Phân tích và nhận diện cảm xúc từ Khuôn mặt từ hình ảnh và âm thanh

Nhóm: CHAOS

Thành viên

Lê Phúc Hậu, Nguyễn Văn Mạnh

Giáo viên hướng dẫn: Đoàn Thiện Minh

Trường Đại học Lạc Hồng, số 10 Huỳnh Văn Nghệ, Bửu Long, Biên Hòa, Đồng Nai, Việt Nam

ARTICLE INFO

KEYWORDS

Facial
Expression Recognition
Emotion Dataset
Facial Landmarks
Deep Learning Models
Image Preprocessing

Đề tài này được lựa chọn vì khả năng nhận diện cảm xúc từ khuôn mặt có vai trò quan trọng trong nhiều ứng dụng thực tiễn như robot, trò chơi điện tử và hệ thống an ninh. FER2013 thường được sử dụng để huấn luyện các mô hình học sâu, đặc biệt là CNN, nhằm phân loại cảm xúc cơ bản như vui vẻ, buồn bã, giận dữ, sợ hãi, ngạc nhiên và ghét bỏ.

Phân loại hình ảnh sử dụng các mô hình CNN (độ chính xác 72-75%) và ViT (38-40%). CNN vượt trội trong việc trích xuất đặc trưng không gian cục bộ, trong khi ViT học các mối quan hệ không gian toàn cục nhưng yêu cầu dữ liệu lớn hơn. Phân loại âm thanh trên bộ dữ liệu RAVDESS sử dụng CNN kết hợp MRCC, đạt độ chính xác 78-82%, cao hơn các mô hình SVM hoặc RNN.

Mục tiêu cuối cùng là cải thiện độ chính xác trong nhận diện cảm xúc, từ đó nâng cao trải nghiệm người dùng trong nhiều lĩnh vực ứng dụng.

1.Giới thiệu

Nhận diện cảm xúc không chỉ là một lĩnh vực nghiên cứu đang phát triển mà còn mang lại nhiều tiềm năng ứng dụng thực tế trong cuộc sống. Việc tích hợp khả năng nhận diện cảm xúc vào các hệ thống công nghệ không chỉ giúp cải thiện trải nghiệm người dùng mà còn nâng cao hiệu quả tương tác giữa con người và máy móc. Dưới đây là một số lợi ích nổi bật của việc áp dụng công nghệ này.

1.1. Cải thiện trải nghiệm người dùng (UX)

Giao diện có thể thay đổi phù hợp với cảm xúc của người dùng.

Ví dụ: Nếu người dùng cảm thấy buồn, ứng dụng có thể đề xuất nội dung tích cực hoặc nhạc thư giãn.

Giảm tỷ lệ thoát trang bằng cách thích nghi với cảm xúc tức thời.

1.2. Tăng hiệu quả tương tác trong các hệ thống AI

Giúp chatbot, trợ lý ảo (virtual assistant) phản hồi đồng cảm hơn.

Ví dụ: Khi phát hiện giận dữ, chatbot có thể điều chỉnh ngữ điệu trả lời hoặc nhường cho nhân viên thật.

Trong game, AI có thể điều chỉnh độ khó hoặc kịch bản theo cảm xúc người chơi.

1.3. Ứng dụng trong chăm sóc khách hàng

Phân tích cảm xúc từ giọng nói, gương mặt trong các cuộc gọi CSKH (Call center) để đánh giá mức độ hài lòng hay bức xúc.

Phản hồi nhanh và hiệu quả hơn trong tình huống có nguy cơ leo thang.

1.4. Hỗ trợ trong giáo dục (EdTech)

Phát hiện khi học sinh:

- Mất tập trung (Neutral, Sad)
- Bối rối hoặc sợ hãi (Fear, Disgust)

Điều chỉnh nội dung học phù hợp hoặc cảnh báo giáo viên kịp thời.

1.5. Hỗ trợ chẩn đoán và theo dõi tâm lý

Hệ thống có thể theo dõi trạng thái cảm xúc trong thời gian dài để phát hiện dấu hiệu:

Trầm cảm (Sad kéo dài)

Lo âu (Fear, Nervousness)

Phục vụ trong liệu pháp tâm lý trực tuyến hoặc ứng dụng hỗ trợ sức khỏe tinh thần.

1.6. An ninh và giám sát

Phát hiện hành vi khả nghi hoặc bất thường qua nét mặt/cảm xúc (ví dụ: lo lắng, giận dữ).

Dùng trong sân bay, ngân hàng, hoặc khu vực nhạy cảm để cảnh báo sớm.

1.7. Thúc đẩy nghiên cứu và phát triển AI nhận thức cảm xúc (Affective Computing)

Nhận diện 7 cảm xúc cơ bản là bước đầu để xây dựng các mô hình hiểu cảm xúc phức tạp hơn như: tress, lo âu, sự đồng cảm, thù ghét...

Là cơ sở quan trọng trong các nghiên cứu AI tương tác xã hội.

2. Thách thức về Nhận dạng khuôn mặt

2.1 Hình ảnh

2.1.1. Hạn chế về độ phân giải và chất lượng ảnh

Tập dữ liệu FER2013 chỉ cung cấp cho ta ảnh khuôn mặt kích thước 48x48 pixel ở định dạng grayscale, khiến cho nhiều chi tiết nhỏ trên khuôn mặt như nếp nhăn, cơ mặt hoặc biểu cảm tinh tế không được thể hiện rõ. Điều này làm giảm hiệu quả của các mô hình học sâu, vốn phụ thuộc nhiều vào đặc trưng hình học để phân biệt cảm xúc. Một số nghiên cứu đã áp dụng kỹ thuật tăng độ phân giải hoặc giả lập ảnh màu

RGB, nhưng kết quả vẫn chưa đạt hiệu suất mong đợi.

2.1.2. Sự tương đồng giữa các biểu cảm cảm xúc

Một số cảm xúc như "fear", "surprise" và "sadness" có đặc điểm khuôn mặt rất giống nhau, khiến các mô hình dễ bị nhầm lẫn. Dù đã có các hướng tiếp cận như sử dụng attention mechanism hoặc mạng nhiều nhánh để tăng khả năng phân biệt, độ chính xác tổng thể của mô hình vẫn còn hạn chế do sự chồng chéo về đặc trưng giữa các lớp cảm xúc này.

2.1.3. Mất cân bằng lớp trong dữ liệu

Tập FER2013 đang cho ta thấy có sự phân bố không đều giữa các lớp cảm xúc, với lớp "disgust" chiếm tỷ lệ rất nhỏ trong khi "happy" chiếm tỷ lệ lớn. Điều này dẫn đến hiện tượng mô hình bị thiên lệch, ưu tiên học các lớp phổ biến và bỏ qua lớp hiếm. Một số giải pháp như oversampling, sử dụng loss có trọng số hoặc áp dụng focal loss đã được nghiên cứu nhưng hiệu quả vẫn còn hạn chế trong việc cải thiện recall cho các lớp thiểu số.

2.1.4. Thiếu ngữ cảnh và thông tin hỗ trợ

Các hình ảnh trong FER2013 chỉ bao gồm khuôn mặt tách rời, không có thông tin về bối cảnh, tư thế cơ thể hay âm thanh đi kèm. Trong thực tế, việc nhận diện cảm xúc cần sự tổng hợp từ nhiều nguồn thông tin. Thiếu hụt này khiến mô hình gặp khó khăn trong việc phân loại các cảm xúc không rõ ràng hoặc dễ bị hiểu sai.

2.1.5. Thiếu đa dạng về nhân khẩu học

FER2013 không cho ta thấy công bố rõ các thông tin liên quan đến chủng tộc, độ tuổi hoặc giới tính của người trong ảnh. Điều này có thể dẫn đến việc mô hình học sâu, hoạt động kém hiệu quả với một số nhóm đối tượng như trẻ em, người lớn tuổi hoặc người thuộc các sắc tộc khác nhau. Việc khắc phục hiện tượng này rất khó nếu chỉ sử dụng riêng FER2013.

