SỰ ẢNH HƯỞNG CÁC YẾU TỐ KIẾN TRÚC MẠNG CONVNEXTV2 ĐẾN NHẬN DẠNG BIỂU CẢM KHUÔN MẶT TỪ DỮ LIỆU THỰC TẾ

Dương Thăng Long*, Vương Thu Trang*, Phạm Quang Huy*
Email: duongthanglong@hou.edu.vn

Ngày tòa soạn nhận được bài báo: 05/04/2024
Ngày phản biện đánh giá: 15/10/2024
Ngày bài báo được duyệt đăng: 28/10/2024

DOI: 10.59266/houjs.2024.465

Tóm tắt: *Thành công của các mô hình Transformer đã cho thấy hiệu suất xuất sắc trong các nhiệm vụ xử lý ngôn ngữ tự nhiên (NLP) đã được mở rộng sang lĩnh vực thị giác máy tính với các kiến trúc Vision Transformers (ViTs), đem lại kết quả tương đương hoặc vượt trội so với các mạng nơron tích chập (CNN) truyền thống trong các nhiệm vụ như nhận dạng hình ảnh và phát hiện đối tượng. Biến thể ConvNeXt V2, một mô hình cải tiến từ kiến trúc ResNet và kế thừa các điểm mạnh của kiến trúc ViTs như cấu trúc phân cấp các lớp nơron và cơ chế mã hóa tự động FCMAF nhằm mang lại hiệu suất cao và mô hình đơn giản hơn. Trong khi đó, nhận dạng biểu cảm khuôn mặt (FER) vẫn là một thách thức đối với các mô hình do hình ảnh trong thực tế bị các yếu tố như che khuất, biến đổi màu sắc và tư thế khuôn mặt. Nghiên cứu này áp dụng ConvNeXt V2 cho bài toán FER với việc điều chỉnh các tham số kiến trúc để đánh giá tác động của chúng trên dữ liệu thực tế của FER từ RAF_DB. Kết quả thử nghiệm cho thấy những yếu tố kiến trúc của ConvNeXt V2 tác động đến độ phức tạp của mô hình và chất lượng nhận dạng cho FER, cung cấp những phân tích ý nghĩa để làm rõ những vận dụng điểm mạnh của mô hình kiến trúc ViTs và kết hợp với các kiến trúc CNN truyền thống nhằm tăng thêm hiệu quả cho mô hình ứng dụng.*

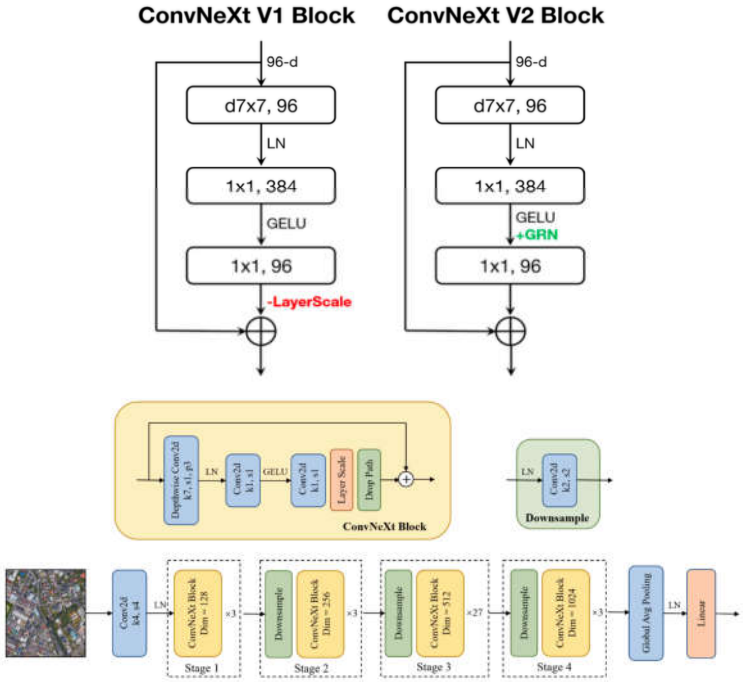
Từ khoá: *Vision transformers, ConvNeXt V2 architecture, facial expression recognition.*

I. Giới thiệu

Transformer là một kiến trúc mạng nơron sâu nổi bật, đạt kết quả xuất sắc trong các nhiệm vụ xử lý ngôn ngữ tự nhiên (NLP), đặc biệt trong dịch máy và phân tích cú pháp. Theo [1], các biến thể như BERT đạt thành tích dẫn đầu trên nhiều nhiệm vụ NLP, trong khi GPT-3, với 175 tỷ tham số, thể hiện hiệu suất mạnh mẽ trên các nhiệm vụ NLP chuyên biệt mà không cần tinh chỉnh. Nhờ những thành công này, Transformer đã được mở rộng sang các nhiệm vụ thị giác máy tính (CV), mang lại hiệu quả đáng kể và tiềm năng lớn.

