một mô hình AI/ML trên giao diện vô tuyến (hoặc các thông số của cấu trúc mô hình đã được biết tại đầu thu hoặc mô hình mới cùng với các thông số) và chuyển phát có thể chứa một mô hình đầy đủ hay một phần mô hình. Lưu rằng ranh giới giữa mức y và mức z được định nghĩa dựa trên định nghĩa là việc chuyển phát mô hình có trong suốt với báo hiệu 3GPP hay không. Cụ thể là, chuyển phát mô hình trong mức độ z không trong suốt với báo hiệu 3GPP, trong khi đó mức y bao gồm các trường hợp không chuyển phát mô hình và

với chuyển phát mô hình trong suốt với báo hiệu 3GPP trên giao diện vô tuyến.



Hình 7. Minh họa các mức độ hợp tác gNB-UE x, y và z

Mục nghiên cứu này được tập trung lên ba trường hợp sử dụng phục vụ như là một chương trình ban đầu để hiểu sâu sắc và toàn diện hơn về không gian giải pháp thông qua các so sánh đánh giá hiệu năng với các thực hiện dựa trên AI/ML liên quan.

5.3. Các trường hợp sử dụng

Trong 3GPP RAN1, đã có ba trương hợp sử dụng dựa trên AI đầu tiên: CSI, phản hồi, BM và các nâng cao độ chính xác định vị. Trong phát hành 19, BM và định vị đã được tiến triển đến công tác chuẩn mực. Tuy nhiên các kết quả đánh giá nhận được từ các công ty tham gia nghiên cứu liên qua đến tăng cường phản hồi CSI không đáp ứng kỳ vọng. Vì thê các công ty tham giá tiếp tục nghiên cứu làn thế nào để cải thiện tăng cao hiệu năng.

BM dựa trên AI bao gồm hai trường hợp sử dụng con (Sub Use Cases): dự báo búp sóng trong miền không gian và dự báo búp sóng trong miền thời gian. Trong miền không gian, một tập con của các búp sóng được chọn, được đặt tên là tập B (Set B), được đo để cung cấp đầu vào cho mô hình để nó dự báo các búp ứng cử cho tập A (Set A) (tập đầy đủ). Khi so sánh với các cơ chế trước đây, phương pháp này giảm kể chi phí và trễ. Trong miền thời gian, lịch sử số liệu đo được dùng là dầu vào của mô hình để cho phép dự báo các búp ứng cử cho K khoảng thời gian

tương lai, trong đó K bằng hoặc lớn hơn 1.Đáng lưu ý là trong các khoảng thời gian nhất định, UE không cần đo nhờ vậy giảm đáng kể chi phí đo.

Các búp sóng được đo nói trên có thể rộng hoặc hẹp vì các tiêu chuẩn hiện này không áp đặt bất kỳ ràng buộc lên chúng.

Các nghiên đối với định vị và phản hồi CSI dựa trên AI được tiến hành như sau. Trong định vị dựa trên AI, có hai phương pháp quan trọng: định vị AI/ML tực tiếp để tao ra các tọa độ vị trí chính xác và định vị AI/MLđược hỗ trợ để cung cấp các kết quả đo trung gian giống như định thời dữ liệu (Timing Data). Phát hành 19 tập trung lên ba trường hợp định vị đặc thù để nhân mạnh sử dụng các mô hình tại các phía khác nhau (UE, nút NG-RAN với lMF và gNB) mà không gây tổn hại tính riêng tư của người dùng.

Đối với phản hồi CSI dựa trên AI, hai trường hợp sử dụng con được định dạng. Trường hợp sử dụng con thứ nhất sử dụng mô hình hai phía để nén CSI trong miền không gian-thời gian nhằm giảm chi phí báo cáo trong khi yêu cầu đào tạo mô hình cộng tác. Tường hợp con thứ hai, sử dụng mô hình một phía UE để dự báo CSI trong miền thời gian cho phép giảm chi phí đo tín hiệu tham chẩn (RS: Reference Signal) và cho phép báo cáo lớp 1 (L1) giống như các phương pháp truyền thống.

Tóm lại mục tiêu của nghiên cứu ba trường hợp sử dụng nói trên như sau.

Phản hồi thông tin trạng thái kênh (CSI: Channel State Inforrmation): mục tiêu của trường hợp sử dụng này là sử dụng các kỹ thuật AI/ML để giảm chi phí CSI, cải thiện độ chính xác phản hồi và cho phép dự báo. Một lĩnh vực được tập trung là nén CSI miền không gian- thời gian bao gồm bộ mã hóa CSI dựa trên AI/ML tại UE và bộ giải mã CSI dựa trên AI/ML tại gNB. Một lĩnh vực được tập trung khác là dự báo CSI miền thời gian tại UE dựa trên các kết quả đo được trong quá khứ.

Quản lý búp sóng (Beam Management): Mục tiêu của trường hợp sử dụng này là sử dụng các kỹ thuật AI/ML để giảm chi phí quản lý búp sóng cũng như cải thiển độ chính xác lựa chọn búp sóng. Một lĩnh vực được tập trung là dự báo búp sóng đường xuống miền không gian với sử dụng một mô hình AI/ML để suy luận búp sóng đường xuống tốt nhất trong tập A (Set A) các búp đường xuống dựa trên các kết quả đo của tập B (Set B) các búp sóng đường xuống tại một thời điểm tương lai dựa trên các kết quả đo tập B các búp sóng đường xuống trong quá khứ.

Định vị (Positioning): Mục đích của trường hợp sử dụng này là sử dụng các kỹ thuật AI/ML để cải thiện độ chính xác định vị cho các kịch bản khác nhau bao gồm, ví dụ: các tình trạng NLoS nặng (Non Line of Sight: không trực xạ). Một lĩnh vực được tập trung là định vị được hỗ trợ bởi AI/ML với sử dụng một mô hình AI/ML để suy luận một thống kê đo trung gian cho định vị.

Các thí dụ của thống kê đo trung gian là xác suất (LoS)/NLoS, thời gian tới, góc tới và góc khởi hành.

Tổng kết lại, hiểu biết toàn diện vai trò của 3GPP đối với hỗ trợ hợp nhất AI/ML vào giao diện vô tuyến nhận được từ các nghiên cứu nói trên sẽ đóng góp cho công việc chuẩn hóa không chỉ trong các phát hành tương lai của 5G mà còn cho các thế hệ của các hệ thống không dây di động tương lai được phát triển bởi 3GPP gồm cả 6G.

5.4. Quản lý mô hình

Quản lý mô hình bao gồm các thao tác khác nhau như: kích hoạt, thôi kích hoạt mô hình, chuyển mạch (Switching), cập nhật, chọn, giám sát, nhận dạng chức năng/mô hình và lùi lại (Fallback) v.v.

