CHƯƠNG I

1.1 **Tại sao lại có Machine Learning? Mục đích của Machine Learning**?

* **Tại sao lại có Machine Learning?**
* **Sự phát triển của dữ liệu lớn (Big Data):**  
  Thế giới hiện nay đang chứng kiến sự bùng nổ dữ liệu. Từ dữ liệu do con người tạo ra như văn bản, hình ảnh, video đến dữ liệu từ các thiết bị IoT, cảm biến hay mạng xã hội. Khối lượng dữ liệu này lớn đến mức con người không thể tự mình phân tích và khai thác giá trị một cách hiệu quả.  
  Ví dụ: Hệ thống thương mại điện tử như Amazon thu thập dữ liệu về hành vi mua sắm của hàng triệu khách hàng mỗi ngày. Để phân tích thói quen mua sắm và đưa ra khuyến nghị sản phẩm phù hợp, cần có một hệ thống tự động và thông minh.
* **Hạn chế của lập trình truyền thống:**  
  Các chương trình máy tính truyền thống hoạt động dựa trên quy tắc được lập trình sẵn. Nhưng nhiều vấn đề không thể mô tả dễ dàng qua các quy tắc này, ví dụ như nhận diện hình ảnh, giọng nói, hay phân tích ngữ nghĩa trong văn bản.  
  Machine Learning xuất hiện để giải quyết vấn đề này, khi máy tính có thể **học** từ dữ liệu mà không cần được lập trình tường minh.
* **Sự tiến bộ của công nghệ tính toán:**  
  Với sự phát triển của phần cứng (GPU, TPU) và khả năng lưu trữ dữ liệu, Machine Learning đã trở nên khả thi và hiệu quả hơn. Việc xử lý hàng triệu hoặc thậm chí hàng tỷ dữ liệu trong thời gian ngắn không còn là trở ngại.
* **Cần tối ưu hóa và tự động hóa:**  
  Các doanh nghiệp và tổ chức cần giải pháp để tối ưu hóa chi phí, tự động hóa quy trình và cải thiện hiệu suất. Machine Learning trở thành công cụ mạnh mẽ để giải quyết các bài toán này, từ tự động hóa dịch vụ khách hàng bằng chatbot đến dự đoán nhu cầu thị trường.
* **Mục đích của Machine Learning**
* **Tự động hóa quyết định và dự đoán:**  
  Machine Learning giúp đưa ra dự đoán hoặc quyết định chính xác hơn dựa trên dữ liệu lịch sử.
  + **Ứng dụng thực tế:**
    - Ngân hàng sử dụng ML để phát hiện giao dịch gian lận.
    - Hệ thống dự báo thời tiết hoặc dự đoán giá cổ phiếu.
* **Tăng cường trí tuệ nhân tạo (AI):**  
  Machine Learning là một phần cốt lõi trong trí tuệ nhân tạo. Nhờ đó, các hệ thống AI có khả năng học hỏi và cải thiện hiệu suất qua thời gian.
  + Ví dụ: Các hệ thống trợ lý ảo như Siri, Google Assistant sử dụng ML để hiểu ngôn ngữ tự nhiên và phản hồi chính xác hơn.
* **Tối ưu hóa và cá nhân hóa trải nghiệm người dùng:**
  + **Cá nhân hóa:** ML giúp các hệ thống như Netflix hoặc Spotify đề xuất nội dung dựa trên sở thích của người dùng.
  + **Tối ưu hóa:** Các công ty logistics như FedEx sử dụng ML để tối ưu hóa lộ trình vận chuyển.
* **Phân tích dữ liệu phức tạp:**  
  Machine Learning có thể phát hiện các mẫu và mối quan hệ phức tạp trong dữ liệu mà con người không thể dễ dàng nhận ra.
  + Ví dụ: Trong y học, ML được dùng để phân tích ảnh chụp X-quang nhằm phát hiện bệnh sớm.
* **Giải quyết các vấn đề phi tuyến tính:**  
  Nhiều bài toán thực tế có tính phi tuyến tính cao, ví dụ như dự đoán thị trường tài chính hay phân loại hình ảnh. Machine Learning, đặc biệt là các phương pháp như Deep Learning, đã chứng minh khả năng vượt trội trong các bài toán này.
* **Thay đổi cách làm việc truyền thống:**  
  ML không chỉ tối ưu hóa các hệ thống hiện có mà còn tạo ra những phương pháp làm việc hoàn toàn mới.
  + Ví dụ: Xe tự lái sử dụng ML để nhận diện đường đi, biển báo, và người đi bộ.
* **Hỗ trợ nghiên cứu khoa học:**  
  Trong các lĩnh vực như vật lý, sinh học, và hóa học, ML giúp các nhà khoa học phân tích dữ liệu lớn và phát hiện ra những kiến thức mới.
  + Ví dụ: ML được dùng để tìm ra các loại thuốc mới hoặc mô hình hóa các phản ứng hóa học.

1.2 Trình bày các kỹ thuật machine learning cơ bản

Các kỹ thuật cơ bản trong Machine Learning thường được chia thành ba nhóm chính: Học có giám sát (Supervised Learning), Học không giám sát (Unsupervised Learning), và Học tăng cường (Reinforcement Learning).

**1. Học có giám sát (Supervised Learning)**

Học có giám sát liên quan đến việc sử dụng một tập dữ liệu đã được gắn nhãn để huấn luyện mô hình. Mô hình học cách ánh xạ từ đầu vào (input) sang đầu ra (output) dựa trên dữ liệu mẫu.

**Các phương pháp phổ biến:**

* **Hồi quy (Regression):**  
  Dùng để dự đoán giá trị liên tục.
  + **Ví dụ:** Dự đoán giá nhà, dự đoán doanh số bán hàng.
  + Thuật toán: Hồi quy tuyến tính (Linear Regression), Hồi quy Ridge, Lasso.
* **Phân loại (Classification):**  
  Dùng để phân loại dữ liệu vào các nhóm cụ thể.
  + **Ví dụ:** Phân loại email là spam hoặc không spam, dự đoán bệnh lý từ triệu chứng.
  + Thuật toán:
    - Hồi quy Logistic (Logistic Regression)
    - K-Nearest Neighbors (KNN)
    - Support Vector Machines (SVM)
    - Random Forest, Gradient Boosting Machines (GBM), XGBoost.

**2. Học không giám sát (Unsupervised Learning)**

Học không giám sát làm việc với dữ liệu chưa gắn nhãn. Mục tiêu là khám phá cấu trúc tiềm ẩn hoặc các mẫu trong dữ liệu.

**Các phương pháp phổ biến:**

* **Phân cụm (Clustering):**  
  Phân nhóm các đối tượng có đặc điểm tương tự.
  + **Ví dụ:** Phân nhóm khách hàng dựa trên hành vi mua sắm.
  + Thuật toán:
    - K-Means
    - Hierarchical Clustering
    - DBSCAN
* **Giảm số chiều (Dimensionality Reduction):**  
  Dùng để giảm bớt số lượng biến (features) trong dữ liệu mà vẫn giữ được thông tin quan trọng.
  + **Ví dụ:** Trực quan hóa dữ liệu, tăng tốc độ xử lý.
  + Thuật toán:
    - Principal Component Analysis (PCA)
    - t-Distributed Stochastic Neighbor Embedding (t-SNE).
* **Phát hiện bất thường (Anomaly Detection):**  
  Tìm ra các điểm dữ liệu không bình thường trong tập hợp.
  + **Ví dụ:** Phát hiện giao dịch gian lận, lỗi trong hệ thống IoT.
  + Thuật toán: Isolation Forest, One-Class SVM.

**3. Học tăng cường (Reinforcement Learning - RL)**

Học tăng cường dựa trên cơ chế học qua thử-sai (trial-and-error). Mô hình (hay tác nhân, agent) học cách đưa ra hành động trong môi trường để đạt được phần thưởng tối đa.

**Các phương pháp phổ biến:**

* **Q-Learning:**  
  Học bảng giá trị Q để tối ưu hóa chính sách hành động.
* **Deep Q-Learning (DQL):**  
  Kết hợp học tăng cường với deep learning, dùng mạng nơ-ron để xấp xỉ giá trị Q.
* **Policy Gradient:**  
  Học trực tiếp chiến lược (policy) thay vì giá trị Q.
* **Ứng dụng:**
  + Chơi game tự động (AlphaGo, Dota 2).
  + Điều khiển robot.

**4. Kỹ thuật bổ trợ và thuật toán nâng cao**

Dưới đây là một số kỹ thuật và thuật toán thường được áp dụng để cải thiện các mô hình Machine Learning:

**Kỹ thuật xử lý trước dữ liệu:**

* Chuẩn hóa dữ liệu (Normalization), tiêu chuẩn hóa (Standardization).
* Xử lý dữ liệu bị thiếu (Missing Data Handling).
* Encoding (One-Hot Encoding, Label Encoding).

**Kỹ thuật điều chỉnh mô hình:**

* **Regularization:** Giảm thiểu overfitting. Ví dụ: L1, L2 Regularization.
* **Cross-Validation:** Đánh giá mô hình bằng cách chia dữ liệu thành nhiều phần nhỏ.

**Học sâu (Deep Learning):**

* **Mạng nơ-ron nhân tạo (Artificial Neural Networks - ANN):** Mô hình cơ bản trong học sâu.
* **Mạng nơ-ron tích chập (Convolutional Neural Networks - CNN):** Dùng trong xử lý ảnh.
* **Mạng nơ-ron hồi tiếp (Recurrent Neural Networks - RNN):** Dùng trong xử lý chuỗi, như dịch máy.

**1.3 Các kỹ thuật của Deep learning**

**Các kỹ thuật của Deep Learning** tập trung vào việc xây dựng và huấn luyện các mô hình mạng nơ-ron sâu (Deep Neural Networks) để giải quyết các bài toán phức tạp liên quan đến dữ liệu lớn, phi cấu trúc như hình ảnh, video, âm thanh và văn bản.

