CÁC ĐỘNG LỰC TÍCH HỢP AI VÀO TRUYỀN THÔNG KHÔNG DÂY

Các hệ thống mạng không dây tương lai sẽ phát triển vào các hạ tầng không dây được tích hợp sâu "truyền thông", "cảm biến", "trí tuệ" và "lưu trữ". Các nền tảng này sẽ có khả năng cung cấp các dịch vụ tùy biến và cá nhân hóa theo yêu cầu, vượt trội các khả năng của các hệ thống truyền thông thông thường dựa trên các kiến trúc cố định và các quy tắc được định trước. Các hệ thống này cần sử dụng các công nghệ AI. Chúng sẽ sử dụng các cơ sở dữ liệu và hiểu biết lớn cho suy luận và đưa ra quyết định bằng cách khai thác các khả năng tổng quát hóa của các mô hình để thích ứng các môi trường và các kịch bản khác nhau và cung cấp ấn định tài nguyên tối ưu và các giải pháp quản lý. Hai động lực chính cho tích hợp truyền thông không dây và AI như sau.

1) **AI đẩy xa giới hạn của các hệ thống truyền thông không dây**. AI có các khả năng phi thường trong xử lý dữ liệu lớn. Dự báo đến năm 2030 dữ liệu được tạo ra hàng tháng toàn cầu đối với truyền thông không dây di động sẽ đạt đến 5016 EB (EB=10^{18} Bytes)). Dữ liệu này bao gồm dữ liệu thiết bị đầu cuối, giao diện vô tuyến, mạng và dịch vụ theo các định dạng khác nhau như văn bản, ngôn ngữ đánh dấu mở rộng (Extencible Marup Language), ngôn ngữ đánh dấu siêu văn bản (Hypertex Markup Language), đồ họa, thông tin Audio/Video. Các nhà mạng có thể sử dụng AI để đào tạo và đưa ra quyết định dựa trên suy luận trên các kiểu số liệu lớn khác nhau để nâng cao hiệu suất và tối ưu hóa mạng theo các mục tiêu và các kích thước khác nhau. AI có thể học từ dữ liệu quá khứ thông qua các giải thuật ML và xây dựng các mô hình để dự báo các yêu cầu mạng và các vấn đề có thể xảy ra như nghẽn mạng, sự cố thiết bị hay sự thay đổi hành vi người dùng. Chẳng hạn AI có thể dự đoán trước các đỉnh lưu lượng tiềm năng trong các ngày nghỉ đặc thù hay các sự kiện diện rộng và đưa ra trước các biện pháp ấn định tài nguyên như tăng dung lượng của BS tạm thời hay tối ưu hóa các chiến lược, nhờ vậy ngăn chặn sự cố hệ thống và tăng cường trải nghiệm người dùng.
   Khả năng thích ứng là một đặc trưng nổi bật của AI. Các môi trường truyền thông không dây có tính biến động cao và chịu ảnh hưởng cao của các nhân tố khác nhau, như thời tiết, địa hình và các thay đổi tại vị trí các thiết bị người dùng. Khả năng thích ứng của AI cho phép nó thực hiện điều chỉnh nhanh chóng theo các biến động môi trường và các thăng giáng trong các điều kiện mạng. Chẳng hạn khi, khi người dùng di chuyển từ trong nhà hay trong một xe ô tô tốc độ cao, cường độ tín hiệu và nhiễu sẽ thay đổi. AI có thể cảm nhận được các thay đổi này theo thời gian thực và tự động điều chỉnh các thông số truyền thông để đảm bảo sự ổn định của chất lượng truyền thông, thích ứng các kịch bản truyền thông khác nhau và cải thiện sự tinh cậy và linh họa của các hệ thống truyền thông không dây.
   Khác với quản lý phân cấp của các mức khác nhau trong các hệ thống truyền thông thông thường, về mặt lý thuyết, AI có thể học các cấu trúc ẩn và các thông số khác nhau để phù hợp các chức năng phức tạp tùy ý. Điều này cung cấp một cách thức hiệu quả hơn để cảm nhận môi trường không dây thay đổi, phức tạp và mô tả đặc điểm không gian trạng thái mạng. Mặt khác, các thiết kế các mo-đun hệ thống truyền thông khác nhau có thể có các mục đích xung đột và có thể có các ràng buộc hiệu năng giữa các mo-đun. Tồn tại các đánh đổi giữa các số đo hiệu năng, như dung lượng kênh và nhiễu, sự tin cậy truyền dẫn và tiêu thụ năng lượng hệ thống. Quá trình tối ưu hóa từng mo-dum riêng lẻ thường không thể đạt được hiệu năng tối ưu tổng thể. Trong các trường hợp này, AI có thể tạo điều kiện thiết kết tối ưu hoá kết hợp giữa các mo-đun.
2) **AI để giải quyết các thách thức trong truyền thông không dây.**Các thách thức chính đối với mạng truyền thông không dây là phải đáp ứng các yêu cầu sau:

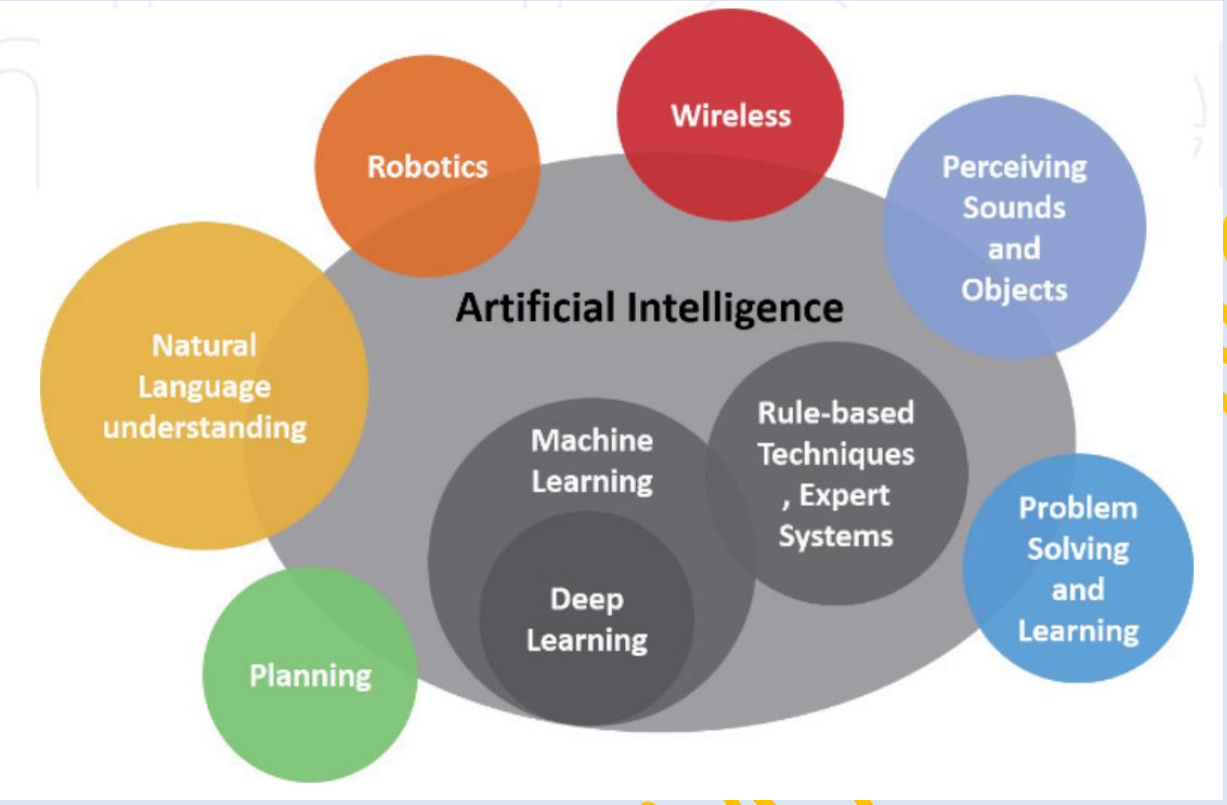
- Sử dụng tài nguyên phổ tốt hơn,
- Vùng phủ không dây carbon thấp,
- O&M (Operation and Maintenance: khai thác và bảo dưỡng) hiệu qủa và chi phí hiệu quả hơn
- Khả năng cá nhân hóa và tùy biến theo yêu cầu khách hàng tốt hơn,
- Truyền dẫn tin cậy và an ninh tốt hơn.

Nguyên nhân cốt lõi hệ thống truyền dẫn không dây hiện thời không đáp ứng được các yêu cầu trên là chúng không đủ trí tuệ. Các hệ thống truyền dẫn không dây tương lai sẽ phát triển vào một hạ tầng không dây tích hợp sâu "truyền thông", "cảm biến", "trí tuệ" và "lưu trữ".

TÓM TẮT AI

KẾT NỐI VÀ CHỒNG LẤN HỌC MÁY, HỌC SÂU VÀ TRÍ TUỆ NHÂN TẠO

AI là sự kết nối của các công nghệ khác nhau cùng làm việc để cho phép các máy cảm nhận, hiểu biết, hành động và học với các cấp độ trí tuệ như con người. Các kỹ thuật dựa trên nguyên tắc cũng như hệ thống chuyên gia là cách tiếp cận đầu tiên đến AI. Các công nghệ như ML (Machine Learning: học máy), DL (Deep Learning: học sâu) và dữ liệu lớn (Big Data) là tất cả các phần của bức tranh AI như minh họa trên hình 1. Vì hầu hết các tiến bộ gần đây trong AI là trong lĩnh vực học máy, mọi người thường nhầm kết hợp AI với ML.

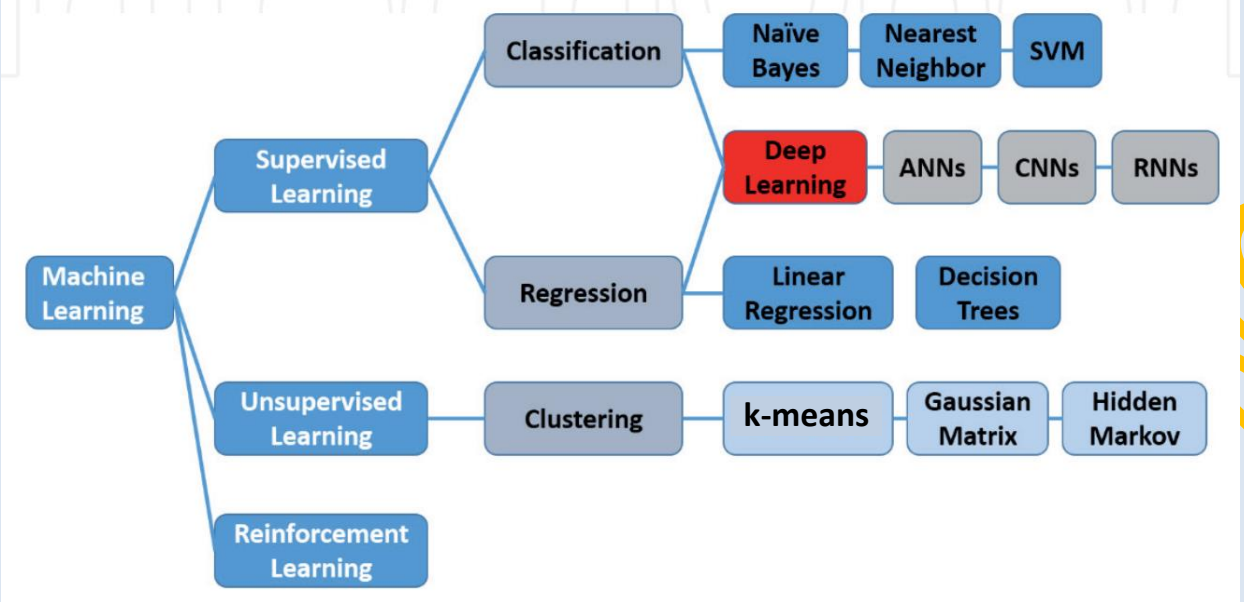


Artificial Intellegence: trí tuệ nhân tạo; Machine Learning: học máy, Rule-based Techniques: các kỹ thuật dựa trên quy tắc; Expert System: hệ thống chuyên gia; Planning: quy hoạch; Natural Language Understanding: hiểu ngôn ngữ tự nhiên; Robotics: Robot; Wireless: không dây; Perceiving Sounds: cảm nhận âm thanh; Problem Solving and Learning: giải quyết vấ đề và học tập.

Hình 1.

PHÂN LOẠI CÁC TIẾP CẬN ML KHÁC NHAU

ML là một tập con với mục đích trao cho một máy tính khả năng thực hiện các nhiệm vụ mà không có các chỉ dẫn tường minh cho trước là làm cách nào để giải quyết nó. Đây là một mô hình hướng đến xây dựng một máy tính có thể học giống như con người. Quá trình học bao gồm cung cấp một giải thuật ML với các thí dụ về nhiệm vụ mà ta muốn giải quyết (dữ liệu) và để máy tính tìm ra các mẫu và thực hiện các suy luận để tối ưu hóa thực hiện quyết định theo mục tiêu được người dùng định nghĩa. Nói chung, ML có thể được sử dụng để thực hiện kiểu các nhiệm vụ khác nhau bao gồm phân loại (Classification), phân cụm (Clustering) và đưa ra quyết định (Making Decision) về dữ liệu.



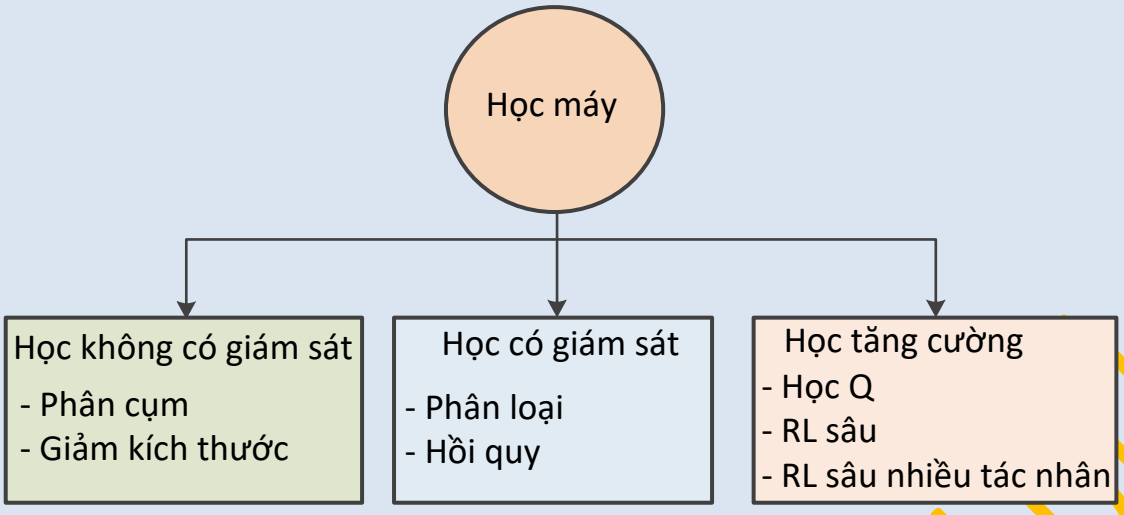
Machine Learning: học máy, Supervised Learning: học có giám sát, Unsupervised Learning: học không có giám sát, Reinforcement Learning: học tăng cường, Classification: ;phân loại, Regression: hồi quy, Clustering: phân cụm, Naïve Bayes: là một thuật toán đơn giản dựa trên định lý Bayes, Deep Learning: học sâu,Linear Repression: hồi quy tuyến tính, K-Means: là một giải thuật đơn giản thường được áp dụng cho các bài toán phân cụm, Nearest Neighbor: thuật toán láng giềng gần nhất, ANN (Artificial Neural Network: mạng tế bào thần kinh nhân tạo, Decision Trees: các cây quyết định, Gaussian Matrix: ma trận Gauss, SVM (Suport Vector Machine): máy vectơ hỗ trợ, CNN (Colvolution Neural Network): Mạng neuron chập, RNN (Recurrent Neural Network): mạng neuron hồi quy, Hidden Markov: Markov ẩn.

Hình 2.

Học có giám sát (Supervised Learning): học với một tập đào tạo được gắn nhãn. Thể loại này bao gồm các nhiệm vụ phân loại và hồi quy (Regression).

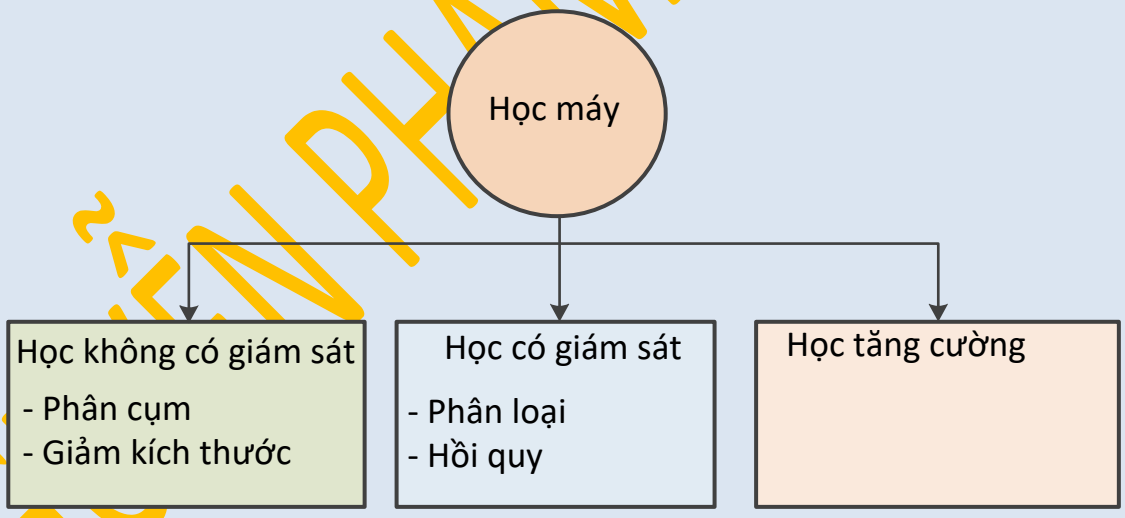
Học không có giám sát (UnsSupervised Learning): Quá trình đào tạo một mô hình sử dụng dữ liệu đào tạo không gắn nhãn. Mô hình phải khám phá các mẫu trong dữ liệu không gắn nhãn. Nhiệm vụ được sử dụng rộng rãi trong học không có giámsát là phân cụm (Clustering).

Học tăng cường (Reinforcement Learning): Quá trình đào tạo một mô hình trên một chuỗi các hành động dẫn đến một kết cục phổ biến, trong đó hệ thống nhận được các phần thưởng cho thực hiện tốt và các trừng phạt cho thực hiện tồi trực tiếp từ môi trường của nó. Học tăng cường được sử dụng trong robot và các trò chơi. Hình 3 cho thấy phân loại ML.



Hình 3.

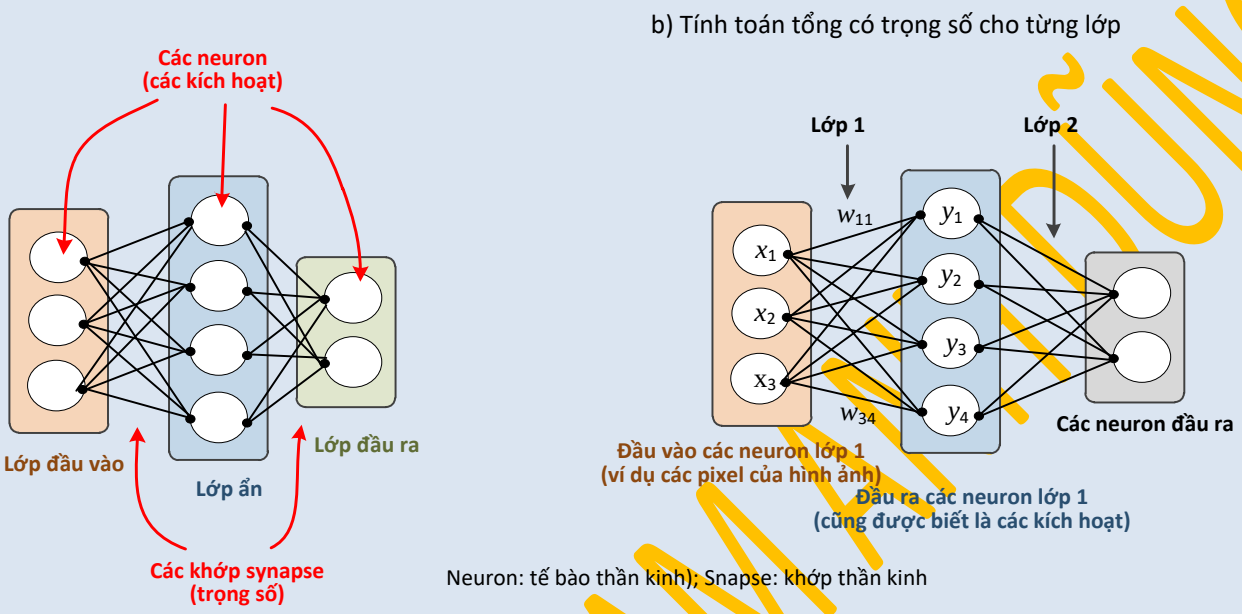
HỌC MÁY TRONG CÁC MẠNG TẾ BÀO THẦN KINH SÂU (DNN: DEEP NEURAL NETWORK)



Hình 3

MẠNG TẾ BÀO THẦN KINH SÂU KẾT NỐI HOÀN TOÀN

Mạng tế bào thần kinh bao gồm các tế bào thần kinh, còn được gọi là perceptron được sắp xấp vào các lớp. Các lớp bao gồm lớp đầu và, lớp đầu ra và nhiều lớp ở gữa được gọi là các lớp ẩn. Mạng tế bào thần kinh nhiều lớp còn được gọi là mạng perceptron đa lớp.