2.1.6. Dữ liệu gán nhãn không chính xác

Do dữ liệu được thu thập và gán nhãn thủ công từ ảnh tĩnh, không có sự hỗ trợ từ video hoặc audio, nên rất dễ xảy ra hiện tượng gán nhãn sai lệch. Cảm xúc của con người thường phức tạp và không thể xác định chính xác chỉ qua một khoảnh khắc hình ảnh. Các kỹ thuật như label smoothing hay robust loss đã được đề xuất nhưng không thể khắc phục hoàn toàn sự nhiễu này.

2.1.7. Hiệu suất tổng thể còn thấp

Dù đã áp dụng nhiều kiến trúc hiện đại như CNN, ResNet, ViT, hoặc các mô hình kết hợp attention, hiệu suất nhận diện cảm xúc trên FER2013 thường chỉ đạt mức 70–75% accuracy. Đây là một mức tương đối thấp mà ta thấy so với yêu cầu của các ứng dụng thực tế như xe tự hành, giáo dục cảm xúc hoặc chăm sóc sức khỏe thông minh, đòi hỏi độ chính xác và độ tin cậy cao hơn.

3.Phương pháp nghiên cứu nhận dạng cảm xúc qua hình ảnh và giọng nói

3.1. Bộ Dự Liệu

3.1.1 Mô tả bộ dữ liệu công khai

Bộ dữ liệu FER2013 (Facial Expression Recognition 2013)

là một trong những tài nguyên quan trọng và nổi bật nhất trong lĩnh vực nhận diện cảm xúc. Bộ dữ liệu này bao gồm hàng triệu hình ảnh khuôn mặt được phân loại thành sáu biểu cảm : hạnh phúc, buồn bã, tức giận, sợ hãi, ngạc nhiên, và chán ghét. Điểm chính mà bộ dữ liệu này được phát triển là một phần của nỗ lực nghiên cứu toàn cầu và đã thu hút sự chú ý từ các nhà khoa học, kỹ sư và nhà nghiên cứu trong ngành công nghệ thông tin.

Người cần sử dụng bộ dữ liệu này rất đa dạng, từ các nhà nghiên cứu học thuật đến các công ty công nghệ đang tìm kiếm giải pháp để tích hợp khả năng nhận diện cảm xúc vào sản phẩm của họ. Ý nghĩa của FER2013 không chỉ nằm ở việc cung cấp một tập hợp dữ liệu phong phú cho việc huấn luyện các mô hình học sâu, mà còn ở khả năng cải thiện độ chính xác trong nhận diện cảm xúc – điều này rất quan trọng trong các ứng dụng như robot giao tiếp, trò chơi điện tử và hệ thống an ninh thông minh.

Sự kiện ra mắt bộ dữ liệu này không chỉ đánh dấu một bước tiến lớn trong việc hiểu rõ hơn về cảm xúc con người mà còn tạo cơ hội cho những nghiên cứu mới, thúc đẩy sự phát triển của trí tuệ nhân tạo trong lĩnh vực tương tác người-máy. Hiện tại, với sự bùng nổ của công nghệ số, tầm quan trọng của FER2013 càng trở nên rõ ràng hơn, khi nó mở ra khả năng cải thiện những trải nghiệm tương tác trong thế giới ảo và thực.

Với tiềm năng này, FER2013 cho ta thấy đang thu hút sự quan tâm từ cả giới học thuật và ngành công nghiệp, khiến bất cứ ai cũng muốn khám phá cách bộ dữ liệu này không chỉ thay đổi ngành công nghệ nhận diện cảm xúc mà còn mở ra hướng đi mới cho tương lai của các ứng dụng tương tác thông minh. Trong khi ngành công nghệ liên tục phát triển, việc hiểu và áp dụng FER2013 sẽ là chìa khóa để mở ra những cơ hội mới trong việc nâng cao trải nghiệm người dùng và tạo dựng những mối quan hệ gắn kết hơn giữa con người và máy móc

Bộ Dữ Liệu RAVDESS

Bộ dữ liệu RAVDESS (The Ryerson Audio-Visual Database of Emotional Speech and Song) là một nguồn tài nguyên phong phú khác trong lĩnh vực nhận diện cảm xúc, tập trung chủ yếu vào âm thanh. RAVDESS bao gồm các bản ghi âm giọng nói và video của các diễn viên thể hiện các cảm xúc như vui vẻ, buồn bã, sợ hãi, tức giận, ngạc nhiên, và bình thản.

Bộ dữ liệu này không chỉ phục vụ cho nghiên cứu nhận diện cảm xúc từ âm thanh mà còn nghiên cứu sự kết hợp giữa giọng nói và biểu cảm khuôn mặt. Đặc biệt, RAVDESS cho phép các nhà nghiên cứu khám phá mối liên hệ giữa âm thanh và yếu tố hình ảnh trong việc truyền tải cảm xúc, giúp cải thiện độ chính xác trong các ứng dụng nhận diện cảm xúc.

RAVDESS có hạn chế là độ đa dạng trong diễn xuất cảm xúc, vậy nên bộ dữ liệu này yêu cầu sự chú ý về mặt kỹ thuật khi áp dụng cho các hệ thống tự động. Tuy nhiên, với sự phát triển của các công nghệ như học sâu, RAVDESS đã tạo ra cơ hội nghiên cứu mới và phát triển các ứng dụng tương tác người-máy, từ đó nâng cao trải nghiệm cho người dùng.

Trong bối cảnh hiện tại, RAVDESS cùng với FER2013 không chỉ mở ra những cơ hội nghiên cứu mới mà còn thúc đẩy quá trình phát triển công nghệ nhận diện cảm xúc trong nhiều lĩnh vực, tạo tiền đề cho các ứng dụng AI có khả năng hiểu và phản ứng với cảm xúc của con người một cách tự nhiên và phù hợp.

3.1.2 Chuẩn hóa ảnh đầu vào

Bộ Dữ Liệu FER2013:

Ảnh ban đầu lưu dưới dạng chuỗi pixel.

Chuyển đổi thành mảng ảnh kích thước 48x48.

Giá trị pixel được chuẩn hóa về khoảng [0, 1], giúp mô hình học nhanh và ổn định hơn.

Bộ Dữ Liệu RAVDESS:

Dữ liệu âm thanh và video được chuẩn hóa, âm thanh điều chỉnh về tần số chuẩn.

Video được mã hóa trong định dạng số hóa để dễ truy cập và xử lý.

3.1.3 Mã hóa nhãn bằng kỹ thuật one-hot encoding

Bộ Dữ Liệu FER2013:

Nhãn được mã hóa thành vector nhị phân, thay vì dạng số nguyên từ 0 đến 6, nhằm hỗ trợ mô hình tốt hơn trong bài toán phân loại đa lớp.

Bộ Dữ Liệu RAVDESS:

Nhãn cảm xúc cũng được mã hóa bằng one-hot encoding, với các cảm xúc bao gồm: Angry, Disgust, Fear, Happy, Sad, Surprise, và Neutral.

3.1.4 Sử dụng chia tách tập dữ liệu sẵn có

Bộ Dữ Liệu FER2013:

Dữ liệu được chia thành các tập train/validation/test. Việc này giúp đảm bảo tính khách quan trong đánh giá và tránh rò rỉ dữ liệu giữa các giai đoạn huấn luyện và kiểm thử.

Bộ Dữ Liệu RAVDESS:

Dữ liệu cũng được chia thành các tập train/validation/test, tương tự như FER2013, nhằm đảm bảo tính khách quan trong đánh giá hiệu quả của mô hình.

