JTE HCMUTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Deep Learning-Based Ensemble Method for Sentiment Analysis on Images

Hoang Nam Do[1], Thi Huyen Trang Phan[2*]
[1]*Nguyen Tat Thanh University, Vietnam*
[2]*Ho Chi Minh City University of Technology and Education, Vietnam*

[*]Corresponding author. *Email: trangpth@hcmute.edu.vn*

ARTICLE INFO

Received: 15/03/2024
Revised: 30/03/2024
Accepted: 12/04/2024
Published: 28/04/2024

KEYWORDS

Image sentiment analysis;
Ensemble model;
VGG19-based CNN;
ResNet50-based CNN;
Convolutional neural network.

ABSTRACT

Sentiment analysis is to identify the polarity of people's emotions toward entities as expressed in their opinions. With the development of science and technology, opinions published on social networks become more diverse in forms, including texts, images, sounds, and videos. Among them, opinions expressed through images increasingly dominate. Many image sentiment analysis methods have been published in recent years. Methods based on convolutional neural networks (CNNs) are notable. However, previous methods based on CNNs often cannot extract features well from low-resolution images. To solve the mentioned problem, in this study, we propose a method to improve the performance of sentiment analysis on images by combining two transfer learning models and a CNN model. Unlike other CNN-based models, our method can better extract local and global features on images. The proposed method was experimented on the FER2013 dataset and shown that it can improve accuracy by up to 9.03% compared to baseline methods.

Phương Pháp Kết Hợp dựa trên Mô Hình Học Sâu cho Phân Tích Tình Cảm trên Hình Ảnh

Đỗ Hoàng Nam[1], Phan Thị Huyền Trang[2*]
[1]*Trường Đại học Nguyễn Tất Thành, Việt Nam*
[2]*Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh, Việt Nam*

[*]Tác giả liên hệ. *Email: trangpth@hcmute.edu.vn*

THÔNG TIN BÀI BÁO

Ngày nhận bài: 15/03/2024
Ngày hoàn thiện: 30/03/2024
Ngày chấp nhận đăng: 12/04/2024
Ngày đăng: 28/04/2024

TỪ KHÓA

Image sentiment analysis;
Ensemble model;
VGG19-based CNN;
ResNet50-based CNN;
Convolutional neural network.

TÓM TẮT

Phân tích tình cảm là quá trình xác định phân cực cảm xúc của con người đối với một thực thể được thể hiện trong các ý kiến của họ. Cùng với sự phát triển của khoa học công nghệ, các ý kiến được đưa lên mạng xã hội trở nên đa dạng hơn về hình thức. Trong đó, các ý kiến thể hiện thông qua các hình ảnh ngày càng chiếm ưu thế. Có nhiều phương pháp phân tích tình cảm trên hình ảnh được công bố trong những năm gần đây. Đáng chú ý phải kể đến các mô hình dựa trên convolutional neural network (CNN). Tuy nhiên, các phương pháp dựa trên mô hình CNN trước đây thường không thể trích xuất tốt các đặc trưng từ hình ảnh có độ phân giải thấp. Để giải quyết vấn đề nêu trên, trong nghiên cứu này, chúng tôi đề xuất phương pháp nâng cao hiệu suất phân tích cảm xúc trên hình ảnh bằng cách kết hợp hai mô hình transfer learning và mô hình CNN. Không giống như các mô hình dựa trên CNN khác, phương pháp của chúng tôi có thể trích xuất tốt hơn các đặc trưng cục bộ và toàn cục trên hình ảnh. Phương pháp đề xuất đã được thử nghiệm trên bộ dữ liệu FER2013 và cho thấy nó có thể cải thiện độ chính xác lên tới 9,03% so với các phương pháp cơ sở.