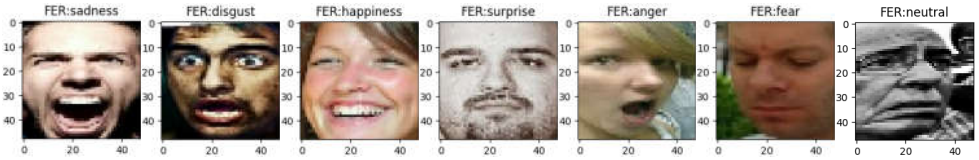
* Trường Đại học Mở Hà Nội

Các mô hình CV dựa trên Transformer (ViTs) đã cho thấy tiềm năng, đạt kết quả tương đương hoặc vượt trội so với CNN trong phân loại hình ảnh, phát hiện đối tượng, phân đoạn ngữ nghĩa, xử lý ảnh và hiểu nội dung video. Nhờ thành công này, nhiều biến thể Transformer đã ra đời, cải tiến về hiệu suất, tổng quát hóa và điều chỉnh mô hình. Trong [2], các tác giả hiện đại hóa kiến trúc ResNet theo hướng của Transformer thị giác phân cấp, dẫn đến sự ra đời của họ mô hình ConvNeXt. ConvNeXt giữ được hiệu suất của CNN tiêu chuẩn và đơn giản hóa quá trình triển khai nhờ tính chất tự nhiên của tích chập. ConvNeXt, đặc biệt mạnh mẽ trong các tình huống có độ phức tạp thấp, đã trở thành kiến trúc chủ đạo trong nhiều ứng dụng CV. Phiên bản ConvNeXt V2 (Hình 1.1) [3] gần đây, với các biến thể từ đơn giản đến phức tạp, có hai đặc điểm chính: bộ mã hóa tự động FCMAE (Fully Convolutional Masked AutoEncoder) và lớp chuẩn hóa toàn cục GRN (Global Response Normalization), tối ưu cho việc học tự giám sát. Nhờ cơ chế FCMAE, hiệu suất của mô hình được cải thiện đáng kể trên nhiều nhiệm vụ, bao gồm phân loại ảnh ImageNet, phát hiện đối tượng COCO và phân đoạn ảnh ADE20K.



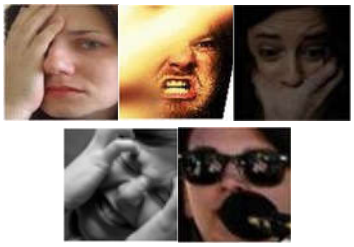
Hình 1.1. Kiến trúc chính ConvNeXt V2 [4]

Bài toán nhận dạng biểu cảm khuôn mặt (FER) đã được nghiên cứu rộng rãi, nhưng vẫn gặp nhiều thách thức khi áp dụng trong thực tế. FER đóng vai trò quan trọng trong giao tiếp giữa con người, giúp hiểu cảm xúc và ý định, làm cho nó trở thành yếu tố thiết yếu trong tương tác xã hội. Ứng dụng của FER rất đa dạng, bao gồm tương tác người-máy, chú thích ảnh, chuyển văn bản video, và giao tiếp xã hội. Theo [5, 6], có sáu biểu cảm khuôn mặt cơ bản: Hạnh phúc (Happiness), Buồn bã (Sadness), Ngạc nhiên (Surprise), Ghê tởm (Digust), Tức giận (Anger) và Sợ hãi (Fear), được coi là phổ quát, không phụ thuộc vào quốc gia hay tôn giáo. Hai biểu cảm khác là Khinh bỉ (Contempt) và Trung tính (Netural) cũng được sử dụng ở trong một số nghiên cứu. Hình 1.2 minh họa một số biểu cảm cơ bản từ bộ dữ liệu RAF_DB [7].



Hình 1.2. Một số biểu cảm khuôn mặt cơ bản

Biểu cảm khuôn mặt gắn liền với sự thay đổi ở các đặc điểm và cơ bắp khuôn mặt, đặc biệt tại mắt, mũi, miệng, là các khu vực quan trọng để giải quyết bài toán FER. Hình ảnh thực tế còn có sự biến đổi về tư thế, che phủ bởi các vật khác, hoặc màu sắc bị thay đổi do điều kiện ánh sáng. Mô hình FER cần không chỉ tập trung vào các khu vực quan trọng mà còn phải bền vững trước các yếu tố như che khuất, màu sắc mờ hay góc nghiêng lớn. Ví dụ, trong Hình 1.3, các ảnh có vùng bị che khuất lớn hoặc tối màu, nghiêng góc mạnh, làm ảnh hưởng đến chất lượng nhận dạng. Đây là những thách thức lớn cho các mô hình CV trong FER. Vì vậy, việc phát triển mô hình cần chú trọng khả năng trích xuất đặc trưng khái quát và chuyên biệt, đặc biệt đối với những đặc điểm quan trọng này..



Hình 1.3. Một số hình ảnh khó khăn cho FER

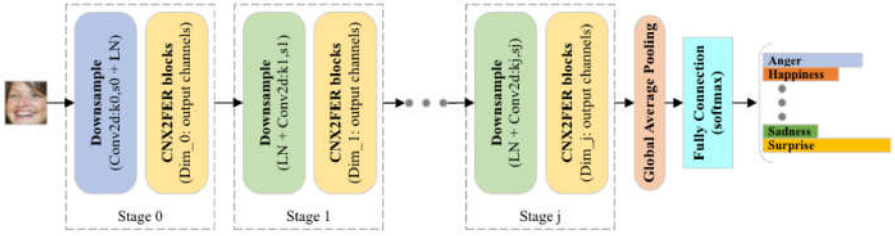
Mặc dù nhiều kiến trúc CNN đạt hiệu quả cao, nhưng chúng phức tạp, tốn kém tài nguyên và yêu cầu hệ thống tính toán lớn cho cả quá trình huấn luyện và triển khai. Vì vậy, một số nghiên cứu đã đề xuất sử dụng các mô hình CNN nhẹ [8], [9], [10], [11], [12], [13], phù hợp với các ứng dụng hạn chế về tài nguyên nhưng vẫn đảm bảo hiệu quả, đặc biệt là trên các nền tảng trực tuyến như ứng dụng web.

Trong nghiên cứu này, chúng tôi áp dụng kiến trúc nhẹ của ConvNeXt V2 cho bài toán FER, với các điều chỉnh tham số và đánh giá tác động của chúng lên chất lượng nhận dạng trên dữ liệu RAF_DB. Phần 2 sẽ giới thiệu chi tiết về ConvNeXt V2 và các điều chỉnh tham số được áp dụng cho FER, cùng với phân tích các yếu tố kiến trúc ảnh hưởng đến kết quả. Các thử nghiệm trên RAF_DB theo nhiều kịch bản khác nhau sẽ được so sánh và đánh giá. Cuối cùng, Phần 3 là nội dung kết luận.