* **Mạng nơ-ron nhân tạo (Artificial Neural Networks - ANN)**
* **Đặc điểm:**  
  ANN là mô hình cơ bản nhất của Deep Learning, bao gồm các lớp đầu vào, ẩn và đầu ra. Mỗi lớp chứa các nút (neuron) liên kết với nhau thông qua trọng số và hàm kích hoạt.
* **Ứng dụng:**
  + Dự đoán giá trị (regression).
  + Phân loại cơ bản (classification).
* **Hạn chế:**  
  Không hiệu quả với dữ liệu hình ảnh, âm thanh hoặc chuỗi thời gian.
* **Mạng nơ-ron tích chập (Convolutional Neural Networks - CNN)**
* **Đặc điểm:**  
  CNN được thiết kế đặc biệt để xử lý dữ liệu hình ảnh. Nó sử dụng các lớp tích chập (convolutional layers) để trích xuất đặc trưng từ ảnh và các lớp gộp (pooling layers) để giảm chiều dữ liệu.
* **Ứng dụng:**
  + Nhận diện khuôn mặt (facial recognition).
  + Phân loại hình ảnh (image classification).
  + Phát hiện vật thể (object detection).
  + Phân đoạn hình ảnh (image segmentation).
* **Các kiến trúc phổ biến:**
  + AlexNet, VGG, ResNet, EfficientNet.
* **Mạng nơ-ron hồi tiếp (Recurrent Neural Networks - RNN)**
* **Đặc điểm:**  
  RNN được thiết kế để xử lý dữ liệu tuần tự hoặc dữ liệu thời gian, chẳng hạn như văn bản, âm thanh, hoặc chuỗi thời gian. Nó sử dụng trạng thái ẩn (hidden state) để ghi nhớ thông tin từ các bước trước.
* **Ứng dụng:**
  + Dịch máy (machine translation).
  + Dự đoán chuỗi thời gian (time series forecasting).
  + Phân tích cảm xúc (sentiment analysis).
* **Hạn chế:**
  + Vanishing gradient (độ dốc mất dần) khiến việc học trên chuỗi dài trở nên khó khăn.
* **LSTM và GRU (Long Short-Term Memory & Gated Recurrent Unit)**
* **Đặc điểm:**  
  LSTM và GRU là cải tiến của RNN, giúp khắc phục vấn đề vanishing gradient. Chúng sử dụng các cổng (gates) để kiểm soát dòng thông tin.
* **Ứng dụng:**
  + Nhận dạng giọng nói (speech recognition).
  + Dự đoán chuỗi dữ liệu dài (long-term dependencies).
* **Mạng nơ-ron tự mã hóa (Autoencoders)**
* **Đặc điểm:**  
  Autoencoders học cách nén dữ liệu vào một không gian nhỏ gọn (latent space) và sau đó tái tạo lại dữ liệu ban đầu.
* **Ứng dụng:**
  + Giảm số chiều dữ liệu.
  + Loại bỏ nhiễu (denoising).
  + Sinh dữ liệu mới (data generation).
* **Biến thể:**
  + Denoising Autoencoders.
  + Variational Autoencoders (VAE).
* **Mạng sinh đối kháng (Generative Adversarial Networks - GANs)**
* **Đặc điểm:**  
  GANs bao gồm hai mạng: một mạng sinh (generator) và một mạng phân biệt (discriminator) hoạt động đối kháng. Generator cố gắng sinh dữ liệu giống thật, trong khi discriminator cố gắng phân biệt dữ liệu thật và giả.
* **Ứng dụng:**
  + Tạo ảnh chân dung giả (deepfake).
  + Tăng cường dữ liệu (data augmentation).
  + Phục hồi ảnh (image restoration).
* **Các kiến trúc nổi bật:**
  + DCGAN, CycleGAN, StyleGAN.
* **Mạng Transformer**
* **Đặc điểm:**  
  Transformer là kiến trúc sử dụng cơ chế attention để tập trung vào các phần quan trọng của dữ liệu. Nó đặc biệt hiệu quả trong xử lý dữ liệu tuần tự như văn bản.
* **Ứng dụng:**
  + Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP).
  + Mô hình GPT, BERT, và các biến thể.
* **Học chuyển giao (Transfer Learning)**
* **Đặc điểm:**  
  Học chuyển giao tận dụng các mô hình đã được huấn luyện trên tập dữ liệu lớn và tái sử dụng chúng cho các bài toán khác với ít dữ liệu hơn.
* **Ứng dụng:**
  + Phân loại ảnh với mô hình ResNet hoặc EfficientNet đã được huấn luyện trước.
  + Xử lý văn bản với BERT hoặc GPT
* **Học sâu tăng cường (Deep Reinforcement Learning)**
* **Đặc điểm:**  
  Kết hợp giữa học tăng cường và mạng nơ-ron sâu, cho phép tác nhân học cách thực hiện hành động trong môi trường phức tạp.
* **Ứng dụng:**
  + Xe tự lái.
  + Chơi game (AlphaGo, Dota 2 AI).
  + Điều khiển robot.
* **Dropout và Batch Normalization**
* **Dropout:** Kỹ thuật để giảm overfitting bằng cách "tắt" ngẫu nhiên một số nút (neuron) trong quá trình huấn luyện.
* **Batch Normalization:** Chuẩn hóa đầu ra của mỗi lớp trong mạng, giúp cải thiện tốc độ huấn luyện và độ ổn định.

**1.**4 Clean và chuẩn hoá Data

1.4.1 Các bước clean data và chuẩn hoá data

**Clean và chuẩn hóa dữ liệu (Data Cleaning and Data Normalization)** là hai bước quan trọng trong quy trình xử lý dữ liệu, đảm bảo dữ liệu có chất lượng tốt và sẵn sàng để sử dụng trong các thuật toán Machine Learning

* **Data Cleaning (Làm sạch dữ liệu)**

**Làm sạch dữ liệu** là bước đầu tiên và rất quan trọng trong quy trình tiền xử lý dữ liệu. Dữ liệu thô thường chứa lỗi, thiếu giá trị, hoặc không đồng nhất. Nếu không làm sạch, các mô hình Machine Learning có thể cho ra kết quả không chính xác hoặc kém hiệu quả.

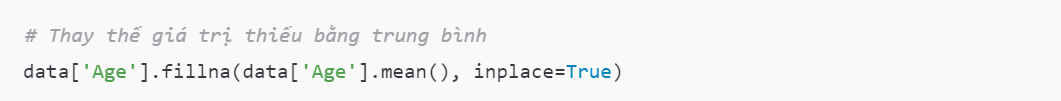
* **Xử lý dữ liệu bị thiếu (Missing Data)**

Dữ liệu bị thiếu là một vấn đề phổ biến, thường xảy ra trong các tập dữ liệu thực tế. Các nguyên nhân có thể bao gồm lỗi thu thập dữ liệu, không trả lời khảo sát, hoặc các vấn đề kỹ thuật.

**Các cách xử lý:**

* **Loại bỏ dữ liệu thiếu:**
  + Áp dụng nếu số lượng dữ liệu bị thiếu không đáng kể.
  + Ví dụ: Nếu chỉ có 1-2% dữ liệu thiếu, bạn có thể xóa các dòng hoặc cột chứa giá trị này.
* **Thay thế dữ liệu:**
  + **Bằng giá trị trung bình:** Tốt với dữ liệu liên tục.
  + **Bằng giá trị trung vị:** Phù hợp khi dữ liệu chứa ngoại lai.
  + **Dự đoán dữ liệu thiếu:** Sử dụng các mô hình Machine Learning như KNN hoặc regression để dự đoán.

**Ví dụ:**



**Những điều cần lưu ý:**

* Không nên xóa dữ liệu thiếu nếu chúng chiếm tỷ lệ lớn, vì điều này có thể làm mất thông tin quan trọng.

Cần kiểm tra sự phân phối của dữ liệu trước và sau khi thay thế.

* **Loại bỏ dữ liệu trùng lặp (Duplicate Data)**

Dữ liệu trùng lặp có thể xảy ra khi kết hợp nhiều nguồn dữ liệu hoặc do lỗi nhập liệu.

**Cách phát hiện và xử lý:**

1. Sử dụng phương pháp thống kê để đếm các bản ghi trùng lặp.
2. Loại bỏ chúng để giảm tải tính toán và tăng hiệu quả mô hình.

**Ví dụ**



**Ứng dụng:**

Trong hệ thống e-commerce, loại bỏ các đơn hàng bị nhập nhiều lần giúp cải thiện chất lượng phân tích.

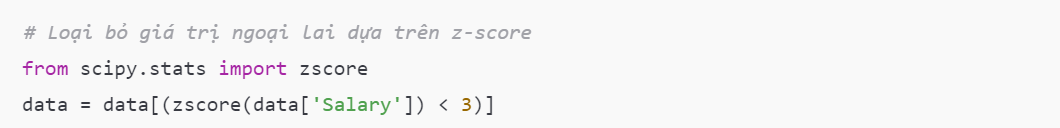
**+ Xử lý giá trị ngoại lai (Outliers)**

**Giá trị ngoại lai** là các điểm dữ liệu không bình thường, thường có giá trị quá lớn hoặc quá nhỏ so với phần còn lại.

**Phát hiện giá trị ngoại lai:**

1. **Boxplot:** Minh họa phân phối dữ liệu và phát hiện các giá trị nằm ngoài IQR (interquartile range).
2. **Z-score:** Xác định các giá trị cách xa trung bình hơn 3 lần độ lệch chuẩn.

**Ví dụ**



**Cách xử lý:**

* Loại bỏ giá trị ngoại lai hoặc biến đổi chúng (log-transform).

**Những lưu ý:**

* Đừng loại bỏ giá trị ngoại lai quá mức nếu chúng có ý nghĩa thực tế (ví dụ: doanh thu rất cao từ một khách hàng đặc biệt).
* **Data Normalization (Chuẩn hóa dữ liệu)**

Chuẩn hóa dữ liệu là bước quan trọng để các đặc trưng có cùng phạm vi hoặc phân phối, đặc biệt khi sử dụng các thuật toán nhạy cảm với tỷ lệ như k-NN hoặc SVM.

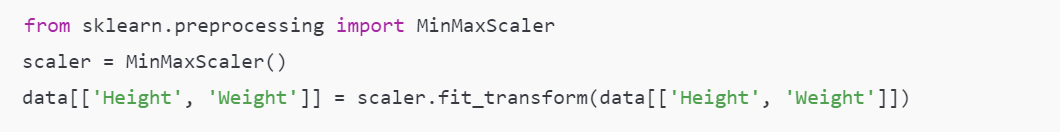
**+ Min-Max Scaling**

* Công thức:

x =

* **Ứng dụng:**
  + Dữ liệu cần đưa về [0, 1].
  + Phù hợp với các bài toán dự đoán giá nhà hoặc phân loại cơ bản.

Ví dụ:



**Ưu điểm:**

* Đơn giản, dễ hiểu.
* Hiệu quả khi không có ngoại lai.

**Nhược điểm:**

* Nhạy cảm với giá trị ngoại lai.

**+ Standardization (Z-Score Scaling)**

* **Công thức:**

x =

**Ứng dụng:**

* Các bài toán cần dữ liệu có phân phối chuẩn.
* Sử dụng trong thuật toán Linear Regression, Logistic Regression.

**+ Log Transformation**

* **Ứng dụng:**
  + Làm giảm tác động của giá trị lớn.
  + Phù hợp với dữ liệu có phân phối lệch phải.

Ví dụ:



**Kết luận**

Data Cleaning và Data Normalization là hai bước không thể thiếu để đảm bảo chất lượng dữ liệu và tối ưu hóa hiệu suất của các mô hình Machine Learning. Khi thực hiện đúng, dữ liệu sẽ trở nên đáng tin cậy, giúp các mô hình học tập hiệu quả hơn, từ đó cải thiện kết quả đầu ra.