- Kích hoạt /thôi kích hoạt mô hình bao gồm cho phép/thôi cho phép một mô hình AI/ML cho một tính năng đạc thù được cho phép bởi AI/ML;
- Cập nhật mô hình dẫn đến quá trình cập nhật các thông số và (hoặc) cấu trúc của một mô hình;
- Nhận dạng chức năng/ mô hình đề cập đến quá trình /phương pháp nhận dạng một chức năng AI/ML cho hểu biết chung giữa mạng và UE;
- Fallback xẩy ra khi hiệu năng của mô hình hiện thời không đáp ứng được kỳ vọng, nhắc nhở UE/mạng quay trở lại cơ chế cũ ;
- Giám sát mô hình là một thủ tục thực hiện giám sát thực hiện suy luận của mô hình AI/ML;
- Chuyển mạch và chọn mô hình luôn thực hiện trong các kích bản có nhiều mô hình phục vụ một chức năng. Chọn mô hình là lựa chọn một mô hình để kích hoạt trong số số nhiều mô hình cho đối với cùng một tính năng được AI/ML cho phép. Chuyển mạch mô hình nghĩa là thôi kích hoạt một mô nhình AI/ML đang hoạt động và kích hoạt một mô hình AI/ML khác cho một tính năng đặc thù được AI/ML cho phép.

Trong phát hành 19, cần thiết phải định nghĩa báo hiệu và các chi tiết mới để tạo thuận tiện cho cho các thao tác nói trên. Tuy nhiên cũng gập phải một số thách thức. Chẳng hạn, đối với nhận dạng, nếu nhận dạng chức năng được áp dụng cho quản lý mô hình, các thao tác nói trên sẽ dựa trên chức năng. Điều này là nẩy sinh một vấn đề nghiêm trọng: có thể có nhiều mô hình cho một chức năng. Nếu độ chi tiết của nhận dạng không được chi tiết hóa đủ, một số mô hình có thể duy trì trong môt trạng thái không hiểu được đối với UE hoặc mạng. Ta xét kịch bản sau, có ba mô nhình được sử dụng cho dự báo tạo búp tại phía UE hoạt động tại tốc độ 3 km/h, 60 km/h và 120 km/h. Tuy nhiên mạng chỉ biết là UE có thể thực hiện dự báo sử dụng mô hình AI và không biết là các mô hình này có thể được phân biệt dựa trên tốc độ chuyển động của UE. Hậu quả là không rõ ràng đối với mạng để quyết định các thao tác quản lý mô hình. Tuy nhiên nếu nhận

dạng mô hình được sử dụng, sẽ không cần giải quyết vấn đề nêu trên, vì quản lý mô hình có thể dựa trên ID của mô hình. Dù sao đi chăng nữa, sử dụng ID mô hình dẫn đến các phức tạp của chính nó. Chẳng hạn định nghĩa ID mô hình có thể là một quá trình phức tạp, nó có thể toàn cầu hoặc địa phương, vật lý hoặc logic. Hơn nữa không thể loại trừ rằng một mô hình có thể phục vụ nhiều chức năng. Cho đến nay, 3GPP vẫn chưa áp đặt các giới hạn lên các phương pháp nhận dạng.

Từ góc nhìn RAN1, có vẻ như cả hai nhận dạng mô hình và nhận dạng chức năng đều có thể chấp nhận.

Đối với các trường hợp sử dụng dựa trên AI, rõ ràng rằng có các ảnh hưởng cần thiết của đặc tả để hỗ trợ quản lý mô hình. Khi mạng quản lý các mô hình phía UE, điều quan trọng là cả mạng và UE phải hiểu giống nhau về thông tin mô hình chẳng hạn chức năng, các thông số và các điều kiện bổ sung. Điều này có thể giúp tận dụng báo cáo khả năng của UE. Liên quan đến các điều kiện bổ sung như các kịch bản, các site và các tập dữ liệu v.v., chúng có thể được phân loại vào hai nhóm: các điều kiện bổ sung phía mạng và các điều kiện bổ sung phía UE. Đồng bộ các điều kiện bổ sung giữa mạng và UE cũng quan trọng để đảm bảo tính nhất quán giữa đào tạo mô hình và suy luận. Ngoài ra các điều kiện bổ sung này có thể gồm chuyển mạch mô hình. Khi đã miêu tả rõ ràng các điều kiện bổ sung cho từng mô hình, việc tạo điều kện chuyển đổi sang mô hình đích dựa trên phát hiện các thay đổi môi trường sẽ trở nên khả thi hơn. Một thách thức quan trọng khác gắn liền với giám sát mô hình. Cho đến này, giám sát mô hình liên quan đến quan sát hiệu năng của mô hình đang hoạt động và đánh giá sự thích hợp của nó với kịch bản hiện thời. Điều này làm nẩy sinh các câu hỏi sau: số đo hiệu năng là gì? Cần thiết kế thủ tục giám sát như thế nào? Theo các nghiên cứu được tiến hành trong phát hành 18, có nhiều phương án khác nhau: chẳng hạn, các số đo hiệu năng có thể bao gồm: độ chính xác của suy luận, phân bố dữ liệu v.v. Trong phát hành 19, chúng ta có thể có thể giảm bớt lựa chọn phương án trong số các phương án nói trên. Những người ủng hộ cũng cho rằng có thể mở rộng cơ chế giám sát đến các mô hình không hoạt động để đánh giá chúng có thể được kích hoạt hay không. Nếu có thể áp dụng giám sát cho các mô hình không hoạt động, thì các mô hình không hoạt động này phải làm việc để tạo ra các kết quả dự báo. Do đó những người không tán thành không hiểu được vì sao các mô hình không hoạt động lại được phép dựa báo. Tuy nhiên nếu các mô hình không hoạt động không làm việc thì vệc chọn mô hình sẽ trở nên không chắc chắn. Các vấn đề này vẫn chưa được giải quyết khiến mọi người quan tâm hơn đối với các kết quả nghiên cứu của phát hành 19. Ngoài ra áp dụng giám sát để đảm bảo sự nhất quán giữa đào tạo và suy luận mô hình cũng đã được đề xuất. Trong các trường hợp sử dụng dựa trên AI, duy trì độ chính xác dự báo bằng cách đồng bộ đào tạo mô hình và suy ulận với các kịch bản giống nhau đóng vai trò rất quan trọng. Giám sát phục vụ như là một phương tiện để cảm nhận các thay đổi kịch bản một cách gián tiếp. Dựa trên các kết quả giám sát, các thao tác quản lý mô hình khác nhau có thể được thực hiện, chẳng hạn: chọn, chuyển mạch, lùi lại (Fallback), kích hoạt hay thôi kích hoạt mô hình.