II. Phương pháp

2.1. Mô hình LC2FER

Phần này trình bày chi tiết mô hình nhẹ dựa trên kiến trúc ConvNeXt V2 cho bài toán FER, ký hiệu là LC2FER (Light ConvNeXt V2 for Facial Expression Recognition). Mô hình được chia thành các giai đoạn chính (stages), mỗi giai đoạn gồm khối giảm độ phân giải đặc trưng (downsample) và khối xử lý chính để trích xuất đặc trưng CX2FER theo kiến trúc ConvNeXt (Hình 2.1). Tính chất nhẹ của mô hình dựa trên các điều chỉnh tham số như số giai đoạn, số lượng khối CX2FER trong mỗi giai đoạn, kích thước không gian đặc trưng, và tỷ lệ giảm độ phân giải giữa các giai đoạn.

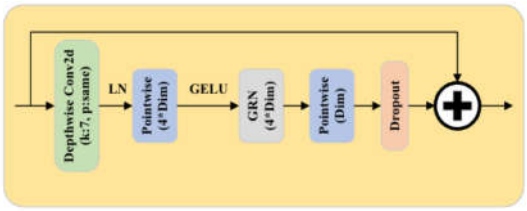


Hình 2.1. Sơ đồ tổng thể mô hình LC2FER

Phần lõi của LC2FER là các khối CX2FER. Mỗi giai đoạn có nhiều khối CX2FER nối tiếp nhau, với số lượng khối này là tham số kiến trúc tạo nên độ sâu của mô hình. Trong mỗi khối CX2FER (Hình 2.2), lớp tích chập theo chiều sâu (depthwise convolution) thực hiện trên từng kênh với bộ lọc riêng, giúp mô hình học đặc trưng cho từng kênh một cách hiệu quả, giảm số tham số và độ phức tạp tính toán. Hai lớp tích chập theo chiều rộng (pointwise convolution) sử dụng phép nhân để biến đổi và kết hợp thông tin từ các kênh, tạo ra số lượng kênh đầu ra mong muốn. Giữa hai lớp này có bộ xử lý chuẩn hóa phản hồi toàn cục (GRN), thực hiện trên các kênh và không gian đặc trưng (công thức (1)).

$$X_{grn} = \gamma.\left(X.\frac{\|X\|_2}{E(\|X\|_2)+\epsilon}\right) + \beta + X \quad (1)$$

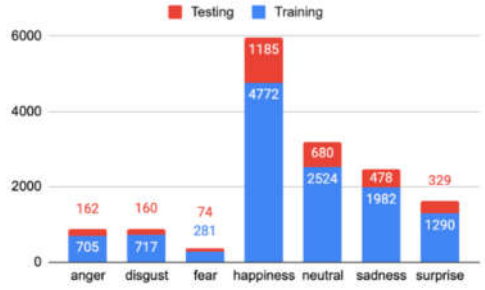
trong đó, γ và β là các tham số học của mô hình, X là bản đồ đặc trưng. GRN giúp ổn định quá trình huấn luyện và cải thiện tính bền vững của mô hình bằng cách chuẩn hóa các phản hồi đặc trưng, ngăn chặn việc giãn nở và dư thừa đặc trưng. Cơ chế này tập trung vào các đặc trưng quan trọng, nâng cao khả năng tổng quát và độ chính xác nhận dạng. Khối CX2FER còn sử dụng cơ chế loại bỏ ngẫu nhiên các kết nối (Dropout) để ngăn ngừa hiện tượng quá khớp trong quá trình huấn luyện.



Hình 2.2. Khối xử lý chính CX2FER

2.2. Ảnh hưởng tham số mô hình LC2FER

Để nghiên cứu tác động của các yếu tố cấu trúc mô hình đến bài toán FER, bài báo sử dụng bộ dữ liệu RAF_DB (Real-world Affective Faces database) [7], với 29,672 hình ảnh khuôn mặt, bao gồm sáu biểu cảm cơ bản và biểu cảm trung tính (Neutral). Dữ liệu được chia thành hai phần: 12,271 hình ảnh cho huấn luyện (Training) và 3,068 hình ảnh cho kiểm tra (Testing, D^{te}). Tập huấn luyện được chia ngẫu nhiên thành 80% để huấn luyện mô hình (D^{tr}) và 20% để thẩm định lựa chọn mô hình (D^{va}). Hình 2.3 thể hiện sự phân bố hình ảnh giữa các biểu cảm, cho thấy sự chênh lệch đáng kể: ví dụ, biểu cảm "fear" chỉ có 281 ảnh huấn luyện, trong khi "happiness" có đến 4,772 ảnh, gấp gần 17 lần. Sự phức tạp và đa dạng của dữ liệu thực tế trong RAF_DB, cùng với mất cân bằng giữa các lớp, tạo ra thách thức cho các mô hình nhận dạng.



Hình 2.3. Sự phân bố hình ảnh của RAF_DB

Nghiên cứu bắt đầu với kiến trúc ConvNeXt V2 phiên bản nhỏ nhất (Atto), bao gồm 4 giai đoạn (n_s=4). Các tham số như số khối trong từng giai đoạn (B), kích thước đầu ra (D), và tỷ lệ giảm độ phân giải mẫu tín hiệu đặc trưng (S) được thiết lập theo [3]. Tham số S được điều chỉnh qua phép tích chập với kích thước bộ lọc kernel (k) và bước trượt stride (a). Bộ tham số cấu hình của mô hình gốc được ký hiệu là C_1. Tiếp theo, các tham số n_s, S, B và D được thay đổi để thu nhỏ cấu trúc mô hình. Chi

tiết các tham số cấu hình (C_i) ở từng giai đoạn được thể hiện trong Bảng 2.1.

