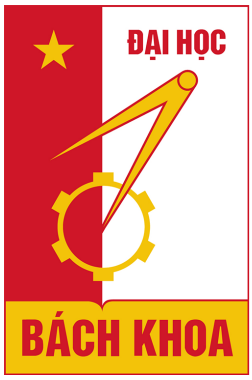
ĐẠI HỌC BÁCH KHOA HÀ NỘI
TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG



NHẬN DẠNG CẢM XÚC KHUÔN MẶT VỚI KIẾN TRÚC MẠNG RESNET-50

MÔN HỌC: PROJECT I

Giáo viên hướng dẫn: **Thầy Nguyễn Duy Tùng**
Sinh viên thực hiện: **Nguyễn Nam - 20220037**

Hà Nội, ngày 3 tháng 1 năm 2025

Mục lục

1 Giới thiệu

Nhận diện cảm xúc khuôn mặt là một bài toán quan trọng trong lĩnh vực Computer Vision, nhằm xác định cảm xúc con người dựa trên biểu cảm khuôn mặt. Các cảm xúc thường được nhận diện bao gồm: vui, buồn, giận, sợ hãi, ngạc nhiên, khinh miệt và bình thường. Trong báo cáo này, tôi sẽ xây dựng một mô hình nhận diện cảm xúc khuôn mặt với độ chính xác cao, sử dụng kiến trúc ResNet-50.

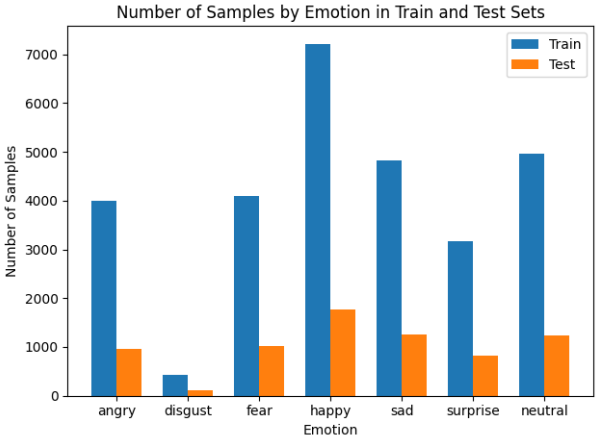
ResNet (Residual Network) được giới thiệu lần đầu tiên trong bài báo "Deep Residual Learning for Image Recognition" của nhóm tác giả Kaiming He, Xiangyu Zhang, Shaoqing Ren và Jian Sun, công bố tại hội nghị CVPR năm 2015. Mạng ResNet đã đạt được nhiều thành tựu quan trọng, đặc biệt là chiến thắng trong cuộc thi ImageNet 2015 với độ chính xác vượt trội. Điều này nhờ vào khả năng giải quyết hiệu quả vấn đề biến mất đạo hàm (vanishing gradient) trong các mạng sâu. Kiến trúc ResNet áp dụng cơ chế residual connection, cho phép các lớp trong mạng học các phép biến đổi dạng nhận diện, từ đó cải thiện hiệu suất và tính ổn định của mô hình.

Tôi lựa chọn ResNet-50 cho bài toán nhận diện cảm xúc khuôn mặt vì đây là một phiên bản tiêu chuẩn, cung cấp sự cân bằng tốt giữa độ phức tạp và hiệu suất. Với 50 lớp, ResNet-50 đủ mạnh để xử lý các đặc trưng phức tạp từ biểu cảm khuôn mặt, đồng thời vẫn đảm bảo khả năng triển khai hiệu quả trong môi trường tính toán thực tế. Việc tái xây dựng và áp dụng kiến trúc này trong dự án không chỉ giúp tôi nâng cao kỹ năng nghiên cứu và phát triển mô hình AI, mà còn mở rộng hiểu biết về cách thiết kế và triển khai các mạng học sâu.

2 Dữ liệu và xử lý dữ liệu

2.1 Bộ dữ liệu FER-2013

Bộ dữ liệu FER-2013 (Facial Expression Recognition 2013) là một trong những tập dữ liệu phổ biến nhất trong lĩnh vực nhận diện cảm xúc khuôn mặt. Được công bố tại hội nghị ICML 2013, FER-2013 bao gồm 35.887 ảnh grayscale với kích thước 48x48 pixel, được gắn nhãn cho bảy cảm xúc chính: Vui (Happy), Buồn (Sad), Giận (Angry), Sợ hãi (Fear), Ngạc nhiên (Surprise), Khinh miệt (Disgust), và Bình thường (Neutral). Dữ liệu được chia thành hai tập: tập huấn luyện (28.709 ảnh) và tập kiểm tra (7.178 ảnh).