Doi: https://doi.org/10.54644/jte.2024.1547

JTE
HCMUTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

1. Giới thiệu

Theo thống kê của trang web backlinko.com, tính đến tháng 10 năm 2023, số người sử dụng mạng xã hội là hơn 4,95 tỷ trên toàn thế giới và con số này tăng 7,07% mỗi năm. Đáng chú ý là mỗi ngày có một số lượng lớn các hình ảnh được chia sẻ lên các trang mạng xã hội, cụ thể: 3,8 tỷ hình ảnh được chia sẽ trên Snapchat, 2,1 tỷ hình ảnh được chia sẻ trên Facebook, 1,3 tỷ hình ảnh được chia sẻ trên Instagram, và 1 tỷ hình ảnh được chia sẻ trên Flickr. Tình cảm mà người đăng muốn gửi gắm thông qua các hình ảnh là một nguồn thông tin có giá trị sử dụng lớn với các hệ thống khuyến nghị, hệ thống hỗ trợ ra quyết định, và hệ thống phát hiện tin tức giả mạo. Các phương pháp phân tích tình cảm trên các hình ảnh ra đời để nắm bắt thông tin cảm xúc đó. Phân tích tình cảm trên hình ảnh là quá trình xác định phân cực cảm xúc của người dùng dành cho các thực thể được thể hiện trên các hình ảnh.

Có nhiều cách tiếp cận khác nhau được công bố để phát triển và cải thiện hiệu suất của các phương pháp phân tích tình cảm trên hình ảnh, trong đó đáng chú ý nhất là các mô hình convolutional neural network (CNN) dựa trên transfer learning, như Inception-V3 và ResNet50. In [1], Gaurav và các cộng sự đã phát triển một cách phát hiện cảm xúc trên khuôn mặt bằng cách sử dụng mô hình CNN được làm giàu bởi mô hình Inception-V3 nhằm giải quyết vấn đề mặc dù ở trong các trạng thái cảm xúc khác nhau nhưng không có quá nhiều sự khác biệt giữa các khuôn mặt. Pranav và các cộng sự [2] đã đề xuất một hệ thống nhận dạng cảm xúc dựa trên mô hình deep CNN nhằm giải quyết vấn đề hạn chế độ sâu của các mô hình CNN. Mehendale [3] đã đề xuất mô hình nhận diện khuôn mặt sử dụng CNNs nhằm giải quyết vấn đề dư thừa các đặc trưng không cần thiết trong quá trình trích xuất đặc trưng. Jaiswual và các cộng sự [4] đã xây dựng phương pháp phân lớp cảm xúc thể hiện trên khuôn mặt dựa trên mô hình CNN nhằm giải quyết vấn đề bỏ qua các đặc trưng đặc biệt trên khuôn mặt do hình ảnh sử dụng một số lượng lớn các bộ lọc. In [5], Modi và các cộng sự đã đề xuất một mô hình nhận diện biểu hiện trên khuôn mặt sử dụng mô hình CNN nhằm giải quyết các vấn đề liên quan đến overfitting do thiếu dữ liệu đào tạo phù hợp và những ảnh hưởng đến việc xác định chính xác cảm xúc trên khuôn mặt như ánh sáng, tư thế cơ thể, và sai hình dạng.

Bảng 1. *So sánh hiệu suất các phương pháp phân tích tình cảm trên hình ảnh.*

Phương pháp	Năm	Dữ liệu	Độ chính xác (%)	Độ lỗi (%)
Inception-v3-based CNN [1]	2023	FER2013	73,09	1,2
FER [11]	2022	FER2013	72,3	--
VGG19-based DCNN [2]	2020	FER2013	65,41	0,7
FERC [3]	2020	FER2013	Từ 70 đến 96	--
Jaiswual và các cộng sự [4]	2020	FER2013	70,14	1,76

Từ các phân tích ở trên và các thông tin trong Bảng 1, chúng ta có thể thấy: (i) Hầu hết các phương pháp phân tích tình cảm trên hình ảnh gần đây đều dựa trên mô hình CNN được tích hợp thêm các mô hình transfer learning. (ii) Nếu xem hiệu suất lý tưởng về độ chính xác là 100% và độ lỗi là 0% thì các phương pháp trên hoàn toàn có thể được cải thiện thêm về mặt hiệu suất. (iii) Có ít phương pháp được xây dựng dựa trên sự kết hợp các mô hình transfer learning-based deep learning. Đây chính là lý do khiến chúng tôi đề xuất phương pháp phân tích tình cảm trên hình ảnh dựa trên việc kết hợp các mô hình học sâu. Mục đích chính của đề xuất này là nhằm giảm lỗi dự đoán do việc trích xuất chưa tốt các đặc trưng ở những hình ảnh có chất lượng thấp của các mô hình đơn lẻ nhằm cải thiện hiệu suất của phân tích tình cảm trên hình ảnh.