5.5. Đánh giá hiệu năng (Performance Evaluation)

Trong phát hành 18, ba trường hợp sử dụng mới được đưa ra trong lớp vật lý, vấn đề đầu tiên và quan trọng là đánh giá các lợi ích của hiệu năng của ba trường hợp sử dụng này. Một số chỉ thị hiệu năng then chốt (KPI: Key Performance Indicators) đóng vai trò quan trọng trong đánh giá là hiệu năng suy luận, độ trễ, độ phức tạp tính toán, chi phí bổ sung và các yêu cầu phần cứng. Các phương pháp đánh giá và các số đo cho từng trường hợp sử dụng cần được thiết kế cho các tính năng và các yêu cầu đặc thù của nó.

Trong trường hợp sử dụng BM dựa trên AI, phương pháp mô phỏng hệ thống được áp dụng làm đường cơ sở. Các KPI quan trọng được xác định, chẳng hạn như, Top-K / 1 (%) là tỷ lệ phần trăm của búp sóng được hỗ trợ bởi thần đèn Top-1 (Top1 genie : một mô hình của AI) là một trong những búp sóng được dự đoán Top-K. Độ chính xác dự đoán búp sóng (%) với lề 1dB (1 dB margin) là tỷ lệ phần trăm của búp sóng được dự đoán Top-1 có L1-RSRP (Layer 1-Reference Signal Received Power: công suất tín hiệu tham chuẩn thu – lớp 1) lý tưởng nằm trong phạm vi 1dB so với L1-RSRP lý tưởng của búp sóng có sự hỗ trợ của thần đèn Top-1.

Nhờ nỗ lực hợp tác của các công ty khác nhau, đối với trường hợp 1 (Case 1), người ta nhận thấy rằng bằng cách sử dụng tập B (Set B) các búp sóng cố định (trình bày gần đúng một phần tư tập A các búp sóng), đã có một sự cải thiện đáng kể độ chính xác búp sóng phát đường xuống top1 (Top1 DLTx Beam). Độ chính xác dự báo có thể đạt được 70% đến 90 %. . Đặc biệt là khi xét búp sóng phát đường xuống Top 1 với lề 1 dB, người ta đã chứng minh rằng độ chính xác vượt quá 90%.

Khi đánh giá định vị dựa trên AI, KPI chính trên tất cả các kịch bản và trường hợp sử dụng xoay quanh phần trăm Hàm phân phối tích lũy (CDF: Commulative Distribution Function) có độ chính xác theo chiều ngang, đặc biệt tập trung lên 90% (Baseline: đường cơ sở) và tùy chọn bao gồm 50%, 67%, 80%. Ngoài ra, báo cáo độ chính xác theo chiều dọc là tùy chọn. Trong bối cảnh định vị được hỗ trợ bởi AI / ML, đầu ra của các mô hình AI / ML có thể bao gồm nhiều loại thông tin khác nhau, chẳng hạn như Thời gian đến (ToA: Time of Arrival), Chênh lệch cường độ tín hiệu nhận được (RSTD: Received Signal Strength Difference), Góc khởi hành (AoD: Angle of Departure), Góc đến (AoA: Angle of Arrival), các chỉ báo LoS / NLoS, trong số những chỉ số khác.

Độ chính xác suy luận trong định vị dựa trên AI bị ảnh hưởng đáng kể bởi môi trường kênh, đặc biệt là trong các tình huống NLoS. Dựa trên đóng góp của các công ty rằng đố với hiệu suất nền tảng không tổng quátt hóa, mô hình AI/ML trải qua quá trình đào tạo và thử nghiệm bằng cách sử dụng các bộ dữ liệu từ cùng một kịch bản triển khai. Định vị dựa trên AI / ML cho thấy những cải tiến đáng kể về độ chính xác của định vị so với các phương pháp định vị phụ thuộc

vào Công nghệ truy cập vô tuyến (RAT: Radio Access Technology) hiện có. Ví dụ: trong kịch bản InF-DH (Indoor Factory-Density High: mật độ cao nhà máy trong nhà) với cài đặt thông số cụm (Clustter Prarrameter) là 60%, 6m, 2m, định vị dựa trên AI / ML đạt được độ chính xác định vị ngang nhỏ hơn 1m ở CDF bằng 90%, trái ngược với >15m đối với các phương pháp định vị thông thường.

Về đánh giá hiệu suất của cải tiến phản hồi CSI dựa trên AI/ML, mô phỏng hệ thống đóng vai trò là cơ sở (Baseline). KPI và số liệu đo lường (Metrics) bao gồm:

- Việc đánh giá nén CSI bao gồm các chỉ số hiệu suất chính như tương đồng Cosin tổng quát bình phương (SGCS: Squared Generalized Cosine Similarity) và / hoặc Sai số bình phương trung bình chuẩn hóa (NMSE: Normalized Mean Square Error) để đánh giá độ chính xác của đầu ra CSI do AI / ML tạo ra. SGCS được tính toán riêng cho từng lớp, khi hạng> 1.
- Đánh giá dự đoán CSI liên quan đến việc tính toán KPI trung gian cho từng trường hợp dự đoán khi mô hình AI/ML tạo ra nhiều dự đoán.

Trong [11], kết quả mô phỏng nén CSI từ các công ty khác nhau được trình bày. Ví dụ: trong các kịch bản triển khai trong đó kịch bản A là UMi (Urban Micro: vùng vi mô trong thành phố) và kịch bản B là UMa (Urban Mcro: vùng vĩ mô trong thành phố), kịch bản A là UMa và kịch bản B là UMi, hoặc kịch bản A là UMa và kịch bản B là InH (Indoor Hospot: điểm nóng trong nhà), sự suy giảm hiệu suất từ -1,69% đến -31,6% được quan sát thấy. Đối với dự đoán CSI, Nếu tốc độ UE B là 30km/h, 60km/h hoặc 120km/h, hoặc nếu tốc độ UE B là 10km/h và tốc độ UE A là 60km/h hoặc 120km/h, hiệu suất giảm từ trung bình đến đáng kể (giảm -2,01% đến -76,85%).

Dựa trên các quan sát nêu trên, kết quả đánh giá hiệu năng nén CSI và dự báo CSI rất không thỏa mãn. Vì thế trong phát hành chuẩn hóa không được tiến hành tiếp, thay vào đó nẫn tập trung duy trì trạng thái của các mục nghiên cứu.