Bảng 2.1. Tham số cấu hình của LC2FER

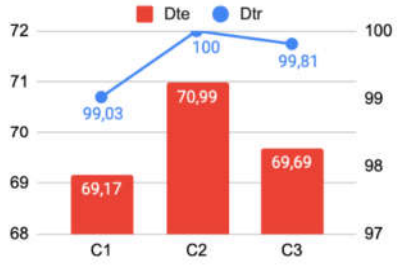
Giai đoạn	Tham số
C_1:4	S=[(4,4),(2,2),(2,2),(2,2)], B=[2,2,6,2], D=[40,80,160,320]
C_2:3	S=[(4,4),(3,3),(2,2)], B=[2,4,2], D=[64,128,256]
C_3:3	S=[(4,4),(3,3),(2,2)], B=[2,4,2], D=[128,256,512]
C_4:3	S=[(4,2),(3,3),(2,2)], B=[2,4,2], D=[64,128,256]
C_5:3	S=[(4,2),(3,3),(2,2)], B=[2,2,2], D=[64,128,256]
C_6:3	S=[(4,2),(3,3),(2,2)], B=[1,1,1], D=[64,128,256]
C_7:3	S=[(4,2),(3,3),(2,2)], B=[1,1,1], D=[128,256,512]
C_8:3	S=[(4,2),(3,3),(2,2)], B=[1,1,1], D=[40,80,160]

Trong kiến trúc ConvNeXt V2, hai giai đoạn đầu tập trung vào việc trích xuất các đặc trưng mức thấp, vì vậy nghiên cứu này xem xét tích hợp chúng thành một, dẫn đến việc các cấu hình từ C_2 trở đi chỉ sử dụng 3 giai đoạn. Số khối ở giai đoạn giữa của mô hình C_2 và C_3 cũng giảm xuống còn 4 thay vì 6 như trong C_1, nhằm giảm khối lượng tính toán và kích thước mô hình. Hơn nữa, tham số S ở giai đoạn thứ hai của cả hai mô hình này được điều chỉnh tăng lên (3,3) thay vì (2,2) trong C_1, để tạo sự chuyển đổi mượt mà giữa độ phân giải của các giai đoạn. Mô hình C_3 tăng số lượng kênh tín hiệu đầu ra lên gấp đôi, D=[128,256,512] thay vì [64,128,256], nhằm đánh giá ảnh hưởng của yếu tố này so với C_2. Quá trình huấn luyện mô hình được thực hiện với bộ tham số giống nhau cho các cấu hình, như thể hiện trong Bảng 2.2.

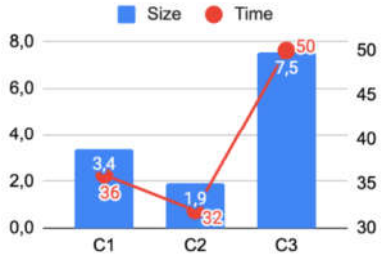
Bảng 2.2. Tham số huấn luyện các mô hình

Stt	Tham số	Giá trị
1	Tốc độ học ban đầu (learning rate)	10^{-3}
2	Kích thước gói dữ liệu (batch size)	128
3	Số lượt huấn luyện tối đa (epochs)	150
4	Phương pháp tối ưu mô hình	AdamW [14]

Kết quả nhận dạng trên các tập dữ liệu huấn luyện và kiểm tra , kích thước (số lượng tham số) và tốc độ (thời gian cho một epoch huấn luyện) của ba mô hình được thể hiện như Hình 2.4 và Hình 2.5. Mô hình cho kết quả tốt nhất cả trên dữ liệu huấn luyện và kiểm tra, mặc dù mô hình này bị giảm một giai đoạn và kích thước mô hình cũng giảm còn 55,8% so với (từ 3,4 triệu xuống 1,9 triệu). Trong khi đó, có kích thước (7,5 triệu) tăng gấp 3,9 lần song kết quả nhận dạng lại giảm (từ 70,99% xuống 69,69% trên), tốc độ của cũng cao gấp 1,56 lần so với , điều này cho thấy việc tăng số lượng kênh tín hiệu đặc trưng đầu ra ở các giai đoạn không đem lại ý nghĩa và chất lượng cho kiến trúc này.

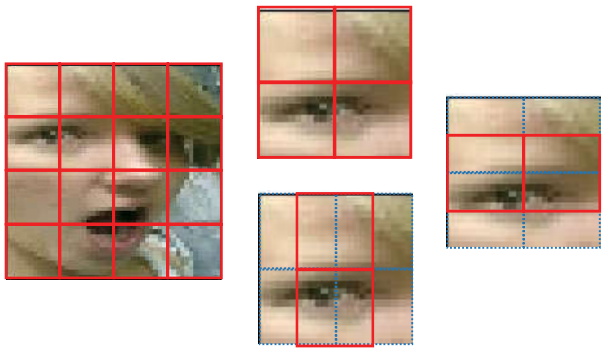


Hình 2.4. Kết quả nhận dạng của



Hình 2.5. Kích thước và tốc độ của

Các mô hình từ đến được điều chỉnh với bước trượt ở giai đoạn thứ nhất giảm một nửa, sử dụng thay vì như các mô hình trước. Điều này tăng gấp đôi độ phân giải không gian tín hiệu trong quá trình trích xuất đặc trưng mức thấp. Ở giai đoạn đầu, ảnh đầu vào được chia thành các phần chồng lấp (overlapped) thay vì các mảnh rời nhau. Như vậy, mỗi điểm ảnh sẽ tham gia vào nhiều cửa sổ trích chọn đặc trưng, giúp cung cấp thêm thông tin và mối liên hệ cục bộ trong các vùng ảnh. Hình 2.6 minh họa rõ sự khác biệt: bên trái là phần không chồng lấp, trong khi bên phải và dưới là phần có chồng lấp.



Hình 2.6. Chia các phần ảnh có chồng lấp theo tham số ở giai đoạn đầu

Kết quả phân lớp của các mô hình đến đều cao hơn các mô hình trước đó sau khi có sự thay đổi tham số ở giai đoạn đầu (Bảng 2.2). Điều này cho thấy mô hình ở kiến trúc này khi được trích chọn đặc trưng trên không gian ảnh đầu vào được mịn hơn sẽ cho kết quả nhận dạng tốt hơn.