Hình 4.

Ví dụ về mạng neuron trong đó lớp đầu vào có ba neuron với các pixel ảnh l là x_1, x_2 và x_3 và lớp ẩn có bốn neuron với các đầu ra là y_1, y_2, y_3 và y_4. Đầu ra của neuron j được tính như sau:

$$y_j = \mathrm{f}\left(\sum_{i=1}^{3} w_{ij} \times x_i + b\right) = \mathrm{f}\left(\boldsymbol{w}_j^{\mathrm{T}} \times \boldsymbol{x} + b\right) = \mathrm{f}\left(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{w}_j \times + b\right) \qquad (3.1)$$

Trong đó

$\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^{\mathrm{T}}$, $\boldsymbol{w}_j = \begin{bmatrix} w_{1,j} & w_{2,j} & w_{3,j} \end{bmatrix}^{\mathrm{T}}$, b là độ lệch

i =1,2,3 và j=1,2,3,4

$\boldsymbol{x}$, $\boldsymbol{w}_j$,và y_j là vectơ kích hoạt đầu vào, vectơ trọng số, kích hoạt đầu ra của neuron j và f (•) là một hàm phi tuyến tính được mô tả trong phần sau.

Tổng quát cho mạng neuron trong đó lớp đầu vào có K neuron và lớp ẩn a có N neuron, đầu ra của neuron j được tính như sau: :

$$y_j = f\left(\sum_{i=1}^{K} w_{ij} \times x + b\right) = f\ \ \boldsymbol{w}_j \times \boldsymbol{x} + b$$

Nếu kết hợp độ lệch *b* vào vectơ trọng số như đã xét trong chương trước, ta có:

$$\boldsymbol{x} = \begin{bmatrix} 1 & x_1 & \dots & x_{K-1} & x_K \end{bmatrix}^T, \boldsymbol{w}_j = \begin{bmatrix} b & w_{1,j} & \dots & w_{K-1,j} & w_{K,j} \end{bmatrix}^T$$

i=1,2,…, K và j=1,2,…,N

Ta có thể biểu diễn phương trình (3.1) vào dạng tích của các ma trận như sau:

$$\boldsymbol{y} = f\ \ \boldsymbol{W}^T \boldsymbol{x}\ \ = f\ \ \boldsymbol{x}^T \boldsymbol{W} \qquad (3.2)$$

Trong đó $\boldsymbol{y} = \begin{bmatrix} y_1, \dots, y_j, \dots, y_N \end{bmatrix}^T$ là vectơ của N đầu ra của lớp ẩn và $\boldsymbol{x} = \begin{bmatrix} 1 & x_1, \dots, x_i, \dots, x_K \end{bmatrix}^T$ là vectơ của K đầu vào của lớp đầu vào, T ký hiệu cho đảo vị và ma trận trọng số:

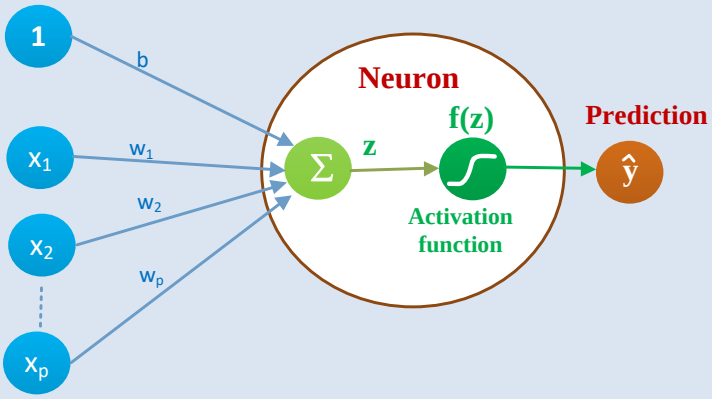
$$\boldsymbol{W} = \begin{bmatrix} w_{0,1} & \dots & w_{0,j} & \dots & w_{0,N} \\ w_{1,1} & & w_{1,j} & & w_{1,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{i,1} & \dots & w_{i,j} & \dots & w_{i,N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{K,1} & \dots & w_{K,j} & \dots & w_{K,N} \end{bmatrix}$$

Trong đó $w_{0,j}$ là độ lệch cho từng đầu vào của lớp ẩn.

Trong lĩnh vực mạng nơ-ron, có một lĩnh vực được gọi là học sâu (Deep Learning), 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.

CÁC HÀM PHI TUYẾN TÍNH

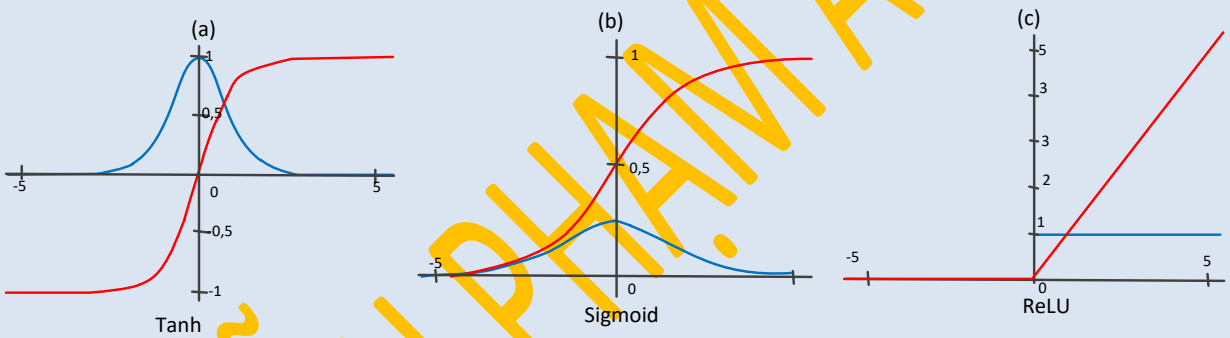
Một trong các phần tử quan trọng nhất của các mạng neuron sâu là các hàm phi tuyến tính còn được gọi là các hàm kích hoạt. Chúng chuyển đổi các đầu vào tuyến tính thành các đầu ra phi tuyến tính để tạo điều kiện cho học các đa thức bậc cao.



Neuron: tế bào thần kinh;
Activation Function: hàm kích hoạt
Prediction: dự đoán

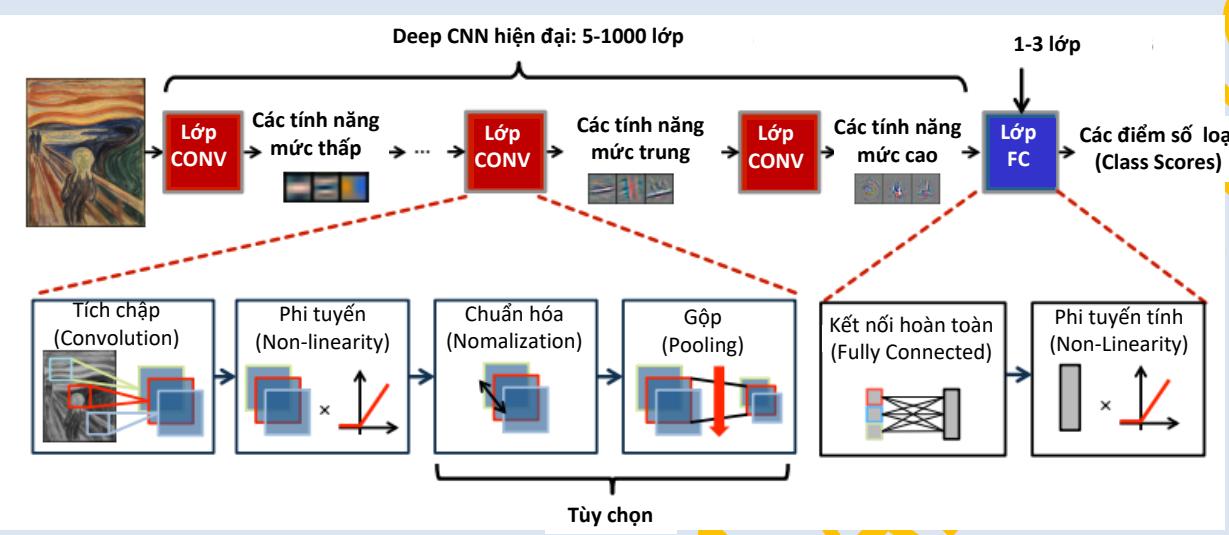
Hình 5.

Thí dụ về một số hàm phi tuyến tính (đỏ) và các đạo hàm bậc một của chúng (xanh da trời). (a) hàm tangent hyperbol, (b) hàm Sigmoid, (c) hàm đơn vị tuyến tính chỉnh lưu (RLU: Rectified Linear Unit)



Hình 6.

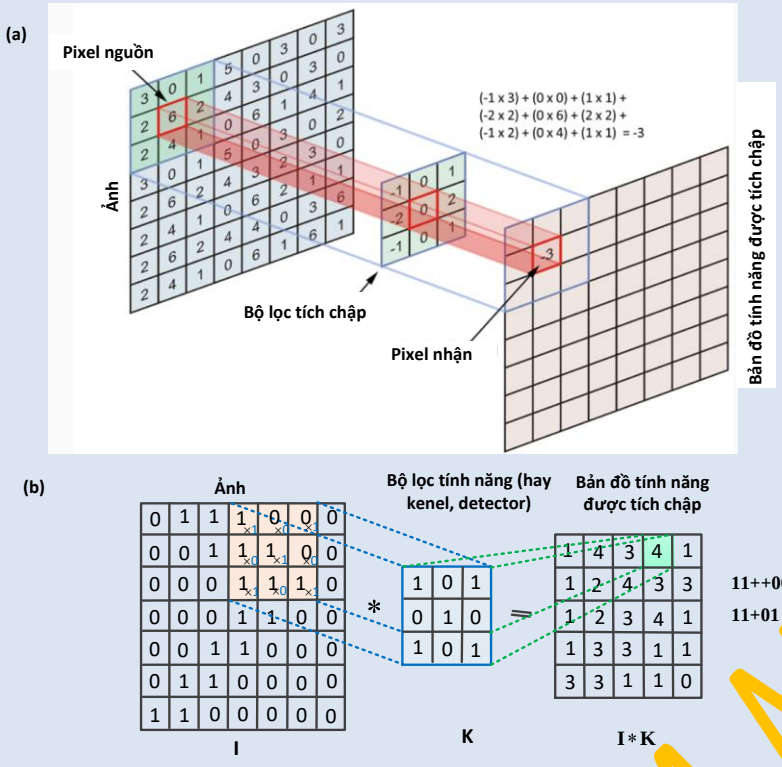
MẠNG TẾ BÀO THẦN KINH TÍCH CHẬP (CONVOLUTIONAL NEURAL NETWORK)



Hình 7.

Tích chập

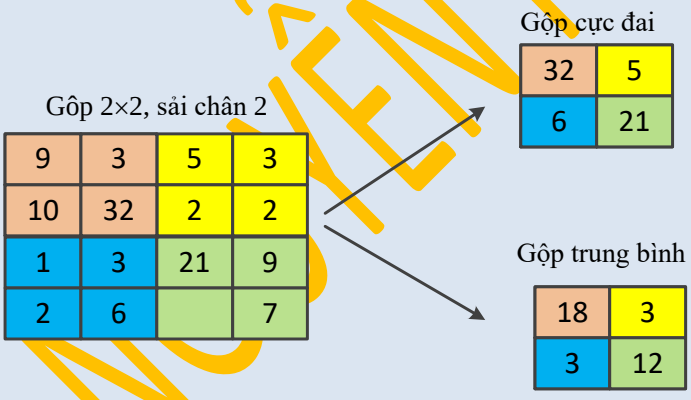
$$h(t) = \int_{-\infty}^{\infty} f(t-\tau)g(\tau)d\tau$$



Hình 8.

Pooling (gộp)

Một số trong số các lớp giảm mẫu phổ biến nhất là gộp cực đại (Maximum Pooling), gộp trung bình (Average Pooling) như được minh họa trên hình vẽ hay gộp trung bình toàn cục (Global Average Pooling).

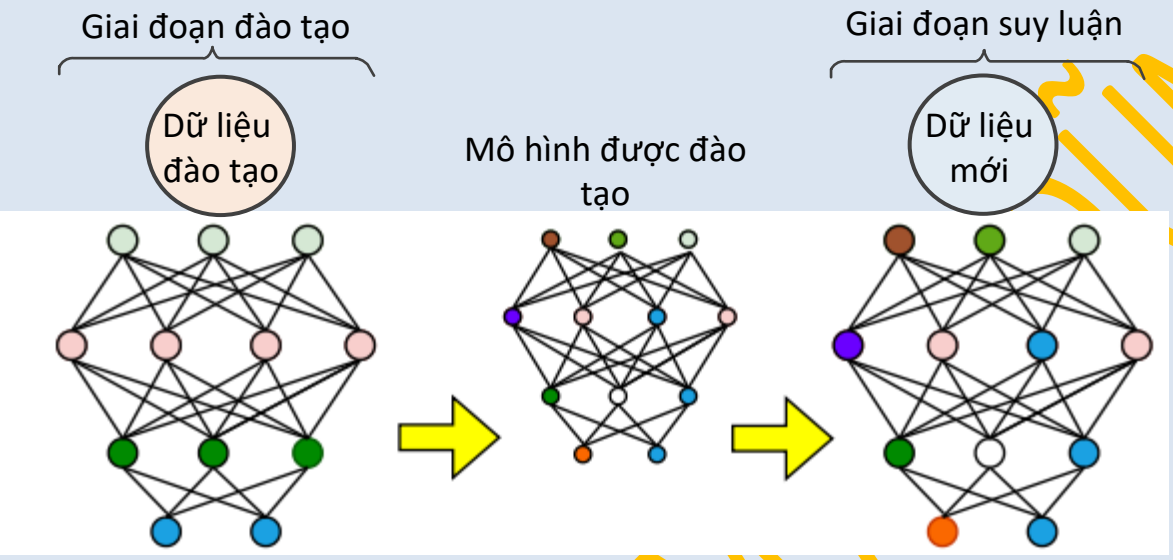


Hình 9.

ĐÀO TẠO VÀ SUY LUẬN (TRAINING AND INFERENCE)

TỔNG QUAN

Các giai đoạn đào tạo và suy luận



Hình 10.

ĐÀO TẠO DNN

Trong trường hợp cụ thể của DNN, việc học này liên quan đến việc xác định giá trị của trọng số (và độ lệch: 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. đào tạo được thực hiện dựa trên cực tiểu hoa hàm mất mát (Loss Function) hay hàm chi phí (Cost Function)

Các hàm mất mát và chi phí (Loss Function and Cost Function)

Hàm này trình bày lỗi đối với một dự đoán cho trước. Để vậy, đối với một mẫu dự đoán cho trước, nó so sánh dự đoán $\mathrm{f}(\boldsymbol{x}_i, \boldsymbol{W})$ với sự thật cơ bản y_i (để đơn giản ta ký hiệu $\boldsymbol{W}$ là tất cả các tham số $\boldsymbol{W}^1$, …. , $\boldsymbol{W}^M$ trong perceptron nhiều lớp đã xét ở trên). Hàm mất mát được ký hiệu như sau $\ell\ y, \mathrm{f}(\boldsymbol{x}, \boldsymbol{W})$. Mất mát trung bình trên tất các các mẫu đào tạo (N) được gọi là hàm chi phí (Cost Function) và được định nghĩa như sau:

$$L(\boldsymbol{W}) = \frac{1}{N}\sum_{i=1}^{N} \ell \ \ y_i, f(\boldsymbol{x}_i; \boldsymbol{W})$$

Quá trình đào tạo

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 mất mát (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 mất mát so với từng trọng số là đạo hàm riêng được sử dụng để cập nhật trọng 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 trọng số cần thay đổi như thế nào

để giảm mất mát. Quá trình đào tạo được lặp đi lặp lại để liên tục giảm mất mát tổng thể. Cho đến khi mất mát dưới ngưỡng xác định trước, DNN với độ chính xác cao sẽ thu được.

Mã giả (pseudocode) cho đào tạo

Algorithm 1 Train Perception

Procedure TRAIN $(\{\boldsymbol{x}_i, y_i\})$

Initialization: initialize randomly the weights **w** and bias b

while $\exists i \in \{1,\ldots,N\}, f(\boldsymbol{x}_i, \boldsymbol{w}, b) \neq y_i$ **do**

Pick i randomly

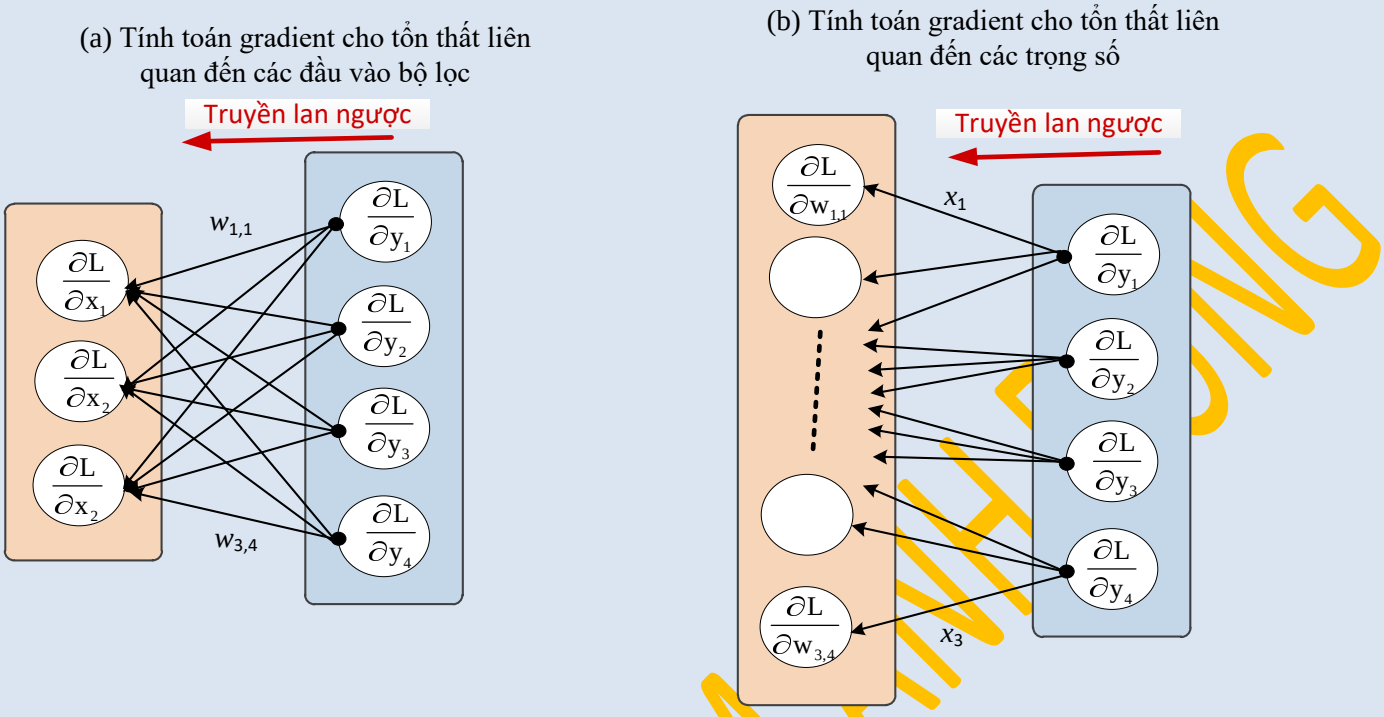
error= $y_i - f(\boldsymbol{x}_i, \boldsymbol{w}, b)$

if error $\neq 0$ **then**

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \alpha \frac{\partial L}{\partial \boldsymbol{w}}$$

$$b \leftarrow b + \text{error}$$

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 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 mất mát đã chịu ảnh hưởng của từng trọng số như thế nào.