6. CÁC TRƯỜNG HỢP SỬ DỤNG AI/ML CHO GIAO DIỆN VÔ TUYẾN CỦA 5G NR

6.1. Trường hợp sử dụng: phản hồi CSI

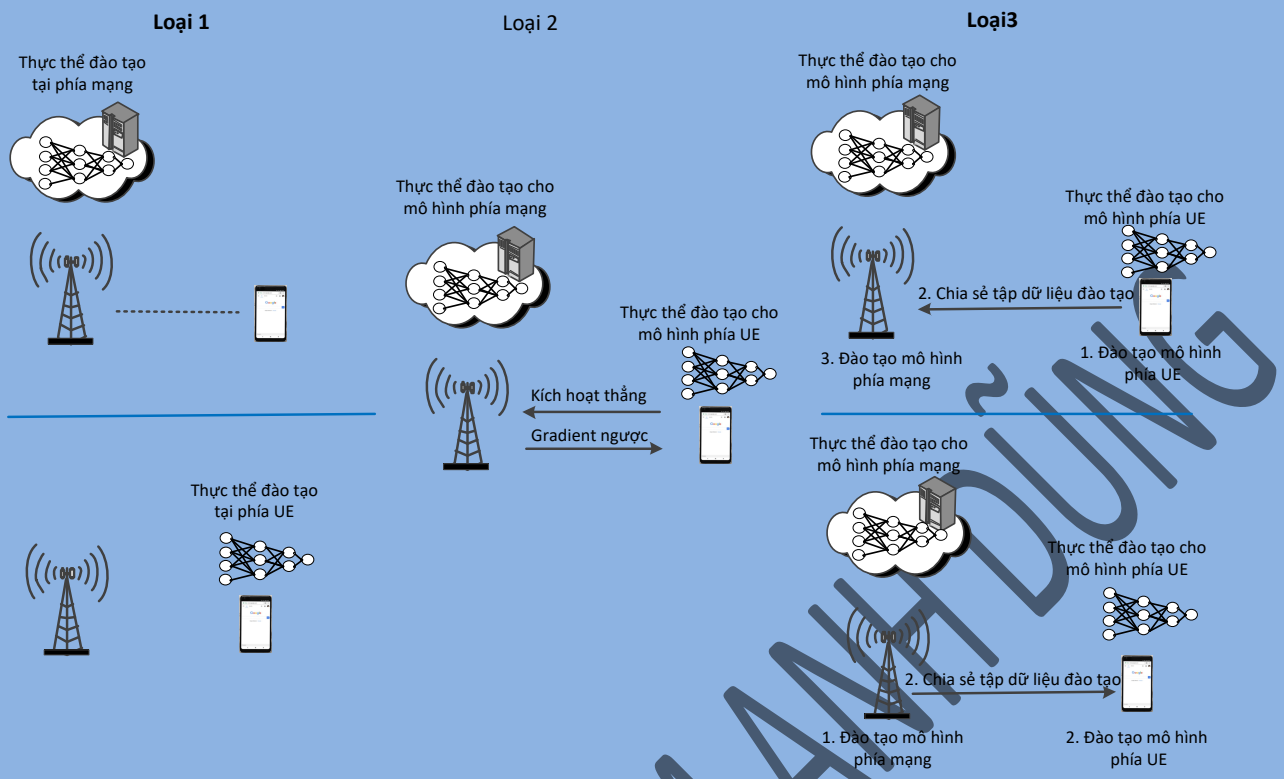
CSI đề cập đến thông tin của kênh không dây đa đường giữa nút 5G gNB và UE. UE có thể đo tín hiệu tham chiếu đường xuống, tính toán CSI đường xuống và cung cấp báo cáo CSI cho gNB, do đó tạo điều kiện thuận lợi cho việc truyền đường xuống. Tận dụng các thuật toán dựa trên AI/ML có thể nâng cao hơn nữa phản hồi CSI, mang lại những lợi thế như giảm chi phí và cải thiện độ chính xác. 3GPP đã nghiên cứu hai trường hợp sử dụng con đại diện của việc sử

dụng các thuật toán dựa trên AI / ML để tăng cường phản hồi CSI: nén CSI miền tần số không gian và dự đoán CSI miền thời gian bằng mô hình phía UE.

1. Nén CSI miền tần số không gian

Trong nén CSI dựa trên AI / ML, UE sử dụng bộ mã hóa CSI dựa trên AI / ML để tạo thông tin phản hồi CSI, trong khi bộ giải mã CSI dựa trên AI / ML tương ứng tại gNB được sử dụng để tái tạo CSI từ dữ liệu phản hồi nhận được. Đây là một ví dụ về mô hình AI/ML hai phía, trong đó hoạt động suy luận được phân chia giữa UE và gNB. Mô hình AI / ML hai phía này có thể được sử dụng để nén ma trận kênh thô do UE ước tính hoặc ma trận mã hóa trước có nguồn gốc từ ma trận kênh thô. Đáng chú ý, nén ma trận mã hóa trước phù hợp với bộ khung phản hồi CSI dựa trên bảng mã hiện có được đặc tả cho giao diện vô tuyến NR, do đó thu hút nhiều sự quan tâm hơn trong quá trình nghiên cứu 3GPP.

Sử dụng mô hình AI/ML hai phía cho giao diện vô tuyến mang lại vô số thách thức. Thách thức đầu tiên liên quan đến việc đào tạo mô hình AI/ML hai phía. Trong bối cảnh này, 3GPP đã điều tra ba loại đào tạo liên quan đến các mức độ hợp tác khác nhau giữa mạng và UE, như được minh họa trong Hình 8. Trong kịch bản đào tạo đầu tiên, được chỉ định là 'Loại 1', bộ mã hóa và bộ giải mã của mô hình hai phía được đào tạo chung ở một phía. Nếu đào tạo này được thực hiện ở phía mạng, mô hình bộ mã hóa kết quả có thể được chuyển sang UE và ngược lại. Chuyển sang 'Loại 2', bộ mã hóa và bộ giải mã của mô hình hai phía trải qua quá trình đào tạo lần lượt ở phía UE và phía mạng. Chính xác, bộ mã hóa và bộ giải mã được đào tạo chung trong cùng một vòng lặp để lan truyền thẳng (Feedforward) và lan truyền ngược bằng cách trao đổi kích hoạt thẳng và gradient lùi giữa UE và mạng. Cuối cùng, 'Loại 3' bao gồm việc đào tạo riêng biệt của bộ mã hóa và bộ giải mã trong mô hình hai phía, được thực hiện trong các phiên đào tạo riêng biệt ở phía UE và mạng. Việc đào tạo riêng biệt có thể bắt đầu ở phía UE hoặc phía mạng. Lấy ví dụ về đào tạo riêng biệt bắt đầu từ phía UE. Trong giai đoạn đầu, một thực thể phía UE đào tạo mô hình bộ mã hóa CSI. Sau đó, thực thể phía UE chia sẻ một bộ dữ liệu đào tạo — bao gồm đầu ra bộ mã hóa và CSI mục tiêu — với một thực thể phía mạng. Tập dữ liệu này sau đó được thực thể phía mạng tận dụng để đào tạo mô hình giải mã CSI của nó.