1.4.2 Ví dụ

Ví dụ 1: Clean và chuẩn hoá dữ liệu điểm thi

Ví dụ demo: *C:\Users\84395\Desktop\Phát triển hệ thống thông minh\Tiểu luận\clean.ipynb*

Ví dụ 3: Clean tập dữ liệu ảnh cảm xúc khuôn mặt

1.5 Kết luận

**Kết luận về một số vấn đề cơ bản của Machine Learning (ML)**

**Tại sao lại có Machine Learning (ML)?**

* **Giải quyết các bài toán phức tạp**: Trước khi có ML, nhiều bài toán yêu cầu phải lập trình thủ công các quy tắc và logic phức tạp. ML xuất hiện như một giải pháp giúp máy tính tự động học từ dữ liệu mà không cần phải lập trình chi tiết mọi quy tắc. Các bài toán như nhận dạng hình ảnh, phân tích văn bản, hay dự đoán hành vi khách hàng có thể được giải quyết hiệu quả hơn bằng cách học từ dữ liệu thực tế.
* **Khả năng xử lý dữ liệu lớn (Big Data)**: ML có thể xử lý khối lượng dữ liệu lớn mà không cần sự can thiệp của con người, giúp phát hiện các mẫu và xu hướng trong dữ liệu mà con người khó có thể nhận ra.
* **Cải thiện hiệu suất theo thời gian**: Mô hình ML có thể tự cải thiện qua từng lần học, giúp tối ưu hóa các quy trình và dự đoán chính xác hơn theo thời gian.

**Mục đích của ML**

* **Tự động hóa các quyết định**: ML giúp tự động hóa các quy trình ra quyết định, đặc biệt là trong các lĩnh vực như tài chính, y tế, và marketing, nơi có rất nhiều dữ liệu và quy trình phức tạp.
* **Dự đoán và phân tích xu hướng**: Các mô hình ML được sử dụng để dự đoán các kết quả trong tương lai (ví dụ: dự báo nhu cầu khách hàng, dự đoán bệnh tật) và phân tích xu hướng từ dữ liệu quá khứ.
* **Cải thiện các dịch vụ và sản phẩm**: ML giúp các doanh nghiệp cải thiện chất lượng dịch vụ và sản phẩm thông qua việc phân tích hành vi người dùng, từ đó cá nhân hóa trải nghiệm và tối ưu hóa quy trình.

**Kỹ thuật cơ bản trong ML**

* **Học giám sát (Supervised Learning)**: Là kỹ thuật học mà trong đó mô hình học từ một bộ dữ liệu có nhãn (ví dụ: phân loại email là spam hay không spam, dự đoán giá nhà dựa trên các đặc trưng).
* **Học không giám sát (Unsupervised Learning)**: Trong học không giám sát, dữ liệu không có nhãn. Mục tiêu là phát hiện các mẫu, cấu trúc hoặc mối quan hệ trong dữ liệu (ví dụ: phân cụm khách hàng, giảm chiều dữ liệu).
* **Học bán giám sát (Semi-supervised Learning)**: Sử dụng cả dữ liệu có nhãn và không có nhãn để cải thiện hiệu quả mô hình, hữu ích khi việc thu thập nhãn là tốn kém.
* **Học tăng cường (Reinforcement Learning)**: Mô hình học từ các phản hồi của môi trường, tối ưu hóa hành động dựa trên thưởng và phạt (ví dụ: tối ưu hóa chiến lược chơi game, robot tự động).

**Deep Learning (Học sâu)**

* **Khái niệm**: Deep Learning là một nhánh con của Machine Learning, sử dụng các mạng nơ-ron nhân tạo với nhiều lớp (deep networks) để học các đặc trưng phức tạp từ dữ liệu. Các mạng nơ-ron này có khả năng học và trích xuất đặc trưng mà không cần sự can thiệp của con người.
* **Ưu điểm**:
  + **Khả năng tự học đặc trưng**: Deep Learning có thể tự động trích xuất đặc trưng mà không cần phải lập trình thủ công.
  + **Xử lý dữ liệu phức tạp**: Đặc biệt mạnh mẽ với các bài toán phức tạp như nhận diện hình ảnh, phân tích văn bản và giọng nói.
* **Nhược điểm**:
  + **Cần nhiều dữ liệu và tài nguyên tính toán**: Deep Learning yêu cầu lượng dữ liệu rất lớn và khả năng tính toán mạnh mẽ như GPU.
  + **Khó giải thích**: Các mô hình Deep Learning thường được coi là "hộp đen", khó để giải thích các quyết định mà chúng đưa ra.

**Làm sạch dữ liệu (Data Cleaning)**

* **Xử lý thiếu dữ liệu**: Dữ liệu thiếu (NaN) cần được xử lý để không ảnh hưởng đến độ chính xác của mô hình. Phương pháp xử lý bao gồm thay thế giá trị thiếu bằng giá trị trung bình, trung vị hoặc mô hình dự đoán.
* **Phát hiện và xử lý giá trị cực trị (Outliers)**: Nhiễu và giá trị bất thường có thể làm giảm hiệu quả của mô hình. Phương pháp xử lý bao gồm loại bỏ hoặc giảm ảnh hưởng của các giá trị cực trị.
* **Chuẩn hoá và mã hóa dữ liệu**: Dữ liệu phân loại cần được mã hóa lại (One-Hot Encoding, Label Encoding). Dữ liệu văn bản cần được vector hóa (TF-IDF, Word2Vec). Việc chuẩn hoá giúp giảm thiểu sự chênh lệch giữa các đặc trưng có đơn vị khác nhau.

**Chuẩn hoá dữ liệu (Data Normalization)**

* **Lý do chuẩn hoá**: Chuẩn hoá dữ liệu giúp các mô hình học máy hoạt động hiệu quả hơn, tránh trường hợp các đặc trưng có phạm vi giá trị lớn chiếm ưu thế trong quá trình huấn luyện.
* **Các phương pháp chuẩn hoá**:
  + **Min-Max Scaling**: Biến đổi giá trị các đặc trưng về một khoảng cố định, thường là [0, 1].
  + **Z-score Standardization**: Biến đổi giá trị các đặc trưng sao cho có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1.
  + **Robust Scaling**: Phương pháp này sử dụng các thông số trung vị và khoảng IQR để chuẩn hoá, giúp giảm ảnh hưởng của các giá trị cực trị.

CHƯƠNG 2: DEEP LEARNING

2.1 Lịch sử phát triển của Deep learning

**Khởi đầu và nền tảng**

Trong những năm 1950, các nghiên cứu về trí tuệ nhân tạo (AI) bắt đầu hình thành, với Alan Turing đưa ra ý tưởng về trí thông minh máy tính qua bài viết nổi tiếng "Computing Machinery and Intelligence". Đồng thời, **Perceptron**, một mô hình mạng nơ-ron đơn giản do Frank Rosenblatt phát triển vào năm 1958, là bước đầu tiên hướng đến việc áp dụng mạng nơ-ron trong nhận diện mẫu, mặc dù chỉ có thể giải quyết các vấn đề rất cơ bản.

**Những thử thách và sự phát triển trong học sâu**

Bước sang những năm 1970 và 1980, nghiên cứu về mạng nơ-ron tiếp tục nhưng gặp phải nhiều hạn chế về tính toán và dữ liệu. Tuy nhiên, vào năm 1986, một đột phá quan trọng xảy ra khi **Geoffrey Hinton** cùng các cộng sự phát triển thuật toán **Lan truyền ngược (Backpropagation)**. Thuật toán này cho phép tối ưu hóa các trọng số trong mạng nơ-ron nhiều lớp, mở ra cơ hội mới cho việc xây dựng các mô hình học máy phức tạp hơn.

**Khởi đầu của Deep Learning**

Năm 2006 đánh dấu một mốc quan trọng trong lịch sử Deep Learning khi **Geoffrey Hinton** và các cộng sự giới thiệu **Deep Belief Networks (DBNs)**, khôi phục lại sự quan tâm vào mạng nơ-ron sâu. Các mô hình này có thể học được các đặc trưng phức tạp từ dữ liệu mà không cần nhãn, đặt nền tảng cho sự phát triển của Deep Learning trong những năm sau.

**Bùng nổ Deep Learning**

Năm 2012, một bước đột phá lớn diễn ra với **AlexNet** — mô hình mạng nơ-ron sâu (CNN) giành chiến thắng trong cuộc thi **ImageNet**. Sự thành công của AlexNet không chỉ làm nổi bật khả năng của Deep Learning trong nhận diện hình ảnh, mà còn đánh dấu sự khởi đầu của thời kỳ bùng nổ công nghệ Deep Learning. Các kỹ thuật như **Generative Adversarial Networks (GANs)** cũng được giới thiệu trong cùng thời gian, mở ra những khả năng mới trong việc tạo dữ liệu giả.

**Sự phát triển vượt bậc trong NLP và các ứng dụng AI**

Vào năm 2014, **GANs** (Generative Adversarial Networks) được **Ian Goodfellow** giới thiệu và nhanh chóng trở thành một trong những công cụ mạnh mẽ trong việc tạo ra dữ liệu giả như hình ảnh, video. Tiếp theo, sự xuất hiện của **ResNet** vào năm 2015 với các mạng nơ-ron sâu hơn đã giải quyết vấn đề "vanishing gradient", giúp các mô hình học sâu ngày càng phức tạp và hiệu quả hơn.

Từ năm 2017, **Transformers** đã thay đổi cách tiếp cận trong xử lý ngôn ngữ tự nhiên (NLP). Kiến trúc này, với cơ chế **chú ý (Attention)**, đã thúc đẩy các mô hình như **BERT** và **GPT**, giúp nâng cao khả năng của AI trong các nhiệm vụ như dịch máy, tạo văn bản và phân tích cảm xúc.

**Hiện tại và tương lai của Deep Learning**

Ngày nay, những mô hình như **GPT-3** với hàng trăm tỷ tham số đã chứng minh khả năng vượt trội trong các ứng dụng NLP. Đồng thời, các nghiên cứu về **AI tổng quát (AGI)** và **đạo đức AI** đang dần trở nên quan trọng. Deep Learning không chỉ tác động mạnh mẽ đến các lĩnh vực như y tế, tài chính, mà còn hứa hẹn mở ra những khả năng chưa từng có trong sáng tạo nội dung, tự động hóa và hiểu biết sâu sắc về thế giới.

2.2 Kiến thức mạng Neuron

**Mạng neuron là gì?**

* **Mô phỏng não người**: Mạng neuron được thiết kế để mô phỏng cách hoạt động của các neuron trong não bộ con người.
* **Tập hợp các nút (nodes)**: Mỗi nút (neuron) được kết nối với nhau qua các liên kết (connections), mỗi liên kết có trọng số (weight) để điều chỉnh mức độ ảnh hưởng giữa các nút.

**Thành phần chính của mạng neuron**

**Neuron**: Là đơn vị cơ bản xử lý thông tin, gồm:

* + **Input**: Dữ liệu đầu vào (x).
  + **Weights**: Trọng số www, điều chỉnh tầm quan trọng của từng đầu vào.
  + **Bias**: Giá trị cố định bbb, thêm linh hoạt cho quá trình học.
  + **Activation Function**: Hàm kích hoạt, quyết định đầu ra của neuron (ví dụ: sigmoid, ReLU).