Phương pháp được đề xuất bao gồm các khối chính sau đây: (i) Hai khối CNN kết hợp với ResNet50 và CNN kết hợp với VGG19 được xây dựng để trích xuất và biểu diễn các đặc trưng về cảm xúc được thể hiện trên hình ảnh; (ii) Khối kết hợp dựa trên cơ chế feature fusion được dùng để kết hợp đầu ra của hai khối đã được xây dựng ở bước (i); và (iii) Mô hình phân loại tình cảm dựa trên việc kết hợp các mô hình được xây dựng để xác định phân cực tình cảm của hình ảnh. Phương pháp đề xuất được thực nghiệm

trên tập dữ liệu FER2013 và được so sánh với các phương pháp cơ sở để chứng minh tính hiệu quả của nó.

Nghiên cứu này có một số đóng góp chính như sau:

- Xây dựng phương pháp phân tích tình cảm trên hình ảnh bằng cách xây dựng mô hình kết hợp dựa trên các thuật toán học sâu nhằm giải quyết vấn đề về việc trích xuất chưa tốt các đặc trưng ở các hình ảnh có chất lượng thấp của các mô hình đơn lẻ.

- Đánh giá kết quả của phương pháp đề xuất bằng cách thực nghiệm nó trên bộ dữ liệu hình ảnh khuôn mặt chuẩn được sử dụng rộng rãi.

- So sánh độ chính xác trong phân tích cảm tính của phương pháp được đề xuất với các phương pháp cơ sở để chứng minh tính hiệu quả của phương pháp.

Chúng tôi tổ chức phần còn lại của bài báo này như sau. Phần 2 trình bày tổng quan các nghiên cứu liên quan. Phần 3 mô tả vấn đề nghiên cứu của đề xuất. Phần 4 đưa ra các mô hình toán học từng bước giải quyết vấn đề nghiên cứu. Việc thu thập dữ liệu, thiết lập, kết quả thực nghiệm và đánh giá đề xuất được trình bày ở Phần 5. Kết luận và hướng nghiên cứu trong tương lai được trình bày ở Phần 6.

2. Các nghiên cứu liên quan

Ở phần này, chúng tôi tiếp tục phân tích, đánh giá chi tiết hơn các phương pháp phân tích tình cảm trên hình ảnh sử dụng các mô hình deep learning và transfer learning đã được giới thiệu ngắn gọn ở phần Giới thiệu.

Bảng 2. *So sánh hiệu suất các phương pháp phân tích tình cảm trên hình ảnh.*

Phương pháp	Cách tiếp cận	Mô hình kết hợp	Biểu diễn đặc trưng	Độ chính xác (%)
Inception-v3-based CNN [1]	Xây dựng phương pháp trích xuất và phân loại cảm xúc dựa trên việc tích hợp Inception-v3 vào mô hình CNN	Có	Có	73,09
VGG19-based DCNN [2]	Phương pháp phân tích tình cảm dựa trên việc trích hợp mô hình VGG19 vào mô hình DCNN.	Có	Có	65,41
FER [11]	Phương pháp phân tích tình cảm dựa vào việc kết hợp các mô hình custom CNN, ResNet50, và InceptionV3	Có	Có	72,30
Jia và cộng sự [12]	Kết hợp các mô hình được huấn luyện trước, AlexNet, VGGNet, và Resnet để trích xuất các đặc trưng, sau đó đưa các đặc trưng đã được kết hợp đó vào mô hình SVM cho phân tích cảm xúc.	Có	Có	71,27
Shengtao và cộng sự [13]	Kết hợp hai mô hình dựa trên ResNet là original ResNet và cropped ResNet	Có	Có	70,74
Jha và cộng sự [14]	Phân tích tình càm bằng cách sử dụng mô hình CNN	Không	Không	69,90
Liu và cộng sự [8]	Xây dựng ba mô hình CNN để trích xuất các đặc trưng khác nhau, sau đó dùng cơ chế fully connected để kết hợp các đặc trưng.	Có	Không	65,03