Hình 8. Các kiểu đào tạo mô hình AI/MI cho nén CSI sử dụng mô hình hai phía

Các phương pháp đào tạo khác nhau mang lại những lợi thế và thách thức khác nhau. Trong số này, đào tạo loại 1 hứa hẹn mang lại hiệu suất vượt trội, nhưng việc triển khai và quản lý các mô hình có thể phức tạp. Lấy loại đào tạo 1 ở phía mạng làm ví dụ. Thực thể phía mạng sẽ liên quan đến nhà cung của loại UE được nhắm mục tiêu trong quá trình đào tạo để mô hình mã hóa CSI được đào tạo tương thích với việc triển khai UE. Tuy nhiên, nỗ lực này đòi hỏi phải tiết lộ thông tin độc quyền của nhà cung cấp UE cho phía mạng, đây có thể là một thách thức đáng kể. Mặt khác, đào tạo loại 2 giữ tính bảo mật của thông tin độc quyền ở cả hai phía mạng và UE, vì thông tin mô hình vẫn nằm trong các miền tương ứng của chúng. Tuy nhiên, hai thực thể đào tạo tương ứng cần phối hợp các lần lặp lại đào tạo của họ để trao đổi các kết quả kích hoạt thẳng và lan truyền ngược, điều này có thể dẫn đến nỗ lực phối hợp đáng kể và chi phí. Đào tạo loại 3 nổi lên như một lựa chọn không chỉ bảo vệ thông tin độc quyền mà còn loại bỏ sự cần thiết phải cộng tác trong quá trình lặp lại đào tạo vì sự phối hợp có thể diễn ra bên ngoài quá trình đào tạo.

Sự phức tạp đào tạo vốn có trong mô hình AI/ML hai phía cho giao diện không vô tuyến càng được kết hợp bởi sự cần thiết phải có khả năng tương tác và tương thích của nhiều nhà cung cấp. Bộ giải mã CSI nằm ở gNB cần tương thích với các bộ mã hóa CSI khác nhau tại UE và ngược lại. Trong các trường hợp mô hình bộ giải mã CSI phổ biến được sử dụng cho nhiều mô hình bộ mã hóa CSI, thực thể đào tạo phía mạng — theo loại đào tạo 1 hoặc 2 — phải phối hợp với các

nhà cung cấp UE để thực hiện các nỗ lực đào tạo chung. Đáng chú ý, việc phát hành một loại UE mới có khả năng kích hoạt đào tạo lại trên tất cả các nhà cung cấp. Những thách thức tương tự tồn tại trong đào tạo loại 1 hoặc 2 khi một mô hình bộ mã hóa CSI được chia sẻ được sử dụng cho nhiều mô hình bộ giải mã CSI. Những vấn đề này có thể được giảm thiểu trong đào tạo loại 3. Đặc biệt, nếu một mô hình bộ giải mã CSI chung được sử dụng cho nhiều mô hình bộ mã hóa CSI (hoặc ngược lại), việc đào tạo lại do sự phát hành một loại UE mới có thể chỉ liên quan đến nhà cung cấp UE liên kết và nhà cung cấp mạng do bản chất đào tạo riêng biệt trong loại đào tạo 3. Tóm lại, đối với phản hồi CSI dựa trên mô hình AI / ML hai phía, đào tạo loại 3 nổi lên như một cách tiếp cận thực tế và khả thi hơn so với đào tạo loại 1 và 2

Một cân nhắc quan trọng khác trong thiết kế nén CSI dựa trên AI/ML là khả năng tổng quát hóa mô hình AI/ML trên nhiều kịch bản và cấu hình. Cụ thể, thiết kế cần xem xét các kịch bản đa dạng, bao gồm các kịch bản triển khai khác nhau (ví dụ: vĩ mô đô thị, vi mô đô thị và trong nhà), phân phối UE, tần số sóng mang, v.v. Mô hình AI / ML cũng phải có thể mở rộng trên nhiều cấu hình khác nhau, chẳng hạn như băng thông kênh khác nhau, độ chi tiết tần số, tải trọng phản hồi CSI, cũng như các biến thể trong bố cục và số cổng ăng-ten. Các cấu hình đa dạng này có thể ảnh hưởng đến thiết kế mô hình AI/ML vì chúng có thể dẫn đến các khía cạnh khác nhau của đầu vào và đầu ra mô hình.

2. Dự đoán CSI miền thời gian với mô hình phía UE

Một thách thức trong bộ khung báo cáo CSI cũ của NR liên quan đến độ trễ thời gian giữa thời gian mà CSI được báo cáo tương ứng với thời điểm mà gNB thực sự sử dụng báo cáo CSI. Sự chậm trễ thời gian này dẫn đến tình huống CSI được báo cáo trở nên lỗi thời, một hiện tượng thường được gọi là lão hóa kênh. Đáng chú ý, tốc độ mà CSI được báo cáo trở nên lỗi thời được khuếch đại bởi tốc độ UE cao hơn. Điều này trở nên đặc biệt rõ rệt trong bối cảnh các kịch bản đa đầu vào nhiều đầu ra nhiều người dùng (MU-MIMO), đặc biệt là trong các triển khai MIMO lớn. Hiệu suất của MU-MIMO đã được quan sát thấy giảm đi khi UE di chuyển ở tốc độ trung bình đến cao. Tận dụng các thuật toán AI/ML để dự đoán CSI nổi lên như một kỹ thuật đầy hứa hẹn để chống lại tác động của CSI lỗi thời.

Trái ngược với nén CSI dựa trên AI / ML, đòi hỏi mô hình hai phía, dự đoán CSI dựa trên AI / ML trong miền thời gian có thể sử dụng mô hình một phía. Đào tạo mô hình một phía này có thể được thực hiện bởi một nhà cung cấp duy nhất và suy luận sau đó có thể được thực hiện bởi một phía (gNB hoặc UE). Xem xét khối lượng công việc trong phát hành 18, 3GPP đã tập trung chiến lược vào mô hình một phía UE để dự đoán CSI trong miền thời gian. Trong thiết lập này, đầu vào cho mô hình một phía bao gồm một chuỗi các kết quả đo CSI quá khứ do UE thực hiện. Kết quả đầu ra của mô hình này là CSI được dự báo cho một trường hợp thời gian trong tương lai, được dự đoán bởi UE.

Từ quan điểm của các tiêu chuẩn 3GPP, dự kiến rằng chúng ta có thể sử dụng phần lớn bộ khung CSI hiện có để hỗ trợ dự đoán CSI. Đặc biệt, mô hình AI/ML LCM để dự đoán CSI phía UE ở một mức độ lớn có thể sử dụng lại những gì được xác định cho các trường hợp sử dụng một phía UE khác vì tác động đặc tả kỹ thuật của các mô hình AI/ML một phía UE đã được nghiên cứu để quản lý búp sóng và định vị, như được mô tả trong các phần sau.

6.2. Trường hợp sử dụng: quản lý búp sóng (Beam Management)

Tạo búp được sử dụng phổ biến trong các hệ thống thông tin di động 5G mặt đất và không gian. Ngoài ra nó cũng được sử dụng cho RIS (Reconfiguration Intellegence Surface; mặt phẳng thông minh khả lập cấu hình) để hướng tới vô tuyến thông minh và khả lập trình trong 6G.