Công thức:



**Layers (Các tầng)**:

* **Input Layer**: Nhận dữ liệu từ bên ngoài.
* **Hidden Layers**: Tầng ẩn thực hiện các phép tính phức tạp.
* **Output Layer**: Tầng đầu ra, trả về kết quả.

**Weights và Biases**:

* Trọng số www: Điều chỉnh tầm quan trọng của từng kết nối.
* Bias bbb: Dịch chuyển đầu ra của hàm kích hoạt, tránh trường hợp giá trị luôn bằng 0.

**Hàm kích hoạt (Activation Function)**:

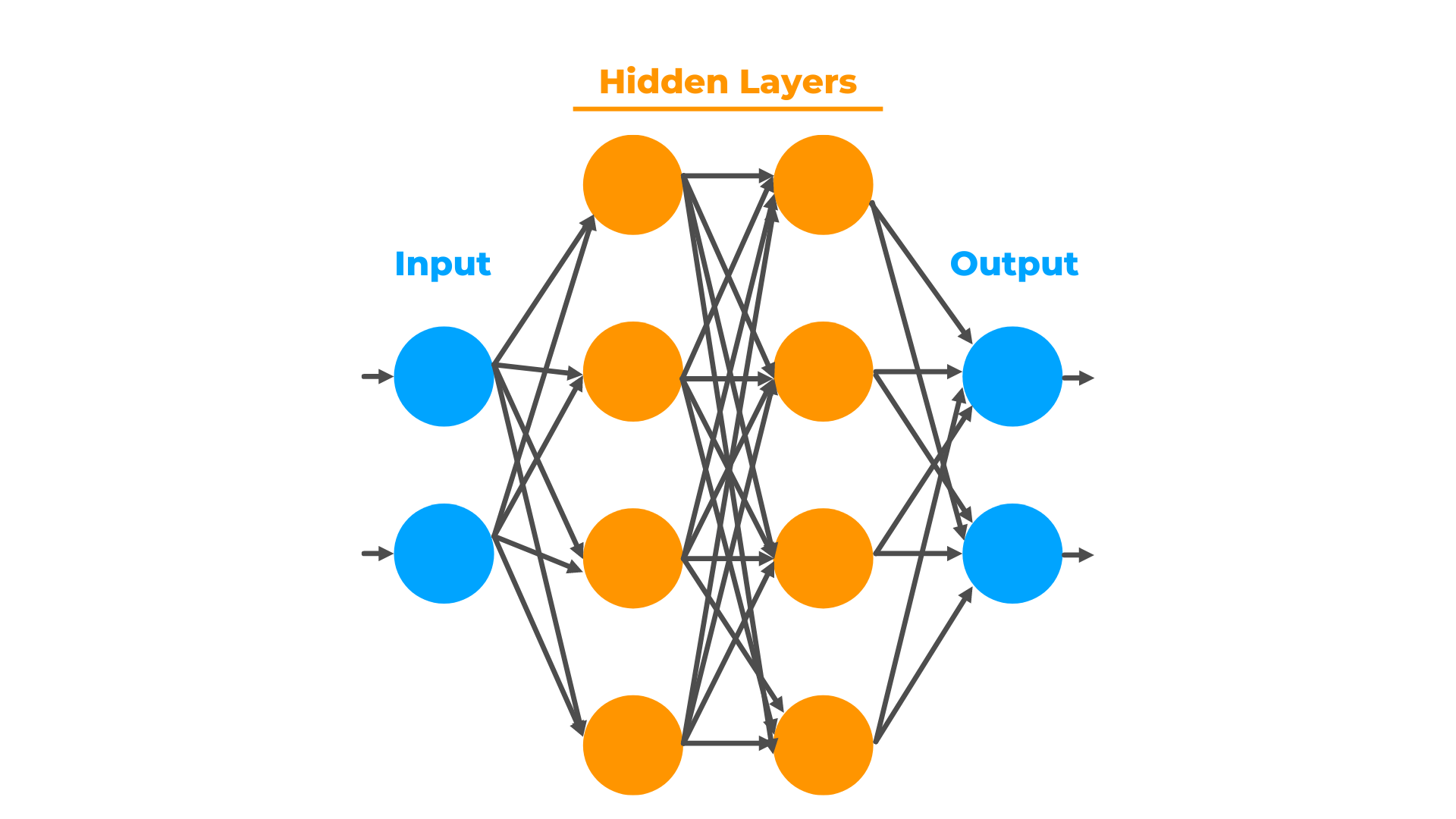
* **ReLU** (Rectified Linear Unit): f(x)=max⁡(0,x)f(x) = \max(0, x)f(x)=max(0,x).
* **Sigmoid**: Chuyển đầu ra về khoảng (0, 1).
* **Tanh**: Đưa đầu ra về khoảng (-1, 1).
* **Softmax**: Thường dùng cho bài toán phân loại đa lớp.

**Cách mạng neuron học**

* **Forward Propagation (Lan truyền tiến)**:
  + Tính toán kết quả đầu ra từ input qua các tầng.
* **Backward Propagation (Lan truyền ngược)**:
  + Dùng để điều chỉnh trọng số dựa trên lỗi (error) giữa đầu ra thực tế và dự đoán.
  + Sử dụng thuật toán **gradient descent**.

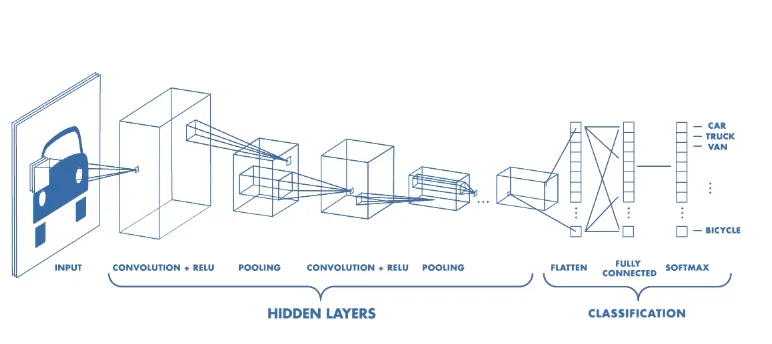
**Các loại mạng neuron phổ biến**

1. **Feedforward Neural Network**:
   * Luồng dữ liệu đi từ đầu vào đến đầu ra.
   * Thường dùng trong các bài toán hồi quy hoặc phân loại cơ bản.



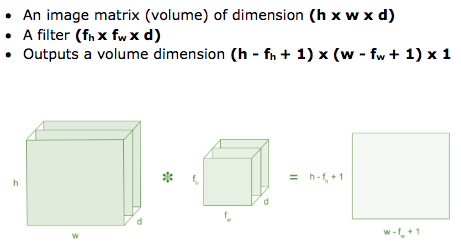
* **Tầng đầu vào (Input Layer):** Đây là tầng đầu tiên trong mạng nơ-ron truyền thẳng, nơi dữ liệu ban đầu được đưa vào hệ thống. Mỗi nơ-ron trong tầng này đại diện cho một đặc trưng duy nhất của tập dữ liệu đầu vào. Ví dụ, nếu chúng ta đang làm việc với dữ liệu hình ảnh, mỗi nơ-ron có thể tương ứng với một giá trị pixel. Tầng đầu vào không thực hiện bất kỳ phép tính nào mà chỉ đóng vai trò là điểm khởi đầu, chuyển thông tin sang các tầng tiếp theo để xử lý.
* **Tầng ẩn (Hidden Layer):** Đây là các tầng trung gian trong mạng, nơi thực hiện các phép tính và xử lý dữ liệu thực sự. Các tầng này rất quan trọng để mạng có thể học được các mẫu phức tạp và biểu diễn các đặc trưng của dữ liệu đầu vào. Một tầng ẩn bao gồm các nơ-ron áp dụng trọng số (weights) lên dữ liệu nhận được và đưa dữ liệu qua một hàm kích hoạt (activation function) để thêm tính phi tuyến. Số lượng tầng ẩn và số lượng nơ-ron trong mỗi tầng có thể ảnh hưởng lớn đến khả năng của mạng trong việc mô hình hóa các hàm phức tạp.
* **Tầng đầu ra (Output Layer):** Đây là tầng cuối cùng chịu trách nhiệm tạo ra dự đoán của mạng hoặc giải pháp cho vấn đề mà dữ liệu đầu vào đặt ra. Số lượng nơ-ron trong tầng này tương ứng với số lượng kết quả hoặc lớp (classes) có thể có trong một bài toán phân loại. Đầu ra được tạo ra thường ở dạng dễ hiểu với người dùng hoặc các hệ thống khác, thường thông qua các loại hàm kích hoạt khác nhau như **softmax** cho bài toán phân loại hoặc các hàm tuyến tính cho bài toán hồi quy.

1. **Convolutional Neural Network (CNN)**:
   * Dùng trong xử lý ảnh và video.
   * Sử dụng các bộ lọc (filters) để trích xuất đặc trưng từ dữ liệu.

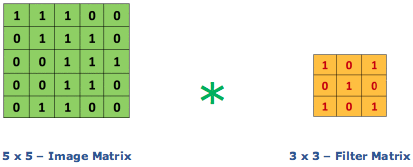


**Lớp tích chập - Convolution Layer**

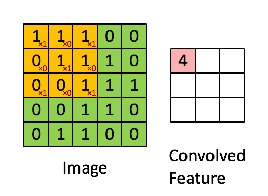
* Tích chập là lớp đầu tiên để trích xuất các tính năng từ hình ảnh đầu vào. Tích chập duy trì mối quan hệ giữa các pixel bằng cách tìm hiểu các tính năng hình ảnh bằng cách sử dụng các ô vương nhỏ của dữ liệu đầu vào. Nó là 1 phép toán có 2 đầu vào như ma trận hình ảnh và 1 bộ lọc hoặc hạt nhân.



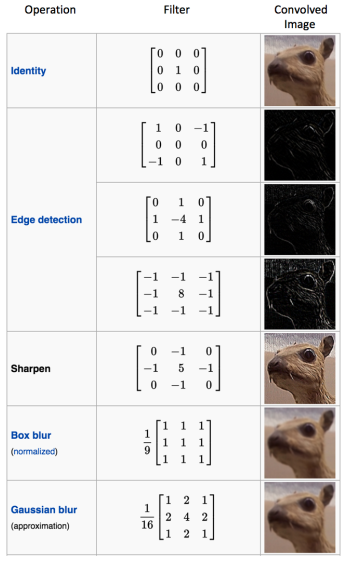
Xem xét 1 ma trận 5 x 5 có giá trị pixel là 0 và 1. Ma trận bộ lọc 3 x 3 như hình bên dưới.



Sau đó, lớp tích chập của ma trận hình ảnh 5 x 5 nhân với ma trận bộ lọc 3 x 3 gọi là 'Feature Map' như hình bên dưới.



Sự kết hợp của 1 hình ảnh với các bộ lọc khác nhau có thể thực hiện các hoạt động như phát hiện cạnh, làm mờ và làm sắc nét bằng cách áp dụng các bộ lọc. Ví dụ dưới đây cho thấy hình ảnh tích chập khác nhau sau khi áp dụng các Kernel khác nhau.