Bảng 2.1. Kết quả của các mô hình

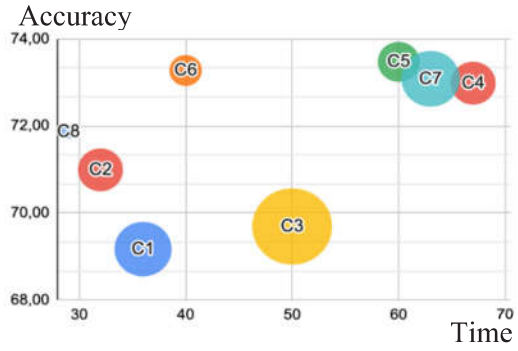
Mô hình	Độ chính xác trên D^{tr}	Độ chính xác trên D^{te}
C_1	99.03	69.17
C_2	100	70.99
C_3	99.81	69.69
C_4	100	72.97
C_5	100	73.47
C_6	100	73.27
C_7	100	73.08
C_8	100	71.87

Mặc dù các mô hình và giảm số khối xử lý CX2FER xuống chỉ còn một khối ở mỗi giai đoạn, nhưng kết quả nhận dạng vẫn cao hơn so với các mô hình trước đó có nhiều khối hơn. Sự giảm nhẹ trong hiệu suất khi giảm số khối CX2FER từ 2 xuống 1 ở và là không đáng kể, cho thấy chất lượng mô hình không bị ảnh hưởng lớn bởi thay đổi này.

Tuy nhiên, ở mô hình , khi giảm kích thước số kênh tín hiệu đầu ra còn thay vì như các mô hình trước, kết quả nhận dạng giảm xuống 71.87%, so với 73.47% của . Đáng chú ý, dù số kênh tín hiệu đầu ra trong tăng gấp đôi, nhưng kết quả chỉ đạt 73.08%, không cao hơn . Điều này cho thấy kích thước số kênh đầu ra ảnh hưởng đến chất lượng mô hình, nhưng không phải lúc nào cũng tỷ lệ thuận.

Kích thước các mô hình giảm khi số khối và kích thước kênh tín hiệu đầu ra giảm. có kích thước thấp nhất với khoảng

0.37 triệu tham số, trong khi có 0.93 triệu tham số. Ngược lại, tăng nhanh lên 3.64 triệu tham số khi số kênh tín hiệu đầu ra tăng gấp đôi. Hình 2.7 so sánh chi tiết kích thước (độ lớn hình tròn), thời gian huấn luyện và kết quả nhận dạng giữa các mô hình. Qua phân tích, là phiên bản ConvNeXt V2 tốt nhất, cân đối giữa kích thước, tốc độ huấn luyện và độ chính xác.



Hình 2.7. So sánh các mô hình

2.3. Phân tích kết quả mô hình tại

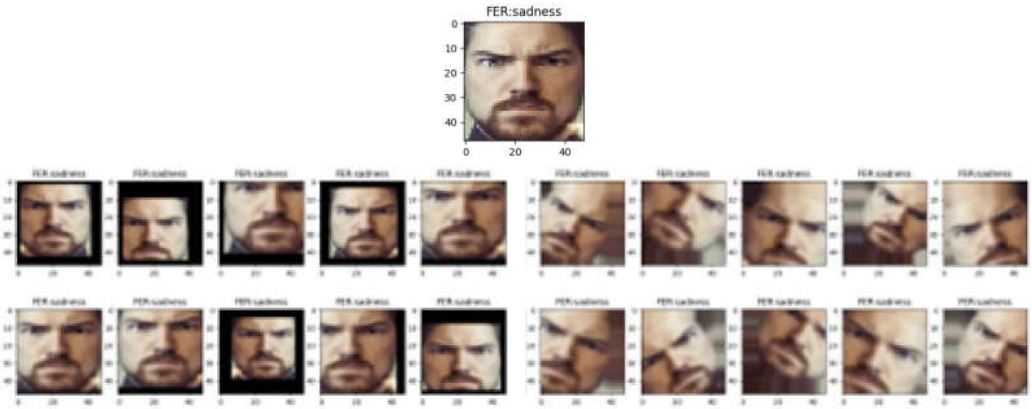
Để có kết qủa tốt hơn tại mô hình C_5, chúng tôi sử dụng phương pháp tăng cương hình ảnh cho tập dữ liệu huấn luyện bằng các phép xử lý ảnh như lật ảnh (flip), tịnh tiến theo chiều ngang (tx) và chiều dọc (ty), thu nhỏ và phóng to, xoay ảnh. Các phép này có thể thực hiện đồng thời trên cùng một ảnh với các tham số khác nhau. Tham số các phép xử lý này lấy ngẫu nhiên trong giới hạn theo Bảng 2.2.

Translation có giá trị âm là tịnh tiến lên (theo chiều dọc) hoặc sang trái (theo chiều ngang), tương tự zoom có giá trị âm là thu nhỏ và dương là phóng to. Các giá trị này theo tỷ lệ với kích thước hình ảnh, ví dụ zoom = -0.1 có nghĩa thu nhỏ hình ảnh xuống 10% theo cả hai chiều ngang và dọc. Phép xoay ảnh được thể hiện là tỷ lệ % của góc π, âm là quay ngược chiều kim đồng hồ và ngược lại. Hình 2.8 thể hiện một số hình ảnh được tăng cường từ ảnh gốc (trên cùng), hai hàng giữa là ảnh chỉ dùng phép tịnh tiến, co giãn và lật, còn hai hàng cuối sử dụng thêm phép xoay ảnh.