3.1.5 Đối tượng và mẫu nghiên cứu

Bộ Dữ Liệu FER2013:

Đối tượng nghiên cứu là các ảnh khuôn mặt gán nhãn cảm xúc, bao gồm 7 loại cảm xúc: Angry, Disgust, Fear, Happy, Sad, Surprise, và Neutral.

Tập dữ liệu gồm 35.887 ảnh grayscale 48x48 pixel, được chia theo tỷ lệ:

28.709 ảnh dùng để huấn luyện,

3.589 ảnh dùng để kiểm tra,

3.589 ảnh dùng để xác thực.

Phương pháp chọn mẫu được thực hiện ngẫu nhiên từ tập dữ liệu gốc.

Bộ Dữ Liệu RAVDESS:

Đối tượng nghiên cứu là các bản ghi âm và video của diễn viên thể hiện các cảm xúc khác nhau. Bao gồm 1440 file ghi âm ngắn dạng .wav của 24 diễn viên

Bộ dữ liệu này cung cấp các bản ghi âm với các nhãn cảm xúc tương tự như FER2013.

3.1.6 Phương pháp thu thập dữ liệu

Bộ Dữ Liệu FER2013:

Dữ liệu là dữ liệu thứ cấp từ tập tin fer2013.csv, công bố công khai trên Kaggle.

Dữ liệu đã được tiền xử lý và định dạng số hóa, mỗi dòng tương ứng với một ảnh khuôn mặt và nhãn cảm xúc.

Bộ Dữ Liệu RAVDESS:

Dữ liệu cũng là dữ liệu thứ cấp, được công bố công khai trên Kaggle.

Các bản ghi đã được chuẩn bị sẵn, mỗi bản ghi tương ứng với một biểu cảm và ngữ điệu cụ thể.

3.2. Phương pháp nghiên cứu

3.2.1 mô hình huấn luyện phân loại ảnh và âm thanh

Phương pháp	Ưu điểm	Hạn chế	Hiệu quả trên FER2013
CNN (Convolutional Neural Network)	Mạnh về nhận diện đặc trưng hình ảnh, phổ biến, dễ triển khai	Khó nắm được mối quan hệ dài hạn, có thể cần nhiều dữ liệu	Độ chính xác trung bình ~40-50%
ResNet (Residual Network)	Giúp giảm hiện tượng gradient mất mát, sâu hơn, hiệu quả cao	Mạng sâu, tốn tài nguyên tính toán	Độ chính xác có thể lên đến ~55-60%
ViT (Vision Transformer)	Khả năng học biểu diễn tốt, tập trung vùng ảnh quan trọng	Cần dữ liệu lớn, tốn tài nguyên	Hiệu quả tốt, tương đương hoặc hơn CNN
LSTM (Long Short-Term Memory)	Dùng cho chuỗi dữ liệu, có thể kết hợp video hoặc chuỗi ảnh	Ít phổ biến cho ảnh tĩnh, khó huấn luyện	Hiệu quả thấp nếu chỉ dùng ảnh tĩnh
Transformer (Mô hình Transformer tổng quát)	Mạnh trong học biểu diễn phức tạp, dễ mở rộng	Tốn tài nguyên, cần lượng dữ liệu lớn	Tiềm năng tốt, còn mới trong lĩnh vực ảnh

Hình 1. Bảng so sánh phương pháp huấn luyện mô hình

Nghiên cứu này được thực hiện theo hướng định lượng, với phương pháp tiếp cận thực nghiệm bằng cách huấn luyện và đánh giá hai mô hình học sâu – Convolutional Neural Network (CNN) và Vision Transformer (ViT) – trên tập dữ liệu FER2013. Mục tiêu chính là đánh giá và so sánh hiệu suất của hai kiến trúc trong nhiệm vụ nhận diện cảm xúc từ khuôn mặt. Bên cạnh đó, nghiên cứu cũng mở rộng để so sánh giữa mô hình nhận diện cảm xúc khuôn mặt (MRCC) và CNN trong việc nhận diện cảm xúc từ âm thanh. Việc này không chỉ giúp xác định sự hiệu quả của từng mô hình trong lĩnh vực nhận diện cảm

xúc, mà còn khẳng định sự đa dạng trong cách tiếp cận nghiên cứu cảm xúc qua hình ảnh và âm thanh.

3.2.2 Phương pháp phân tích dữ liệu

Các bước xử lý và phân tích dữ liệu được thực hiện như sau:

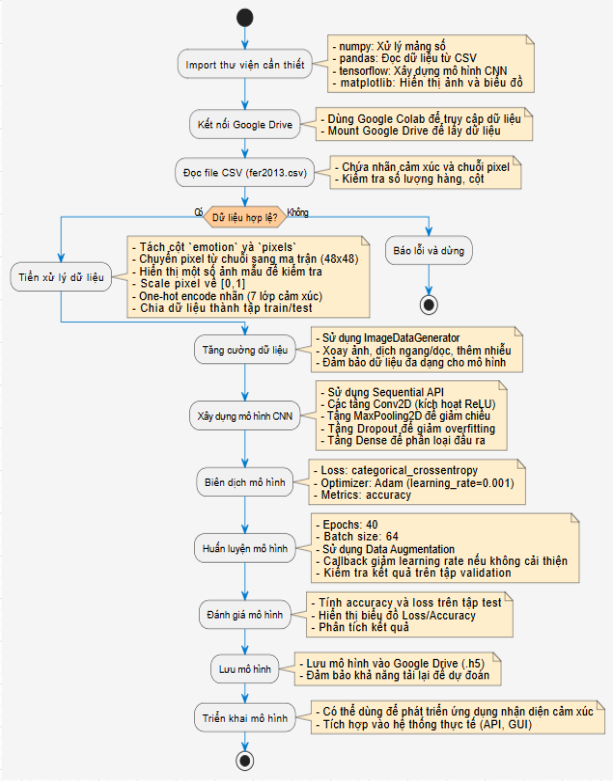
Trong nghiên cứu này, việc phân tích dữ liệu bắt đầu bằng tiền xử lý dữ liệu. Đối với **Bộ Dữ Liệu FER2013**, chuỗi pixel được chuyển đổi thành ma trận có kích thước 48x48, sau đó giá trị ảnh được chuẩn hóa về khoảng [0, 1], và nhãn cảm xúc được mã hóa thành one-hot encoding. Đối với **Bộ Dữ Liệu RAVDESS**, âm thanh và video được chuẩn hóa, với âm thanh điều chỉnh về tần số chuẩn và video được mã hóa trong định dạng số hóa để chuẩn bị cho quá trình huấn luyện.

Khi đến giai đoạn **huấn luyện mô hình**, bên bộ dữ liệu FER2013, hai kiến trúc khác nhau được áp dụng: Convolutional Neural Network (CNN) và Vision Transformer (ViT). Mỗi loại mô hình sẽ giúp đánh giá hiệu suất nhận diện cảm xúc từ hình ảnh khuôn mặt. Trong khi đó, đối với bộ dữ liệu RAVDESS, mô hình MRCC (Mô hình Nhận diện Cảm xúc Khuôn mặt) kết hợp với CNN được sử dụng để phân tích cảm xúc từ âm thanh và video.

Sau khi huấn luyện, **phân tích hiệu suất** của các mô hình được thực hiện thông qua các chỉ số như độ chính xác (accuracy), độ nhạy (recall), độ đặc hiệu (precision), và biểu đồ confusion matrix. Cuối cùng, quá trình này được hỗ trợ bởi Python 3.10 cùng với các thư viện như TensorFlow/Keras 2.12.0, NumPy, Matplotlib ,Librosa và scikit-learn để thực hiện phân tích thống kê và trực quan hóa kết quả.