Hình 1: Biểu đồ số lượng mẫu

Đặc điểm nổi bật:

- **Quy mô hợp lý:** Gần 36.000 ảnh đủ lớn để đào tạo các mô hình học sâu như ResNet-50.
- **Chuẩn hóa kích thước:** Tất cả các ảnh có kích thước cố định (48x48 pixel), giúp đơn giản hóa quá trình tiền xử lý.
- **Tính đa dạng:** Hình ảnh đến từ nhiều góc chụp và mức độ phức tạp khác nhau, giúp mô hình học được các đặc trưng biểu cảm phong phú.

Nhờ những đặc điểm này, bộ dữ liệu FER-2013 đã được lựa chọn cho dự án này, đặc biệt khi tài nguyên tính toán còn hạn chế.

2.2 Tiền xử lý dữ liệu

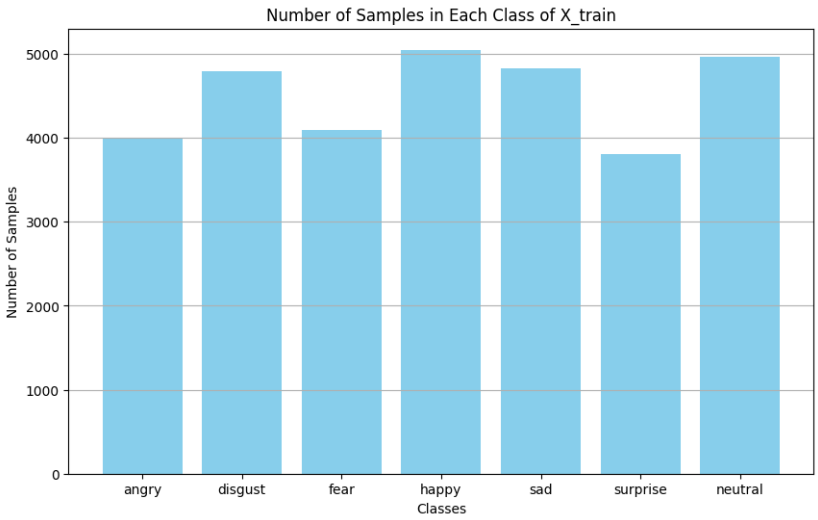
Trong Hình 1, có thể rút ra một số nhận xét như sau:

- Số lượng mẫu của các lớp `disgust` và `surprise` thấp hơn so với các lớp khác. Đặc biệt, lớp `disgust` chỉ chiếm khoảng 1/6 so với số lượng trung bình của các lớp còn lại.
- Lớp `happy` có số lượng mẫu cao hơn so với các lớp khác.

Để xử lý dữ liệu cho mô hình, có thể triển khai hai hướng tiếp cận chính:

- **Cân bằng dữ liệu:** Đối mặt với sự mất cân bằng nghiêm trọng giữa các lớp, tôi áp dụng các kỹ thuật cân bằng dữ liệu như:

- *Tăng cường dữ liệu:* Tăng số lượng mẫu cho lớp `disgust` lên 6 lần và lớp `surprise` thêm 20% thông qua các kỹ thuật tăng cường như lật, xoay, và thay đổi độ sáng.
- *Giảm dữ liệu:* Giảm 20% số lượng mẫu của lớp `happy` bằng cách ngẫu nhiên loại bỏ một phần dữ liệu.



Hình 2: Số lượng mẫu các lớp trong tập train sau khi được cân bằng

- **Thêm trọng số cho các lớp:** Trong quá trình huấn luyện mô hình, tôi bổ sung trọng số cho các lớp dựa trên tỷ lệ mẫu của từng lớp. Các lớp có số lượng mẫu nhỏ hơn được gán trọng số cao hơn để giảm thiểu ảnh hưởng của sự mất cân bằng dữ liệu đến quá trình tối ưu hóa mô hình.

Quá trình xử lý này đóng vai trò quan trọng trong việc đảm bảo chất lượng dữ liệu đầu vào cho mô hình, từ đó cải thiện hiệu suất nhận diện cảm xúc. Ngoài ra, việc kết hợp cả hai hướng tiếp cận trên giúp tăng độ chính xác và giảm thiểu sự thiên vị đối với các lớp có ít dữ liệu.