Hình 11.

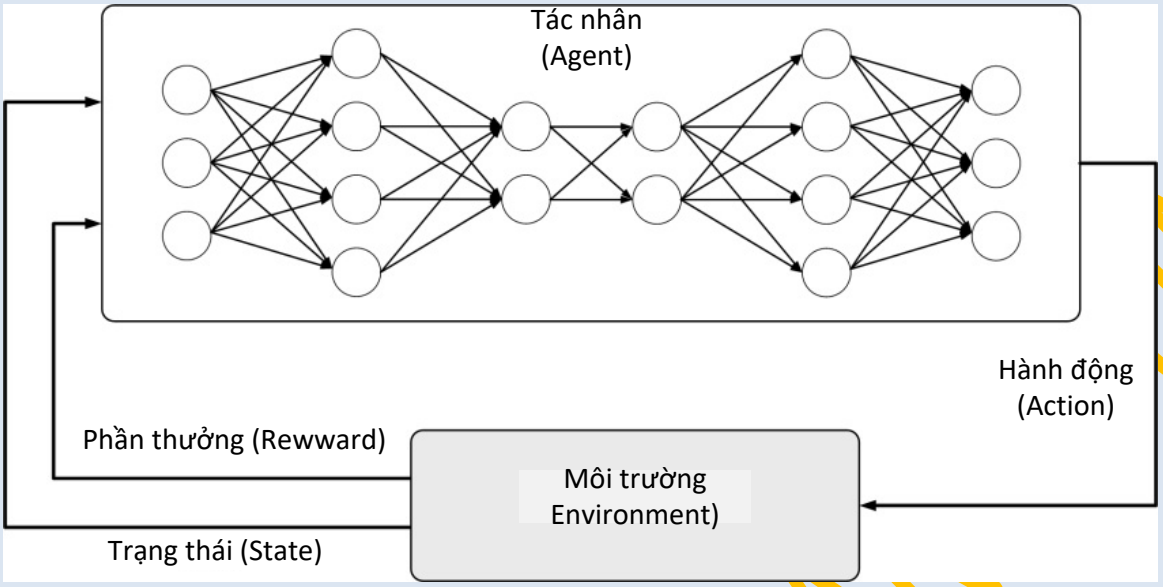
Có nhiều cách để đào tạo mạng (đào tạo trọng số) cho các mục tiêu khác nhau. Giới thiệu ở trên là học có giamsát (supervised learning) sử dụng các mẫu đào tạo được gắn nhãn để tìm đầu ra chính xác cho một nhiệm vụ. Học không giám sát (unsupervised learning) sử dụng các mẫu đào tạo không được gắn nhãn 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 cục) 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.

Đào tạo tăng cường sâu (RL)

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 dưới 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

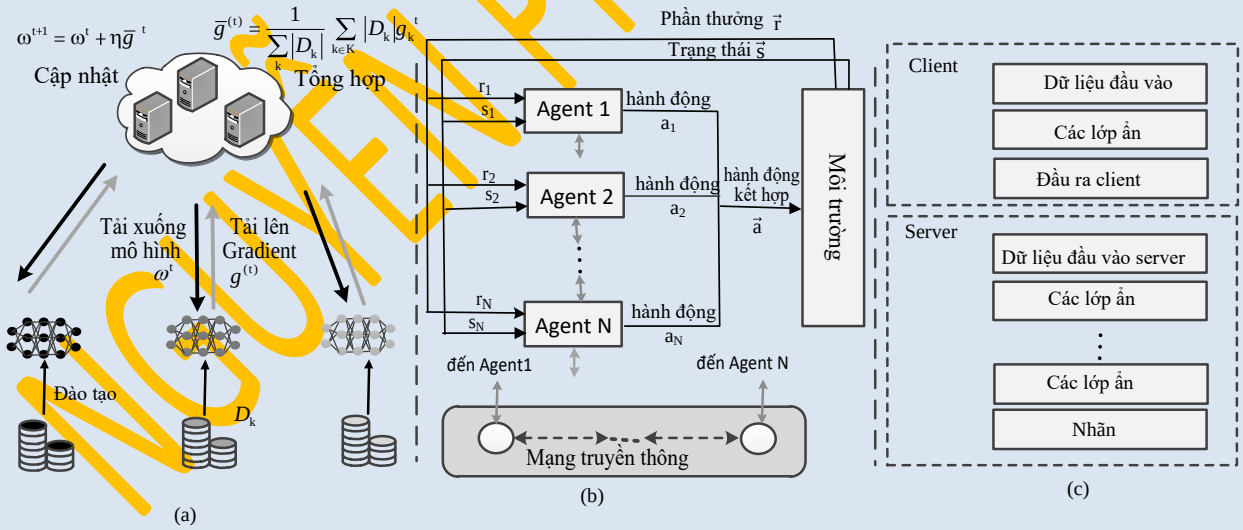
đề lập lịch khác nhau, chẳng hạn như các vấn đề quyết định trong trò chơi, lựa chọn tốc độ truyền video, v.v.



Hình 12.

HỌC PHÂN TÁN

Học phân tán dựa trên trí tuệ phân tán gồm ba thành phần: (a) FL (Federated Learning: học liên kết), (b) MARL (Multi-Agent Reinforcement Learning: học tăng cường đa tác nhân) và (ic) SL (Split Learning: học chia sẻ), như được minh họa trong hình dưới.



Hình 13

1) **FL**. FL là một mô hình của trí tuệ phân tán cho phép tăng cường tính riêng tư bằng cách đào tạo các mô hình AI tại các thiết bị cục bộ và chỉ chia sẻ các thông số mô hình thay vì dữ liệu thô. Trong quá trình đào tạo, từng nút client đào tạo cục bộ và tải lên các trọng số mô hình đều đặn. Một nút trung tâm tổng hợp các trọng số này và hồi tiếp các trọng số được tổng hợp đến client cho vòng đào tạo tiếp theo hay cho suy luận. Trên hình (a) qúa trình bắt đầu với tải xuống mô hình toàn cục $\omega^{(t)}$ cho vòng t đến các thiết cục bộ. Từng thiết bị k đào tạo mô hình trên tập dữ liệu D_k riêng của mình, tạo ra một cập nhật gradient $g_k^{(t)}$ và sau đó nó được cập nhật đến server thông số. Server này áp dụng giải thuật trung bình bình hóa liên kết (FedAvg: Federated Averaging) để tổng hợp các cập nhật này và gradient toàn cục $\bar{g}^{(t)} = \frac{1}{\sum_k |D_k|} \sum_k |D_k| g_k^{(t)}$ sau đó cập nhật mô hình toàn cục ω^{t+1}cho vòng tiếp theo.

2) **Học tăng cường đa tác nhân (MARL: Multi-Agent Reinforcement Learning)**

Trong môi trường chia sẻ, các tác nhân học tìm kiếm các giải thưởng riêng cho mình. Hệ thống đa tác nhân (MAS) bao gồm một tập hợp các tác nhân ra quyết định tồn tại trong một môi trường chung, như minh họa trong hinh (b). Các tác nhân này quan sát môi trường của chúng và giao tiếp với nhau để thực hiện các hành động phù hợp với mục tiêu của chúng. Thiết lập đa tác nhân liên quan đến việc quản lý các quá trình ra quyết định và học tập của nhiều thực thể có thể tương tác với nhau và môi trường chung của chúng.

3) Học học phân chia (Split learning)

Do các hạn chế của các thiết bị bị ràng buộc tài nguyên không thể hỗ trợ các mô hình học sâu hay cộng tác FL. SL (Split Learning: học phân chia) mỏi lên như là một giải pháp khả thi. SL phân chia các mô hình ML bằng cách phân tán các phần khác nhau giữa client và server để đảm bảo rằng dữ liệu vẫn được an ninh. Trong khi các nhiệm vụ tính toán nặng được trao cho các nút trung tâm mạnh mẽ, thì các lớp xử lý dữ liệu nhẹ hơn được vẫn được giữ nguyên tại đầu cuối nơi lưu trữ dữ liệu này. Như là một giải pháp thay thế hay bổ sung cho FL, SL giảm gánh nặng xử lý trên các thiết bị bị ràng buộc bởi tài nguyên

CÁC YÊU CẦU VỀ HIỆU SUẤT ĐỐI VỚI HỌC PHÂN TÁN/ LIÊN KẾT

- Mất mát đào tạo

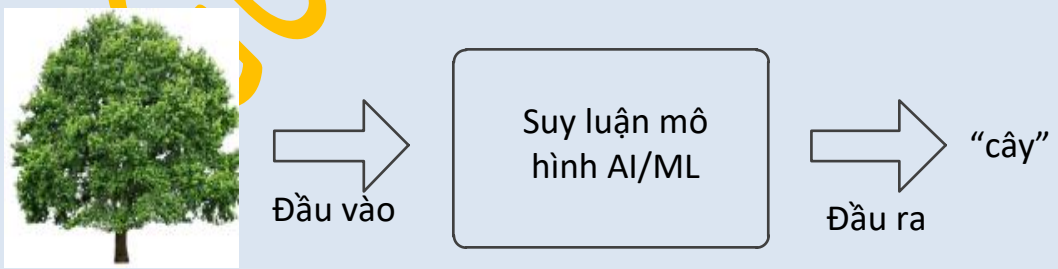
  Mất mát đào tạo là khoảng cách giữa đầu ra chính xác và đầu ra được tính toán bởi mô hình DNN, cho biết mô hình DNN được đào tạo phù hợp với dữ liệu đào tạo như thế nào. Mục đích của nhiệm vụ đào tạo là giảm thiểu mất mát đào tạo. Mất mát đào tạo chủ yếu bị ảnh hưởng bởi chất lượng của dữ liệu đào tạo 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, mất mát đào tạo mới có thể được giảm thiểu.

- 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.
- 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.
- Quyền 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.

SUY LUẬN MÔ HÌNH AI/ML AI/ML (AI/ML MODEL INFERENCE)

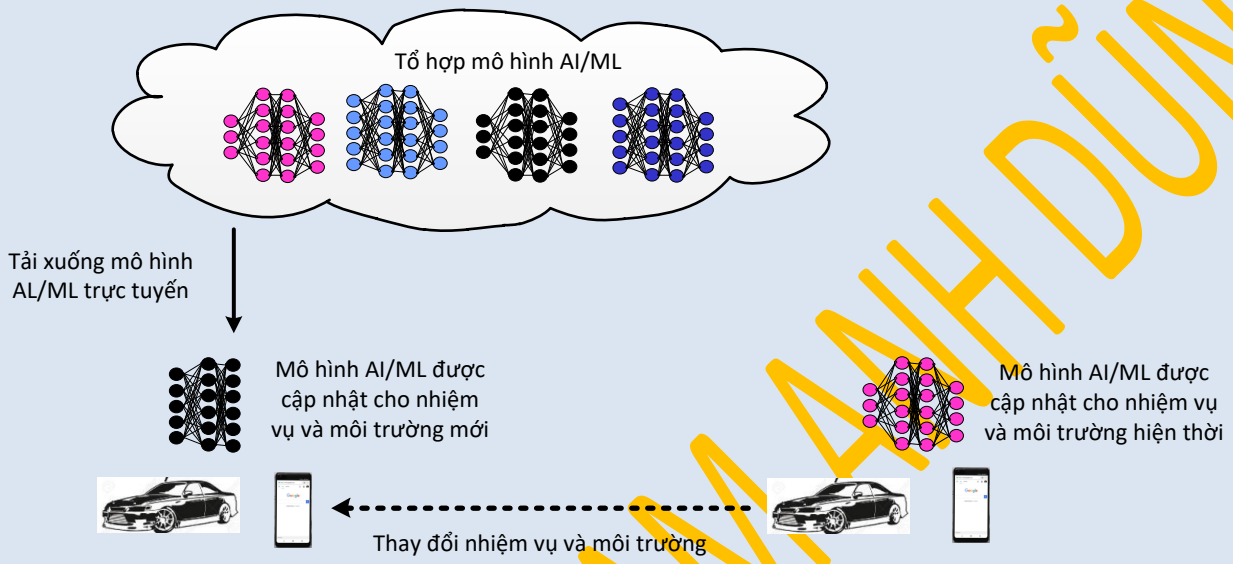
Sau khi DNN được đào tạo, nó có thể thực hiện nhiệm vụ của mình bằng cách tính toán đầu ra của mạng bằng cách sử dụng trọng số được xác định trong quá trình đào tạo, được gọi là suy luận (Inference). Trong quá trình suy luận mô hình, đầu vào từ thế giới thực được truyền qua DNN. Sau đó, dự đoán (Prediction) cho nhiệm vụ là đầu ra, như trong hình dưới đây. 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 14

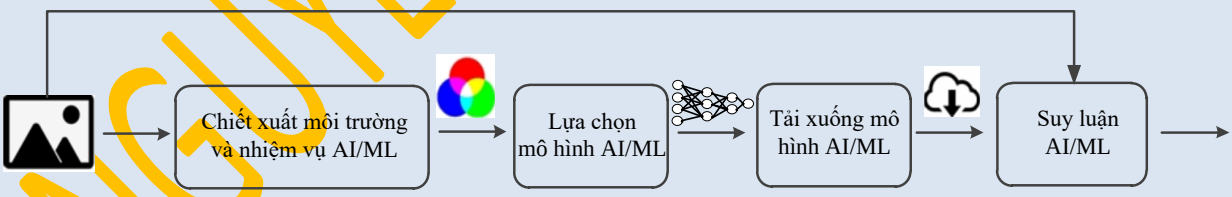
PHÂN PHỐI VÀ LỰA CHỌN MÔ HÌNH THÍCH ỨNG CÓ SẴN CHO SUY LUẬN

Như minh họa trong Hình dưới đây, mô hình AI/ML có thể được phân phối 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 15.

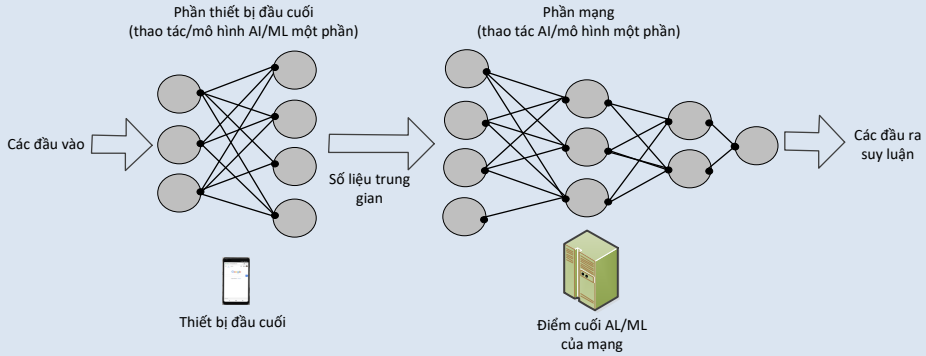
Mô hình được phân phối có thể được xác định theo hai cách: được yêu cầu bởi một thiết bị hoặc được điều khiển bởi một máy chủ mạng. Điều kiện của cơ chế đầu tiên là thiết bị có thể đưa ra quyết định lựa chọn / lựa chọn lại mô hình dựa trên sự hiểu biết về nhiệm vụ AI/ML sắp tới, môi trường và danh sách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



Hình 16.

SUY LUẬN PHÂN CHIA

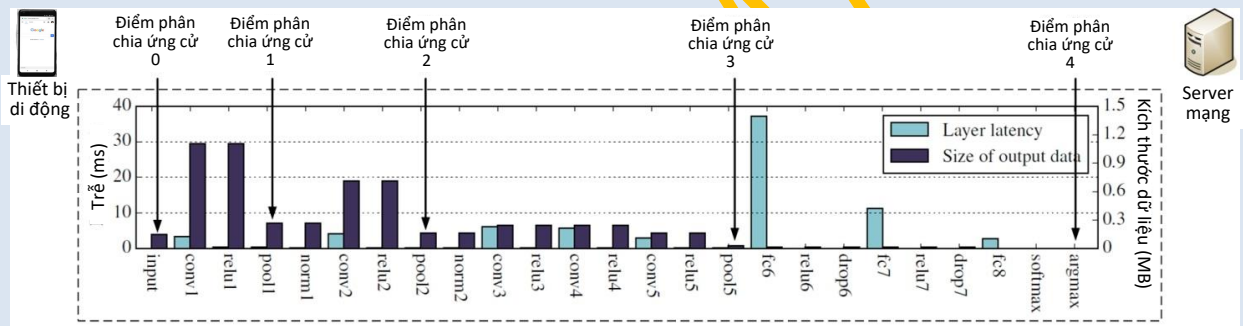
Nguyên lý suy luận phân chia (Split Inferrence)



Hình 17

Chọn điểm phân chia

Hình dưới đây minh họa chọn điểm phân chi dựa trên kích thước dưa liệu đầu ra của phần suy lận tại UE cho nhận dạng ảnh trên mạng AlexNET.

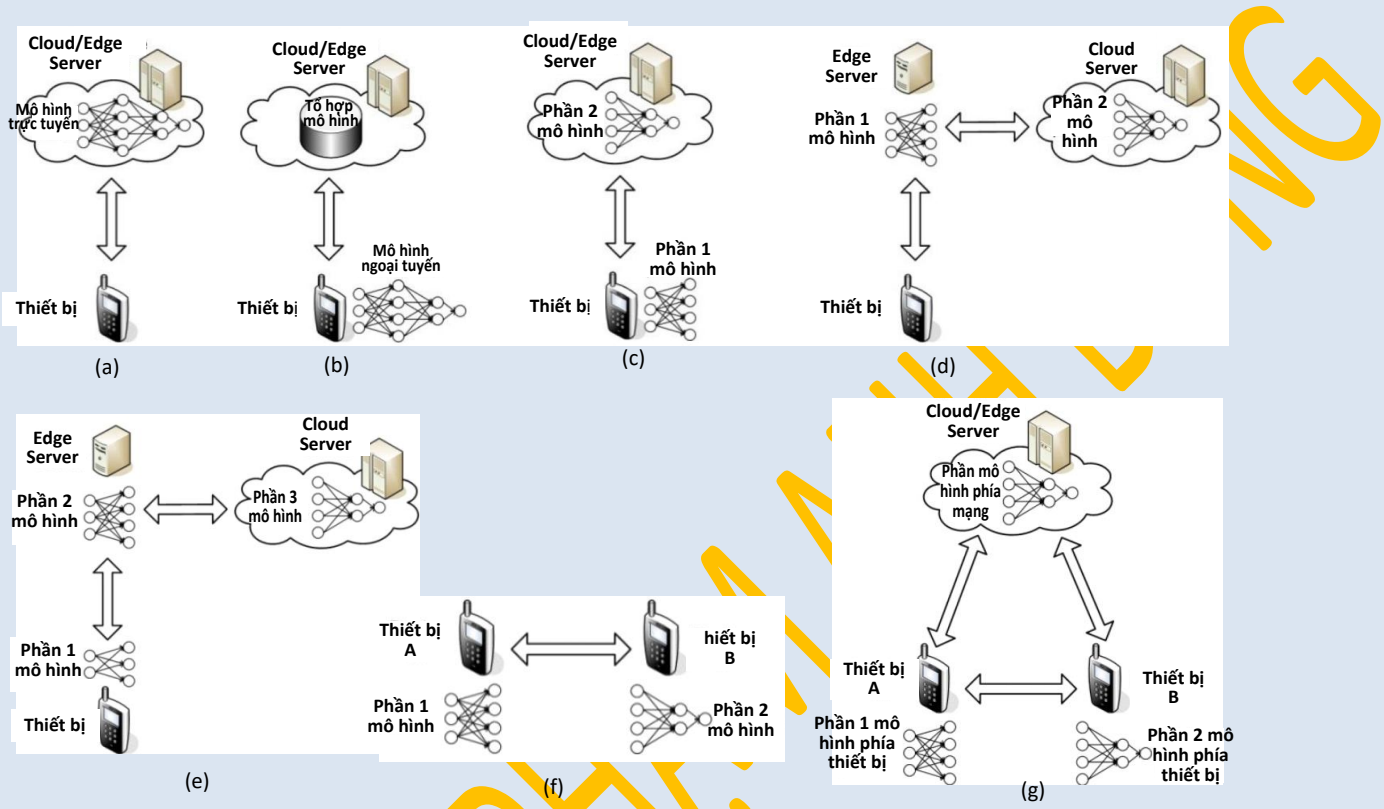


CNN cho nhận dạng hình ảnh (Alex NET)

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

Hình 18

Các chế độ phân chia



Hình 19

TÍCH HỢP AI/ML VÀO MẠNG TRUYỀN THÔNG DI ĐỘNG 6G

Để khắc phục các hạn chế của 5G nhằm hỗ trợ các yêu cầu mới, cần phát triển một hệ thống không dây thế hệ 6 (6G) với các tính năng hấp dẫn mới. Các động lực chính của 6G sẽ là sự hội tụ của tất cả các tính năng cũ như mật độ mạng cao, thông lượng cao, độ tin cậy cao, tiêu thụ công suất thấp và kết nối số đông. 6G cũng sẽ tiếp tục xu thế của các thế hệ trước như đưa ra các dịch vụ mới cùng với các công nghệ bổ sung mới. Các dịch vụ mới bao gồm AI, các thiết bị đeo thông minh, các thiết bị cấy trên người, giao thông tự động, các thiết bị thực tế ảo, cảm biến và lập bản đồ 3D. Yêu cầu quan trọng nhất

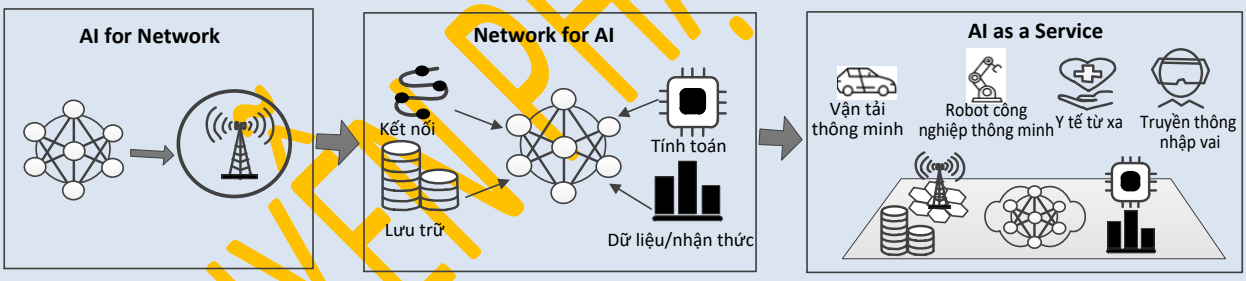
đối với các mạng 6G là khả năng xử lý khối lượng lớn số liệu và kết nối tốc độ số liệu cao trên một thiết bị.