Chức năng quản lý búp sóng (Beam Management) trong NR được sử dụng để hỗ trợ tạo búp (Beamforming). Nó đặc biệt cần thiết cho các hệ thống sóng milimet 5G dựa trên tạo búp tia tương tự (Analog Beamforming). Trong quy trình cơ bản để quản lý tạo búp đường xuống, UE đo tín hiệu tham chiếu được liên kết với búp phát của gNB và kiểm tra các búp thu khác nhau của UE đối với mỗi búp phát của gNB để tìm một cặp búp đường xuống phù hợp. Quá trình này có thể tốn thời gian và kéo theo chi phí đáng kể về các tín hiệu tham chiếu. Các thuật toán dựa trên AI / ML mang lại tiềm năng nâng cao chức năng quản lý búp sóng, mang lại những lợi thế bao gồm giảm chi phí, giảm thiểu độ trễ và cải thiện độ chính xác trong việc lựa chọn búp.

3GPP đã nghiên cứu hai trường hợp sử dụng con đại diện liên quan đến việc áp dụng các thuật toán dựa trên AI / ML để quản lý búp sóng. Chúng được gọi là 'dự đoán búp sóng đường xuống miền không gian' và 'dự đoán búp sóng đường xuống miền thời gian'.

Dự đoán búp sóng đường xuống miền không gian tận dụng kết quả đo lường từ một tập hợp được chỉ định', được ký hiệu là 'tập B', để dự đoán búp sóng tốt nhất trong một tập hợp búp sóng đường xuống khác, được gọi là tập A’ vào thời điểm hiện tại.

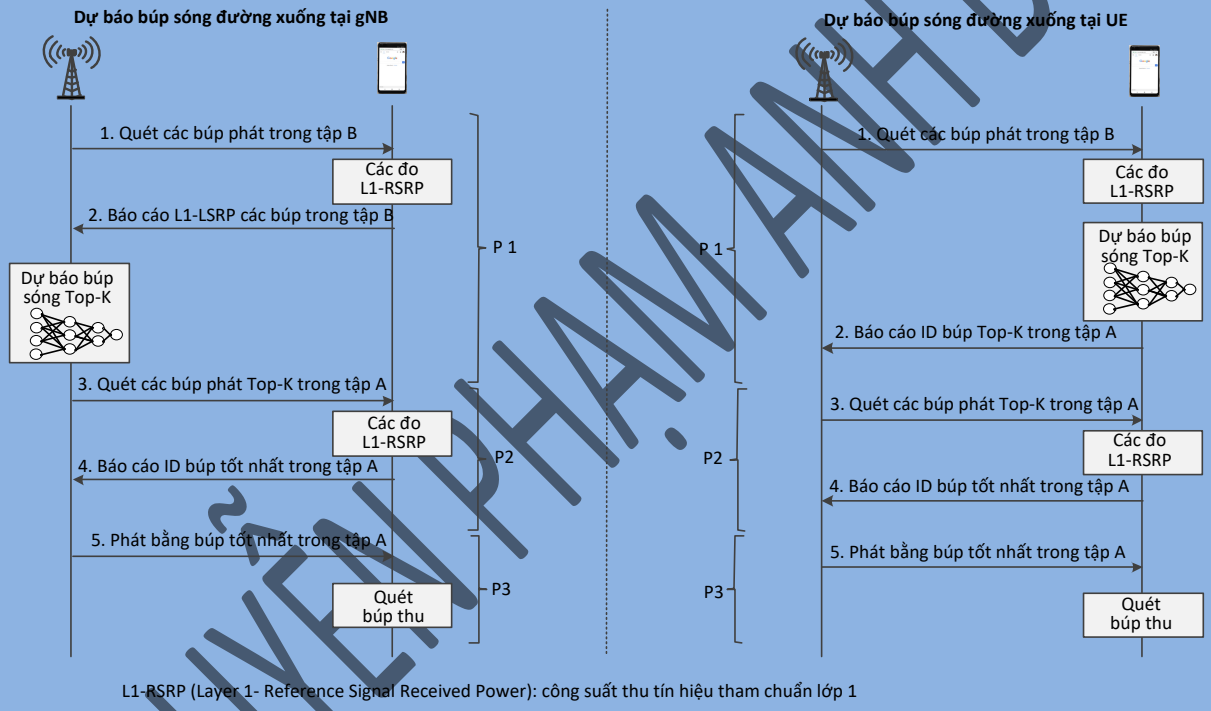
Dự đoán búp sóng đường xuống miền thời gian khai thác các kết quả đo lịch sử bắt nguồn từ 'tập B' để dự đoán búp sóng tốt nhất trong 'tập A' cho một hoặc nhiều trường hợp thời gian trong tương lai.

Đáng chú ý, 'Tập B' có thể tạo thành một tập hợp con của 'Tập A' hoặc hai tập hợp có thể khác nhau (ví dụ: 'Tập B' bao gồm các búp rộng trong khi 'Tập A' bao gồm các búp hẹp). Ngoài ra, trong bối cảnh dự đoán búp sóng đường xuống miền thời gian, 'Tập A' và 'Tập B' cũng có thể giống nhau.

Đầu vào điển hình cho mô hình AI / ML để dự đoán búp sóng đường xuống miền không gian hoặc miền thời gian là các kết quả đo công suất nhận được tín hiệu tham chiếu lớp 1 (L1-RSRP)

của các búp sóng trong 'tập B.' Đầu ra điển hình từ mô hình AI / ML là các búp sóng topK được dự đoán trong ' tập A.' Đào tạo và suy luận mô hình AI / ML có thể nằm ở phía gNB hoặc ở phía UE. Trong các tình huống suy luận AI/ML xảy ra ở phía UE, UE cần báo cáo (các) búp sóng dự đoán của nó cho gNB. Ngoài ra, khi suy luận AI / ML diễn ra ở phía gNB, UE được yêu cầu báo cáo các phép đo L1-RSRP của nó cho các búp sóng trong ' tập B' cho gNB.

Để đảm bảo sự rõ ràng về dự đoán búp sóng đường xuống dựa trên AI / ML, bây giờ chúng ta trình bày chi tiết về cách dự đoán búp sóng có thể được tích hợp vào khuôn khổ quản lý búp sóng của NR hiện có. Nhớ lại rằng bộ khung hiện tại bao gồm ba quy trình được gọi là P1 (thiết lập cặp búp sóng ban đầu), P2 (tinh chỉnh búp sóng phát) và P3 (tinh chỉnh búp sóng nhận). Hình 9 cung cấp hình minh họa về các quy trình quản lý búp sóng đường xuống với dự đoán búp sóng dựa trên AI / ML ở phía gNB và ở phía UE.