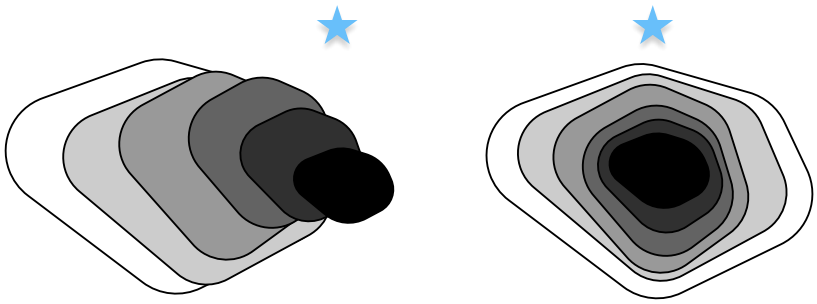
3 Mô hình

3.1 Mạng phần dư ResNet

Xét $\mathcal{F}$ là một lớp các hàm mà một kiến trúc mạng cụ thể (cùng với tốc độ học và các siêu tham số khác) có thể đạt được. Nói cách khác, với mọi hàm số $f \in \mathcal{F}$, luôn tồn tại tập tham số W có thể tìm được bằng việc huấn luyện trên một tập dữ liệu phù hợp. Giả sử f^* là hàm cần tìm. Ta cố gắng tìm hàm số $f^*_{\mathcal{F}}$ tốt nhất trong $\mathcal{F}$ mà xấp xỉ f^* tốt nhất. Bằng các thuật toán học, ta thường tìm hàm trên bằng cách giải bài toán tối ưu:

$$f^* = \text{argmin}_{f \in \mathcal{F}} L(X, Y, f) \quad \text{đối với} \quad f \in \mathcal{F}.$$

Một cách hiển nhiên rằng, nếu ta thiết kế được một kiến trúc $\mathcal{F}'$ mạnh mẽ hơn $\mathcal{F}$, cụ thể hơn là $\mathcal{F} \subset \mathcal{F}'$, thì chắc chắn sẽ thu được hàm $f^*_{\mathcal{F}'}$ tốt hơn $f^*_{\mathcal{F}}$. Tuy nhiên, việc thêm các tầng không phải lúc nào cũng tăng tính biểu diễn của mạng mà đôi khi còn tạo ra những thay đổi rất khó lường.

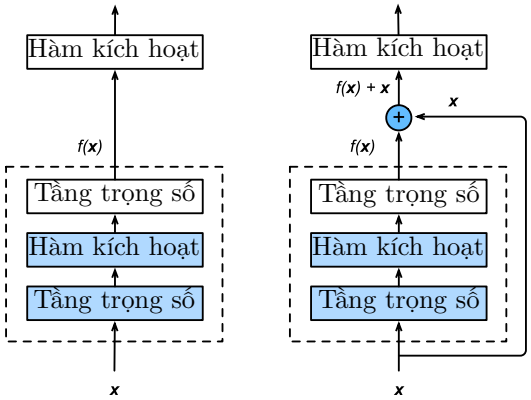


các lớp hàm số tổng quát các lớp hàm số lồng nhau

Hình 3: Hình trái: Các lớp hàm số tổng quát. Khoảng cách đến hàm cần tìm f^* (ngôi sao), trên thực tế có thể tăng khi độ phức tạp tăng lên. Hình phải: với các lớp hàm số lồng nhau, điều này không xảy ra.

Chỉ khi các lớp hàm lớn hơn chứa các lớp nhỏ hơn, thì mới đảm bảo việc tăng thêm các tầng sẽ tăng khả năng biểu diễn của mạng. Ý tưởng trọng tâm của ResNet là mỗi tầng được thêm vào nên có thành phần là hàm số đồng nhất. Có nghĩa rằng, nếu ta huấn luyện tầng mới được thêm vào một ánh xạ đồng nhất $f(\mathbf{x}) = \mathbf{x}$, thì mô hình mới sẽ hiệu quả ít nhất bằng mô hình ban đầu.