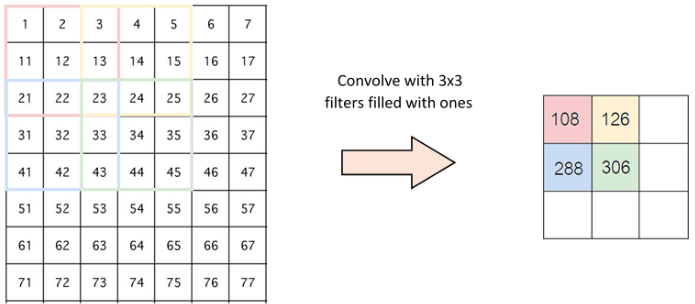


Ví dụ nhận diện chữ viết tay: *C:\Users\84395\Desktop\Phát triển hệ thống thông minh\Tiểu luận\xla.ipynb*



**Bước nhảy - Stride**

Stride là số pixel thay đổi trên ma trận đầu vào. Khi stride là 1 thì ta di chuyển các kernel 1 pixel. Khi stride là 2 thì ta di chuyển các kernel đi 2 pixel và tiếp tục như vậy. Hình dưới là lớp tích chập hoạt động với stride là 2.

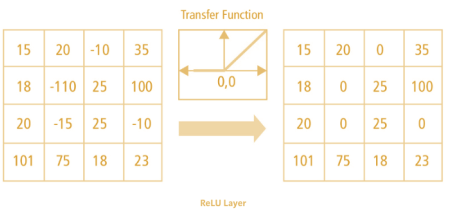


**Đường viền - Padding**

* Đôi khi kernel không phù hợp với hình ảnh đầu vào. Ta có 2 lựa chọn:
* Chèn thêm các số 0 vào 4 đường biên của hình ảnh (padding).
* Cắt bớt hình ảnh tại những điểm không phù hợp với kernel.

**Hàm phi tuyến - ReLU**

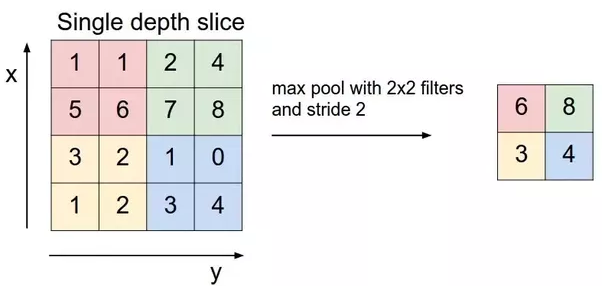
* ReLU viết tắt của Rectified Linear Unit, là 1 hàm phi tuyến. Với đầu ra là: ƒ (x) = max (0, x).
* Tại sao ReLU lại quan trọng: ReLU giới thiệu tính phi tuyến trong ConvNet. Vì dữ liệu trong thế giới mà chúng ta tìm hiểu là các giá trị tuyến tính không âm.



* Có 1 số hàm phi tuyến khác như tanh, sigmoid cũng có thể được sử dụng thay cho ReLU. Hầu hết người ta thường dùng ReLU vì nó có hiệu suất tốt.

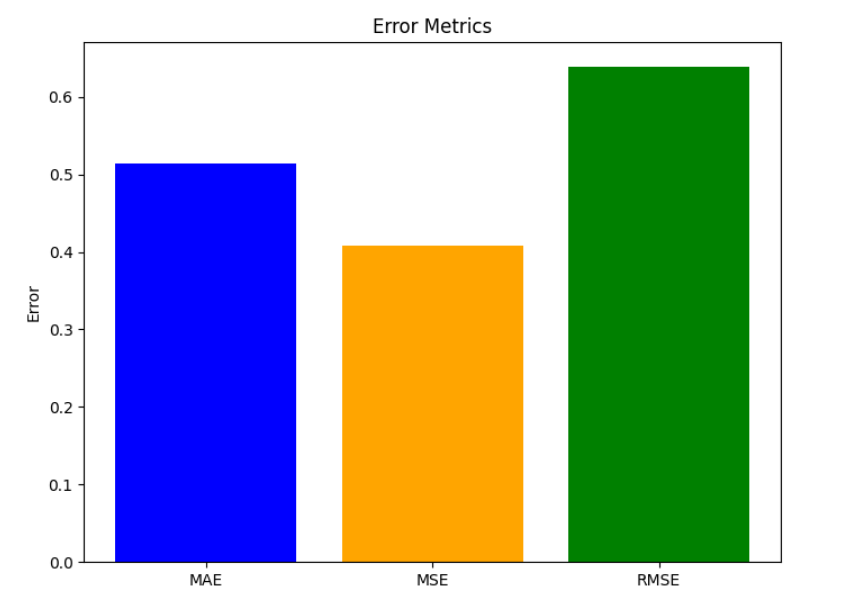
**Lớp gộp - Pooling Layer**

* Lớp pooling sẽ giảm bớt số lượng tham số khi hình ảnh quá lớn. Không gian pooling còn được gọi là lấy mẫu con hoặc lấy mẫu xuống làm giảm kích thước của mỗi map nhưng vẫn giữ lại thông tin quan trọng. Các pooling có thể có nhiều loại khác nhau:
* Max Pooling
* Average Pooling
* Sum Pooling
* Max pooling lấy phần tử lớn nhất từ ma trận đối tượng, hoặc lấy tổng trung bình. Tổng tất cả các phần tử trong map gọi là sum pooling

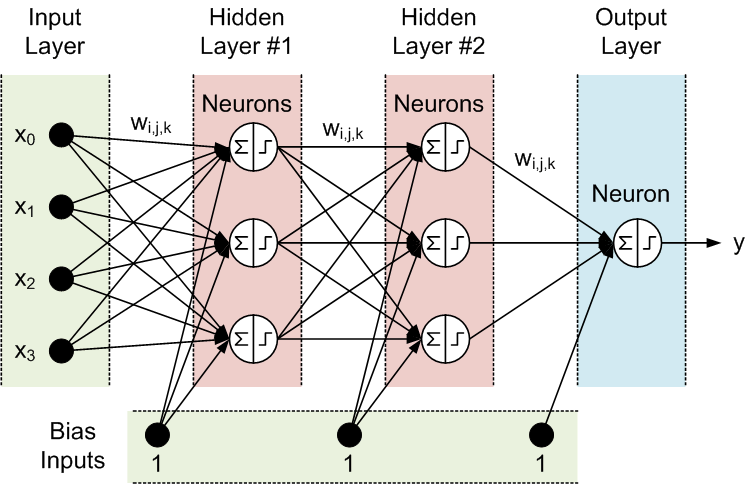


* Biểu đồ Accuracy và Loss trên hai tập validate và train khá sát nhau cho thấy model học khá tốt

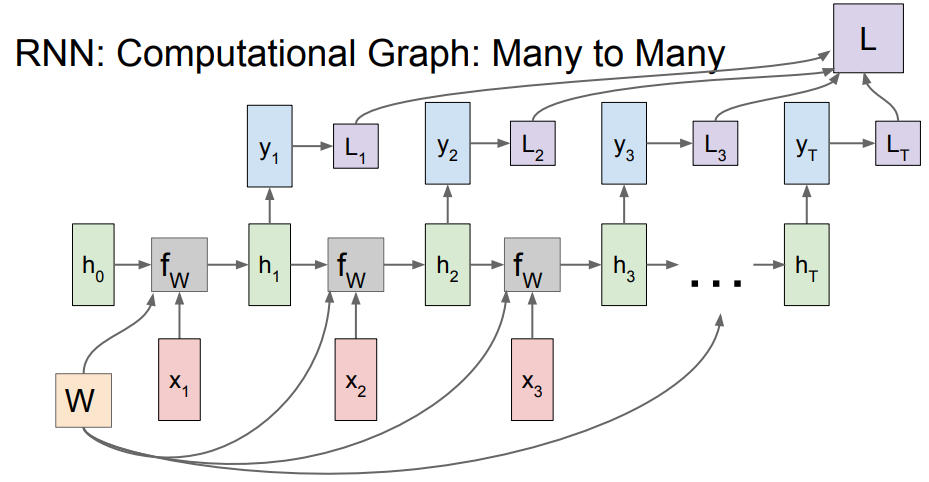
Biểu đồ so sánh giá trị MAE, MSE, RMSE:



1. **Recurrent Neural Network (RNN)**:
   * Dùng cho dữ liệu tuần tự, như xử lý ngôn ngữ tự nhiên.
   * Có kết nối ngược để lưu trạng thái trước đó.



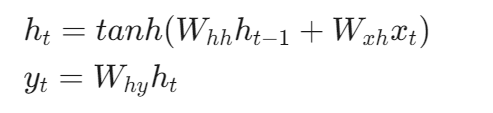
Như đã biết thì Neural Network bao gồm 3 phần chính là Input layer, Hidden layer và Output layer, có thể thấy là đầu vào và đầu ra của mạng neuron này là độc lập với nhau. Như vậy mô hình này không phù hợp với những bài toán dạng chuỗi như mô tả, hoàn thành câu, ... vì những dự đoán tiếp theo như từ tiếp theo phụ thuộc vào vị trí của nó trong câu và những từ đằng trước nó.  
Và như vậy RNN ra đời với ý tưởng chính là sử dụng một bộ nhớ để lưu lại thông tin từ từ những bước tính toán xử lý trước để dựa vào nó có thể đưa ra dự đoán chính xác nhất cho bước dự đoán hiện tại. Nếu các bạn vẫn chưa hiểu gì thì hãy cùng xem mô hình mạng RNN sau và cùng phân tích để hiểu rõ hơn:



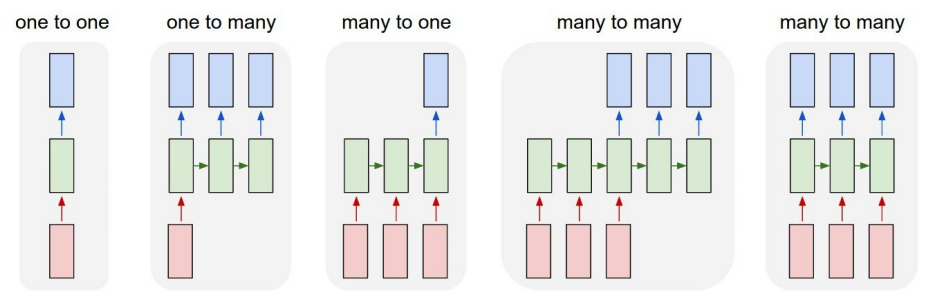
Nếu như mạng Neural Network chỉ là input layer x*x* đi qua hidden layer h*h* và cho ra output layer y*y* với ***full connected*** giữa các layer thì trong RNN, các input xt*xt*​ sẽ được kết hợp với hidden layer ht−1*ht*−1​ bằng hàm fW*fW*​ để tính toán ra hidden layer ht*ht*​ hiện tại và output yt*yt*​ sẽ được tính ra từ ht*ht*​, W*W* là tập các trọng số và nó được ở tất cả các cụm, các L1,L2,...,Lt*L*1​,*L*2​,...,*Lt*​ là các hàm mất mát sẽ được giải thích sau. Như vậy kết quả từ các quá trình tính toán trước đã được "nhớ" bằng cách kết hợp thêm ht−1*ht*−1​ tính ra ht*ht*​ để tăng độ chính xác cho những dự đoán ở hiện tại. Cụ thể quá trình tính toán được viết dưới dạng toán như sau:



Hàm fw​ chúng ta sẽ xử dụng hàm **tanh**, công thức trên sẽ trở thành :

  
  
  
Đến đây có 3 thứ mới xuất hiện: Wxh,Whh,Why*Wxh*​,*Whh*​,*Why*​. Đối với mạng NN chỉ sử dụng một ma trận trọng số W duy nhất thì với RNN nó sử dụng 3 ma trận trọng số cho 2 quá trình tính toán: Whh*Whh*​ kết hợp với "bộ nhớ trước" ht−1*ht*−1​ và Wxh *Wxh*​ kết hợp với xt*xt*​ để tính ra "bộ nhớ của bước hiện tại" ht*ht*​ từ đó kết hợp với Why*Wh*​*y* để tính ra yt*yt*​.