Hình 9. Quy trình quản lý búp sóng đường xuống với dự đoán búp sóng dựa trên AI / ML ở phía gNB (trái) và ở phía UE (phải)

Lấy dự đoán búp sóng ở phía gNB làm ví dụ. Đầu tiên, gNB quét qua các búp sóng phát khác nhau trong 'tập B', được đo bởi UE. Sau đó, UE truyền các kết quả đo L1-RSRP cho các búp sóng trong 'tập B' trở lại gNB. Sau khi nhận được báo cáo, gNB sử dụng các giá trị L1-RSRP nhận được làm đầu vào cho mô hình AI / ML của nó, sau đó đưa ra dự đoán về các búp sóng top-K trên cùng trong 'tập A.' Dựa trên kết quả suy luận, gNB sau đó quét các búp sóng top- K trên cùng được dự đoán trong 'tập A', UE đo các búp sóng này để xác định búp sóng phát tốt nhất mang lại giá trị L1-RSRP cao nhất. Sau đó, UE báo cáo ID của búp sóng tốt nhất trở lại

gNB. Nếu có nhu cầu, gNB có tùy chọn kích hoạt quy trình P3. Trong giai đoạn này, gNB tiếp tục sử dụng búp sóng phát tốt nhất, trong khi UE thăm dò các búp sóng thu được khác nhau để xác định búp sóng phù hợp nhất.

6.3. Trường hợp sử dụng: định vị (positioning)

Các phương pháp định vị 5G NR hiện tại thường dựa trên hình học, bao gồm hai bước chính: 1) tiến hành đo tín hiệu vô tuyến và 2) tính toán ước tính vị trí bằng cách giải một hệ phương trình phi tuyến tính thiết lập mối quan hệ giữa vị trí của UE và các phép đo. Độ chính xác của các phương pháp định vị dựa trên hình học phụ thuộc rất nhiều vào tính khả dụng của các phép đo được liên kết đường trực xạ (LOS). Trong các tình huống liên quan đến điều kiện LOS yếu hoặc môi trường đa đường dày đặc, chẳng hạn như lắp ráp thiết bị bên trong nhà máy trong nhà (Indoor Factory), độ chính xác của các phương pháp dựa trên hình học có xu hướng giảm sút. Các thuật toán dựa trên AI / ML có thể nâng cao độ chính xác định vị trong một loạt các tình huống, bao gồm cả những tình huống được đặc trưng bởi các điều kiện không LOS (NLOS) phổ biến.

3GPP đã nghiên cứu hai trường hợp sử dụng con đại diện liên quan đến việc áp dụng các thuật toán dựa trên AI / ML để định vị. Chúng được gọi là 'định vị AI / ML trực tiếp' và 'định vị hỗ trợ AI / ML'.

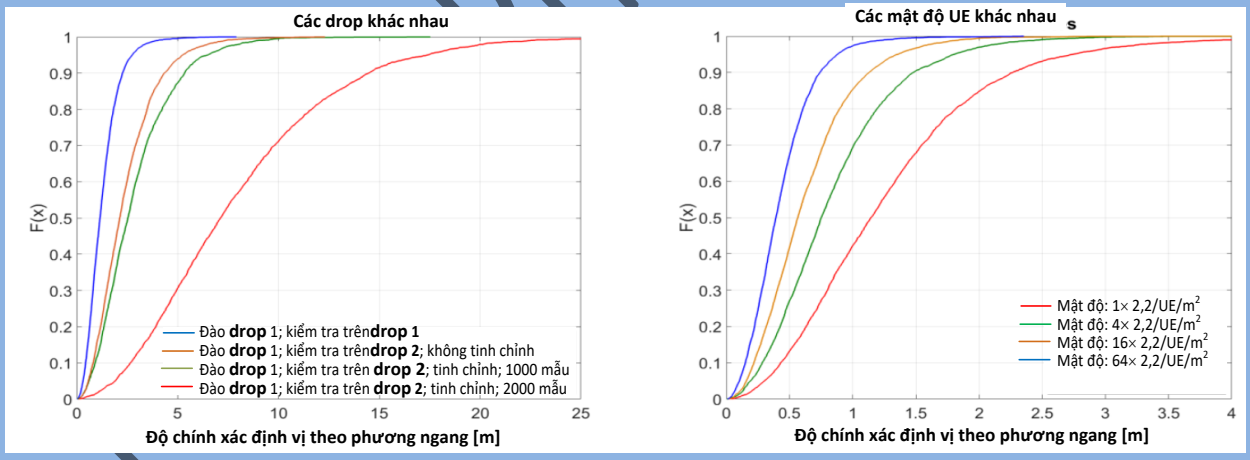
Định vị AI / ML trực tiếp sử dụng mô hình AI / ML để xác định trực tiếp vị trí của UE. Ví dụ: điều này có thể bao gồm định vị dựa trên dấu vân tay (Fingerprinting) bằng cách sử dụng các quan sát kênh, chẳng hạn như phản ứng xung kim kênh (CIR: Channel Impulse Response) hoặc hồ sơ độ trễ công suất (PDP: power delay profile), làm đầu vào cho mô hình AI / ML.

Định vị được hỗ trợ bởi AI / ML liên quan đến việc tận dụng mô hình AI / ML để tạo ra thống kê đo lường trung gian, là công cụ trong việc định vị. Điều này có thể bao gồm các phép đo như xác suất LOS / NLOS, góc đến/ góckhởi hành hoặc thời gian đến.

Đào tạo và suy luận mô hình AI/ML có thể nằm ở phía UE, phía chức năng quản lý vị trí (LMF: Location Management Function) hoặc phía gNB. Tùy thuộc vào vai trò của UE, LMF và gNB trong các quy trình định vị, 3GPP tập trung vào ba loại phương pháp định vị. Loại đầu tiên là định vị dựa trên UE, trong đó bản thân UE thực hiện định vị AI / ML trực tiếp hoặc định vị có sự hỗ trợ của AI / ML. Loại thứ hai là định vị dựa trên LMF có sự hỗ trợ của UE, trong đó UE cung cấp hỗ trợ cho LMF trong việc ước tính vị trí của UE. Trong kịch bản này, LMF có thể thực hiện định vị AI / ML trực tiếp hoặc UE có thể tham gia vào định vị có sự hỗ trợ của AI / ML. Loại thứ ba là định vị có sự hỗ trợ của gNB, trong đó gNB cung cấp hỗ trợ cho LMF trong

việc ước tính vị trí của UE. Trong trường hợp này, LMF có thể thực hiện định vị AI / ML trực tiếp hoặc gNB có thể tham gia vào định vị được AI / ML hỗ trợ.