Bảng 2.2. Tham số của phép tăng cường ảnh

Stt	Tham số	Giá trị
1	Lật ảnh đối xứng ngang	True
2	Góc quay (rotation) tối đa so với ảnh gốc (radian, âm là quay sang trái)	±0.05π
3	Hệ số dịch chuyển (translation) tối đa so với kích thước ảnh gốc (âm là dịch sang trái)	±0.1
4	Hệ số co dãn (zoom) tối đa so với kich thước ảnh gốc (giá trị âm là thu nhỏ)	±0.1
5	Tốc độ học ban đầu (theo phương pháp AdamW)	10^{-3}
6	Kích thước mỗi gói (batch) dữ liệu	128
7	Số lượt học tối đa của mô hình (epoch)	150

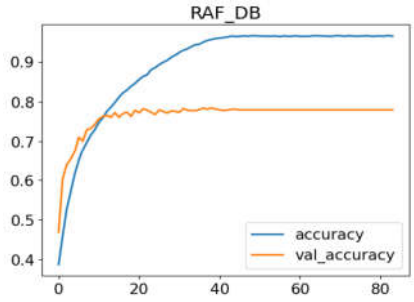
Hình 2.8. Một số hình ảnh tăng cường bằng các phép xử lý cơ bản



Mô hình C_5 được huấn luyện trên tập dữ liệu tăng cường cho kết quả cao hơn so với huấn luyện trên dữ liệu ban đầu, độ chính xác trên tập kiểm tra D^{te} ở đây đạt 79,56% so với 73.47% ban đầu (Bảng 2.1), chứng tỏ mô hình đáp ứng tốt hơn trên tập dữ liệu đa dạng hơn.

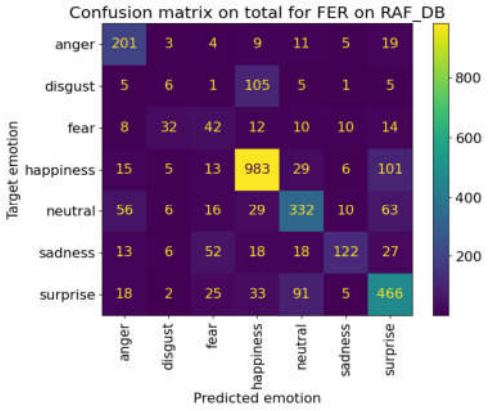
Kết quả quá trình huấn luyện thể hiện trong Hình vẽ 2.9 gồm độ chính xác trên tập huấn luyện D^{tr} (accuracy) và trên tập chọn mô hình D^{va} (val_accuracy). Ở đây quá trình huấn luyện thể hiện đến bước thứ 86 bởi vì quá trình sau đó không cho kết quả tốt hơn trên D^{va}.

Hình 2.9. Độ chính xác quá trình huấn luyện



Để minh hoạ kết quả tổng thể của mô hình LC2FER, chúng tôi xây dựng ma trận nhầm lẫn trên toàn bộ tập dữ liệu kiểm tra D^{te} (Hình 2.10). Mỗi hàng trong ma trận là một nhãn biểu cảm của hình ảnh trong dữ liệu (target emotion), mỗi cột tương ứng là một nhãn biểu cảm được mô hình nhận dạng (predicted emotion). Tổng số hình ảnh trong ma trận này đúng bằng tổng số hình ảnh của tập dữ liệu kiểm tra D^{te}. Trong ma trận này, biểu cảm "disgust" bị nhận dạng sai nhiều nhất thành "happiness" (105 ảnh), thứ hai có 101 ảnh của "happiness" bị nhận dạng thành "surprise". Tất cả các ô của ma trận đều có giá trị chứng tỏ sự nhầm lẫn của mô hình trong tập dữ liệu là đa dạng, có ít nhất một hình ảnh nhầm lẫn giữa biểu cảm "disgust" và "fear", giữa "disgust" và "sadness". Xét tỷ lệ nhầm lẫn trong tổng số ảnh của từng biểu cảm thì "disgust" có tỷ lệ cao nhất lên đến 95,3%, các biểu cảm có tỷ lệ này thấp gồm "happiness" có 14,7%, "anger" có 20,2% và "surprise" có 27,2%. Cả ba biểu cảm này đều có hình ảnh rất nhiều trong tập dữ liệu huấn luyện, và nó làm cho mô hình được huấn luyện để đáp ứng tốt ở các biểu cảm nhiều hình ảnh là cơ chế bình thường trong học máy. Vì vậy chúng ta có thể áp dụng các kỹ thuật để cân bằng dữ liệu trong trường hợp như vậy.

Hình 2.10. Ma trận nhầm lẫn tại C_5

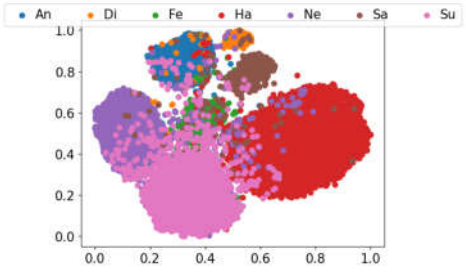


Một số trường hợp nhầm lẫn được thể hiện trong Hình 2.11, tiêu đề trên ảnh thể hiện hai chữ cái đầu tên biểu cảm trong tập dữ liệu (Target) và biểu cảm được mô hình nhận dạng (Predicted), ở giữa có ký hiệu ">". Có thể thấy rằng các hình ảnh nhầm lẫn này cũng rất khó phân biệt bằng trực quan của chúng ta. Riêng một số trường hợp cũng có thể thực sự không rõ ràng của việc gán nhãn trong tập dữ liệu. Chẳng hạn, ảnh thứ 3 hàng đầu bị biến dạng rất khó nhận biết, ảnh thứ hai ở hàng giữa được gán nhãn "happiness" nhưng biểu cảm không thực sự rõ ràng, hoặc ảnh cuối cùng ở hàng giữa được gán nhãn "happiness" nhưng đây cũng không rõ ràng, thậm chí trực quan xác định nghiêng về "surprise".