Ngoài mô hình Many to Many như ở trên thì RNN còn rất nhiều dạng khác như sau:



Ví dụ 1:

Ví dụ 1: Xây dựng model dự đoán điểm thi

Link: *C:\Users\84395\Desktop\Phát triển hệ thống thông minh\ 5-2\_DL.ipynb*

Code model:

*# Chuẩn hóa dữ liệu*

scaler = MinMaxScaler()

X\_mon1 = scaler.fit\_transform(mon1\_df[["TP1", "TP2", "TP3"]])

*# Không cần chuẩn hóa y nếu giá trị CK không cần thiết*

y\_mon1 = mon1\_df[["CK"]]

*# Chia dữ liệu thành tập huấn luyện và kiểm tra*

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_mon1, y\_mon1, test\_size=0.2, random\_state=42)

*# Định dạng lại dữ liệu cho RNN (thêm trục thời gian)*

X\_train = np.expand\_dims(X\_train, axis=1) *# (samples, timesteps, features)*

X\_test = np.expand\_dims(X\_test, axis=1)

*# Xây dựng mô hình RNN với 7 lớp*

model = Sequential([

SimpleRNN(128, activation='relu', input\_shape=(1, 3), return\_sequences=True),

Dropout(0.2), *# Dropout để giảm overfitting*

SimpleRNN(128, activation='relu', return\_sequences=True),

Dropout(0.2),

SimpleRNN(128, activation='relu', return\_sequences=True),

Dropout(0.2),

SimpleRNN(128, activation='relu', return\_sequences=True),

Dropout(0.2),

SimpleRNN(128, activation='relu', return\_sequences=True),

Dropout(0.2),

SimpleRNN(128, activation='relu', return\_sequences=True),

Dropout(0.2),

SimpleRNN(128, activation='relu'), *# Lớp cuối cùng không trả về chuỗi*

Dense(1) *# Đầu ra là một giá trị*

])

Kết quả chạy code:

Epoch 1/50

**50/50** ━━━━━━━━━━━━━━━━━━━━ **15s** 40ms/step - loss: 18.1617 - mae: 3.6225 - val\_loss: 3.0583 - val\_mae: 1.4745

Epoch 2/50

**50/50** ━━━━━━━━━━━━━━━━━━━━ **1s** 11ms/step - loss: 3.4923 - mae: 1.5418 - val\_loss: 3.5047 - val\_mae: 1.5388

Epoch 3/50

**50/50** ━━━━━━━━━━━━━━━━━━━━ **0s** 7ms/step - loss: 3.1111 - mae: 1.4541 - val\_loss: 3.2856 - val\_mae: 1.4959

Epoch 4/50

**50/50** ━━━━━━━━━━━━━━━━━━━━ **0s** 8ms/step - loss: 3.1066 - mae: 1.4551 - val\_loss: 2.9174 - val\_mae: 1.4352

Epoch 5/50

**50/50** ━━━━━━━━━━━━━━━━━━━━ **1s** 11ms/step - loss: 2.8250 - mae: 1.4160 - val\_loss: 2.5447 - val\_mae: 1.3808

Epoch 6/50

**50/50** ━━━━━━━━━━━━━━━━━━━━ **0s** 8ms/step - loss: 2.9747 - mae: 1.4547 - val\_loss: 3.2306 - val\_mae: 1.4861

Epoch 7/50

**50/50** ━━━━━━━━━━━━━━━━━━━━ **1s** 9ms/step - loss: 2.7273 - mae: 1.3847 - val\_loss: 2.7132 - val\_mae: 1.4004

Epoch 8/50

**50/50** ━━━━━━━━━━━━━━━━━━━━ **1s** 9ms/step - loss: 2.7797 - mae: 1.4014 - val\_loss: 2.5789 - val\_mae: 1.3855

Epoch 9/50

**50/50** ━━━━━━━━━━━━━━━━━━━━ **0s** 8ms/step - loss: 2.7702 - mae: 1.4000 - val\_loss: 2.7369 - val\_mae: 1.4069

Epoch 10/50

**50/50** ━━━━━━━━━━━━━━━━━━━━ **0s** 8ms/step - loss: 2.8048 - mae: 1.4023 - val\_loss: 2.7824 - val\_mae: 1.4099

Epoch 11/50

**50/50** ━━━━━━━━━━━━━━━━━━━━ **1s** 10ms/step - loss: 2.7107 - mae: 1.3859 - val\_loss: 2.6139 - val\_mae: 1.3881

Epoch 12/50

**50/50** ━━━━━━━━━━━━━━━━━━━━ **0s** 8ms/step - loss: 2.9089 - mae: 1.4511 - val\_loss: 2.8124 - val\_mae: 1.4155

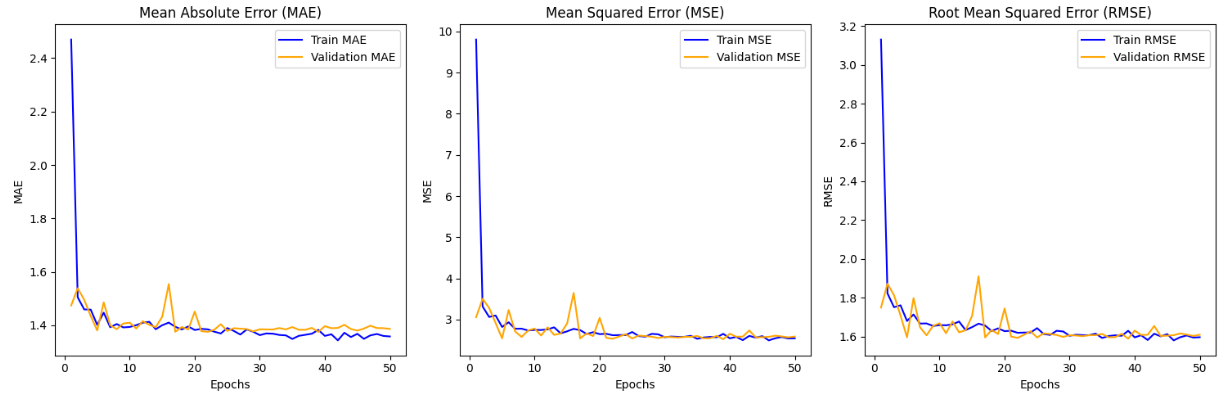
Epoch 13/50

...

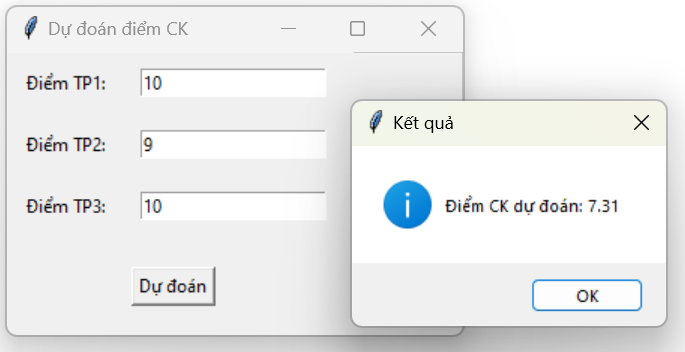
**13/13** ━━━━━━━━━━━━━━━━━━━━ **0s** 3ms/step - loss: 2.6820 - mae: 1.4008

Test Loss (MSE): 2.5885772705078125, Test MAE: 1.3866147994995117

**13/13** ━━━━━━━━━━━━━━━━━━━━ **3s** 106ms/step



Giao diện:



Ví dụ 3: Xây dựng model dự đoán chỉ số BMI

Link: *C:\Users\84395\Desktop\Phát triển hệ thống thông minh\ 5-1\_DL.ipynb*

*# Chia dữ liệu thành X và y*

X = df\_bmi[['Height', 'Weight', 'BMI']].values

y = df\_bmi['Category'].values

*# Chia dữ liệu thành tập huấn luyện và kiểm tra*

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

*# Định dạng lại dữ liệu cho RNN (tạo thêm trục thời gian)*

X\_train = np.expand\_dims(X\_train, axis=1) *# (samples, timesteps, features)*

X\_test = np.expand\_dims(X\_test, axis=1)

*# Chuyển đổi nhãn thành dạng one-hot encoding*

y\_train = to\_categorical(y\_train, num\_classes=4)

y\_test = to\_categorical(y\_test, num\_classes=4)

*# Xây dựng mô hình RNN với 7 lớp*

model = Sequential([

SimpleRNN(128, activation='relu', input\_shape=(1, 3), return\_sequences=True),

SimpleRNN(128, activation='relu', return\_sequences=True),

SimpleRNN(128, activation='relu', return\_sequences=True),

SimpleRNN(128, activation='relu', return\_sequences=True),

SimpleRNN(128, activation='relu', return\_sequences=True),

SimpleRNN(128, activation='relu', return\_sequences=True),

SimpleRNN(128, activation='relu') *# Lớp cuối cùng không cần return\_sequences*

])

*# Lớp Dense cuối cùng để phân loại*

model.add(Dense(4, activation='softmax')) *# 4 lớp phân loại (Underweight, Normal, Overweight, Obesity)*

*# Biên dịch mô hình, thêm 'mae' vào chỉ số*

model.compile(optimizer='adam', loss='categorical\_crossentropy', metrics=['accuracy', 'mae'])

*# Huấn luyện mô hình*

history = model.fit(X\_train, y\_train, epochs=50, validation\_data=(X\_test, y\_test), batch\_size=32)

Kết quả:

Epoch 1/50

**25/25** ━━━━━━━━━━━━━━━━━━━━ **14s** 78ms/step - accuracy: 0.3612 - loss: 1.2595 - mae: 0.3472 - val\_accuracy: 0.6550 - val\_loss: 0.7006 - val\_mae: 0.2110

Epoch 2/50

**25/25** ━━━━━━━━━━━━━━━━━━━━ **0s** 9ms/step - accuracy: 0.7907 - loss: 0.5291 - mae: 0.1713 - val\_accuracy: 0.9200 - val\_loss: 0.1803 - val\_mae: 0.0708