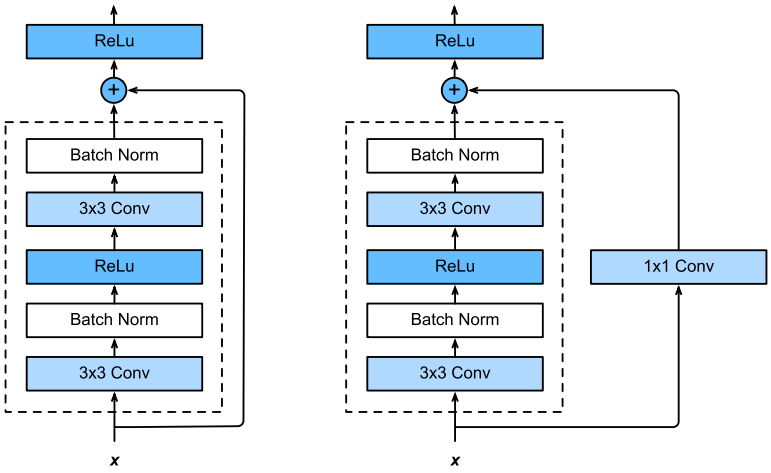
3.1.1 Khối phần dư (residual block)



Hình 4: Sự khác biệt giữa một khối thông thường (trái) và một khối phần dư (phải). Trong khối phần dư, các tích chập có thể được nối tắt.

Thay vì học trực tiếp ánh xạ $F(\mathbf{x})$ từ đầu vào $\mathbf{x}$, mô hình sẽ học phần dư, tức là $F(\mathbf{x}) - \mathbf{x}$. Trên thực tế, ánh xạ phần dư thường dễ tối ưu hơn, vì chỉ cần đặt $F(\mathbf{x}) - \mathbf{x} = 0$. Cách tiếp cận này giúp gradient truyền ngược dễ dàng hơn, giảm nguy cơ gặp phải hiện tượng *vanishing gradient* (độ dốc biến mất).

ResNet được thiết kế với các tầng tích chập kích thước 3×3, tương tự như kiến trúc VGG. Mỗi khối phần dư bao gồm hai tầng tích chập 3×3 với số kênh đầu ra tương đương. Sau mỗi tầng tích chập, có một tầng chuẩn hóa theo batch và một hàm kích hoạt ReLU. Đầu vào được đưa qua khối phần dư và sau đó cộng với chính nó trước khi đi qua hàm kích hoạt ReLU cuối cùng. Thiết kế này yêu cầu đầu ra của hai tầng tích chập phải có cùng kích thước với đầu vào, nhằm đảm bảo có thể thực hiện phép cộng.



Hình 5: Trái: khối ResNet thông thường; Phải: Khối ResNet với tầng tích chập 1×1.

Nếu cần thay đổi số lượng kênh hoặc kích thước bước trong khối phần dư, một tầng tích chập 1×1 sẽ được thêm vào để điều chỉnh kích thước đầu vào ở nhánh ngoài cho phù hợp.

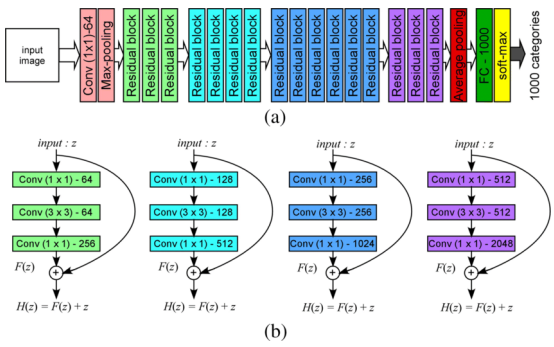
3.1.2 Kiến trúc mạng ResNet-50

ResNet-50 là một mạng nơ-ron sâu thuộc họ ResNet được thiết kế để giải quyết vấn đề suy giảm gradient trong các mạng nơ-ron sâu thông qua việc sử dụng các khối residual. Mạng bao gồm 50 lớp và được chia thành các thành phần chính như sau:

- **Lớp tích chập và pooling ban đầu:**
  - Lớp đầu vào là một lớp tích chập với kernel kích thước 7×7, stride là 2, và 64 bộ lọc, tiếp theo là một lớp Batch Normalization (chuẩn hóa batch) và hàm kích hoạt ReLU.
  - Một lớp Max Pooling với kernel kích thước 3×3 và stride là 2 được áp dụng để giảm kích thước của đặc trưng đầu ra.
- **Các khối residual:** Kiến trúc ResNet-50 bao gồm 4 giai đoạn chính, mỗi giai đoạn chứa một số khối residual với số lượng bộ lọc tăng dần:
  - *Giai đoạn 1:* Gồm 3 khối residual, mỗi khối có các lớp tích chập 1×1, 3×3, và 1×1 với số lượng bộ lọc lần lượt là 64, 64, và 256.
  - *Giai đoạn 2:* Gồm 4 khối residual, mỗi khối có các lớp tích chập 1×1, 3×3, và 1×1 với số lượng bộ lọc lần lượt là 128, 128, và 512.
  - *Giai đoạn 3:* Gồm 6 khối residual, mỗi khối có các lớp tích chập 1×1, 3×3, và 1×1 với số lượng bộ lọc lần lượt là 256, 256, và 1024.
  - *Giai đoạn 4:* Gồm 3 khối residual, mỗi khối có các lớp tích chập 1×1, 3×3, và 1×1 với số lượng bộ lọc lần lượt là 512, 512, và 2048.

  Mỗi khối residual sử dụng liên kết tắt (*shortcut connection*) để truyền trực tiếp đầu vào qua các tầng, giúp giải quyết vấn đề mất mát thông tin khi mạng trở nên sâu hơn.
- **Lớp Global Average Pooling và Fully Connected:**
  - Sau khi qua các khối residual, một lớp Global Average Pooling được sử dụng để giảm chiều không gian, biến đầu ra thành một vector đặc trưng có kích thước nhỏ hơn.
  - Vector này được đưa qua lớp fully connected với số đầu ra bằng số lớp cần phân loại (7 lớp trong bài toán nhận diện cảm xúc).
- **Hàm softmax:** Lớp đầu ra sử dụng hàm softmax để tính xác suất dự đoán cho mỗi lớp.

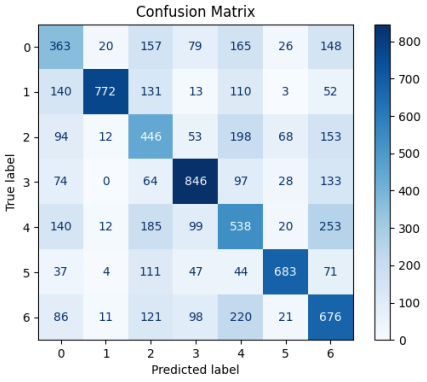
Hình 6 minh họa cấu trúc chi tiết của mạng ResNet-50, bao gồm các khối residual và các thành phần chính.



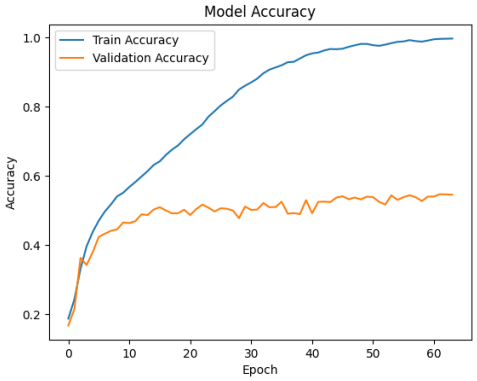
Hình 6: Cấu trúc chi tiết mạng ResNet-50

3.2 Triển khai mô hình vào bài toán

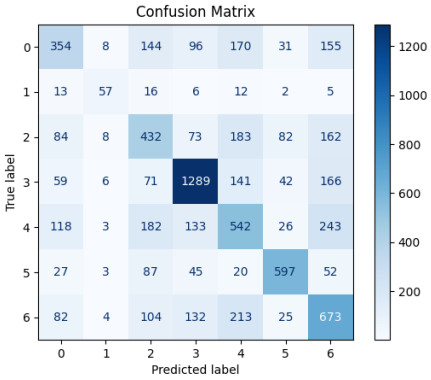
Với việc triển khai kiến trúc mạng ResNet-50, kết quả thu được qua hai biểu đồ sau:



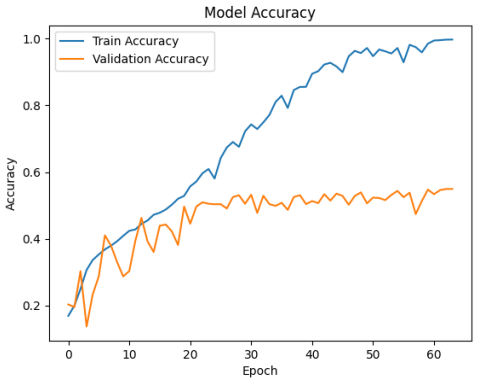
Hình 7: Ma trận nhầm lẫn với dữ liệu được cân bằng.