Hình 2.11. Một số hình ảnh nhầm lẫn

Kỹ thuật t-SNE [15] được sử dụng nhằm thể hiện trực quan hóa sự phân loại của đặc trưng trích xuất bởi mô hình theo các lớp biểu cảm (Hình 2.12). Kết quả t-SNE được tính toán theo cách trong [16] cho toàn bộ tập dữ liệu, bao gồm cả , và . Các nhãn của biểu cảm khuôn mặt được thể hiện bằng hai chữ cái đầu cho phù hợp kích thước. Hình ảnh trực quan này cho thấy mô hình tại tạo thành các cụm đặc trưng tương đối rõ rệt, đáng chú ý là đặc trưng của biểu cảm "surprise" có lẫn lộn vào nhiều cụm đặc trưng ở các biểu cảm khác và nó cũng tương ứng với ma trận nhầm lẫn (cột phải cùng trong Hình 2.10). Hình ảnh t-SNE này cung cấp trực quan về khả năng trích xuất đặc trưng từ hình ảnh có tính phân loại của mô hình LC2FER. Nó cũng cho thấy được khả năng trích xuất đặc trưng nhằm giảm sự khác biệt giữa các hình ảnh trong cùng lớp biểu cảm (intra-class) trong khi tăng cường sự khác biệt về đặc trưng của các hình ảnh giữa các lớp biểu cảm (inter-class).

Hình 2.12. t-SNE trực quan của mô hình

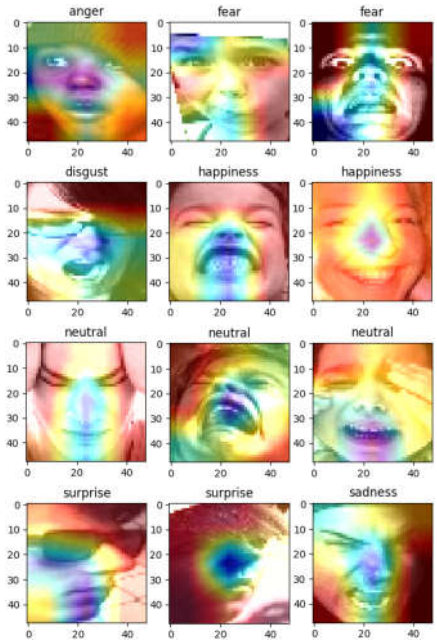


Ngoài ra, để thấy rõ hơn sự tác động của mô hình LC2FER trong việc trích xuất đặc trưng hình ảnh, chúng tôi sử dụng kỹ thuật bản địa hóa dựa trên gradient (gradient-based localization) [17] hay được gọi là bản đồ nhiệt (heatmap) để hiển thị các hình ảnh biểu diễn trực quan về các vùng trên ảnh được quan tâm và tập trung trong trích chọn đặc trưng của mô hình. Trong Hình 2.13, bản đồ nhiệt của mô hình được thể hiện trên một số hình ảnh với các biểu cảm khác nhau. Những hình ảnh này minh họa rõ ràng bản đồ nhiệt chủ yếu tập trung vào các khu vực quan trọng để thể hiện nét mặt, chẳng hạn như mũi, miệng và mắt. Chẳng hạn, ba ảnh đầu ở hàng cuối có phần mắt bị che khuất bởi tóc, bàn tay hoặc kính đen thì mô hình chỉ tập trung vùng mắt còn lại và vùng mũi, vùng miệng ở phần nhìn thấy.

Bản đồ nhiệt này cũng trực quan nhấn mạnh rằng mô hình LC2FER ưu tiên các vùng hình ảnh quan trọng để trích xuất các đặc trưng mô tả cho FER. Ngược lại, khi những vùng hình ảnh này không được xem xét, việc xác định chính xác biểu cảm khuôn mặt sẽ trở nên khó khăn. Ở đây, một lần nữa thấy rằng dữ liệu gán nhãn có chỗ không rõ ràng, chẳng hạn ảnh

cuối hàng giữa và ảnh đầu hàng cuối có nhãn "neutral" nhưng thực tế trực quan có thể nghiêng về biểu cảm "sadness" hoặc "fear" thì phù hợp hơn.

Hình 2.13. Heatmap của mô hình trên ảnh



III. Kết luận

Trong nghiên cứu này, chúng tôi đã đề xuất sử dụng mô hình mạng nơron tích chập LC2FER theo kiểu kiến trúc Vision Transformers (ViTs) cho bài toán nhận dạng biểu cảm khuôn mặt (FER). Kiến trúc của mô hình này dựa trên chuẩn kiến trúc của ConvNeXt V2, trong đó các tham số kiến trúc của mô hình được điều chỉnh để phù hợp và đáp ứng tốt cho bài toán FER với dữ liệu từ thực tế và có nhiều thách thức đối với độ chính xác của các mô hình nhận dạng.

Nghiên cứu cũng phân tích các yếu tố kiến trúc của LC2FER tác động đến việc trích chọn đặc trưng cho FER và phân lớp nhận dạng biểu cảm khuôn mặt. Mô hình này cũng có độ phức tạp (độ sâu số lớp nơron, kích thước bản đồ đặc trưng) ở mức thấp, tạo nên mô hình ở thể nhẹ và có số lượng tham số khá thấp (chỉ mức 0.9 triệu) hơn nhiều các mô hình khác trong [8] [18] [6] [16] [5]. Kết quả nhận dạng cho thấy khả quan trên tập dữ liệu đa dạng từ thực tế RAF_DB, đạt mức cao nhất là 79.56% trên dữ liệu kiểm tra đối với phiên bản mô hình LC2FER tại . Đây là phiên bản có độ phức tạp thấp hơn cả mô hình gốc nhỏ nhất ConvNeXt V2 trong [4] nhưng lại cho kết quả cao hơn trên 10% (so với mức 69,17% của mô hình trong [4]). Mô hình LC2FER có thể được tiếp tục nghiên cứu và áp dụng trong việc kết hợp với các kiến trúc khác nhằm đem lại kết quả tốt hơn cho FER. Đặc biệt, các mô hình ở thể nhẹ nên dễ dàng tích hợp ứng dụng trên các hệ thống có năng lực tính toán không đòi hỏi quá cao, phù hợp với đa dạng điều kiện trong thực tế nhưng vẫn cho kết quả tốt đối với các bài toán thực tế.