Epoch 3/50

**25/25** ━━━━━━━━━━━━━━━━━━━━ **0s** 8ms/step - accuracy: 0.8926 - loss: 0.2321 - mae: 0.0774 - val\_accuracy: 0.9450 - val\_loss: 0.1507 - val\_mae: 0.0544

Epoch 4/50

**25/25** ━━━━━━━━━━━━━━━━━━━━ **0s** 9ms/step - accuracy: 0.9299 - loss: 0.1780 - mae: 0.0566 - val\_accuracy: 0.9250 - val\_loss: 0.1489 - val\_mae: 0.0516

Epoch 5/50

**25/25** ━━━━━━━━━━━━━━━━━━━━ **0s** 11ms/step - accuracy: 0.8991 - loss: 0.2702 - mae: 0.0648 - val\_accuracy: 0.9400 - val\_loss: 0.1704 - val\_mae: 0.0561

Epoch 6/50

**25/25** ━━━━━━━━━━━━━━━━━━━━ **0s** 9ms/step - accuracy: 0.9443 - loss: 0.1634 - mae: 0.0556 - val\_accuracy: 0.9550 - val\_loss: 0.1026 - val\_mae: 0.0378

Epoch 7/50

**25/25** ━━━━━━━━━━━━━━━━━━━━ **0s** 9ms/step - accuracy: 0.9474 - loss: 0.1428 - mae: 0.0445 - val\_accuracy: 0.9650 - val\_loss: 0.1025 - val\_mae: 0.0369

Epoch 8/50

**25/25** ━━━━━━━━━━━━━━━━━━━━ **0s** 10ms/step - accuracy: 0.9609 - loss: 0.1156 - mae: 0.0369 - val\_accuracy: 0.9900 - val\_loss: 0.0625 - val\_mae: 0.0265

Epoch 9/50

**25/25** ━━━━━━━━━━━━━━━━━━━━ **0s** 10ms/step - accuracy: 0.9647 - loss: 0.0955 - mae: 0.0306 - val\_accuracy: 0.9500 - val\_loss: 0.0948 - val\_mae: 0.0316

Epoch 10/50

**25/25** ━━━━━━━━━━━━━━━━━━━━ **0s** 11ms/step - accuracy: 0.9481 - loss: 0.1343 - mae: 0.0378 - val\_accuracy: 0.9800 - val\_loss: 0.0666 - val\_mae: 0.0260

Epoch 11/50

**25/25** ━━━━━━━━━━━━━━━━━━━━ **0s** 10ms/step - accuracy: 0.9692 - loss: 0.0944 - mae: 0.0315 - val\_accuracy: 0.9600 - val\_loss: 0.0794 - val\_mae: 0.0273

Epoch 12/50

**25/25** ━━━━━━━━━━━━━━━━━━━━ **0s** 9ms/step - accuracy: 0.9478 - loss: 0.1789 - mae: 0.0391 - val\_accuracy: 0.9800 - val\_loss: 0.0679 - val\_mae: 0.0259

Epoch 13/50

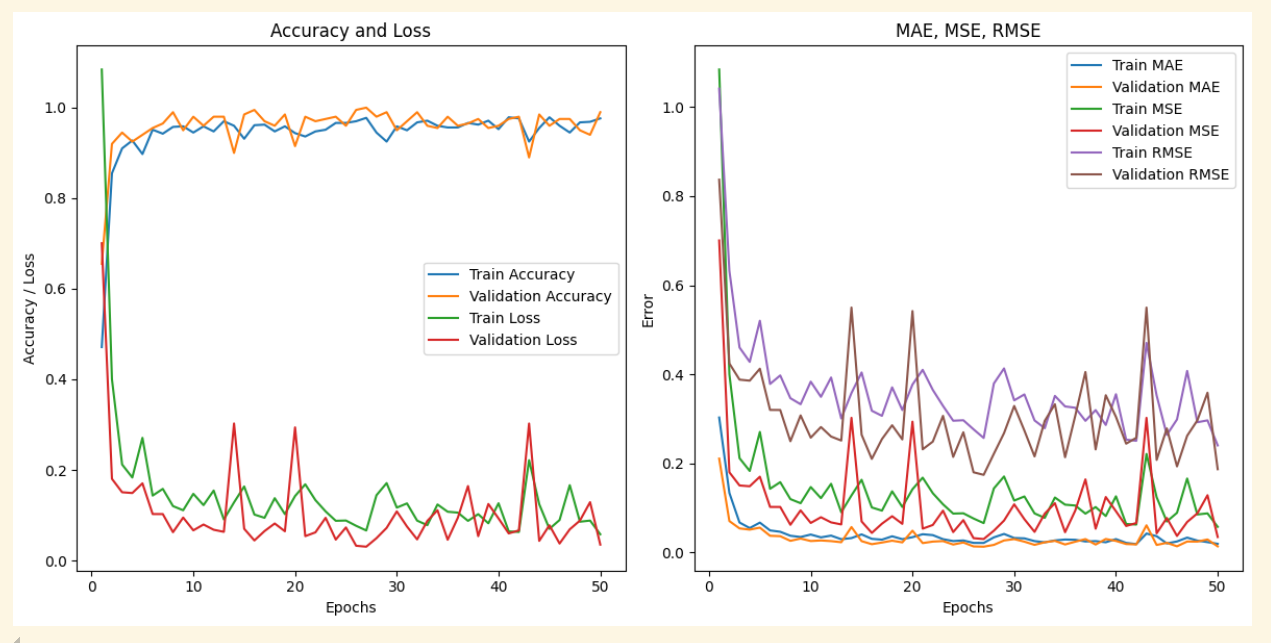
...

**25/25** ━━━━━━━━━━━━━━━━━━━━ **0s** 9ms/step - accuracy: 0.9704 - loss: 0.0858 - mae: 0.0234 - val\_accuracy: 0.9400 - val\_loss: 0.1288 - val\_mae: 0.0295

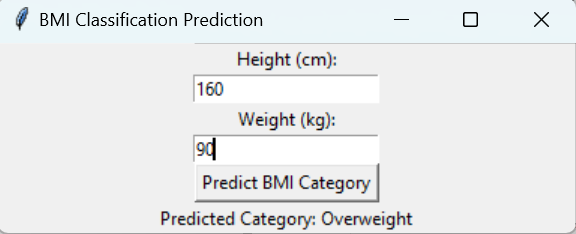
Epoch 50/50

**25/25** ━━━━━━━━━━━━━━━━━━━━ **0s** 11ms/step - accuracy: 0.9700 - loss: 0.0594 - mae: 0.0194 - val\_accuracy: 0.9900 - val\_loss: 0.0350 - val\_mae: 0.0140

**7/7** ━━━━━━━━━━━━━━━━━━━━ **0s** 5ms/step - accuracy: 0.9842 - loss: 0.0418 - mae: 0.0163



Giao diện demo:



1. **LSTM (Long Short-Term Memory)**:
   * Biến thể của RNN, giúp ghi nhớ thông tin trong thời gian dài.+ Nhận xét:

**CNN**

* **Điểm mạnh**:
  + Xử lý dữ liệu không gian rất hiệu quả (đặc biệt là hình ảnh).
  + Khả năng tự động học đặc trưng qua các tầng tích chập (*convolution*).
  + Tính toán nhanh và dễ mở rộng mô hình.
* **Điểm yếu**:
  + Không phù hợp cho dữ liệu tuần tự, chuỗi thời gian.
  + Yêu cầu định dạng dữ liệu đầu vào đặc biệt (hình ảnh hoặc tín hiệu).
* **Khi nào sử dụng**:
  + Phân loại, nhận diện đối tượng từ hình ảnh hoặc tín hiệu 2D/3D.

**RNN**

* **Điểm mạnh**:
  + Phù hợp với các bài toán tuần tự ngắn.
  + Ít tốn tài nguyên tính toán hơn so với LSTM.
* **Điểm yếu**:
  + Hiệu suất giảm sút với chuỗi dữ liệu dài do hiện tượng *vanishing gradient*.
  + Không thể học hiệu quả các mối quan hệ dài hạn trong dữ liệu.
* **Khi nào sử dụng**:
  + Bài toán đơn giản với chuỗi dữ liệu ngắn (ví dụ: phân tích cảm xúc từ đoạn văn bản ngắn).

**LSTM**

* **Điểm mạnh**:
  + Hiệu quả trong việc học và lưu trữ thông tin dài hạn.
  + Giải quyết được vấn đề vanishing gradient nhờ kiến trúc với các cổng nhớ (*forget*, *input*, *output*).
* **Điểm yếu**:
  + Đòi hỏi tài nguyên tính toán nhiều hơn so với RNN.
  + Huấn luyện lâu hơn.
* **Khi nào sử dụng**:
  + Bài toán cần lưu giữ mối quan hệ dài hạn trong dữ liệu (dự báo chuỗi thời gian dài, dịch máy, nhận dạng giọng nói).

CHƯƠNG 3: DỮ LIỆU TRONG DỰ ĐOÁN VÀ XỬ LÝ ẢNH

3.1 Dữ liệu ảnh, dữ liệu điểm

3.1.1 Dữ liệu ảnh

**Biểu diễn dữ liệu ảnh**

Dữ liệu ảnh thường được biểu diễn dưới dạng các tensor (mảng nhiều chiều) tùy thuộc vào định dạng và màu sắc của ảnh.

**a. Dạng ma trận (Grayscale Image)**

* Ảnh đen trắng được biểu diễn dưới dạng ma trận 2D.
* Kích thước: (height, width)
* Mỗi phần tử trong ma trận là một giá trị pixel, thường từ 0 (đen) đến 255 (trắng).

**b. Dạng tensor (RGB Image)**

* Ảnh màu được biểu diễn dưới dạng tensor 3D.
* Kích thước: (height, width, channels)
  + channels = 3 cho các kênh màu đỏ (Red), xanh lá (Green), và xanh dương (Blue).

**c. Các định dạng khác**

* **Ảnh RGBA**: Bao gồm thêm kênh alpha (độ trong suốt), kích thước là (height, width, 4).
* **Batch ảnh**: Khi xử lý nhiều ảnh cùng lúc, dữ liệu được biểu diễn dưới dạng tensor 4D, kích thước (batch\_size, height, width, channels).

**Chuẩn hóa dữ liệu ảnh**

Chuẩn hóa giúp giảm thiểu sai lệch trong dữ liệu và cải thiện hiệu quả của mô hình.

**a. Chia giá trị pixel cho 255**

* Thông thường, giá trị pixel nằm trong khoảng [0, 255].
* Chuẩn hóa bằng cách chia cho 255 để giá trị nằm trong [0, 1]

**b.Trừ trung bình và chia độ lệch chuẩn**

* Phương pháp này đặt giá trị pixel về dạng chuẩn hóa với trung bình bằng 0 và độ lệch chuẩn bằng 1.
  + Trung bình mean và độ lệch chuẩn std được tính từ tập dữ liệu.