Hệ thống truyền thông không dây 6G sẽ tăng QoS nhiều lần so với 5G cùng với một số tính năng hiện có. Nó sẽ bảo vệ hệ thống và đảm bảo an ninh số liệu người dùng. Nó sẽ cung cấp các dịch vụ tiện lợi. Hệ thống truyền thông không dây 6G được kỳ vọng là một phương tiện truyền thông toàn cầu. Nó được dự kiến đạt tốc độ vào khoảng 1Tbit/s trong nhiều trường hợp. 6G được kỳ vọng đảm bảo kết nối đồng thời 1000 lần cao hơn 5G.

Ngoài ra nó cũng được kỳ vọng đảm bảo kết nối cự ly rất xa với độ trễ chỉ 1 ms. Tính năng hấp dẫn nhất của 6G là đảm bảo hỗ trợ hoàn toàn AI cho các hệ thống tự lái. Trong truyền thông 6G, dự tính lưu lượng kiểu video sẽ nổi trội. Các công nghệ quan trọng nhất đóng vai trò động lực cho 6G sẽ là băng tần terahertz (THz), AI, truyền thông không dây quang (OWC: optical wireless communications), nối mạng 3D, thiết bị bay không người lái (UAV: unmanned aerial vehicle) và chuyển công suất không dây.

TẦM NHÌN TÍCH HỢP AI VÀ MẠNG TRUYỀN THÔNG 6G

Có thể mường tượng mức độ tích hợp giữa 6G và AI từ ba góc nhìn: AI cho mạng (AI4N: AI for Network), mạng cho AI (NET4AI: Network for AI) và AI như là một dịch vụ (AIaaS: AI as a Service).



AI for Network: AI cho mạng, Network for AI: mạng cho AI, AI as a Service: AI như là một dịch vụ.

Hình 20

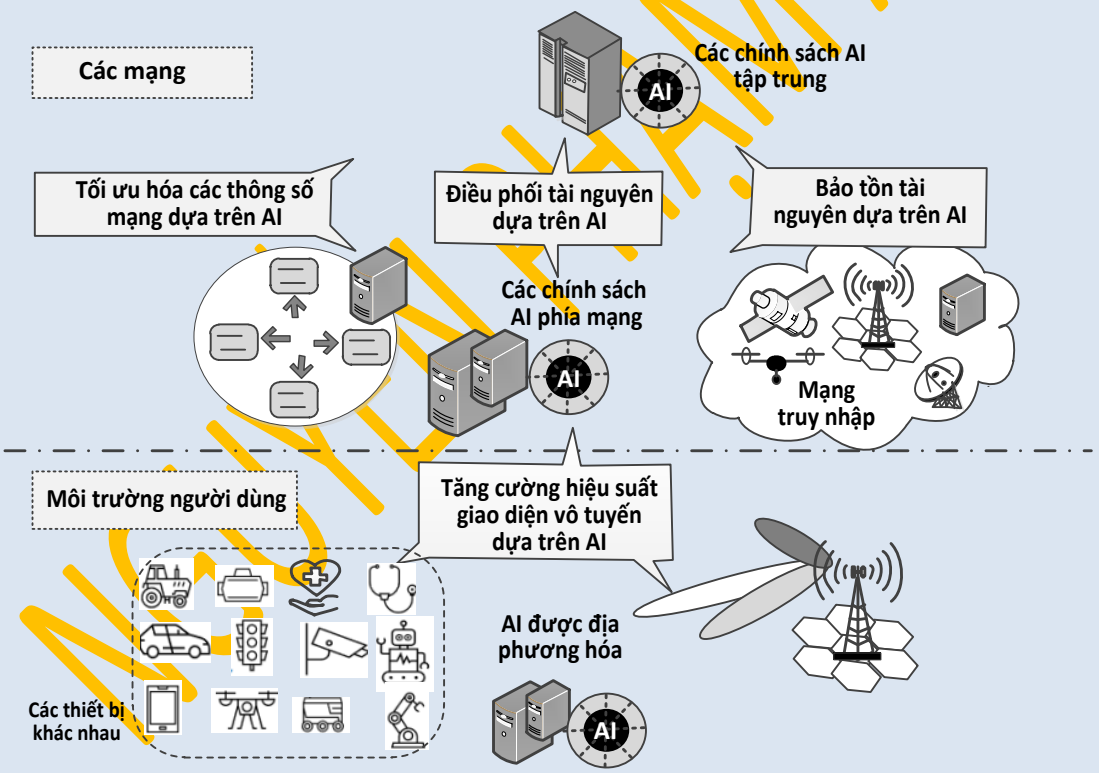
AI4NET

AI4NET đề cập đến kịch bản trong đó trí tuệ nhân tạo/học máy (AI/ML) được sử dụng để tăng cường sức mạnh cho mạng. Trong bối cảnh này, các công nghệ AI được khai thác để nâng cao và tối ưu hóa hiệu suất và quản lý

mạng không dây. Mục tiêu tổng thể của AI4NET là tận dụng trí tuệ nhân tạo để tăng cường hiệu quả, tính ổn định của mạng và cuối cùng là hiệu suất mạng và trải nghiệm người dùng.

Tại giai đoạn này, chỉ tập trung là làm thế nào sử dụng các giải thuật AI để tối ưu hóa hiệu suất truyền thông và các chức năng mạng. Chẳng hạn, AI có thể tối ưu hóa quá trình điều chế và giải điều chế để truyền dẫn tín hiệu chính xác và hiệu quả hơn; với sự hỗ trợ của AI, tài nguyên mạng có thể được ấn định thông minh để đạt được cân bằng tải, và và có thể xây dựng quản lý O&M tự động để cải hiện hiệu suất O&M và giảm chi phí. Việc đưa vào AI không được kỳ vọng có ảnh hưởng đáng kể lên kiến trúc mạng gốc. Thay vào đó, nó cải thiện các vấn đề truyền thông trong việc nhận biết đặc thù bằng các đào tạo các mô hình giải huật AI. Qúa trình này tương tự như cấy ghép chính xác các miếng vá thông minh vào mạng. Nó không chỉ duy trì sự ổn định kiến trúc mạng mà còn từ từ tăng cường mức độ thông minh của mạng.

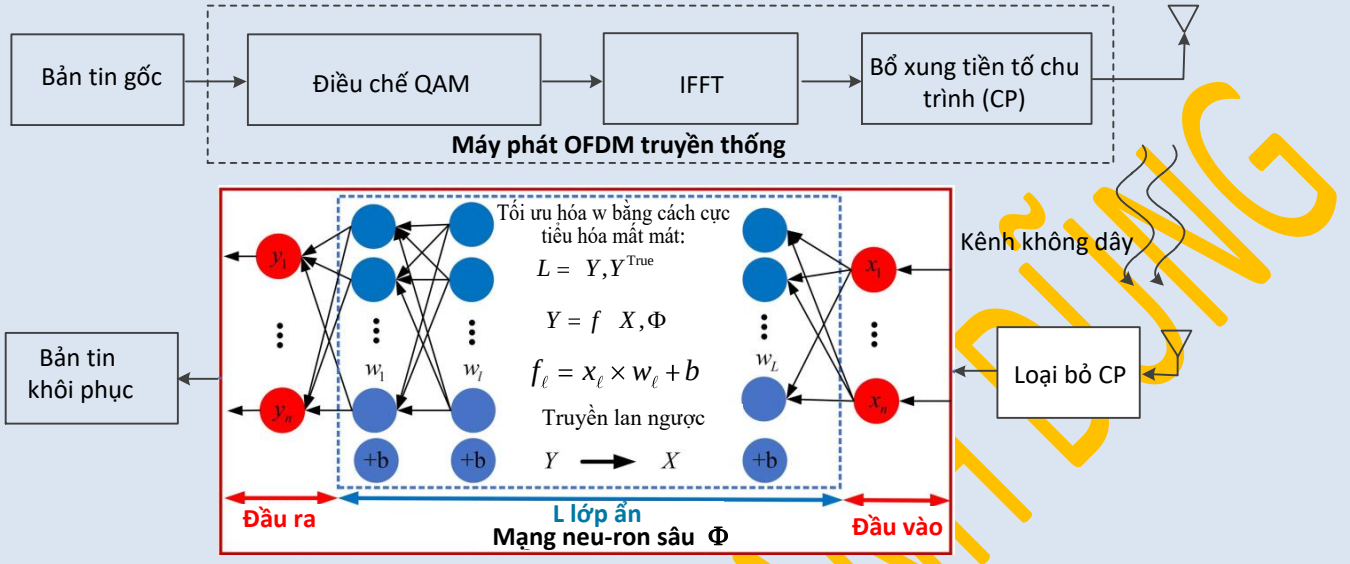
Khái niệm AI4NET được trình bày trong hình dưới đây, trong đó AI không dây có các khả năng như chiết xuất tính năng, dự đoán thích ứng, tối ưu hóa, xử lý thời gian thực, đánh giá tương quan và phân cụm cảnh hiện trường (Scene Clustering). Các khả năng này tạo lập nền tảng của các mạng không dây được AI hỗ trợ và có thể được triển khai trực tiếp trên BS và CN để tạo điều kiện cho truyền thông không dây.



HÌnh 21.

Tăng cường hiệu suất giao diện vô tuyến:

- Giải thuật phản hồi CSI dựa trên AI
- Máy thu OFDMA dựa trên AI

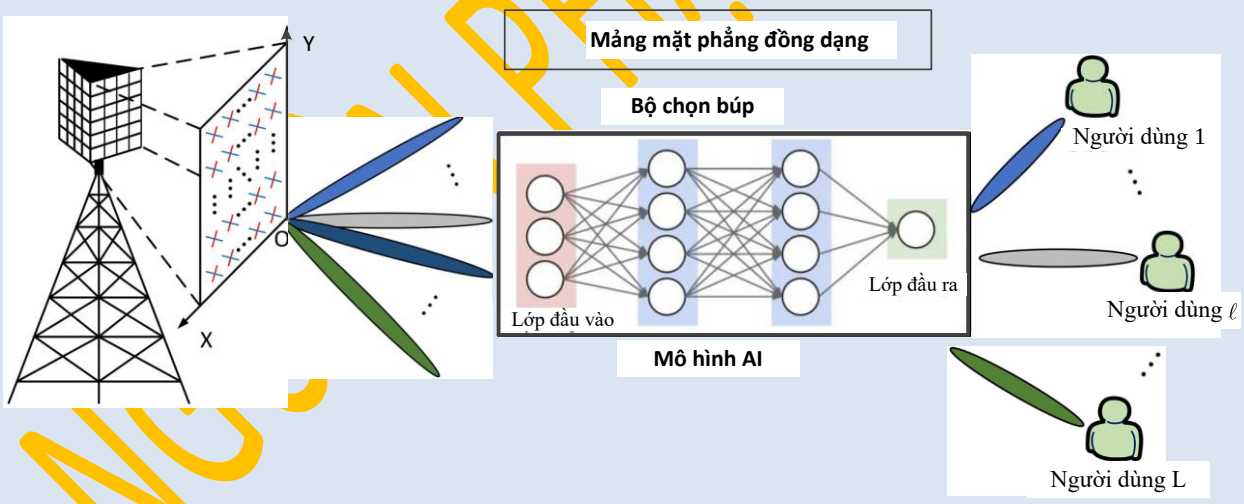


QAM (Quuadrature Amplitude Modulation): điều chế biên độ cầu phương, IFFT (Inverse Fast Fourier Transformation): biến đổi Fourier nhanh đảo, CP (Cyclic Prefix): tiền tố chu trình.

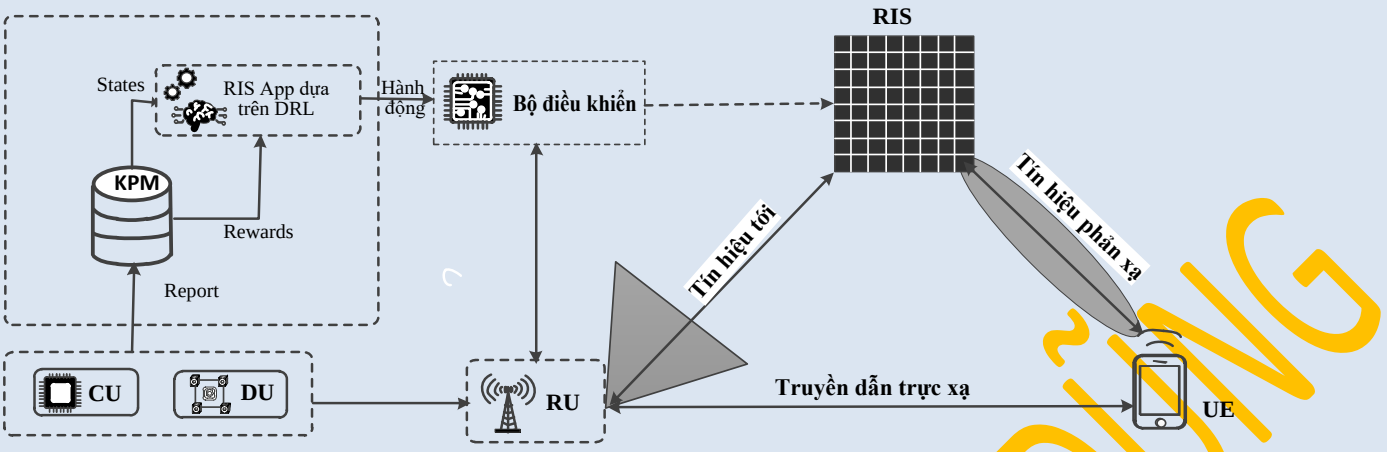
Hình 22.

- Quản lý búp sóng dựa trên AI

a) Điều khiển búp sóng hướng đến người dùng



b) Điều khiển RIS (Reconfigurable Intellegence Surface)



Hình 24.

- Định vị dựa trên AI

Cải thiện hiệu năng O&M mạng:

- Dự báo lưu lượng dựa trên AI: Trong các phương pháp dự đoán lưu lượng, một BS thông thường cần thu thập dữ liệu từ các vị trí địa lý khác nhau, điều này không thể tránh khỏi đưa vào chi phí truyền thông gián tiếp bổ sung và các rủi do an ninh tiềm năng. Kết quả là, các phương pháp dự đoán dựa trên học liên kết (FL: Fedeerated Learning) đã nổi lên. Phương pháp FeDDA đã được đề xuất, nó cho phép giảm đáng kể trễ và chi phí băng thông gián tiếp liên kết với truyền dẫn dữ liệu vì chỉ phát các thông số mô hình thay vì dữ liệu lưu lượng thô.
- Bảo tồn năng lượng BS dựa trên AI: Chiến lược tiết kiệm năng lượng cho BS dựa trên học tăng cường (RL: Reinforcement Learning) giữa tác nhân và môi trường khai thác, thay đổi động các thông số để thích ứng các thay đổi tải dịch vụ và các giao động trong các điều kiện kênh.

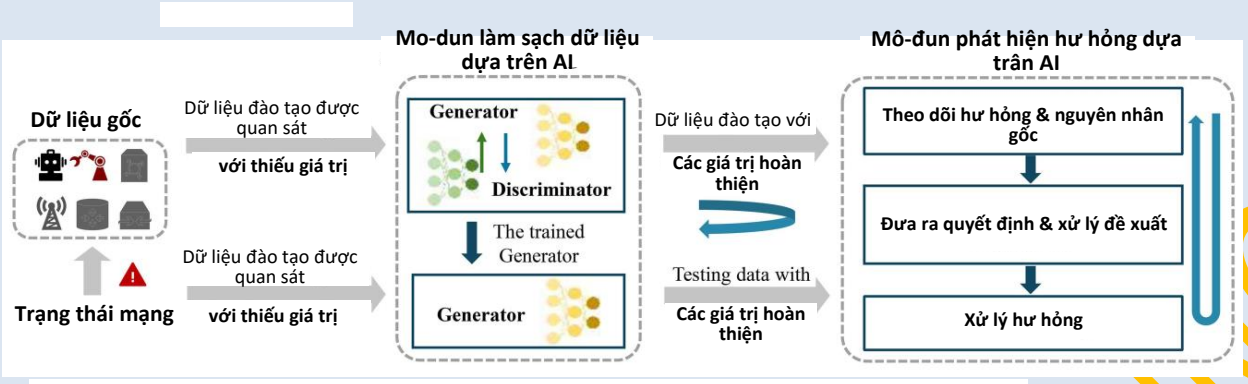
- Tối ưu hóa các thông số mạng dựa trên AI:

Các thông số khác nhau trong các mạng hiện thời

Kiểu thông số	Tên thông số	Lĩnh vực áp dụng
Thông số thiết bị đầu cuối	Tỷ lệ lỗi khối Thông tin lưu lượng Báo cáo đo đạc	Tối ưu hóa lập lịch Giám sát lưu lượng Mã hóa thích ứng
Thông số giao diện vô tuyến	Trễ thời gian đa đường Nhiễu giữa các ô Thông tin trạng thái kênh	Lập mô hình kênh Loại bỏ nhiễu Dự báo trạng thái kênh
Thông số CN	Nhật ký hoạt động Cấu hình topo mạng Tiêu thụ năng lượng mạng	Xử lý lỗi Cân bằng tải Tiết kiệm năng lương
Thông số dịch vụ	Thời gian trực tuyến Giao thức vận tải Lịch sử tiêu thụ	Dịch vụ được cá nhân hóa Điều khiển truyền dẫn Giám sát dịch vụ

Các mô hình AI đã chứng tỏ các ưu điểm không thể dự đoán trước trong điều chỉnh thông số quy mô lớn, khớp phi tuyến tính (Nonlinear Fitting) và tối ưu hóa thời gian thực

- *Quản lý di động dựa trên AI*: AI có thể dự đoán các hành vi của người dùng, bao gồm tốc độ và phương di động, bằng cách đó cho phép chọn và chuyển giao chủ động giữa các ô để giảm chi phí báo hiệu. Thông qua RL và các kỹ thuật khác, AI có thể tối ưu hóa động các chiến lược quản lý di động dựa trên môi trường mạng và hành vi người dùng, xử lý các môi trường mạng tốt hơn
- *Điều phối dựa trên AI*: Với hàng trăm tỷ thiết bị được kết nối, dữ liệu được tạo ra không ngừng tăng trưởng bùng nổ. Chỉ đơn giản tăng cường các khả năng truyền thông là không đủ để đáp ứng các yêu cầu xử lý dữ liệu thời gian thực trong các kịch bản ứng dụng điều khiển. Bằng cách khai thác các khả năng truyền thông và tính toán tại biên mạng và tại phía thiết bị, toàn bộ quá trình thu thập, xử lý, phân tích dữ liệu và đưa ra quyết định được thực hiện gần phía thiết bị hơn nhờ vậy tránh được nhược điểm trễ gây ra bởi nghẽn trong CN. Điều này dẫn đến điều phối phải dựa trên AI. Điều phối dựa trên AI đối mặt nhiều thách thức và các cơ hội trong tương lai. Chẳng hạn, cần nhanh chóng cải thiện các giải thuật DRL để đối phó các môi trường có khả năng thay đổi. Cần ra sức tăng cường sự thích ứng với các kịch bản phức tạp để đáp ứng yêu cầu tài nguyên đa dạng trong các kich bản đặc biệt của các nghành công nghiệp khác nhau, cải thiện đánh giá hiệu suất và tăng cường tính tương tác của giải thuật.
- *Nhận thức tình trạng và phát hiện hư hỏng dựa trên AI*: Công nghệ nhận thức tình trạng cho phép các mạng truyền thông khả năng giám sát, dự đoán và các khả năng đáp ứng thời gian thực. Các ứng dụng của nó bao gồm quản lý và tối ưu hóa, an ninh mạng cũng như phát hiện và dự báo hư hỏng. Nhận thức tình trạng thu thập, lưu trữ và phân tích các khối lượng lớn dữ liệu mạng thông qua các nền tảng dữ liệu lớn và các kỹ thuật DL, qua đó cung cấp tình báo mối đe dọa (Threat Intellegence) chính xác và kiểm tra lưu lượng toàn diện. Công nghệ này sử dụng DL để nhận dạng chính xác các trạng thái của các phần tử khác nhau bao gồm lưu lượng mạng, các nhật ký và các chỉ số hiệu suất quan trọng (KPI), dự đoán các xu thế phát triển mạng và chính thức hóa các chiến lược đáp ứng chính xác thông qua thông tin phân tích thu nhận được và các trạng thái được dự đoán để thực hiện chuyển từ "phòng vệ thụ động" sang "phòng vệ chủ động". Phân tích hư hỏng mạng cấu thành một trong các phần tử thiết yếu của O&M. Vì số lượng các thiết bị tăng và quy mô mạng mở rộng, số lượng các cảnh báo mạng tăng bùng nổ và quan hệ giữa chúng cực kỳ phức tạp. Áp dụng AI để xây dựng một mô hình phân tích phát hiện hư hỏng cho ta mô hình có khả năng phân tích dữ liệu đa chiều cho phép nó phục vụ như là một công cụ thông minh và hiệu quả hơn cho phân tích hư hỏng. Hình dưới đây minh họa luồng công việc của phát hiện hư hỏng dựa trên AI.