Trong những nghiên cứu tiếp theo, chúng tôi sẽ cải tiến tích hợp lai ghép giữa các kiến trúc ViTs hiện đại với các kiến trúc CNN truyền thống nhằm đạt chất lượng cao hơn trích chọn đặc trưng cho FER và đồng thời mô hình có độ phức tạp vừa phải để phù hợp rộng rãi trong ứng dụng thực tế.

Lời cảm ơn: Nghiên cứu này được tài trợ bởi đề tài cấp Trường Đại học Mở Hà Nội, mã số MHN2024-01.21

Tài liệu tham khảo

[1] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang và D. Tao, "A Survey on Vision Transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* tập 45, pp. 87-110; DOI: 10.1109/TPAMI.2022.3152247, 2023.

[2] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell và S. Xie, "A convnet for the 2020s," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. pp. 11966-11976, doi: 10.1109/CVPR52688.2022.01167, 2022.

[3] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon và S. Xie, "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 16133-16142, doi: 10.1109/CVPR52729.2023.01548., 2023.

[4] S. Chen, Y. Ogawa, C. Zhao và Y. Sekimoto, "Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach," *ISPRS Journal of Photogrammetry and Remote Sensing,* tập 195, pp. 129–152, https://doi.org/10.1016/j.isprsjprs.2022.11.006, 2023.

[5] D. T. Long, "Efficient Multi-Task CNN for Face and Facial Expression Recognition Using Residual and Dense Architectures for Application in Monitoring Online Learning," *International Journal of Fuzzy Logic and Intelligent Systems,* tập 23, số 3, pp. 229-243. http://doi.org/10.5391/IJFIS.2023.23.3.229, 2023.

[6] D. T. Long, "EFFICIENT CNN MODEL BASED ON COMBINING RESIDUAL NETWORK AND DENSE-CONNECTED NETWORK ARCHITECTURES FOR FACIAL EXPRESSION RECOGNITION," *International Journal of Innovative Computing, Information and Control,* tập 19, số 5, p. 1661–1678. DOI: 10.24507/ijicic.19.05.1661, 2023.

[7] S. Li, W. Deng và J. Du, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2584-2593, doi: 10.1109/CVPR.2017.277, 2017.

[8] D.T.Long, "A Lightweight Face Recognition Model Using Convolutional Neural Network for Monitoring Students in E-Learning," *I.J. Modern Education and Computer Science,* tập 6, pp. 16-28, 2020.

[9] R. R. Devaram và A. Cesta, "LEMON: A Lightweight Facial Emotion Recognition System for Assistive Robotics Based on Dilated Residual Convolutional Neural Networks," *Sensors,* tập 22, số 3366, pp. 1-20, 2022.

[10] N. Zhou, R. Liang và W. Shi, "A Lightweight Convolutional Neural Network for Real-Time Facial Expression Detection," *IEEE Access,* tập 9, pp. 5573 - 5584, 2020.

[11] P. N. R. Bodavarapu và P. Srinivas, "An Optimized Neural Network Model for Facial Expression Recognition over Traditional Deep Neural Networks," *International Journal of Advanced Computer Science and Applications,* tập 12, số 7, pp. 443-451, 2021.

[12] Y. Nan, J. Ju, Q. Hua, H. Zhang và B. Wang, "A-MobileNet: An approach of facial expression recognition," *Alexandria Engineering Journal,* tập 61, p. 4435–4444, 2022.

[13] S.-C. Lai, C.-Y. Chen và J.-H. Li, "Efficient Recognition of Facial Expression with Lightweight Octave Convolutional Neural Network,"

Journal of Imaging Science and Technology, pp. 040402.1-9, 2022.

[14] I. Loshchilov và F. Hutter, "Decoupled Weight Decay Regularization," *International Conference on Learning Representations, ICLR2019,* pp. 1-8; https://doi.org/10.48550/arXiv.1711.05101, 2019.

[15] L. v. d. Maaten và G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research,* tập 9, số 86, pp. 2579-2605, 2008.

[16] D. T. Long, "Efficient DenseNet Model with Fusion of Channel and Spatial Attention for Facial Expression Recognition," *CYBERNETICS AND INFORMATION TECHNOLOGIES,* tập 24, số 1, pp. 171-189, 2024.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh và D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *IEEE International Conference on Computer Vision (ICCV),* pp. 618-626; doi: 10.1109/ICCV.2017.74, 2017.

[18] D.T.Long, "A Facial Expressions Recognition Method Using Residual Network Architecture for Online Learning Evaluation," *Journal of Advanced Computational Intelligence and Intelligent Informatics,* tập 25, số 6, pp. 1-10, 2021.

THE EFFECTS OF CONVNEXTV2 NETWORK ARCHITECTURAL FACTORS ON FACIAL EXPRESSION RECOGNITION FROM REAL-WORLD DATA

***Duong Thang Long*[†]*, Vuong Thu Trang*[†]*, Pham Quang Huy*[†]**

Abstract: *The success of Transformer models, which have shown excellent performance in natural language processing (NLP), has extended to computer vision (CV) with Vision Transformer (ViT) architectures, achieving results comparable to or surpassing traditional convolutional neural networks (CNNs) in CV like image recognition and object detection. The ConvNeXt V2 model is an improved ResNet architecture and inherits strengths from ViTs, including hierarchical structures and the fully convolutional masked autoencoder (FCMAE) mechanism; this is to provide a high-performance and simpler model. Meanwhile, facial expression recognition (FER) remains a challenge due to real-world image factors like occlusion, color variation, and facial pose. This study applies ConvNeXt V2 to FER, adjusting architectural parameters to evaluate their impact on real-world images from the RAF_DB dataset. Experimental results show how ConvNeXt V2's architectural factors affect FER model complexity and recognition quality. These provide meaningful analyses to clarify how leveraging strengths from ViT architectures combined with traditional CNN architectures can enhance model application efficiency.*

Keywords: *Vision Transformers, ConvNeXt V2 architecture, Facial Expression Recognition.*

[†] Hanoi Open University