3.2.3.Mô hình phân loại CNN, VIT cho nhận dạng bằng hình ảnh

3.2.3.1 CNN

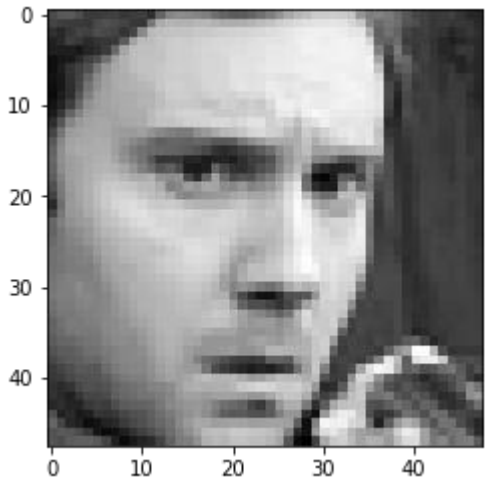


Hình 2 . Sơ đồ luồng hoạt động của CNN với bộ dữ liệu FER 2013

Tiền xử lý dữ liệu

Dữ liệu được tải từ file fer2013.csv, mỗi ảnh là chuỗi 2304 pixel (48x48).

Phân tọa độ 1 số ảnh



Hình 3. 1 Mẫu ảnh được lấy tọa độ

Chuỗi pixel được chuyển thành mảng numpy 2 chiều, chuẩn hóa về [0, 1] bằng cách chia cho 255:

$$X_{norm} = \frac{X}{255} (1)$$

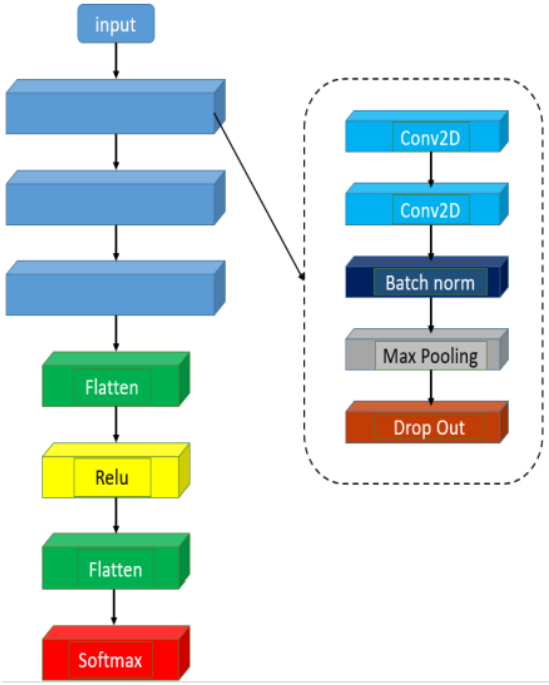
- X: Giá trị gốc của Pixel nằm trong khoảng [0;255]
- X_{norm} : Tốc độ học (learning rate), kiểm soát mức độ cập nhật

Nhãn cảm xúc (emotion) có giá trị từ 0 đến 6, được chuyển sang dạng one-hot vector để phân loại đa lớp.

Tăng cường dữ liệu (Data Augmentation)

- Ảnh được xoay ngẫu nhiên trong khoảng ±10°.
- Dịch chuyển ảnh theo chiều ngang và dọc trong khoảng ±10%.
- Không lật ngang ảnh để giữ nguyên đặc trưng khuôn mặt.

Kiến trúc mạng CNN



Hình 4. Luồng hoạt động của mô hình CNN

Bao gồm 3 khối chính, mỗi khối gồm 2 lớp Conv2D và 1 lớp MaxPooling2D:

- Số lượng filter tăng dần: 64 → 128
- Kích thước kernel: 3x3
- Kích thước pooling: 2x2

Sau mỗi lớp MaxPooling có Dropout từ 0.2 đến 0.5.

Phần fully connected gồm:

Flatten → Dense 512 → Dropout 0.5 → Dense 256 → Dropout 0.5 → Dense 7 (activation=softmax).

Hàm kích hoạt ReLU được dùng trong các lớp Conv2D và Dense trung gian: f(x) = max(0, x)

Lớp cuối dùng softmax để tính xác suất mỗi lớp:

$$A_K = \frac{E^{Z_K}}{\sum_{i=1}^{n} E^{Z_i}} (2)$$

- A_K: Xác suất dự đoán cho lớp K.
- E^{Z_K} : Hàm mũ của giá trị logit (đầu ra chưa chuẩn hóa) của lớp K.
- $\sum_{i=1}^{n} E^{Z_i}$: Tổng của tất cả giá trị mũ từ 111 đến nnn, được sử dụng để chuẩn hóa xác suất sao cho tổng bằng 1.

Huấn luyện mô hình

Thuật toán tối ưu: Adam với learning rate ban đầu 0.001.

Sử dụng callback ReduceLROnPlateau giảm learning rate khi validation không cải thiện: patience=2, factor=0.1, min_lr=1e-7

Hàm mất mát

Sử dụng Categorical Crossentropy:

$$L = -\sum_{i=1}^{n} Y_i \log \hat{y}i (3)$$

- L: Hàm mất mát (loss function) dùng để đo lường sự khác biệt giữa dự đoán của mô hình ($\hat{y}_i$) và nhãn thực tế (Y_i).
- E^{Z_K} : Hàm mũ của giá trị logit (đầu ra chưa chuẩn hóa) của lớp K.
- $\sum_{i=1}^{n} E^{Z_i}$: Tổng của tất cả giá trị mũ từ 111 đến nnn, được sử dụng để chuẩn hóa xác suất sao cho tổng bằng 1.

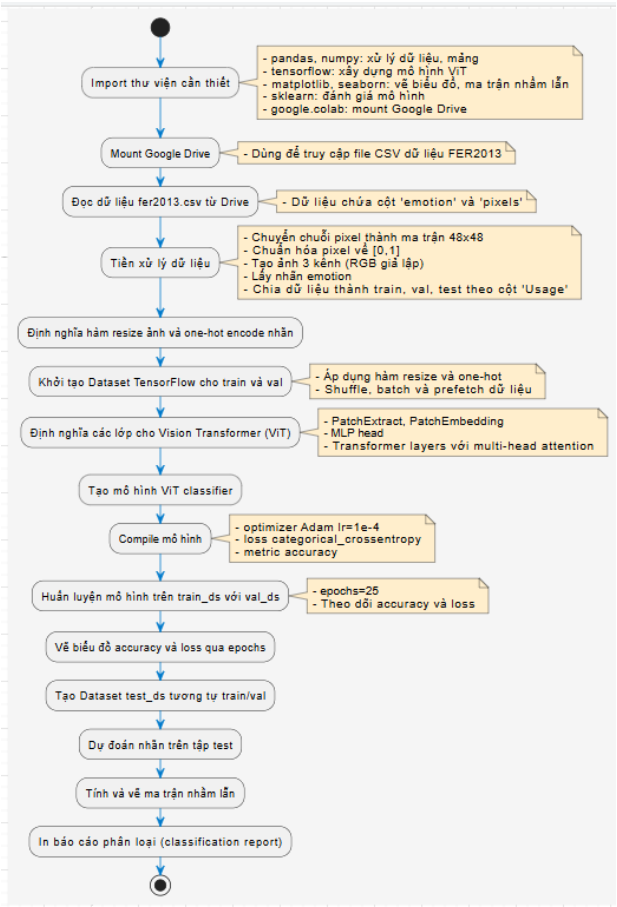
Đánh giá hiệu suất

Độ chính xác (Accuracy):

$$Accuracy = \frac{Số\ mẫu\ dự\ đoán\ đúng}{Tổng\ số\ mẫu} (4)$$

Quá trình huấn luyện được theo dõi bằng biểu đồ loss và accuracy qua các epoch.