Generator: bộ tạo sinh, Discriminator: bộ phân biệt, The trained Generator: bô tạo sinh được đào tạo

Hình 25

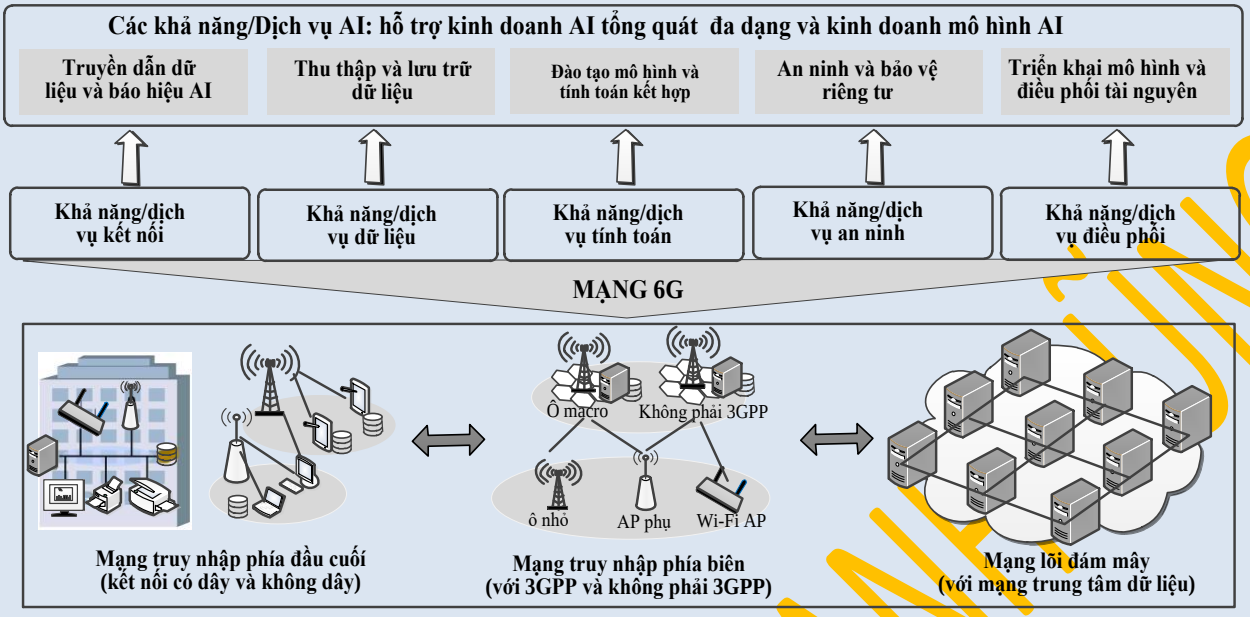
Bằng cách tích hợp công nghệ AI, ta có thể thiết lập phân tích hư hỏng mạng có tệ thống, thông minh và sơ đồ đào tạo ngược để có thể cung cấp các khả năng phát hiện hư hỏng tự động, chính xác, linh hoạt và theo dõi ngược, liên tục tăng cường hiệu quả chẩn đoán và xử lý. Trong khi đó, nó cũng có thể thích ứng các thay đổi trong môi trường mạng và các yêu cầu kinh doanh.

NET4AI (NETWORK FOR AI)

NET4AI đề cập đến trường hợp mạng đóng vai trò hỗ trợ AI/ML. Ở đây, mạng không dây được sử dụng để hỗ trợ và nâng cao hoạt động và hiệu suất của các dịch vụ AI/ML. Thông qua việc tối ưu hóa kiến trúc mạng, công nghệ thu thập và truyền dữ liệu, mục tiêu là cung cấp hỗ trợ mạng mạnh mẽ cho các dịch vụ AI/ML. Hai khía cạnh này sẽ định hướng các nguyên tắc thiết kế của mạng 6G tích hợp AI/ML. Bản chất của NET4AI là cung cấp AI với các khả năng khác nhau để cho phép đào tạo/suy luận AI thời gian thực và hiệu quả hơn, và tăng cường an ninh cũng như bảo vệ tính riêng tư của dữ liệu.

NET4AI sẽ cung cấp hỗ trợ kiến trúc cho thực hiện 6G AI gốc để đạt được tích hợp sâu giữa mạng và AI như minh họa trong hình dưới đây.

Mạng 6G về mặt vật lý bao gồm: CN đám mây (gồm các trung tâm dữ liệu tập trung/tính toán), mạng truy nhập phía biên (gồm các phương pháp kết nối 3GPP và không phải 3GPP), mạng truy nhập phía thiết bị đầu cuối (gồm các phương pháp kết nối có dây và không dây) được mô tả ở phần dưới của hình vẽ. NET4AI thể hiện vai trò ánh xạ và hỗ trợ kết nối truyền thông, tính toán dữ liệu, an ninh và quản lý và điều phối trong AI, trong số sáu khả năng và dịch vụ (nghĩa là: kết nối truyền thông, dữ liệu, tính toán, an ninh, AI và quản lý và điều phối) được trừu tượng hóa từ các dịch vụ chức năng của mạng 6G được chỉ thị bởi các mũi tên trong phần dưới hình vẽ.



Hình 26

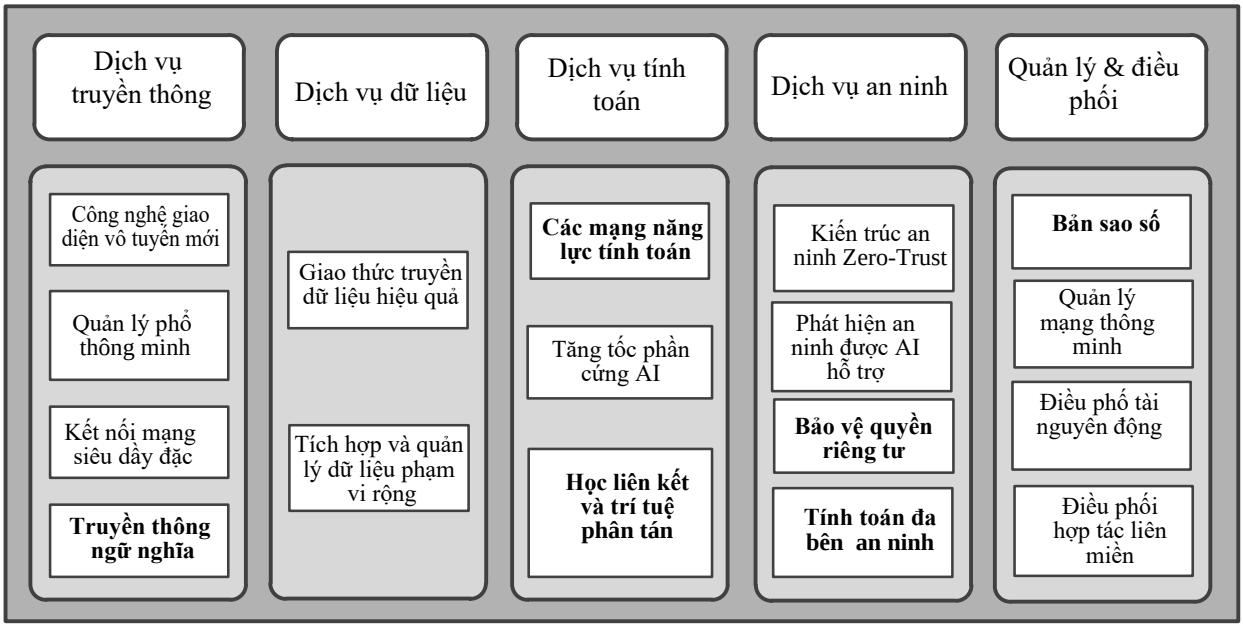
- **Kết nối cho AI:** Kết nối cho AI bao hàm cung cấp hỗ trợ kết nối cho báo hiệu AI và dữ liệu AI. *Báo hiệu AI* được sử dụng để phát đi các bản tin điều khiển liên quan đến AI, chẳng hạn: các bản tin yêu cầu /trả lời dịch vụ AI liên quan đến năng lực tính toán cho phân tích AI, v.v… Nó cũng có thể bao hàm cả thông tin bổ sung cho phân tích AI phối hợp giữa nhiều phần tử. *Dữ liệu AI* được sử dụng để chuyển giao dữ liệu AI đầu vào một giải thuật AI, thường được đề cập như là các vectơ để yêu cầu mô hình/giải thuật AI. Nó cũng được sử dụng để phát đi các mô hình AI, chẳng hạn phát đi một mạng nơ-ron, thường có nghĩa là phát đi các trọng số thông số (chũng có thể được trình bày như là các vectơ) và cấu trúc mạng của nó (có thể được trình bày như là các vô hướng(Scalar)). Nó có thể được sử dụng hơn nữa cho phát đi dữ liệu đào tạo AI. Dữ liệu này thường rất lớn và dành riêng cho giai đoạn đào tạo các mô hình trí tuệ.
- **Dữ liệu cho AI (Data for AI):** Dữ liệu cho AI (Data for AI) nhắm đến hỗ trợ hiệu quả thu thập, truyền dẫn, lưu trữ và chia sẽ dữ liệu từ đầu đến cuối cho phép giải quyết làm sao tạo điều kiện, đẩy nhanh và cung cấp an ninh dữ liệu cho các chức năng AI bên trong hoặc bên ngoài mạng 6G. Phụ thuộc vào phạm vi tiềm năng, hỗ trợ dữ liệu được cung cấp bởi dữ liệu cho AI phải bao gồm năm thành phần: (i) thu thập/phân phối dữ liệu, (ii) An ninh và quyền riêng tư, (iii) Phân tích dữ liệu, . (iv) Xử lý dữ liệu và (v) Lưu trữ dữ liệu.
- **Tính toán cho AI (Computation for AI):** Sự phát triển của 6G sẽ đem lại các tiến bộ đáng kể trong tính tóan, bao gồm cảm biến tính toán (Computational Sensing), điều khiển tính toán (Computational Control) và thực hiện tính toán (Computational Execution). Các khả năng tính toán phải kết hợp với kết nối truyền thông để tối ưu hóa hiệu quả tiêu thụ năng lượng của các dịch vụ mới nổi như AI. Tích hợp này cần thiết phát triển một môi trường thống nhất để đảm

bảo hoạt động liên tục của các dịch vụ tính toánn trên các thiết bị, các nút biên và các đám mây. Môi trường này sẽ cho phép chọn động các nút tính toán biên cho các nhiệm vụ tính toán có xét đến các nhân tố như trễ, băng thông, năng lực tính toán và hiệu quả năng lượng.

- **An ninh cho AI (Security for AI):** Mạng 6G cần thiết lập một hệ thống an ninh tự lực và có tính tin cậy cố hữu tại lõi của hệ thống. Các tiến bộ nhanh chóng của điện toán đám mây, dữ liệu lớn và các công nghệ AI cung cấp hỗ trợ kỹ thuật để xây dựng một hệ thống an ninh mạng 6G. Trong những năm gần đây, an ninh nội sinh tin cậy đã nổi lên như là một cách tiếp cận mới cho 6G, được đặc trưng bởi bốn tính năng: hợp tác, phòng vệ chủ động thông minh, sự đáng tin cậy và bảo vệ quyền riêng tư. Khái niệm an ninh nội sinh có tin cậy cũng được áp dụng cho AI.
- **Điều phối cho AI (Orchestration for AI):** Điều phối cho AI có nghĩa là hỗ trợ hiệu quả cho triển khai, hoạt động và tối ưu hóa các dịch vụ AI trong mạng thông qua điều phối. Nó lập cấu hình tự động và điều phối các tài nguyên trong mạng dựa trên các yêu cầu cụ thể của các dịch vụ AI để đảm bảo các ứng dụng AI nhận được năng lực tính toán, lưu trữ, băng thông mạng cần thiết và các tài nguyên khác. Ngoài ra, khả năng điều phối cho phép giám sát hiệu suất AI thời gian thực bằng cách điều chỉnh động ấn định tài nguyên để đáp ứng các thăng giáng tải dịch vụ, qua đó đảm bảo sự ổn định và hiệu quả trong mạng và cung cấp các đảm bảo mạng vững bền cho các ứng dụng thông minh khác nhau.

CÁC CÔNG NGHỆ HỖ TRỢ CHỦ CHỐT

Chine Mobile đề xuất thiết kế một hệ thống 6G tập trung trên khái niệm “ba lớp và năm mặt phẳng”. “ba lớp” đề cập đến lớp vật lý, lớp mạng và lớp ứng dụng và dịch vụ. “năm mặt phẳng” bao gồm mặt phẳng truyền thông, mặt phẳng dữ liệu , mặt phẳng tính toán, mặt phẳng trí tuệ, mặt phẳng an ninh, mặt phẳng quản lý và điều phối. Năm mặt phẳng khả năng/dịch vụ trong kiến trúc mạng 6G tương lai được trình bày trong dưới đây.



Hình 27

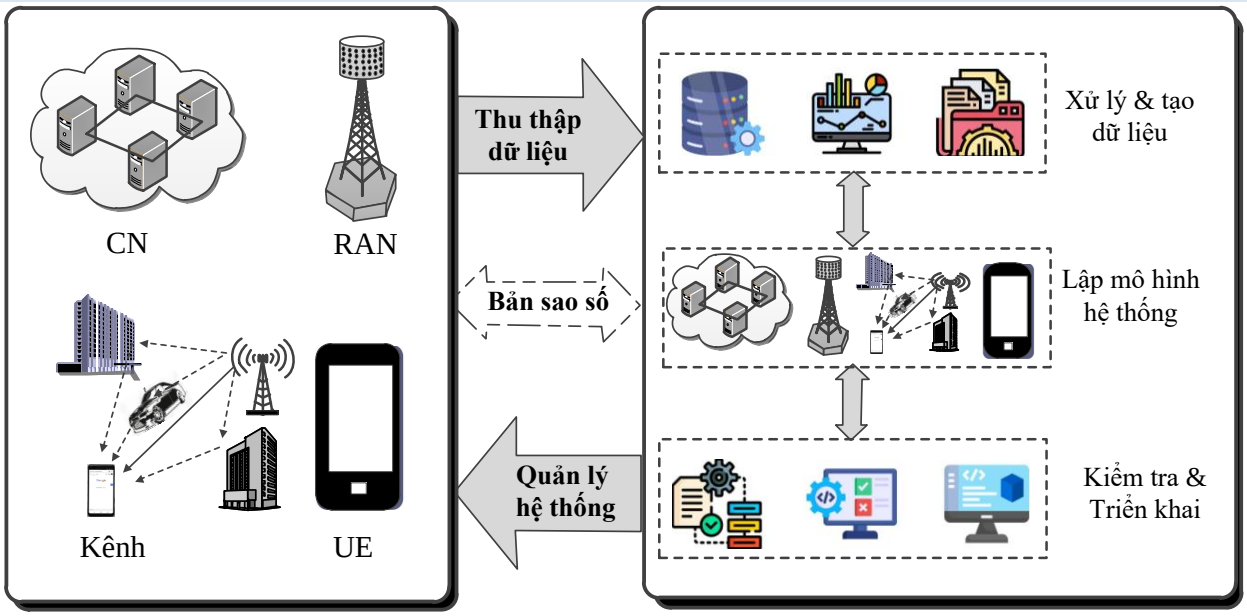
- **Trí tuệ phân tán và FL:** Trí tuệ phân tán và FL (Federated Learning: học liên kết) là các công nghệ chủ chốt có thể giải quyết các thách thức về trễ và an hinh. Trong các mạng phân tán, dữ liệu được lưu trữ trên nhiều nút và các nhiệm vụ tính toán được xử lý hợp tác giữa chúng. Cách tiếp cận lưu trữ và tính toán không tập trung cải thiện sự khả định cỡ mạng, tăng tốc đào tạo và suy luận của các LAM (Large-Scale AI Model: mô hình AI quy mô lớn) và tăng cường hiệu suất toàn mạng. AI phân tán sẵn sàng để tạo ra một hệ sinh thái thông minh mới cho 6G, thúc đẩy một mô hình mạng khả định cỡ, hiệu quả, bảo vệ quyền riêng tư hơn. Trí tuệ phân tán gồm ba thành phần: (i) FL (Federated Learning: học liên kết), (ii) MARL (Multi-Agent Reinforcement Learnig: học tăng cường đa tác nhân) và (iii) SL (Sprit Learning: học chia sẻ), như được minh hoàn trong hình 13.

CÁC BẢN SAO SỐ GỐC AI CỦA MẠNG 6G

Tài liệu của ITU định nghĩa các yêu cầu và các kiến trúc của các mạng DT, nhấn mạnh vai trò cốt lõi của nó là trình bày ảo các mạng vật lý. Các mạng DT cho phép phân tích, hội chẩn, mô phỏng và điều khiển mạng vật lý.