Hình 8: Biểu độ độ chính xác của mô hình.



Hình 9: Ma trận nhầm lẫn với các class được đánh trọng số.



Hình 10: Biểu độ độ chính xác của mô hình.

Mặc dù độ chính xác (tính theo số dự đoán đúng trên tổng dự đoán) của hai hướng triển khai là như nhau, xấp xỉ khoảng 55%. Tuy nhiên, khi nhìn vào ma trận nhầm lẫn, ta có thể thấy rằng mô hình triển khai với các class được đánh trọng số có vẻ tốt hơn. Cụ thể, bảng đánh giá chi tiết với các lớp của mô hình này được tính toán như sau:

Class	Precision	Recall	F1-score	Support
Angry	0.48	0.37	0.42	958
Disgust	0.64	0.51	0.57	111
Fear	0.42	0.42	0.42	1024
Happy	0.73	0.73	0.73	1774
Sad	0.42	0.43	0.43	1247
Surprise	0.74	0.72	0.73	831
Neutral	0.46	0.55	0.50	1233
Accuracy			0.55	7278
Macro avg	0.56	0.53	0.54	7178
Weighted avg	0.55	0.55	0.55	7178

Bảng 1: Bảng thống kê precision, recall, f1-score và support cho các lớp cảm xúc.

4 Kết luận

Mặc dù độ chính xác tổng thể của cả hai hướng triển khai đều chỉ đạt xấp xỉ 55%, kết quả này vẫn có tiềm năng được cải thiện trong tương lai. Khi phân tích chi tiết qua ma trận nhầm lẫn và các chỉ số đánh giá như precision, recall và F1-score, mô hình triển khai với các class được đánh trọng số đã thể hiện sự vượt trội hơn, đặc biệt ở các lớp có số lượng mẫu nhỏ như `disgust` và `surprise`. Điều này cho thấy việc sử dụng trọng số cho các lớp là một phương pháp hiệu quả để giảm thiểu ảnh hưởng của sự mất cân bằng dữ liệu.

Tuy nhiên, kết quả hiện tại vẫn còn hạn chế bởi các yếu tố sau:

- **Dữ liệu huấn luyện:** Bộ dữ liệu hiện tại chưa đủ lớn và đa dạng để mô hình có thể học được các đặc trưng phức tạp của từng cảm xúc.
- **Tài nguyên tính toán:** Quá trình huấn luyện mô hình bị giới hạn bởi tài nguyên tính toán, dẫn đến việc không thể thử nghiệm các kiến trúc mô hình phức tạp hơn hoặc áp dụng các kỹ thuật tối ưu hóa hiệu quả hơn.

Trong tương lai, các giải pháp cải thiện có thể bao gồm:

1. Thu thập và mở rộng bộ dữ liệu với nhiều mẫu hơn cho các lớp ít dữ liệu, đồng thời bổ sung dữ liệu đa dạng về ánh sáng, góc chụp và biểu cảm khuôn mặt.
2. Sử dụng các mô hình sâu hơn như Transformer hoặc các kiến trúc hiện đại khác trong nhận diện cảm xúc.
3. Tối ưu hóa quy trình huấn luyện thông qua việc tận dụng các tài nguyên tính toán mạnh mẽ hơn, ví dụ như GPU/TPU, hoặc áp dụng kỹ thuật huấn luyện phân tán.
4. Kết hợp thêm các kỹ thuật tăng cường dữ liệu tiên tiến và sử dụng phương pháp học chuyển giao (transfer learning) từ các mô hình tiền huấn luyện trên các bộ dữ liệu lớn hơn.

Kết quả thu được từ dự án này cung cấp nền tảng ban đầu để khám phá và phát triển các mô hình nhận diện cảm xúc. Với những cải thiện trong tương lai, mô hình có thể đạt được độ chính xác cao hơn, từ đó ứng dụng hiệu quả hơn trong các lĩnh vực thực tiễn như chăm sóc khách hàng, phân tích tâm lý học, ...

5 Tài liệu tham khảo

References

[1] Đắm mình vào học sâu (bản dịch tiếng việt của cuốn sách Dive to Deep Learning)

https://d2l.aivivn.com/index.html

[2] Deep Residual Learning for Image Recognition

https://arxiv.org/pdf/1512.03385

[3] FER-2013 Dataset

https://www.kaggle.com/datasets/msambare/fer2013/data