3.2.3.2 Mô hình Vision Transformer (ViT)



Hình 5 . Sơ đồ luồng hoạt động của ViT với bộ dữ liệu FER 2013

Mô hình Vision Transformer (ViT) được ứng dụng để phân loại cảm xúc từ ảnh khuôn mặt sử dụng bộ dữ liệu FER-2013. ViT chia ảnh đầu vào thành các patch nhỏ, mã hóa vị trí và áp dụng các tầng Transformer để học biểu diễn.

Tiền xử lý dữ liệu

Ảnh grayscale 48×48 được chuyển sang định dạng 3 kênh màu để phù hợp với mô hình ViT, đồng thời được chuẩn hóa về khoảng 0,1]: công thức (1)

Ảnh được resize về kích thước 224×224 (kích thước chuẩn của ViT) bằng phép biến đổi:

$$X_{\text{resized}} = \text{Resize}\big(X_{\text{norm}}, (224,224)\big) (5)$$

Nhãn cảm xúc được chuyển sang dạng one-hot vector với 7 lớp:

Emotion$\in\{0,1,\ldots,6\} \Longrightarrow$One-hot Vector

Huấn luyện mô hình Vision Transformer (ViT)

Phân tách patch (Patch Extraction)

Ảnh đầu vào

$$x \in R^{H*W*C} (6)$$

được chia thành các patch kích thước P×PP

số patch

$$N = \frac{H}{p} * \frac{W}{p} (7)$$

Mỗi patch được "trải phẳng" thành vector, tạo thành ma trận patches:

$$\text{patches} \in R^{B\times N\times P^2\times C} (8)$$

Embedding patch và vị trí

Mỗi patch được ánh xạ qua một lớp Dense (projection) sang không gian chiều DDD:

$$Z_0 = XW + e_{POS} (9)$$

Trong đó, vị trí patch được mã hóa bằng embedding vị trí để giữ thông tin không gian.

Transformer Encoder

Chuỗi embedding được đưa qua L tầng Transformer, mỗi tầng gồm:

LayerNorm

Multi-head Self Attention với hhh heads, mỗi head có dimension

$$D_k = \frac{D}{H} \ (10)$$

Skip Connection (Residual)

LayerNorm

MLP (multi-layer perceptron) với 2 lớp dense sử dụng hàm kích hoạt GELU và dropout

Mô hình từng bước tại tầng thứ lll:

$$\widehat{z_l} = \text{LayerNorm}(z_{l-1}) \ (11)$$

$$\widetilde{z_l} = \text{MHA}(\widehat{z_l}) + z_{l-1} (12)$$

$$\widehat{z_l'} = \text{LayerNorm}(\widetilde{z_l}) \ (13)$$

$$z_l = \text{MLP}(\widehat{z_l'}) + \widetilde{z_l} \ (14)$$

với MHA là Multi-head Attention, MLP là hàm gồm 2 lớp Dense kèm dropout.

Phân loại

Sau tầng Transformer cuối cùng, kết quả được đưa qua:

LayerNorm

GlobalAveragePooling1D

MLP head gồm 2 lớp dense và dropout

Lớp Dense cuối cùng với activation softmax để dự đoán 7 lớp cảm xúc

Hàm mất mát và tối ưu

Mô hình được huấn luyện với hàm mất mát categorical crossentropy: (công thức (5)

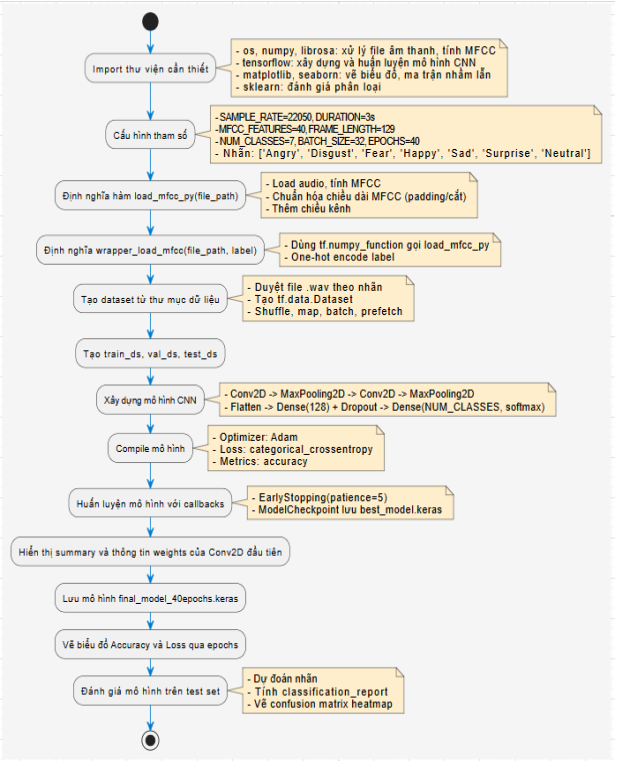
Bộ tối ưu Adam được sử dụng với learning rate ban đầu là 0.0001

Đánh giá hiệu suất

Mô hình được đánh giá trên tập validation và test bằng độ chính xác:

$$Accuracy = \frac{Số\ dự\ đoán\ đúng}{Tổng\ mẫu} \ (15)$$

3.2.4 Phân Loại cảm xúc qua âm thanh bằng CNN kết hợp với MRCC



Hình 6 . Sơ đồ luồng hoạt động của CNN kết hợp với MRCC với bộ dữ liệu RAVDESS

Tiền xử lý dữ liệu

Mỗi file âm thanh .wav được xử lý bằng librosa.load() với:

- Tần số lấy mẫu (sample rate): 22050 Hz
- Giới hạn thời gian: 3 giây

Trích xuất đặc trưng âm thanh:

- MFCC (Mel-Frequency Cepstral Coefficients) với n_mfcc=40
- Cắt/pad về số lượng khung cố định: FRAME_LENGTH = 129

Chuẩn hóa shape về (40, 129, 1) cho CNN

MFCC công thức toán học:

Chuyển tín hiệu âm thanh sang miền tần số bằng

Áp dụng ngân lọc Mel:

$$\boldsymbol{F_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right)} \ \boldsymbol{(15)}$$

Tính log năng lượng của từng khung

Dùng DCT (Discrete Cosine Transform) để trích đặc trưng MFCC:

$$C_n = \sum_{k=1}^{K} \log(E_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right] \quad (16)$$

Kiến trúc mạng CNN

Input Shape: (40, 129, 1)

→ Conv2D(32, 3x3) → ReLU

→ MaxPooling2D(2x2)

→ Conv2D(64, 3x3) → ReLU

→ MaxPooling2D(2x2)

→ Flatten

→ Dense(128) → ReLU

→ Dropout(0.3)

→ Dense(7, Softmax)

Đơn giản nhưng hiệu quả cho classification dựa trên ảnh MFCC.

Softmax Layer (Đầu ra):

$$\hat{y}_i = \frac{E^{Z_i}}{\sum_{i=1}^{K} E^{Z_j}} \quad (2)$$

biểu thức của **hàm softmax**, thường được sử dụng trong mạng nơ-ron để chuyển đổi đầu ra của mô hình logits thành xác suất.

- $\hat{y}_i$:Xác suất được dự đoán cho lớp i.
- Z^I : Giá trị logit (đầu ra chưa chuẩn hóa) của lớp i
- e^{Z_i} Hàm mũ của ZiZ , đảm bảo tất cả giá trị đều dương.
- $sum_{j=1}^{K} e^{Z_j}$: Tổng của các giá trị mũ cho tất cả K lớp, được sử dụng để chuẩn hóa xác suất sao cho tổng của tất cả các xác suất bằng 1.