**c. Chuẩn hóa theo miền giá trị [-1, 1]**

* Một số kiến trúc mạng (như ResNet, MobileNet) yêu cầu giá trị đầu vào trong khoảng [-1, 1].
* Chuyển đổi: normalized\_image = (image / 127.5) – 1

**Tăng cường dữ liệu (Data Augmentation)**

Trong nhiều trường hợp, trước khi chuẩn hóa, dữ liệu ảnh thường được tăng cường để cải thiện hiệu suất và khả năng tổng quát hóa của mô hình:

* **Các phương pháp tăng cường**:
  + Xoay (rotation), lật (flipping), dịch chuyển (translation).
  + Thay đổi độ sáng, độ tương phản.
  + Cắt (cropping) và thay đổi kích thước (resizing).
* Thư viện phổ biến:
  + TensorFlow/Keras (ImageDataGenerator)
  + PyTorch (torchvision.transforms)
  + Albumentations

**Chú ý khi chuẩn hóa**

* **Consistency**: Phải đảm bảo cách chuẩn hóa giống nhau giữa dữ liệu huấn luyện, kiểm tra và dự đoán.
* **Reverse Normalization**: Nếu cần hiển thị hoặc lưu ảnh, cần khôi phục giá trị pixel về dạng ban đầu.

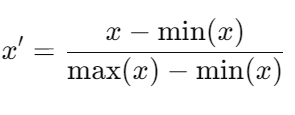
3.1.2 Dữ liệu điểm

**Chuẩn hóa dữ liệu điểm**

Việc chuẩn hóa rất quan trọng để đảm bảo rằng tất cả các cột dữ liệu có giá trị trong cùng một phạm vi, tránh cột có giá trị lớn lấn át các cột khác.

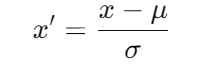
**a. Min-Max Scaling (Phổ biến nhất)**

* Đưa dữ liệu về khoảng [0, 1] hoặc một phạm vi tùy ý như [0, 10].
* Công thức



**b. Standardization (Z-Score Scaling)**

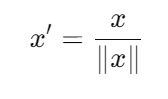
* Đưa dữ liệu về trung bình 0 và độ lệch chuẩn 1



* + μ: trung bình
  + σ: độ lệch chuẩn

**c. Normalization (L2 Norm)**

* Dùng để chuẩn hóa độ lớn của dữ liệu, hữu ích trong một số bài toán cụ thể như tính toán độ tương đồng:



3.2 Áp dụng cho tập dữ liệu ảnh:

Chuẩn hoá đưa giá trị các pixel về trong khoảng từ [0,1]:

3.3 Áp dụng cho tập dữ liệu điểm:

Chuẩn hoá data dùng công thức min-max

*# Chuẩn hoá data -> [0,1]*

cols\_to\_normalize = ["assignment\_1", "assignment\_2", "project", "final\_score"]

df\_normalized = finnal\_data.copy()

df\_normalized[cols\_to\_normalize] = (finnal\_data[cols\_to\_normalize] - finnal\_data[cols\_to\_normalize].min()) / (

finnal\_data[cols\_to\_normalize].max() - finnal\_data[cols\_to\_normalize].min()

)

*# Kết quả*

print(df\_normalized)

Kết quả thu được sau khi chuẩn hoá:

student\_id assignment\_1 assignment\_2 project final\_score

0 b1 0.900 0.777778 0.631579 0.600

1 b2 0.900 0.777778 0.631579 0.600

2 b3 1.000 1.000000 0.736842 0.725

3 b4 0.900 0.777778 0.631579 0.600

4 b5 1.000 0.888889 0.736842 0.750

.. ... ... ... ... ...

305 31\_151 0.825 0.861111 0.868421 0.800

306 31\_152 0.350 0.361111 0.631579 0.650

307 31\_153 0.725 0.722222 0.631579 0.650

308 31\_154 0.775 0.388889 0.789474 0.750

309 31\_155 0.000 0.000000 0.000000 0.000

3.4 Kết luận

**Dữ liệu ảnh**:

* **Biểu diễn**:  
  Dữ liệu ảnh được biểu diễn dưới dạng tensor, với các chiều thể hiện chiều cao, chiều rộng, và số kênh màu (grayscale: 2D; RGB/RGBA: 3D). Khi làm việc với nhiều ảnh cùng lúc, dữ liệu được biểu diễn thành tensor 4D với batch size.
* **Chuẩn hóa**:
  + Giá trị pixel thường nằm trong khoảng [0, 255], cần được chuẩn hóa về [0, 1] bằng cách chia cho 255 hoặc về [-1, 1] khi cần.
  + Chuẩn hóa giúp giảm thiểu sai lệch và tăng tốc độ hội tụ của mô hình, đặc biệt trong học sâu.
* **Tăng cường dữ liệu**:  
  Các kỹ thuật như xoay, lật, cắt ảnh, thay đổi độ sáng, và dịch chuyển giúp cải thiện khả năng tổng quát hóa và giảm overfitting.

**Dữ liệu điểm**:

* **Biểu diễn**:  
  Dữ liệu điểm thường là dữ liệu số, được biểu diễn dưới dạng mảng (array), ma trận, hoặc bảng (dataframe). Các cột dữ liệu đại diện cho các đặc trưng (ví dụ: bài tập, bài kiểm tra, thi giữa kỳ) và các dòng đại diện cho các mẫu dữ liệu.
* **Chuẩn hóa**:
  + **Min-Max Scaling**: Chuẩn hóa dữ liệu về khoảng [0, 1] để đảm bảo tính đồng nhất giữa các đặc trưng.
  + **Standardization (Z-score)**: Đưa dữ liệu về trung bình 0 và độ lệch chuẩn 1, phù hợp với các mô hình nhạy cảm với tỷ lệ.
  + **Trọng số**: Với dữ liệu điểm có trọng số, cần tính tổng điểm dựa trên tỷ lệ phần trăm đóng góp của từng cột vào tổng điểm.

**So sánh chung**:

* **Mục tiêu của chuẩn hóa**:  
  Cả dữ liệu ảnh và dữ liệu điểm đều cần được chuẩn hóa để cải thiện hiệu quả xử lý của mô hình, giúp các đặc trưng có phạm vi giá trị nhất quán, giảm tác động của các đặc trưng có giá trị lớn hơn.
* **Độ phức tạp**:  
  Dữ liệu ảnh phức tạp hơn do kích thước lớn (đặc biệt với ảnh độ phân giải cao) và có nhiều kênh thông tin. Trong khi đó, dữ liệu điểm thường nhỏ hơn và dễ xử lý.
* **Ứng dụng**:
  + Dữ liệu ảnh thường được sử dụng trong các bài toán như phân loại ảnh, nhận dạng đối tượng, hoặc phát hiện bất thường.
  + Dữ liệu điểm phù hợp với các bài toán dự đoán, phân loại (ví dụ: dự đoán điểm số, xếp loại).

**CHƯƠNG 4: DEEP LEARNING TRONG XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

#### **4.1.1. Text Generation (Sinh Văn Bản)**

Text generation là một ứng dụng phổ biến của Deep Learning trong xử lý ngôn ngữ tự nhiên (NLP), với khả năng tạo ra các đoạn văn bản có ý nghĩa dựa trên dữ liệu đầu vào. Các nội dung chính trong phần này gồm:

1. **Khái niệm về Text Generation**
   * Là quá trình mô phỏng khả năng sáng tạo văn bản của con người.
   * Dựa trên các mô hình học sâu như RNN (Recurrent Neural Networks), LSTM (Long Short-Term Memory), và Transformers.
2. **Phương pháp tiếp cận**
   * **Mô hình chuỗi thời gian (Sequential Models)**: Xử lý các chuỗi văn bản và dự đoán từ tiếp theo dựa trên ngữ cảnh trước đó.
   * **Các kiến trúc phổ biến**:
     + RNN và các biến thể như LSTM, GRU.
     + GPT (Generative Pre-trained Transformer) và các mô hình transformer khác (như BERT, T5).
3. **Ứng dụng của Text Generation**
   * Tạo nội dung tự động (viết báo, truyện, thơ).
   * Tạo hội thoại tự nhiên cho chatbot.
   * Hoàn thành văn bản hoặc hỗ trợ người dùng viết tài liệu.
4. **Cách triển khai mô hình Text Generation**
   * **Huấn luyện dữ liệu**: Dữ liệu văn bản được xử lý trước (tokenization, embedding).
   * **Sinh văn bản**: Mô hình dự đoán từ tiếp theo với các kỹ thuật như Greedy Search, Beam Search hoặc Sampling.
5. **Thách thức**
   * Hiểu ngữ nghĩa và ngữ cảnh phức tạp.
   * Kiểm soát nội dung sinh ra để tránh thông tin sai lệch hoặc không phù hợp.

4.1.2. DeepDream là một kỹ thuật xử lý hình ảnh độc đáo, sử dụng mạng nơ-ron học sâu để khuếch đại và trực quan hóa các mẫu đặc trưng mà mạng học được trong quá trình huấn luyện. Mặc dù chủ yếu liên quan đến thị giác máy tính (Computer Vision), DeepDream cũng có thể được ứng dụng trong nghệ thuật và phân tích hình ảnh.

1. **Khái niệm DeepDream**
   * Được phát triển bởi Google, DeepDream sử dụng các mạng CNN (Convolutional Neural Networks) để tạo ra hình ảnh giống như "ảo giác" từ các đặc trưng đã học.
   * Phương pháp này hoạt động bằng cách tối ưu hóa một hình ảnh đầu vào để khuếch đại các mẫu đặc trưng mà các lớp nơ-ron nhận diện.
2. **Nguyên lý hoạt động**
   * **Chọn một lớp trong mạng nơ-ron**: Ví dụ, lớp đầu hoặc lớp giữa của một mạng CNN.
   * **Kích hoạt lại hình ảnh đầu vào**: Hình ảnh ban đầu được điều chỉnh để tăng cường các mẫu mà lớp đã chọn "thích nghi".
   * **Kết quả cuối cùng**: Hình ảnh được khuếch đại, tạo ra các họa tiết và cấu trúc nghệ thuật kỳ lạ.
3. **Ứng dụng của DeepDream**
   * **Nghệ thuật số (Digital Art)**: Tạo ra các tác phẩm nghệ thuật trừu tượng và độc đáo.
   * **Hiểu mạng học sâu**: Trực quan hóa những gì một mạng nơ-ron học được từ dữ liệu hình ảnh.
   * **Khuếch đại các đặc trưng hình ảnh**: Hỗ trợ nghiên cứu và phân tích hình ảnh trong lĩnh vực thị giác máy tính.
4. **Thực thi DeepDream**
   * Sử dụng các thư viện như TensorFlow hoặc PyTorch để tạo DeepDream.
   * Cần có mô hình CNN được huấn luyện trước (như Inception).
5. **Thách thức và hạn chế**
   * Hình ảnh đầu ra có thể khó hiểu hoặc không liên quan đến nội dung gốc.
   * Quá trình tối ưu hóa có thể yêu cầu tài nguyên tính toán lớn.