DT cung cấp một bản sao ảo hóa cho các mạng 6G, cho phép giám sát lưu lượng và phân tích tích mạng toàn diện. Sử dụng phản hồi từ mạng ảo hóa, các hệ thống 6G có thể tăng cường an ninh của chúng bằng cách chuẩn bị trước các mối đe dọa. Ngoài ra, DT cho phép nhận dạng yêu cầu và cung cấp dịch vụ bằng cách phân tích dữ liệu truyền thông để phân biệt các mẫu và các quy tắc sử dụng. Hiểu biết dự đoán các nhu cầu truyền thông cho phép các mạng 6G dành trước tài nguyên như phổ, dự

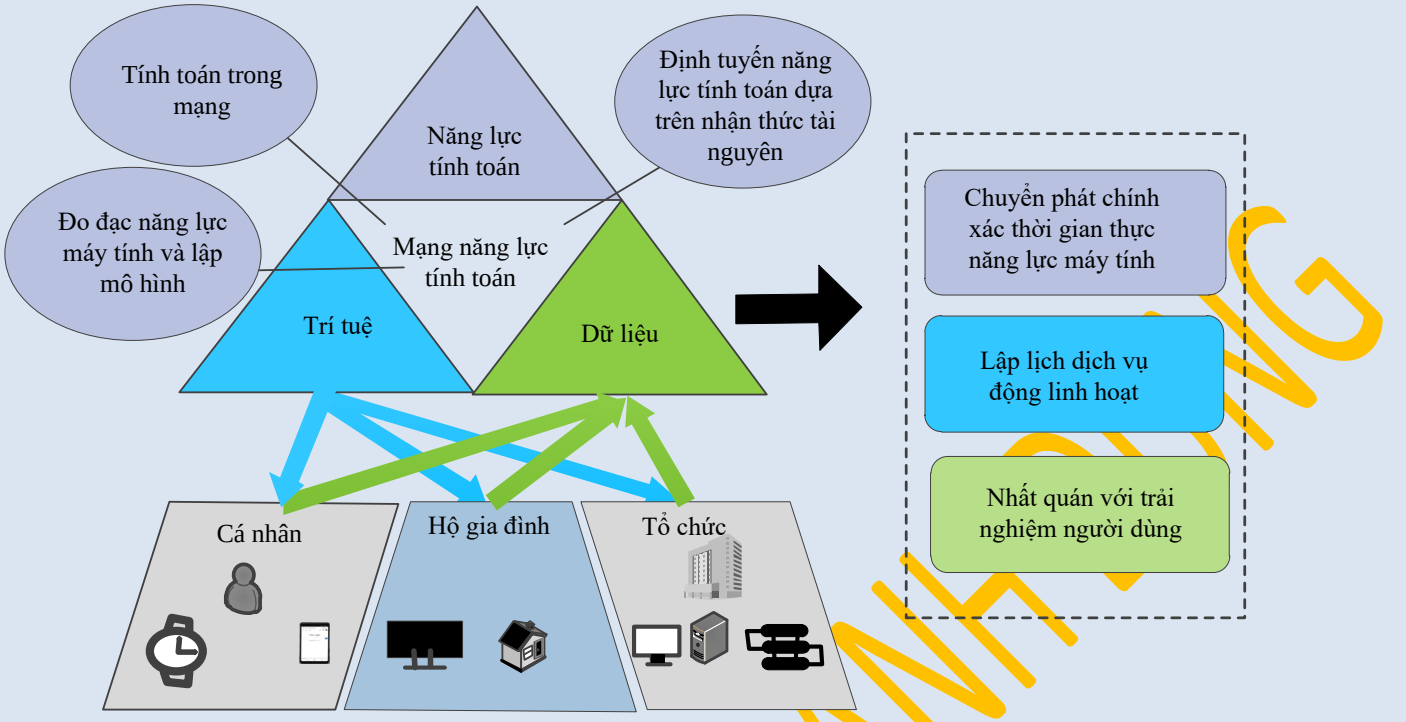
báo trước các nhu cầu tương lai. Ngoài ra tích hợp công nghệ DT cho phép 6G hỗ trợ các dịch vụ đổi mới bao gồm AR/VR và xe tự hành. Với việc giải quyết các vấn đề an ninh, hiệu suất phổ, trí tuệ và tùy chỉnh, DT định nghĩa lại và tăng tốc phát triển các mạng 6G.



Hình 28

MẠNG NĂNG LỰC TÍNH TOÁN

Kiến trúc NET4AI đòi hỏi các khả năng dịch vụ tính toán mạnh mẽ, vì các mạng không dây rất quan tâm đến các nhiệm vụ tính toán liên quan đến tính toán hiệu suất cao như xử lý dữ liệu phạm vi siêu lớn và DL. Mạng năng lực tính toán (Computing Power Network) là công nghệ hỗ trợ quan trọng cho các dịch vụ tính toán. Nó đạt được liên kết nối tính toán mọi nơi thông qua cảm biến và hợp tác tương hỗ giữa mạng và các tài nguyên tính toán, hòa hợp các tài nguyên đám mây, biên, đầu cuối và thực hiện các nhiệm vụ tính toán trong mạng, đặc biệt là các dịch vụ tính toán liên quan đến AI.

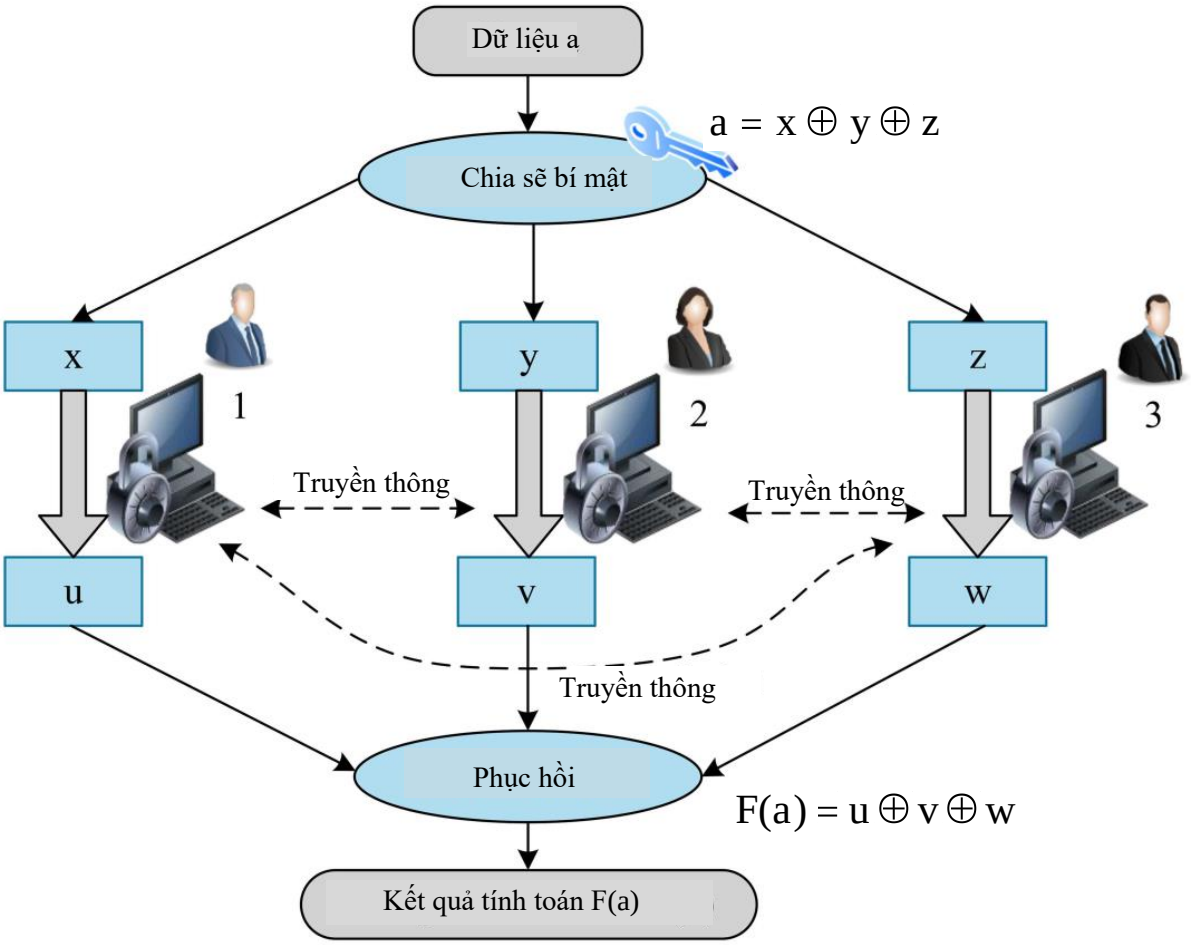


Hình 29

1) **Đo năng lực tính toán và lập mô hình.** Đây là một công nghệ nền tảng để cung cấp các dịch vụ năng lực tính toán. Trong tương lai, các nhà cung cấp dịch vụ năng lực tính toán trong các mạng năng lực tính toán sẽ không bị giới hạn đến các trung tâm dữ liệu đặc thù hay các cụm tính toán. Họ có năng lực tính toán mọi nơi từ đám mây, biên các các thiết bị đầu cuối. Chia sẻ hiệu quả năng lực tính toán này thông qua các kết nối mạng đòi hỏi cảm nhận chính xác dung lượng tính toán của các chip đa tạp, các kiểu kinh doanh phù hợp cho các chip hiện thời và vị trí của chúng trong mạng, cũng như quản lý và giám sát hiệu quả.
2) **Định tuyến năng lực tính toán dựa trên nhận thức.** Trong mạng năng lực tính toán, sau khi đo và lập mô hình các tài nguyên tính toán, thông tin này được mã hóa và được đóng vào các gói của mạng dựa trên thông tin tài nguyên tính toán được chia sẻ, hướng dẫn định tuyến kinh doanh đến các tổ hợp tài nguyên (Resource pool) khác nhau thông qua cộng tác giữa các tổ hợp tài nguyên tính toán để xử lý kinh doanh, điều này cho phép nhận thức của mạng về các tài nguyên tính toán để hướng dẫn định tuyến toàn cục.
3) **Tính toán trong mạng (INC: In-Network Computing).** Sử dụng việc triển khai các công nghệ mạng khả lập trình, INC xử lý các gói trong mạng. Chia xẻ năng lực INC với việc sử dụng các tài nguyên đa tạp mở và khả lập trình sẽ tăng tốc xử lý dữ liệu gần nguồn mà không cần thay đổi chế độ hoạt động kinh doanh gốc, giảm trễ trễ phản hồi của ứng dụng và đơn giản hóa các xử lý triển khai ứng dụng. Các mạng năng lực tính toán hỗ trợ các người dùng điều chỉnh định cỡ tài nguyên theo yêu cầu để thích ứng đến các ứng dụng AI có các kích cỡ và các mức độ phức tạp khác nhau.

TÍNH TOÁN NHIỀU BÊN AN NINH (SERCURE MULTI-PARTY COMPUTATION)

Tính toán an ninh nhiều bên (SMPC: Security Multi-Party Computation) là một giao thức mật mã hóa cho phép nhiều bên kết hợp tính toán một hàm trong khi đó vẫn duy quyền riêng tư dữ liệu đầu vào của họ. Mỗi bên nhận được kết quả tính toán đầu ra đúng mà không biết các đầu vào riêng tư của các bên khác.

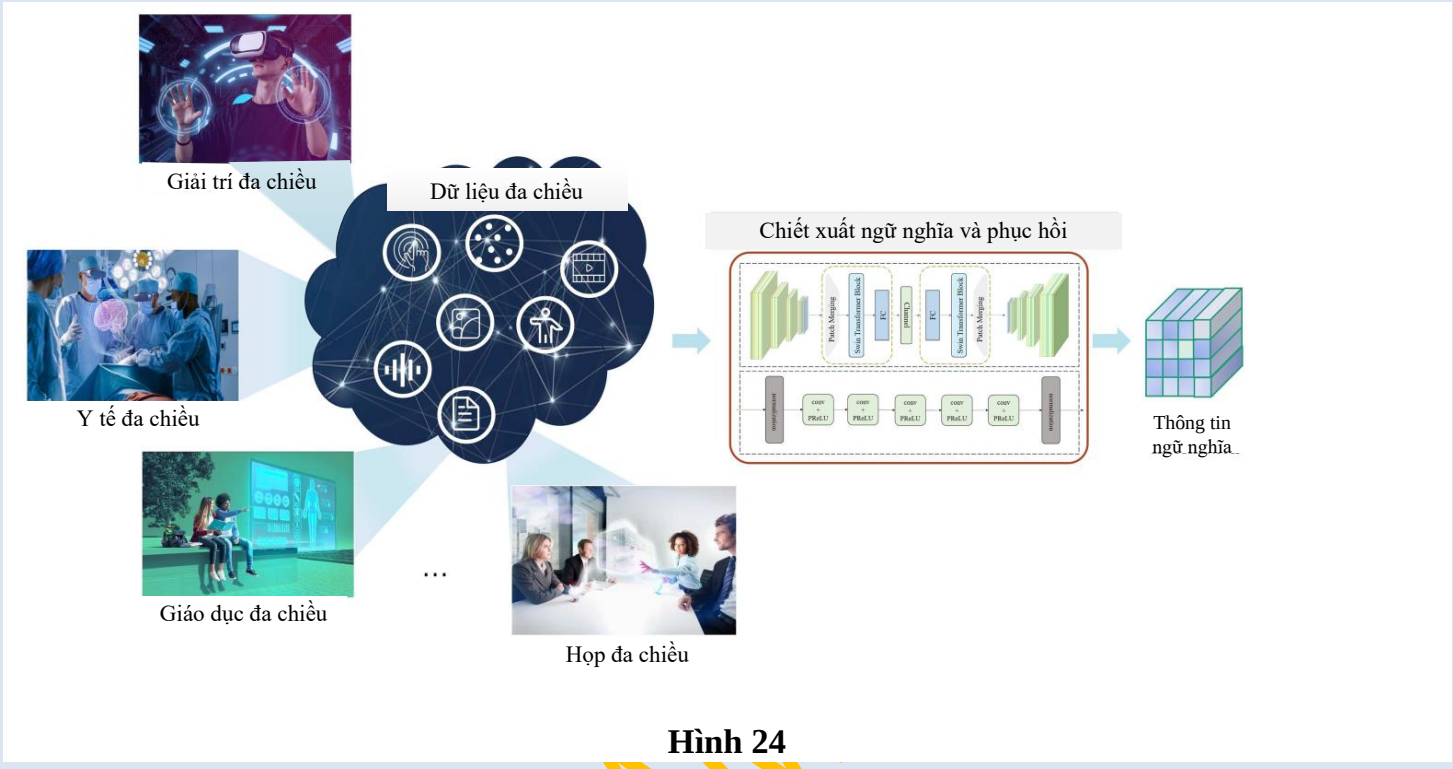


Hình 24

TRUYỀN THÔNG NGỮ NGHĨA (SEMATIC COMMUNICATION)

Truyền thông ngữ nghĩa, thường dựa trên các mạng DNN, được coi là một công nghệ hứa hẹn trong 6G. So với truyền thông cú pháp (Syntactic Communication) tập trung lên truyền dẫn chính xác dữ liệu theo các bit, truyền thông ngữ nghĩa chỉ phát đi các ngữ nghĩa của bản tin. Nó có thể bảo toàn các quan hệ ngữ nghĩa khi phát thông tin, nhờ vậy hỗ trợ tải xuống các nhiệm vụ. Các đặc tính này hỗ trợ đáp

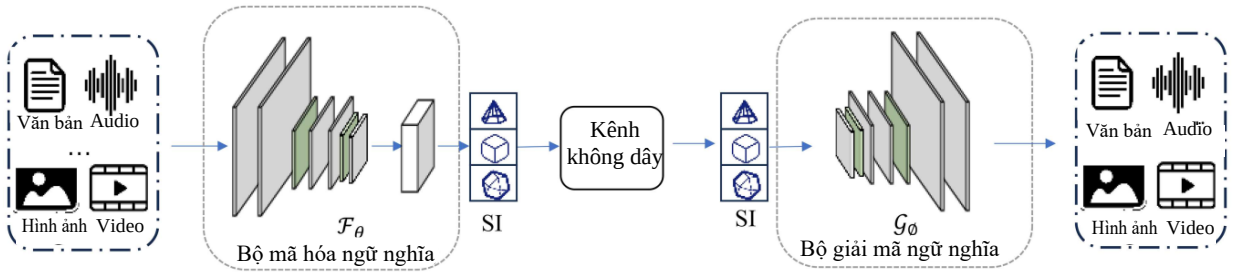
ứng các yêu cầu của các ứng dụng 6G tiêu tốn dữ liệu, chẳng hạn truyền thông đa chiều (Holographic Communication) và XR



Hình 24

Mô hình truyền thông ngữ nghĩa cơ sở

Tại phía máy phát, bộ mã hóa ngữ nghĩa chiết xuất và mã hóa SI (Semantic Information) của dữ liệu nguồn. Quá trình này bao gồm chiết suất SI và nén hoặc loại bỏ thông tin không không quan trọng. Để đạt được điều này dữ liệu thô được mã hóa bởi mạng nơ-rôn ký hiệu là $\mathcal{F}_\theta$, đầu ra duy trì được ý nghĩa quan trọng trong khi xóa bỏ các chi tiết không quan trọng. Các ngữ nghĩa súc tích được phát đi trên kênh một vật lý bị tạp âm, chẳng hạn kênh không dây. Tại máy thu, bộ giải mã ngữ nghĩa giải mã dữ liệu thu. Quá trình này bao gồm "sự hiểu biết" và "suy luận" nội dung ngữ nghĩa được gửi đến từ máy phát. Để hoàn thành điều này, bộ giải mã ngữ nghĩa , được kỹ hiệu là $\mathcal{G}_\theta$, xử lý dữ liệu thu với mục đích ánh xạ nó ngược về dữ liệu nguồn.



SI (Semantic Information): thông tin ngữ nghĩa

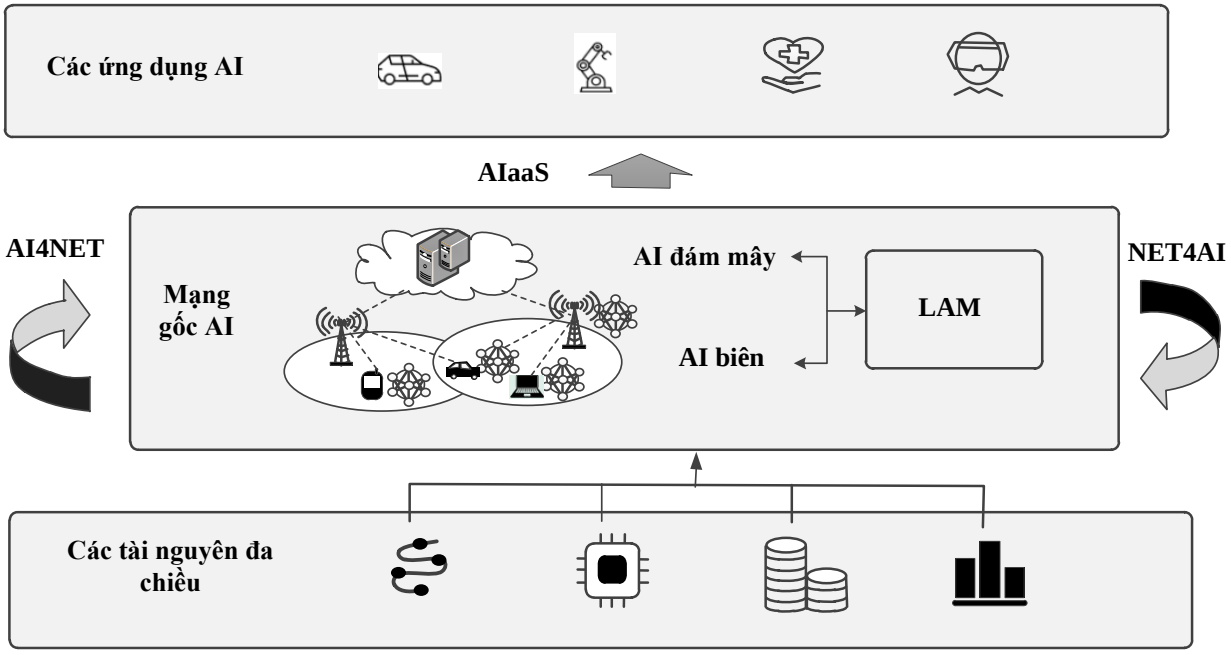
Hình 24

CÁC MÔ HÌNH NGÔN NGỮ LỚN (LLM: LARGE LANGUAGE MODEL)

Các mô hình ngôn ngữ lớn (LLM: Large Language Model) đã thể hiện một bước đột phá đáng kể trong NLP (Natural Language Processing: xử lý ngôn ngữ tự nhiên) và có thể đóng góp tiềm năng cho phát triển 6G AI. So với các mô hình thông số nhỏ hơn truyền thống, các LLM thể hiện hiểu biết ngữ cảnh (Contextual Understanding) mạnh, tạo văn bản kết hợp (Coherent Text Generation), lý luận logic (Logical Reasoning) và các khả năng tổng quát hóa. Các mô hình lớn mục đích chung (General Purpose Large Model) hiện có như GPT mã nguồn đóng (Closed Source Generative Pretrained Transformer: bộ chuyển đổi tiền huấn luyện tạo sinh được phát triển bởi OpenAI và mô hình Llama mã nguồn mở (Open source Llama Model) được phát triển bởi Meta (trước đây là Google) có thể xử lý các dải rộng các vấn đề chuyên biệt miền/ lĩnh vực. Mới đây, DeepSeek LLM mã nguồn mở được phát triển bởi đội ngũ DeepSeek tại Trung Quốc đã trở thành người thay đổi tiềm năng, chứng minh khả năng đào tạo các mô hình AI quy mô lớn (LAM: Large Scale AI Model) với chi phí thấp không ngờ. Các LLM có thể tiềm năng đóng vai trò then chốt trong các hệ thống truyền thông, chẳng hạn: trong quá trình xử lý nhiệm vụ và các dịch vụ thông minh.