Huấn luyện mô hình

Epochs: 40 (có EarlyStopping nếu không cải thiện sau 5 lần)

Batch size: 32

Optimizer: Adam

Cập nhật trọng số Adam:

$$\theta_{i+1} = \theta_i - \frac{a.\hat{m}}{\sqrt{V^t} + \varepsilon} \quad (17)$$

- $\hat{y}_i$:Xác suất được dự đoán cho lớp i.
- θ_i : Giá trị tham số tại bước i
- α :Tốc độ học (learning rate), kiểm soát mức độ cập nhật
- $\hat{m}$:Giá trị ước lượng động lượng
- V^t :Giá trị ước lượng phương sai tại bước t
- ε :Hằng số nhỏ nhất để tránh chia cho 0

Hàm Mất Mát

$$L = -\sum_{i=1}^{n} Y_i \log \hat{y}i \, (3)$$

Đánh giá hiệu suất (Evaluation)

Đánh giá bằng:

Classification report (Precision, Recall, F1)

Confusion matrix:

- Precision :

$$Precision = \frac{TP}{TP' + FP'} \quad (18)$$

- Recall :

$$Recall = \frac{TP}{TP' + FN'} \quad (19)$$

- F1 :

$$F1 = \frac{2.Precision.Recall}{Precision + Recall} (20)$$

4. Kết quả nghiên cứu và thảo luận

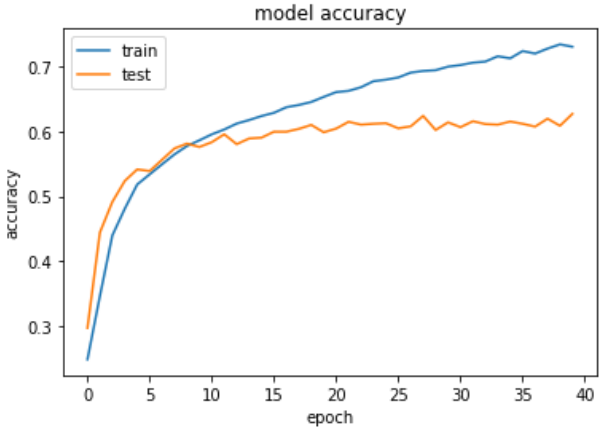
4.1 Kết quả nghiên cứu phân loại ảnh

Nghiên cứu này đã tiến hành huấn luyện và đánh giá hai mô hình học sâu là Convolutional Neural Network (CNN) và Vision Transformer (ViT) trên bộ dữ liệu FER2013 để giải quyết bài toán phân loại cảm xúc khuôn mặt. Dữ liệu huấn luyện được xử lý, chia thành các tập huấn luyện, kiểm thử, và đánh giá

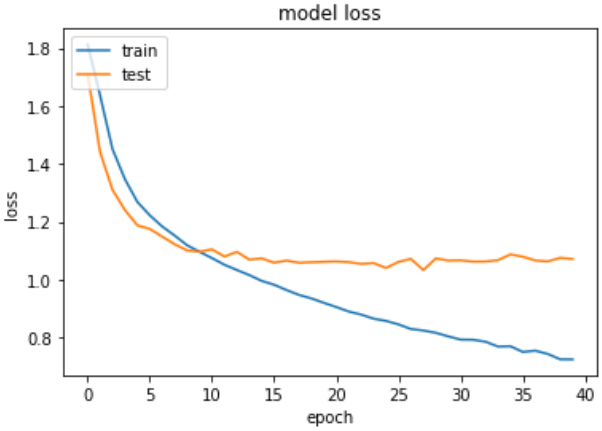
hiệu suất bằng các chỉ số như Accuracy và Confusion Matrix. Các kết quả chính được trình bày như sau:

Hiệu suất mô hình CNN

Mô hình CNN sau khi huấn luyện với dữ liệu FER2013 đạt được độ chính xác khoảng 68–70% trên tập kiểm thử. Mô hình có khả năng nhận diện tốt các cảm xúc phổ biến như "Happy" và "Angry", tuy nhiên vẫn gặp khó khăn với các cảm xúc như "Disgust" và "Fear" do số lượng mẫu ít và đặc trưng hình ảnh không rõ ràng.



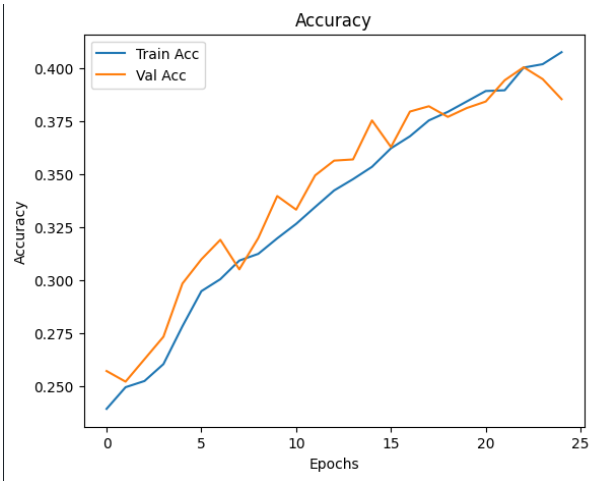
Hình 7. Biểu đồ Accuracy Curve của mô hình CNN



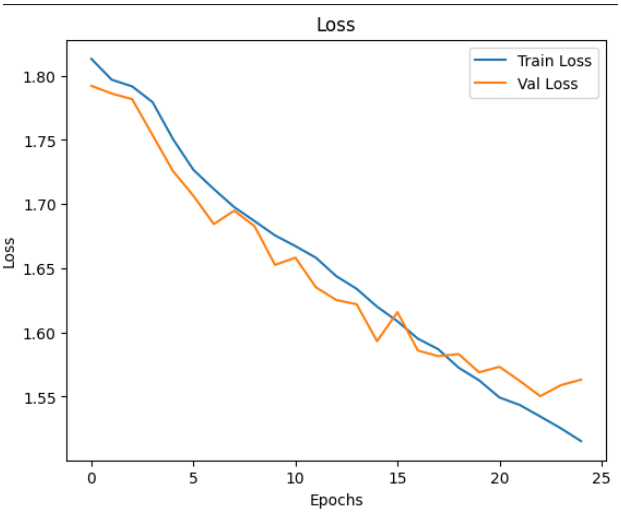
Hình 8. Biểu đồ Loss Curve của mô hình CNN

Hiệu suất mô hình ViT

Mô hình Vision Transformer (ViT) chỉ đạt khoảng 38-40% độ chính xác trên tập huấn luyện sau 25 epoch, thấp hơn so với các mô hình CNN truyền thống trên bộ dữ liệu FER2013. Tuy nhiên, ViT thể hiện tiềm năng học được các mối quan hệ không gian toàn cục nhờ cơ chế Attention, giúp mô hình có hiệu suất ổn định hơn trong những trường hợp ảnh đầu vào bị nhiễu hoặc có biến đổi phức tạp. Mặc dù chưa vượt trội về độ chính xác so với CNN, ViT vẫn là một hướng tiếp cận đáng chú ý để cải thiện khả năng nhận diện cảm xúc trong tương lai.