Khám phá khả năng tổng quát hóa của các mô hình AI/ML là một lĩnh vực nghiên cứu chính trong các trường hợp sử dụng trong phát hanhd 18 của 3GPP. Liên quan đến định vị dựa trên AI / ML, các khía cạnh khác nhau đã được xem xét để điều tra khả năng tổng quát hóa mô hình. Ví dụ: tập dữ liệu đào tạo và thử nghiệm được tạo theo các lần thả (Drop) khác nhau, các thông số lộn xộn khác nhau hoặc lỗi thời gian riêng biệt. Để giúp hiểu vấn đề này, phần bên trái của Hình 10 trình bày một nghiên cứu điển hình minh họa khả năng khái quát hóa định vị AI / ML trực tiếp, sử dụng thiết lập mô phỏng 3GPP [7]. Kịch bản triển khai mô phỏng là nhà máy trong nhà (Indoor Factory) với điều kiện NLOS nặng. Mô hình AI/ML mô phỏng có dạng mạng nơ-ron tích chập (CNN), trong đó CIR đóng vai trò là đầu vào mô hình và vị trí UE dự đoán tạo thành đầu ra mô hình. Hai tập dữ liệu riêng biệt, được ký hiệu là Drop 1 1' (thả 1) và 'Drop 1 2' (thả 2), được tạo bằng cách sử dụng các giá trị hạt giống ngẫu nhiên khác nhau và được sử dụng để đào tạo và thử nghiệm mô hình tương ứng. Về bản chất, hai Drop này có thể được hình dung là đại diện cho nhà máy trong nhà khác nhau có cùng sắp đặt lộn xộn. Kết quả cho thấy sự sụt giảm đáng kể về độ chính xác của định vị AI/ML trực tiếp khi đào tạo và thử nghiệm được thực hiện trên các Drop khác nhau. Một biện pháp khắc phục hợp lý cho vấn đề này là sử dụng tinh chỉnh mô hình AI / ML, như cũng được thể hiện trong phần bên trái của Hình 10. Điều đáng chú ý là mô hình ban đầu được đào tạo với tập dữ liệu gồm 16.000 mẫu. Do đó, việc sử dụng 1.000 (hoặc 2.000) mẫu tinh chỉnh tương đương với 6,25% (hoặc 12,5%) trong số 16.000 mẫu cần thiết để đào tạo mô hình AI/ML ban đầu.



Hình 10. Độ chính xác định vị của định vị AI / ML trực tiếp dưới các 'drop' khác nhau (trái) và mật độ UE khác nhau (phải)

Thu thập dữ liệu là một lĩnh vực khám phá quan trọng khác trong các trường hợp sử dụng được nghiên cứu trong phát hành 18 của3GPP. Khi xem xét định vị dựa trên AI / ML, có một kỳ vọng trực quan rằng mật độ dữ liệu thu thập được tăng lên sẽ dẫn đến độ chính xác của định vị được

nâng cao. Trực giác này được minh họa định lượng trong phần bên phải của Hình 10. Xu hướng rõ ràng cho thấy rằng việc cải thiện độ chính xác của định vị đi kèm với yêu cầu cao về thu thập dữ liệu. Điều này nhấn mạnh tầm quan trọng của các chiến lược thu thập dữ liệu để thực hiện hiệu quả định vị dựa trên AI/ML.

Tài liệu tham khảo

1. Nguyễn Phạm Anh Dũng, "5G và lộ trình phát triển lên 6G". Nhà xuất bản Thông tin và Truyền thông, 12/2022
2. Nguyễn Phạm Anh Dũng, **"**IoT (Internet vạn vật): *Kiến trúc IoT, IoT công ghiệp và công nghiệp 4.0, IoT tổ ong"*. Nhà xuất bản Thông tin và Truyền thông, 12/2022
3. Nguyễn Phạm Anh Dũng, **"**Truyền thông vệ tinh và truyền thông 3D trong các hệ thống 5G, 6G**".** Nhà xuất bản Thông tin và Truyền thông, 2/2024
4. Nguyễn Phạm Anh Dũng "Tích hợp RIS vào các mạng truyền thông không dây 5G và 6G". Sách điện tử, 4/2025
5. Vivienne Sze and others "Efficient Processing of Deep Neural Networks: A Tutorial and Survey". arXiv:1703.09039v2 [cs.CV] 13 Aug 2017
6. J. Woodhouse "Big, big, big data: higher and higher resolution video surveillance". Technology.ihs.com, January 2016.
7. Yiping Kang and others "Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge". ACM. ISBN, 2017
8. Mai Le and other "Applications of Distributed Machine Learning for the Internet-of-Things: A Comprehensive Survey". arXiv:2310.10549v1 [cs.NI] 16 Oct 2023
9. Chen Sun and others "On the Combination of AI and Wireless Technologies: 3GPP Standardization Progress". arXiv:2407.10984v1 [cs.NI] 17 Jun 2024
10. Xingqin Lin NVIDIA "An Overview of the 3GPP Study on Artificial Intelligence for 5G New Radio". 2013
11. Xingqin Lin NVIDIA "Artificial Intelligence in 3GPP 5G-Advanced: A Survey". 2024
12. Omar Nassef and others "A survey: Distributed Machine Learning for 5G and beyond". Computer Networks 207 (2022) 108820
13. Maria Vakalopoulou and others "Deep learning: basics and convolutional neural networks (CNN)". HAL. 2023
14. Tsung-Yi Lin and others "Feature Pyramid Networks for Object Detection". arXiv:1612.03144v2 [cs.CV] 19 Apr 2017
15. 3GPP TS 23.288 "Architecture enhancements for 5G system (5GS) to support network data analytics services" V17.8.0, March 2023

16. 3GPP TR 37.817 “Study on enhancement for data collection for NR and EN-DC” V17.0.0, April 2022
17. 3GPP TS 28.533 “Management and orchestration; Architecture framework” V17.2.0, March 2022
18. 3GPP TR 22.874 “Study on traffic characteristics and performance requirements for AI/ML model transfer” V18.2.0., December 2021
19. 3GPP TS 22.261 “Service requirements for the 5G system” V18.6.0, March 2022
20. 3GPP TR 23.700-80, “Study on 5G system support for AI/MLbased services” V18.0.0, December 2022
21. 3GPP TR 33.898 “Study on security and privacy of AI/ML-based services and applications in 5G” V0.6.0., May 2023
22. 3GPP TR 26.927 “Study on artificial intelligence and machine learning in 5G media services” V0.3.1, February 2023
23. 3GPP TS 28.105 “Artificial intelligence/machine learning (AI/ML) management” V17.3.0., March 2023
24. 3GPP TR 28.908, “Study on artificial intelligence/machine learning (AI/ML) management,” V0.2.1., March 2023
25. RP-213602 “Artificial intelligence (AI)/machine learning (ML) for NG-RAN,” 3GPP TSG RAN Meeting #94e, Dec. 2021
26. 3GPP TR 38.843 “Study on artificial intelligence (AI)/machine learning (ML) for NR air interface” V0.0.0., June 2022
27. Huaijiang Zhu, Manali Sharma, Kai Pfeiffer, Marco Mezzavilla, Jia Shen, Sundeep Rangan, and Ludovic Righetti, “Enabling Remote Whole-body Control with 5G Edge Computing”, to appear, in Proc. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Available at: https://arxiv.org/pdf/2008.08243.pdf
28. B. Taylor, V. S.Marco, W. Wolff, Y. Elkhatib, and Z. Wang, “Adaptive deep learning model selection on embedded systems,” in Proc. ACM LCTES, 2018, pp. 31–43.