Sự nổi lên của các LLM có thể đem đến một mô hình tích hợp 6G và AI mới. KIến trúc AI bản năng trong 6G có thể cung cấp các dịch vụ liên kết, tính toán, phân tách mô hình (chia nhỏ mô hình) và phân phối mô hình cho đào tạo và suy luận của các LLM. Các LLM có thể tăng thêm sức mạnh các lĩnh vực khác nhau của các mạng 6G (chẳng hạn: giao diện vô tuyến, phía mạng, an ninh mạng, O&M mạng, v.v..). Trong 6G, các mô hình lớn đặc thù hóa tùy biến có thể được triển khai trên các lớp mạng khác nhau. Các mô hình đặc thù hóa này có thể cộng tác xử lý các vấn đề mạng thông qua phân tách và kết hợp nhiệm vụ (Task Decomposition and Composition), cộng tác liên lớp (Cross-Layer) và cộng tác đám mây biên (Edge Cloud), để tăng cường mức trí tuệ của mạng.

Ngoài việc sử dụng các mô hình AI tùy chỉnh được thiết kế riêng cho các chức năng và các nhiệm vụ đặc thù, các mô hình lớn thể hiện hoạt động tổng quát nổi trội trên nhiều nhiệm vụ và các tính năng đột phá trên các nhiệm vụ phức tạp mà các mô hình nhỏ không thể đạt được. Các chức năng chính của các mô hình lớn ngày càng phù hợp với các đặc tính đa kịch bản và đa nhiệm vụ của 6G, giảm bớt hiệu quả khối lượng công việc trong nhận thức mô hình và cài đặt số đo hiệu suất trong các mạng 6G cho thấy các triển vọng ứng dụng rộng lớn. Một dạng ttích hợp 6G với AI có thể có được minh họa trong hình dưới đây. Tận dụng dữ liệu từ đa dạng các mạng không đồng nhất trong 6G, các mô hình lớn này có thể tổng hợp các kịch bản và các dịch vụ 6G khác nhau, đặt nền tảng trọng yếu cho tăng cường AI trong các mạng 6G. Chẳng hạn đề xuất 6GGANA trong sách trắng 6GNETGPT đã xây dựng một LLM giống như ChatGPT, cung cấp một mô hình mới cho các hoạt động và quản lý nâng cao các mạng 6G.

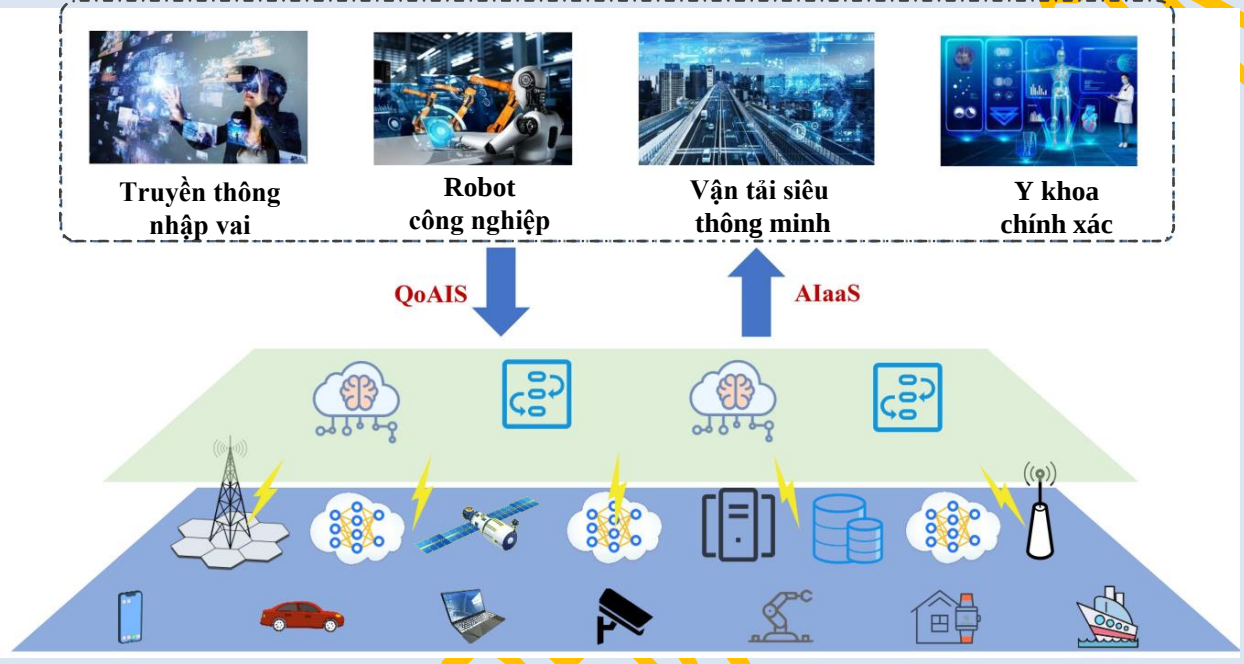


LAM (Large Scale AI Model): mô hình AI quy mô lớn.

Hình 25

AI NHƯ LÀ MỘT DỊCH VỤ (AIaaS)

Để hỗ trợ hiệu quả hơn "trí tuệ gốc" (Native Intellegence) và đạt được "trí tuệ nhân tạo toàn năng" (Artificial Universal Intellegence) mọi nơi, các mạng 6G sẽ xử lý trí tuệ nhân tạo như là một dịch vụ (AIaaS: AI as a Service), dẫn đến xuất hiện khái niệm 6G AIaaS như được trình bày trong hình 26.



Hình 26

6G AIaaS sử dụng các tài nguyên và các chức năng trong mạng (bao gồm 6G CN, các mạng truy nhập không dây và các đầu cuối), chẳng hạn kết nối, tính toán, dữ liệu và các mô hình. 6G AIaaS nhắm đến xây dựng một hệ sinh thái dịch vụ AI phân tán, hiệu suất, hiệu suất năng lượng và an ninh, bao gồm đào tạo mô hình, suy luận AI, triển khai, và các chức năng khác trong một môi trưởng mở carbon thấp . Nó không chỉ định nghĩa lại hệ sinh thái của đám mây biên mà còn xây dựng các mô hình kinh doanh mới thông qua các mạng di động 6G, cho phép chuyển đổi từ các mạng định hướng kết nối quá khứ sang các mạng định hướng dịch vụ, cuối cùng đạt được trí tuệ mọi nơi. Các kịch bản điển hình của AIaaS bao gồm (nhưng không phải tất cả) các thành phố thông minh, nông nghiệp thông minh giáo dục toàn cầu và công nghiệp thông minh. Các ứng dụng điển hình bao gồm các dịch vụ taxi không người lái, kiểm tra lưới điện thông minh, giám sát y tế tại nhà và các lớp học ảo.

QoAIS (Quality of AI Service: Chất lượng dịch vụ AI) đề cập đến khung các số đo toàn diện để đánh giá các dịch vụ AI trong mạng. Nó là một khung đa chiều bao gồm các chiều (hay các đại lượng) then chốt như hiệu suất, kết nối, tính toán, dữ liệu , an ninh và điều phối. Bảng 2 cho thấy các chỉ báo cmà một hệ thống QoAIS cần có.

Chiều chỉ báo	Các chỉ báo của QoAIS
Hiệu suất	Biên các số đo hiệu suất, thời gian đào tạo, tổng quát hóa, khả năng tái sử dụng, tính bền vững, khả năng tương tác, tính nhất quán giữa hàm mất mát và các đối tượng tối ưu hóa, sự công bằng, v.v..
Kết nối	Băng thông và jitter, trễ liên kết và jitter, tỷ số bit lỗi và jitte, v.v…
Dữ liệu	Sự dư thừa tính năng, sự hoàn thiện, sự chính xác dữ liệu, chuẩn bị dữ liệu tốn thời gian, cân bằng không gian mẫu, động lực học phân phối dữ liệu, v.v…
Tính toán	Sự chính xác tính toán, khoảng thời gian, hiệu quả, v.v….
An ninh	Bảo mật thông tin, các mức độ quyền riêng tư dữ liệu/giải thuật,
Điều phối	Hoàn toàn tự quản, có khả năng điều khiển một phần bởi con người, hoàn toàn có khả năng điều khiển bởi con người, v.v…

Các kịch bản điển hình của AIaaS

Truyền thông nhập vai (Immersive Communication). Trong truyền thông tương lai, các công nghệ truyền thông nhập vai như VR, AR và các báo cáo đo đạc sẽ tăng cường các trải nghiệm tương tác. AaaS sẽ hỗ trợ các công nghệ này với phiên dịch ngôn ngữ thời gian thực, phân tích cảm xúc và các đề xuất được cả nhân hóa trong các môi trường ảo.

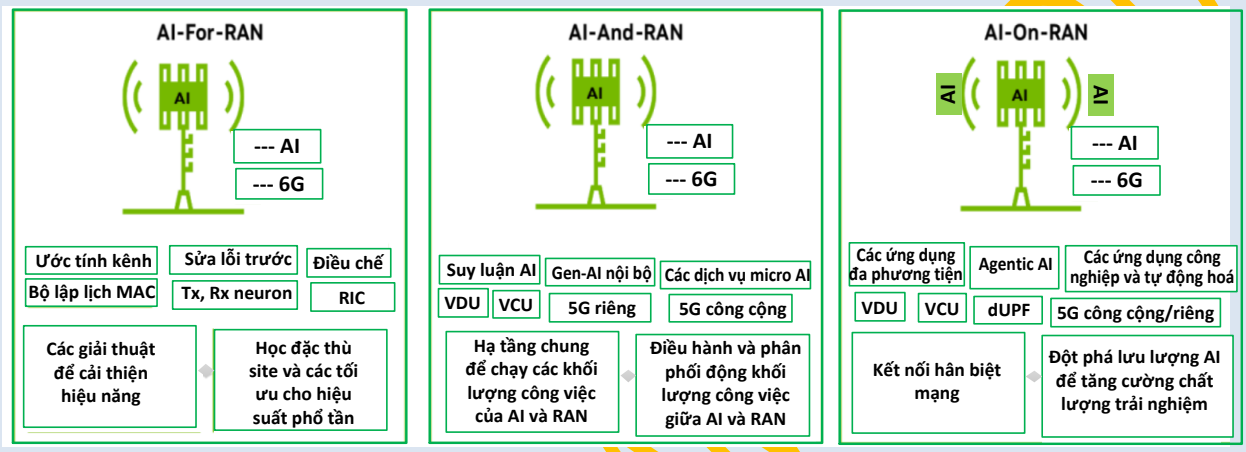
Các robot công nghiệp thông minh (Intellegence Indusstrial Robots). 6GaaaS cung cấp các dịch vụ đào tạo AI cho các robot, chẳng hạn thu thập dữ liệu thông qua các robot, áp dụng đào tạo mô hình cho học đa tác nhân và phân tán các mô hình đã được đào tạo cho các robot.

Y khoa chính xác (Precision Medical). Y tế thông minh sẽ bao gồm các khía cạnh khác nhau của phòng ngừa, dự báo, hội chẩn, suy luận, giám sát bệnh tật, phẫu thuật lâm sàng, chăm sóc bệnh nhân, phát triển thuốc và và vắc -xin trong suốt vòng đời đời. Hệ thống 6G mới sẽ hỗ trợ tốt hơn truyền dẫn khối lượng lớn thông tin và đồng bộ cần thiết cho y tế thông minh và trực tiếp cho phép xử lý và đưa ra quyết định về thông tin

AI-RAN MẠNG TRUY NHẬP VÔ TUYẾN AI)

AI-RAN Alliance (Liên minh AI-RAN) vạch ra ba lĩnh vực đặc thù để tích hợp AI vào RAN:

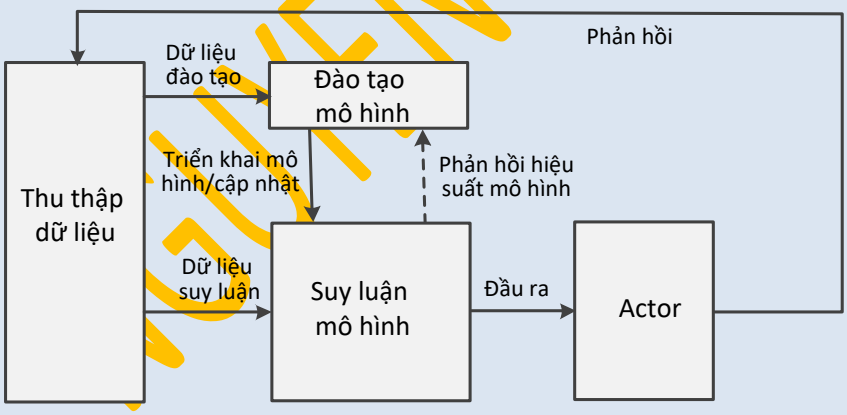
- AI and RAN (AI và RAN), (còn được đề cập là AI with RAN): sử dụng một hạ tầng chung chia sẻ để chạy cả AI và RAN nhằm cực đại hóa sử dụng, giảm tổng chi phí quyền sở hữu (TCO: Total Cost of Onership) và tạo ra các cơ hội lợi nhuận mới bởi vận hành AI.
- AI for RAN (AI for RAN): Nâng cao các khả năng của RAN thông qua nhúng các các giải thuật, các mô hình AI/ML và các mạng tế bào thần kinh vào lớp xử lý tín hiệu vô tuyến để cải thiện hiệu suất phổ tần, vùng phủ vô tuyến, dung lượng và hiệu suất.
- AI on RAN (AI trên RAN): cho phép các dịch vụ AI trên RAN tại biên mạng tăng hiệu năng khai thác và cung cấp các dịch vụ mới cho các người dùng di động. Đến lượt mình RAN từ một trung tâm chi phí trở thành một nguồn lợi nhuận.



MAC (Medium Access Control); điều khiển truy nhập môi trường; RIC (RAN Intellegence Controller): bộ điều khiển trí tuệ RAN; VCU (Virtual Central Uni)t: đơn vị trung tâm ảo; VDU (Virtual Distributed Unit): đơn vị phân tán ảo; dUPF (Differentiated User Plane Function): chức năng mặt phẳng người dùng phân biệt.

Hình 27

Khung cấu trúc cho trí tuệ RAN



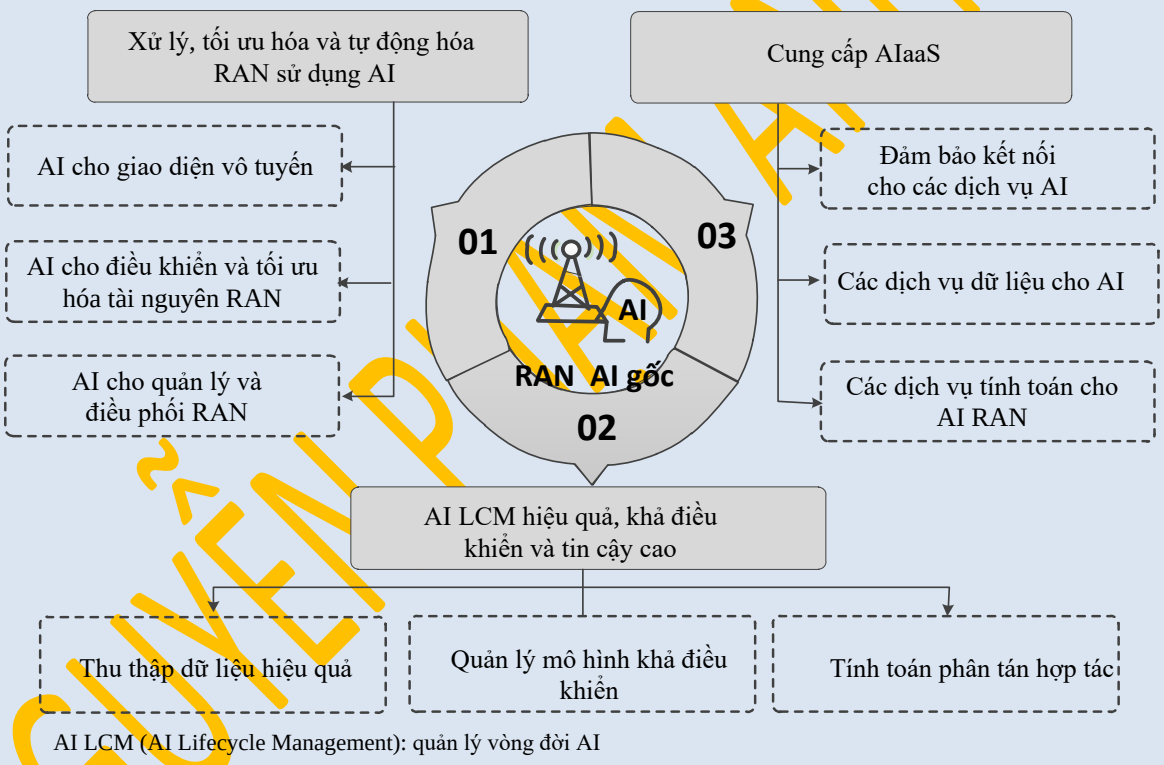
Actor: tác nhân tương tác

Hình 28

CÁC YÊU CẦU VÀ CÁ YẾU TỐ CỦA AI-RAN

- **Hạ tầng tính toán tăng tốc.**
- **Thiết kế điện toán đám mây gốc (Native Cloud) được, định nghĩa bằng phần mềm.**
- **Điều phối kết hợp các tài nguyên truyền thông và tính toán.**
- **Hỗ trợ AI gốc (Native AI).** Mô hình AI RAN mới nổi phải hỗ trợ AI gốc bằng cách nhúng trực tiếp các khả năng AI vào hạ tầng RAN để cho phép đưa ra quyết định thông minh thời gian thực, tự động hóa và tối ưu hóa trên cả các lớp mạng và các lớp điện toán.

BA KHẢ NĂNG QUAN TRỌNG CỦA RAN AI GỐC (AI NATIVE RAN)



Hình 29

XỬ LÝ TỐI ƯU HÓA VÀ TỰ ĐỘNG HÓA SỬ DỤNG AI

Tổng thể, các khả năng này đặt AI vào vị trí như là một yếu tố mạnh mẽ trên lớp vật lý (L1), lớp liên kết (L2) và lớp mạng (L3). Bảng 13.2 cho thấy các trường hợp sử dụng của AI cho giao diện vô tuyến (L1) và AI cho điều khiển tà tối ưu hóa tài nguyên (L2/L3) từ góc nhìn hệ thống.