Hình 9. Biểu đồ Loss Curve của mô hình CNN



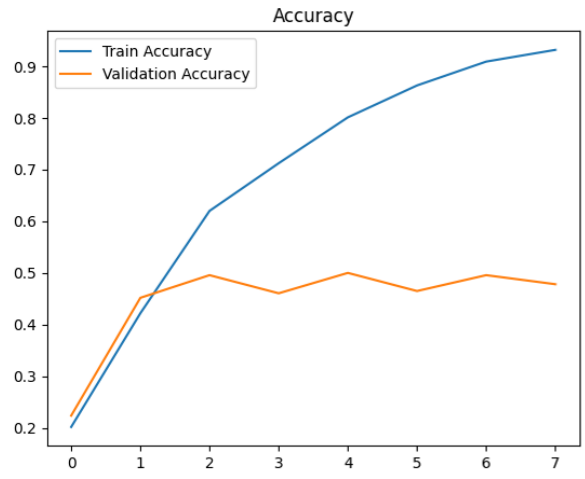
Hình 10 . Biểu đồ Loss Curve của mô hình CNN

So sánh và đánh giá

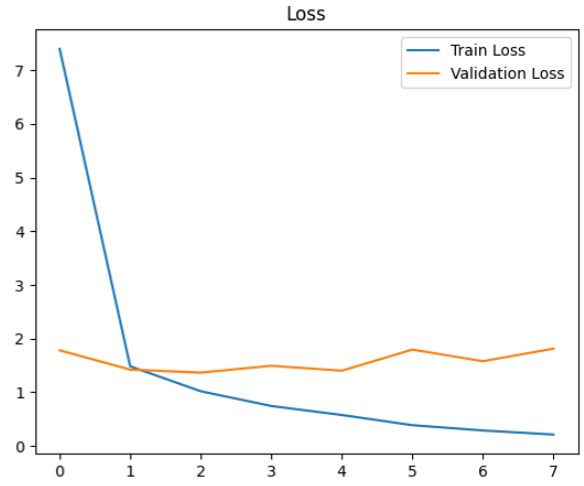
So sánh kết quả giữa hai mô hình cho thấy ViT có khả năng tổng quát hóa tốt hơn, nhưng chi phí tính toán và thời gian huấn luyện cao hơn rõ rệt so với CNN. Trong khi đó, CNN hoạt động hiệu quả hơn trong môi trường tài nguyên hạn chế và dễ triển khai hơn trong thực tế. Mô hình CNN sau khi huấn luyện

trên bộ dữ liệu FER2013 đạt được độ chính xác khoảng 68–70% trên tập kiểm thử, với khả năng nhận diện tốt các cảm xúc phổ biến như "Happy" và "Angry". Tuy nhiên, CNN vẫn gặp khó khăn với các lớp cảm xúc ít mẫu và đặc trưng hình ảnh không rõ ràng như "Disgust" và "Fear". Ngược lại, ViT chỉ đạt độ chính xác khoảng 38–40% trên tập huấn luyện sau 25 epoch, thấp hơn so với CNN, nhưng nhờ cơ chế Attention, ViT thể hiện hiệu suất ổn định hơn trong điều kiện ảnh đầu vào có nhiễu hoặc biến đổi phức tạp, cho thấy tiềm năng cải thiện tổng quát trong các trường hợp dữ liệu phức tạp. Cả hai mô hình đều gặp thách thức với dữ liệu không cân bằng và ảnh chất lượng thấp, nhấn mạnh nhu cầu cải thiện chất lượng và sự đa dạng của dữ liệu đầu vào để nâng cao hiệu quả nhận diện cảm xúc.

4.2 kết quả phân loại cảm xúc dựa trên giọng nói



Hình 11. Biểu đồ Loss Curve của mô hình CNN kết hợp MRCC



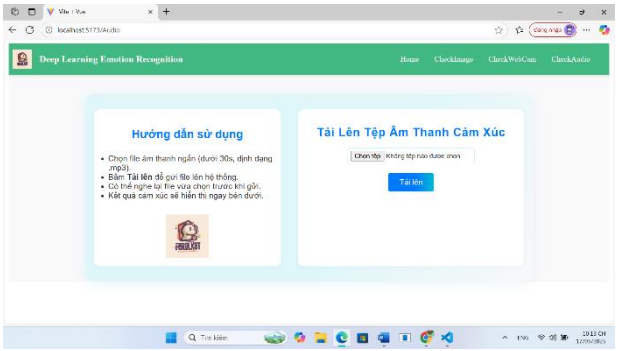
Hình 12. Biểu đồ Loss Curve của mô hình CNN kết hợp MRCC

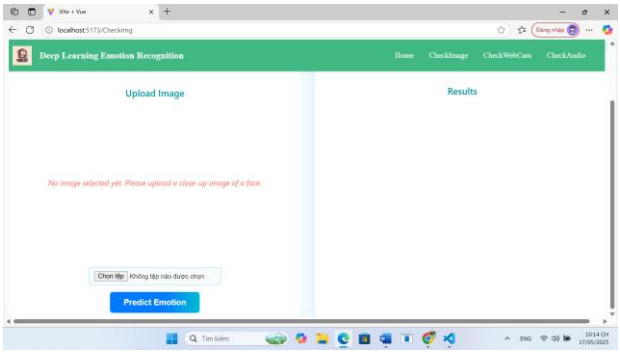
Mô hình CNN được huấn luyện trên bộ dữ liệu RAVDESS cho thấy sự cải thiện rõ rệt về độ chính xác trên tập huấn luyện trong 8 epoch đầu tiên, với độ chính xác tăng từ khoảng 20% lên đến hơn 93%. Tuy nhiên, độ chính xác trên tập kiểm tra (validation) duy trì ở mức khoảng 45-50%, cho thấy mô hình có dấu hiệu overfitting.

Kiến trúc của mô hình gồm hai lớp Conv2D kèm theo MaxPooling, tiếp theo là lớp Flatten và hai lớp Dense với số lượng tham số lên tới gần 2 triệu, cho phép mô hình học các đặc trưng phức tạp từ dữ liệu âm thanh của RAVDESS. Lớp Conv2D đầu tiên sử dụng kernel kích thước (3x3), đầu vào 1 kênh (âm thanh dạng spectrogram), và 32 bộ lọc, cùng với 32 biases.

Mặc dù mô hình CNN này đạt độ chính xác cao trên tập huấn luyện, hiệu suất chưa thực sự ổn định trên tập kiểm tra, điều này đặt ra thách thức trong việc cải thiện khả năng tổng quát hóa khi nhận diện cảm xúc từ dữ liệu RAVDESS trong các điều kiện thực tế đa dạng hơn. Đây vẫn là một hướng phát triển tiềm năng trong nhận dạng cảm xúc dựa trên dữ liệu âm thanh.

4.3 giao diện chương trình phân loại





4.4 Thảo luận

Kết quả nghiên cứu đã làm rõ tính khả thi và hiệu quả của việc áp dụng mô hình học sâu để nhận diện cảm xúc từ hình ảnh khuôn mặt. Việc so sánh hai kiến trúc CNN và ViT trong cùng một điều kiện thực nghiệm đã cung cấp một cái nhìn toàn diện về ưu và nhược điểm của từng mô hình trong bài toán nhận diện cảm xúc.

Về lý thuyết, nghiên cứu khẳng định rằng ViT – một mô hình vốn nổi bật trong thị giác máy tính hiện đại – hoàn toàn có thể thích ứng với các tác vụ phân loại cảm xúc khuôn mặt, vốn trước đây được thống trị bởi CNN. Điều này mở ra tiềm năng mở rộng ứng dụng ViT trong các hệ thống cảm xúc học đa phương thức và thời gian thực.

Trong khi đó, dữ liệu âm thanh từ RAVDESS cho thấy tiềm năng lớn trong việc nhận diện cảm xúc thông qua phân tích giọng nói, đặc biệt khi sử dụng các đặc trưng như MRCC (Mel-frequency cepstral coefficients), giúp tăng độ chính xác trong nhận diện cảm xúc dựa trên âm thanh.

Về thực tiễn, kết quả đạt được góp phần tạo nền tảng cho việc xây dựng các hệ thống nhận diện cảm xúc phục vụ trò chơi tương tác, robot xã hội, hoặc công cụ hỗ trợ tư vấn tâm lý tự động. CNN có thể triển khai trên thiết bị di động hoặc IoT, trong khi ViT thích hợp cho hệ thống máy chủ hoặc nền tảng đám mây.

Tuy nhiên, nhóm cũng gặp phải một số khó khăn trong quá trình thực hiện, bao gồm:

(1) Chất lượng dữ liệu ảnh thấp (48x48 pixel grayscale), hạn chế khả năng học đặc trưng sâu.

(2) Phân bố nhãn không đồng đều, gây mất cân bằng khi huấn luyện.

(3) Thiếu thông tin ngữ cảnh hoặc âm thanh, khiến mô hình khó nhận diện các cảm xúc có biểu hiện tương đồng (ví dụ: "Sad" vs "Fear").

(4) Khó khăn trong việc tối ưu siêu tham số, đặc biệt với ViT vì kiến trúc này yêu cầu dữ liệu lớn và ổn định.

(5) Khả năng khái quát của mô hình còn hạn chế, do chưa tích hợp các kỹ thuật tăng cường dữ liệu mạnh mẽ hoặc đa mô thức.

Hướng nghiên cứu tương lai đề xuất:

Mở rộng nghiên cứu trên bộ dữ liệu lớn hơn như AffectNet hoặc RAF-DB để tăng tính khái quát.

Kết hợp đặc trưng âm thanh (ví dụ từ RAVDESS) để xây dựng mô hình đa mô thức.

Tích hợp kỹ thuật cân bằng dữ liệu như SMOTE hoặc focal loss để cải thiện khả năng phân loại với các lớp hiếm.

Tối ưu mô hình ViT theo hướng nhẹ hơn (MobileViT) nhằm áp dụng thực tiễn trên thiết bị di động.

Tóm lại, nghiên cứu này không chỉ cung cấp minh chứng thực nghiệm cho hiệu quả của CNN và ViT trong nhận diện cảm xúc, mà còn tạo tiền đề cho các nghiên cứu ứng dụng sâu hơn trong tương lai.

5.Kết luận

Nghiên cứu này đã triển khai và đánh giá hiệu quả của hai mô hình học sâu là Convolutional Neural Network (CNN) và Vision Transformer (ViT) trên bộ dữ liệu FER2013 trong bài toán nhận diện cảm xúc từ hình ảnh khuôn mặt. Kết quả thực nghiệm cho thấy cả hai mô hình đều có khả năng phân loại cảm xúc với độ chính xác tương đối tốt, trong đó ViT cho thấy hiệu suất vượt trội hơn so với CNN ở một số lớp cảm xúc phức tạp nhờ cơ chế Attention toàn cục.

Đóng góp chính của nghiên cứu là cung cấp cái nhìn so sánh rõ ràng giữa hai kiến trúc học sâu phổ biến trong lĩnh vực thị giác máy tính hiện nay, từ đó giúp định hướng lựa chọn mô hình phù hợp với từng bài toán cụ thể trong nhận diện cảm xúc. Bên cạnh đó, việc áp dụng bộ dữ liệu FER2013 – một tập dữ liệu mang tính học thuật cao nhưng nhiều thách thức – cũng giúp làm nổi bật những giới hạn hiện tại của mô hình học sâu khi áp dụng vào các tình huống thực tế. Về mặt thực tiễn, nghiên cứu mở ra hướng ứng dụng tiềm năng trong các lĩnh vực như giáo dục, y tế, chăm sóc khách hàng, trò chơi tương tác và robot xã hội, nơi cảm xúc người dùng đóng vai trò quan trọng trong tương tác. Về mặt lý thuyết, nghiên cứu góp phần bổ sung vào kho tư liệu ứng dụng ViT trong các bài toán phi truyền thống so với mục tiêu ban đầu của nó.

Tuy nhiên, nghiên cứu vẫn còn một số hạn chế như chưa tích hợp dữ liệu đa mô thức (âm thanh, văn bản), mô hình chưa được tối ưu hóa đầy đủ về siêu tham số, và bộ dữ liệu FER2013 vẫn còn hạn chế về độ phân giải và sự cân bằng lớp.

Trong tương lai, nhóm nghiên cứu đề xuất mở rộng mô hình sang các tập dữ liệu lớn hơn như AffectNet hoặc RAF-DB, đồng thời xây dựng hệ thống nhận diện cảm xúc đa mô thức kết hợp cả hình ảnh và âm thanh (ví dụ từ RAVDESS) nhằm tăng độ chính xác và độ tin cậy của mô hình trong môi trường thực tế.

Tài liệu tham khảo

[1] Bledsoe, W. W (1964). “The Model Method in Facial Recognition”, Technical Report PRI 15. Panoramic
Research, Inc., Palo Alto, California.

[2] Matsumoto, David, and Hyi Sung Hwang (2011). "Reading facial expressions of emotion", Psychological Science
Agenda, Vol 25, No5, pp. 10-18.

[3] K. Mase, A. Pentland (1991), “Recognition of facial expression from optical flow”, IEEE TRANSACTIONS on
Information and Systems, Vol E74-D, No10, pp. 3474-3483.

[4] I Goodfellow, D Erhan, PL Carrier, A Courville, M Mirza, B Hamner, W Cukierski, Y Tang, DH Lee, Y Zhou, C
Ramaiah, F Feng, R Li, X Wang, D Athanasakis, J Shawe-Taylor, M Milakov, J Park, R Ionescu, M Popescu, C
Grozea, J Bergstra, J Xie, L Romaszko, B Xu, Z Chuang, and Y. Bengio (2013). "Challenges in Representation
Learning: A report on three machine learning contests." arXiv 2013.

[5] Paul Viola and Michael Jones (2001). "Rapid Object Detection using a Boosted Cascade of Simple Features",
Conference on Computer vision and Pattern recognition 2001.

[6] Docs, OpenCV. “Face Detection Using Haar Cascades.”, OpenCV: Face Detection Using Haar Cascades, 4 Aug.
2017.

[7] François Chollet. “Xception: Deep Learning with Depthwise Separable Convolutions”. arXiv 2017.

[8] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset
(CK+): A complete dataset for action unit and emotion-specified expression", Proceedings of IEEE on CVPR for
Human Communicative Behavior Analysis, San Francisco, USA, 2010.

[9] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., & van Knippenberg, A. (2010). Presentation
and validation of the Radboud Faces Database. Cognition & Emotion, 24(8), 1377-1388. DOI: 10.1080/02699930903485076.

[10]Bengio, Yoshua. "Learning Deep Architectures for AI". Foundations and Trends in Machine Learning: Vol. 2: No.
1, pp 1-127, (2009).

[11]Krizhevsky, Alex. "ImageNet Classification with Deep Convolutional Neural Networks". Retrieved 17 November
2013.

[12]Y LeCun, L Bottou, Y Bengio, P Haffner (1998). "Gradient-based learning applied to document recognition",
Proceedings of the IEEE 86 (11), p2278-2324.

[13]Honglak Lee, Roger Grosse, Rajesh Ranganath and Andrew Y. Ng, "Convolutional Deep Belief Networks for
Scalable Unsupervised Learning of Hierarchical Representations", ICML 2009.

[14]C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the
automatic annotation of a million facial expressions in the wild," in Proceedings of IEEE International Conference
on Computer Vision & Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016.

[15]A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal
computing in the wild," IEEE Transactions on Affective Computing, vol. PP, no. 99, pp. 1-1, 2017.

[16]S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C¸ . Gulc¸ehre, R. Memisevic, P. Vincent, A. Courville, Y.
Bengio, R. C. Ferrari et al., "Combining modality specific deep neural networks for emotion recognition in video,"
in Proceedings of the 15th ACM on International conference on multimodal interaction. ACM, 2013, pp. 543-550.

[17]X. Liu, B. Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression
recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2017, pp. 522-531.

[18]V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in
Proceedings of the 19th ACM International Conference on Multimodal Interaction. ACM, 2017, pp. 569-576.