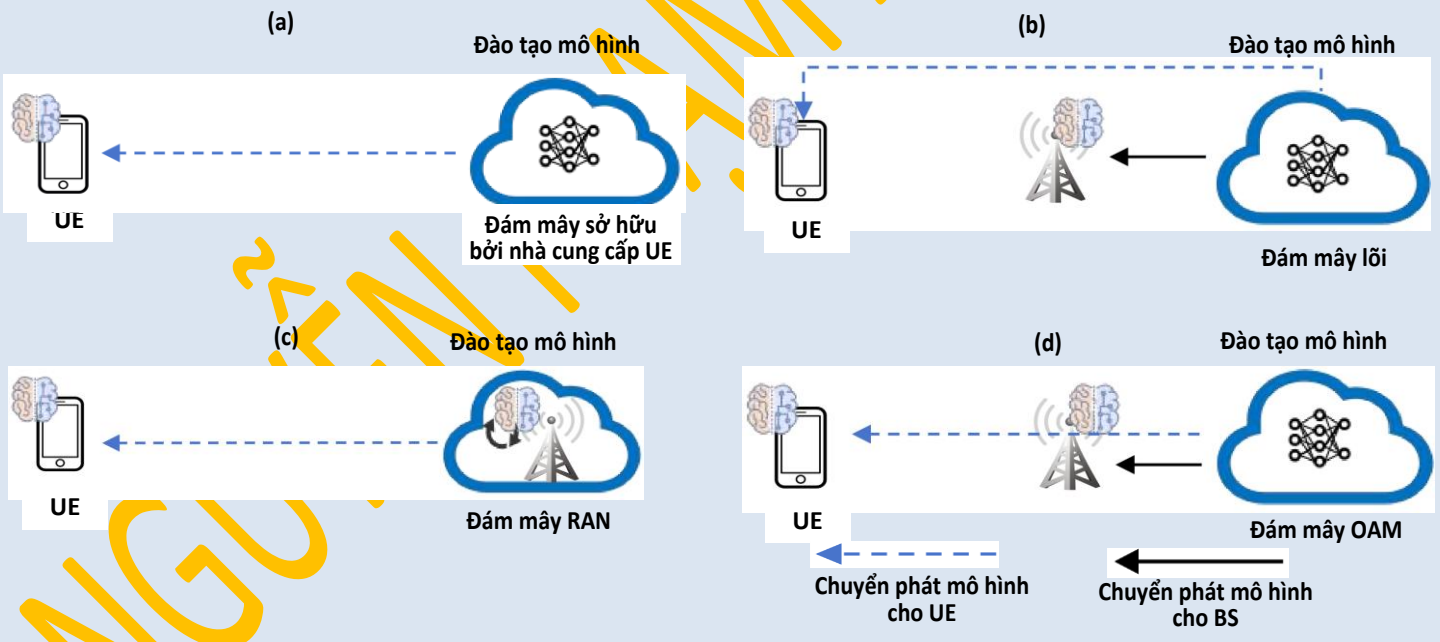
Các kịch bản	Trường hợp sử dụng	Các thách thức chính cho các giải pháp dựa trên không AI	Vị trí của suy luận AI	Các phương pháp AI tiềm năng & quy mô thông số	Trạng thái nghiên cứu
AI cho điều khiển và tối ưu hóa tài nguyên RAN	Cân bằng tải	Tối ưu hóa thụ động giảm cấp hiệu năng; đa chiều, phức tạp và thường gặp các vấn đề tối ưu hóa không lồi; khó tìm được giải pháp tối ưu thời gian thực	Phía BS	DNN (1M) LSTM (10M) DRL (<10M)	5G-A 3GPP Release 18, O-RAN
	Quản lý di động		Phía UE, phía BS	Rừng cây Tăng cường gradient LSTM (<1M)	5G-A 3GPP Release 18/19, O-RAN
	Tiết kiệm năng lượng mạng		Phía BS	DNN (1M) LSTM (10M) DRL (<10M)	5G-A 3GPP Release 18, thử nghiệm hiện trường trong mạng thương mại 5G-A, O-RAN
	Cắt lát mạng			LSTM (10M) GRU (<10M) Transformer[1] (10M-1B) Dubling Duel DQN[2] (1M)	5G-A 3GPP Release 19, O-RAN
	Tối ưu hóa dung lượng và vùng phủ			DRL(<10M) GNN (10M) LSTM (<10M) Transformer (10-1 B)	
	Ấn định tài nguyên và lập biểu đa người dùng			DNN (1M) DRL (10M)	Trường hợp sử dụng ứng viên 6G , O-RAN
	Tối ưu hóa QoS/QoE và đảm bảo thỏa thuận mức dịch vụ (SLA)			k-means [3](<1M) DNN (10M) CNN (1M) DRL (10M)	Thử nghiệm hiện trường trong mạng thương mại 5G-A, trường hợp sử dụng ứng viên 6G, O-RAN
AI cho giao diện vô tuyến	Định vị	Chính xác bị giới hạn	Phía UE, phía BS Phía LMF[5]	DNN (<30M) CNN (<30M)	5G-A 3GPP Release 18/19
	Quản lý búp sóng	Chính xác bị giới hạn trong kịch bản tốc độ trung bình và cao	Phía UE, phía BS	CNN (<10M) LSTM (<10M) Transformers (<10M)	
	Nén CSI	Khó lập mô hình/lập công thức về mặt lý thuyết	Mô hình hai phía	CNN (<10M) LSTM (<10M) Transformers (<10M)	
	Dự đoán CSI	Chính xác bị giới hạn	Phía UE, phía BS	MLP-Mixer (<10M) 2D-FCN (1M) CNN (<1M)	
	Máy phát/ thu dựa	Khó lập mô		ESN (10M)	

	trên AI	hình/lập công thức về mặt lý thuyết		Transformers (<10M)	dụng ứng cử 6G
	Bù méo phi tuyến tính khuếch đại công suất	Khó lập mô hình/lập công thức về mặt lý thuyết		ESN (<10M)	
	Dự đoán nhiễu và xử lý	Chính xác bị hạn chế và lỗi thời		LSTM (<10M) Transformers (<10M)	
	Giảm chi phí gián tiếp RS[4] (gồm cả hoa tiêu được chồng lên)	Chi phí gián tiếp cho tính hiệu tham chiếu cao	phía BS	CNN (<10M) LSTM (<10M) Transmers (<10M)	

Transformer[1] : bộ chuyển đổi, Double Dueling DQN[2] (Double Dueling Deep Q-Network: mạng Q sâu đối đầu nhân đôi), k-means[3]: thuật toán phân cụm k-mean, RS[4] (Reference Signal): tính hiệu tham chiếu, LMF[5] (Location Management Function): chức năng quản lý vị trí.

Các mô hình đào tạo, kiểm tra và thẩm định

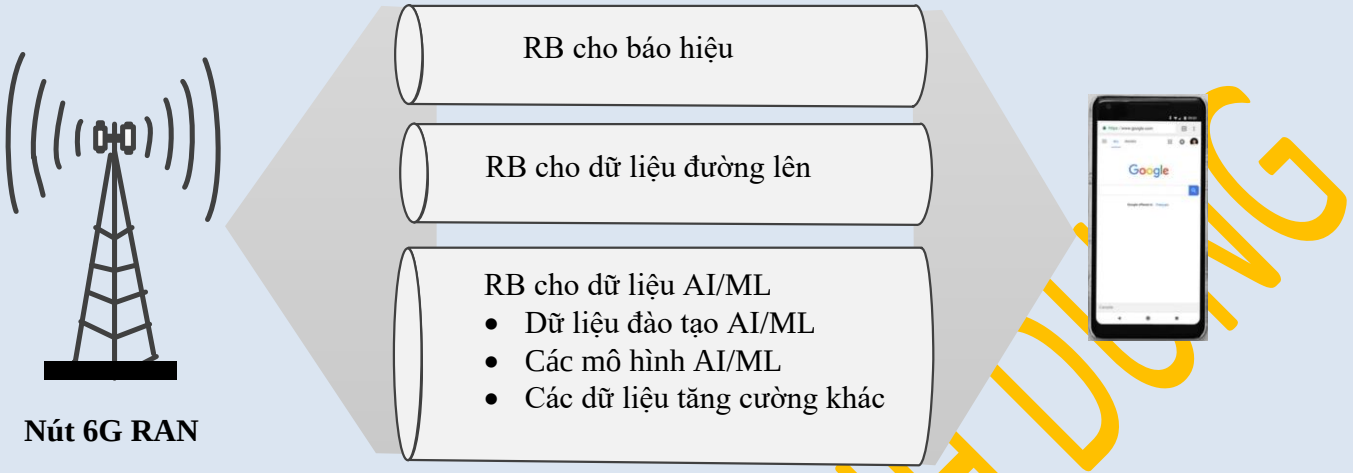
Các mô hình được áp dụng cho các trường hợp khác nhau trước hết phải trải qua đào tạo, kiểm tra và thẩm định để đảm bảo hiệu suất xử lý, tối ưu hóa và tự động hóa RAN dựa trên AI trong các hệ thống thực tế.



Hình 30

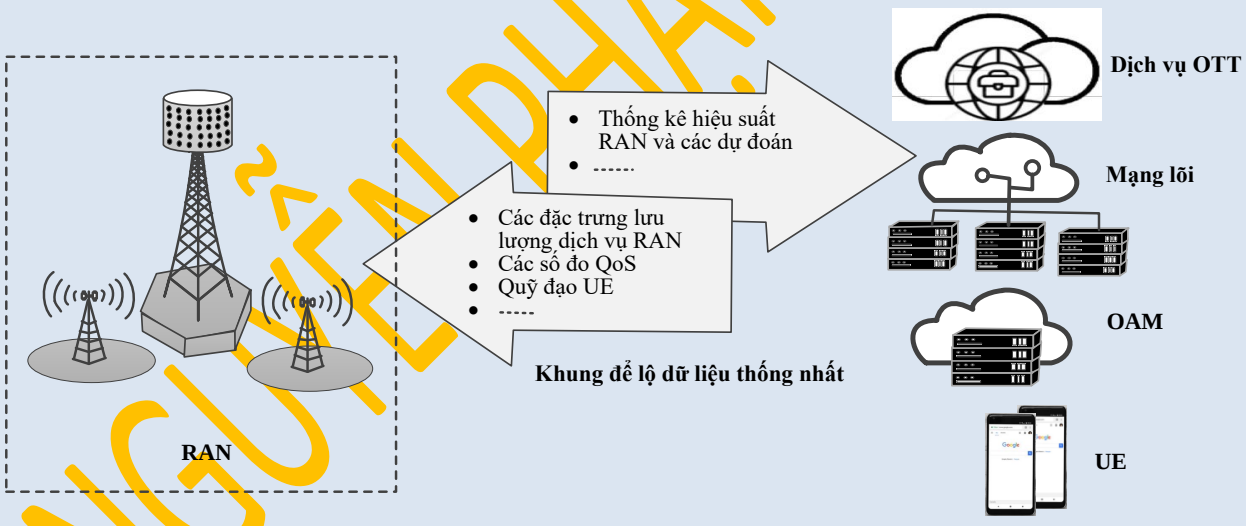
Kênh mang vô tuyến dành riêng (Dedicated Radio Bearer. Về phần tải tin, kênh mang vôt tuyến có thể hỗ trợ nhiều kiểu dữ liệu liên quan đến AI/ML bao gồm (nhưng không phải tất cả) dữ liệu đào tạo AI/ML, các mô hình AI/ML và các kiểu dữ liệu tăng cường khác có thể hỗ trợ các luồng công

việc AI/ML (chẳng hạn: các số đo hiệu suất của các mô hình AI/ML). Các kênh mang vô tuyến mới sẽ đảm bảo cả khả năng linh hoạt lẫn hiệu suất trong các luồng công việc AI/ML trong 6G RAN.



Hình 31

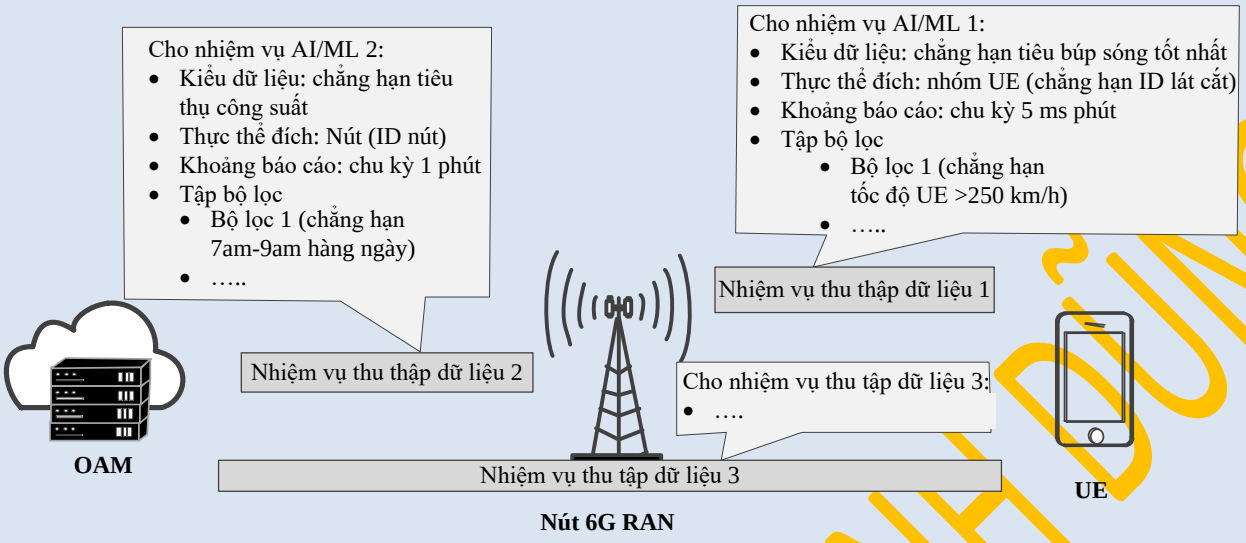
Thu thập dữ liệu liên miền. RAN AI gốc của 6G được kỳ vọng cho phép AI/ML liên miền (Cross Domain AI/ML) để đảm bảo dịch vụ từ đầu đến cuối, dẫn đến thu thập dữ liệu liên miền (Cross Domain).



Hình 32

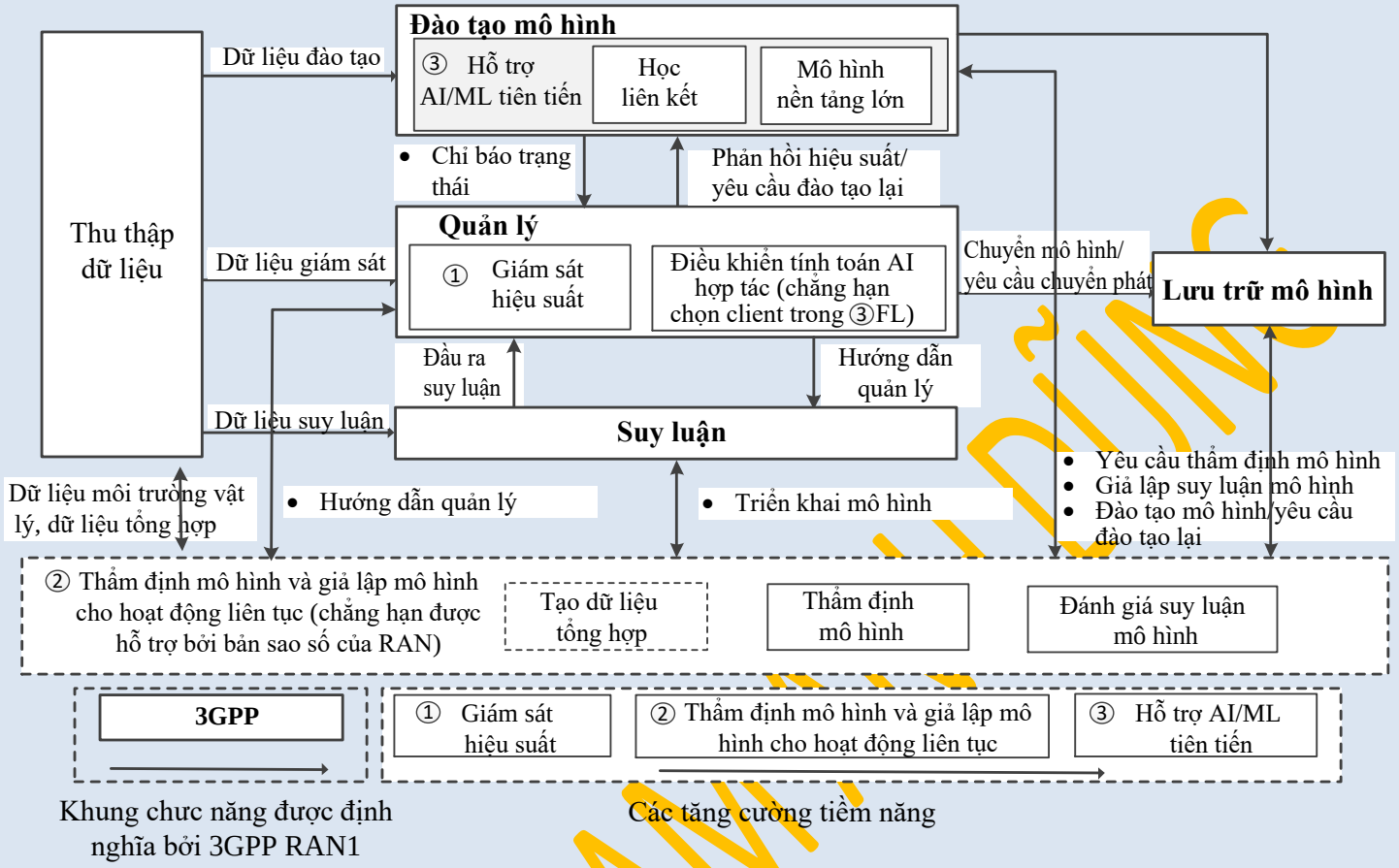
Cơ chế thu thập dữ liệu RAN khả tùy thỉnh dựa trên nhiệm vụ. Vượt trội hơn MDT, các đo đạc hiệu suất (PM: Performance Meassurements) và các báo cáo đo đạc (Meassurement Reports), một cơ

chế thu thập dữ liệu mới là yếu tố thiết yếu để thực hiện các yêu cầu thu thập dữ liệu đa dạng trong 6G RAN.



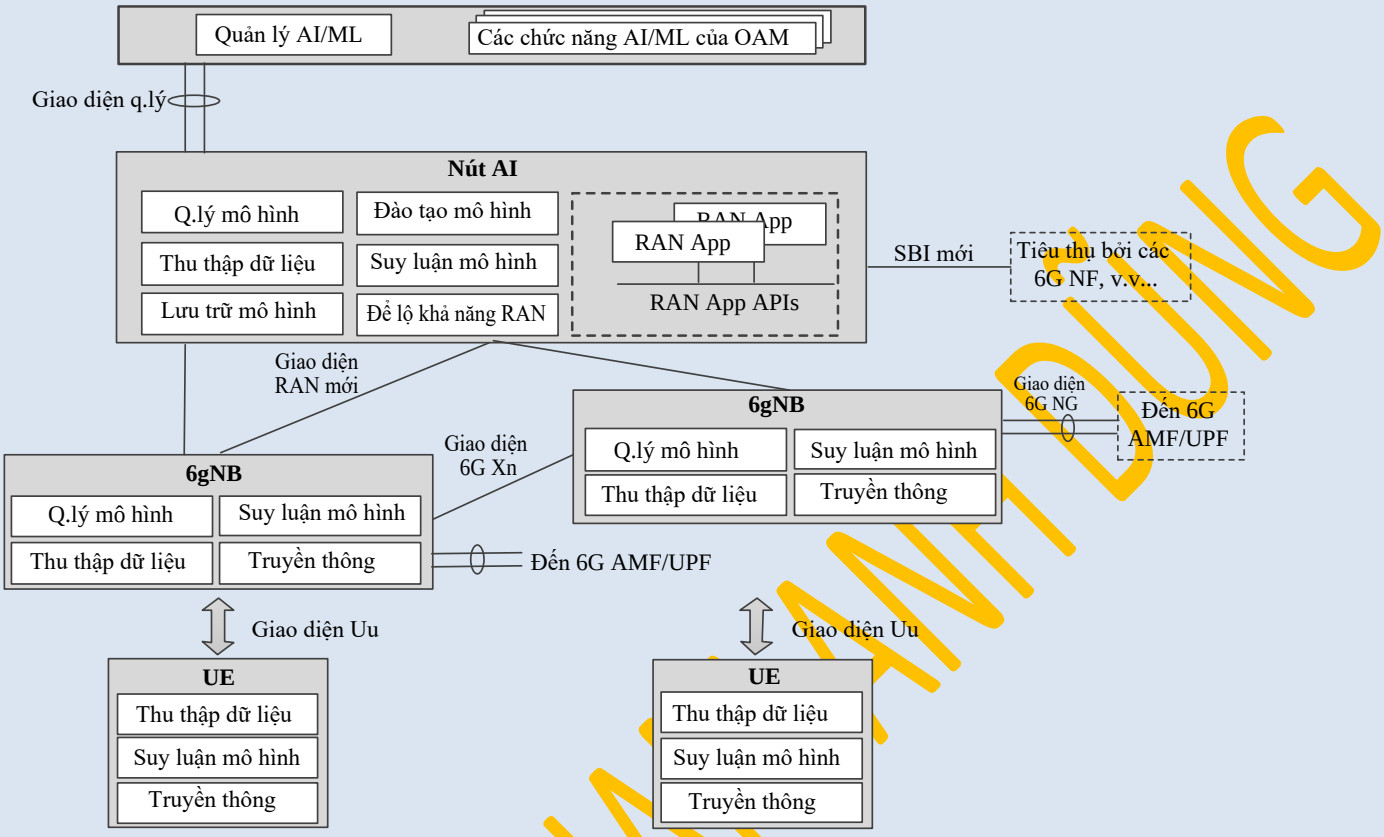
Hình 33

Quản lý mô hình AI/ML khả điều khiển và tin cậy cao hơn. Các tiến bộ cho quản lý mô hình được kỳ vọng để để tăng cường quản lý 6G RAN AI gốc.



Hình 33

KIẾN TRÚC THAM KHẢO CHO RAN AI GỐC



RAN App (RAN Application): ứng dụng của RAN, RAN App API (RAN Application Application Programming Interfac): Giao diện lập trình ứng dụng của ứng dụng RAN, SBI (Service Based Interface): giao diện dựa trên dịch vụ, 6G NG (6G Next Generation) thế hệ sau 6G, 6G AMF/UPF (6G Access Management Function/ User Plane Function: chức năng quản lý truy nhập/ chức năng mặt phẳng người dùng của 6G, 6gNB: trạm gốc 6G, UE (User Equipment): thiết bị người dùng, q.lý: quản lý, Giao diện Uu: giao diện vô tuyến., Giao diện Xn: giao diện giữa các 6gNB.