Nhìn vào Bảng 2 chúng ta thấy: (i) phương pháp phân tích tình cảm dựa vào việc kết hợp các mô hình đang được quan tâm và đạt được hiệu suất đáng chú ý. (ii) các phương pháp kết hợp có sử dụng các mô hình được huấn luyện trước để biểu diễn các đặc trưng đạt hiệu suất tốt nhất (iii) các phương pháp có kết hợp các mô hình nhưng không sử dụng mô hình được huấn luyện để trích xuất đặc trưng có hiệu suất không đáng kể. (iv) mô hình chính được dùng để phân tích tình cảm ở hầu hết các phương pháp là mô hình CNN và mô hình CNN khi kết hợp với một mô hình được huấn luyện trước cho việc biểu diễn các đặc trưng thường đạt kết quả tốt hơn mô hình CNN đơn lẻ.

JTE
HCMUTE
ISSN: 2615-9740

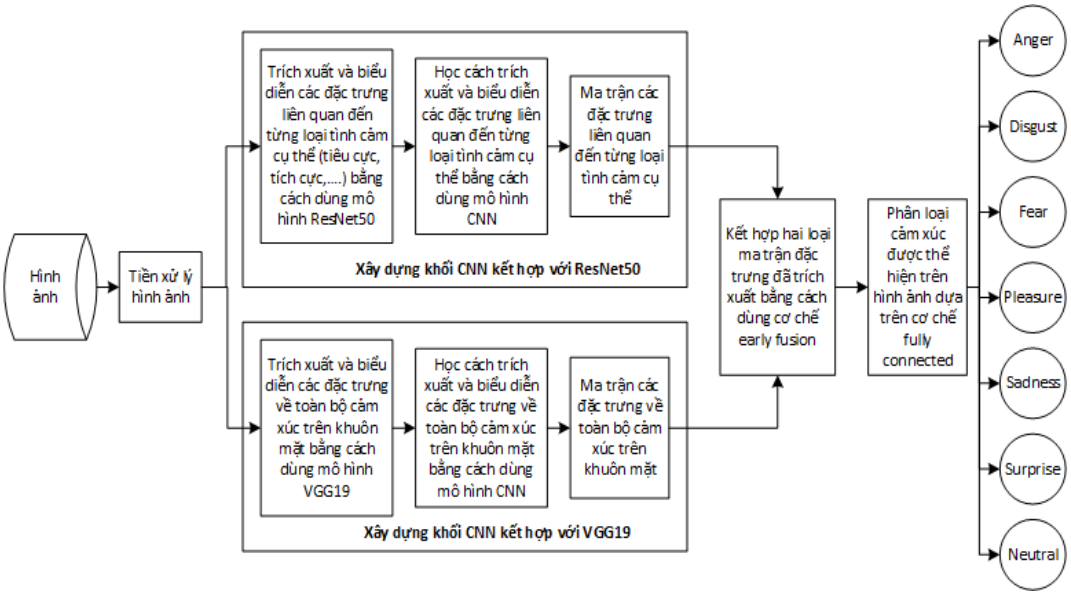
JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

3. Vấn đề nghiên cứu

Mục đích chính của đề xuất này là nhằm giảm các lỗi dự đoán do việc trích xuất chưa tốt các đặc trưng khi đầu vào là các hình ảnh có chất lượng thấp của các mô hình đơn lẻ nhằm cải thiện hiệu suất của phân tích tình cảm trên hình ảnh. Câu hỏi nghiên cứu đặt ra đó là *"Làm sao chúng ta có thể cải thiện hiệu xuất của việc phân tích tình cảm trên hình ảnh bằng cách xây dựng mô hình có khả năng trích xuất các đặc trưng cấp cao từ các hình ảnh có độ phân giải thấp?"* Từ câu hỏi nghiên cứu đặt ra, chúng tôi đưa ra vấn đề nghiên cứu như sau:

Cho một tập dữ liệu đầu vào I bao gồm n hình ảnh về khuôn mặt con người đang thể hiện các biểu hiện cảm xúc khác nhau, và tập I được biểu diễn như sau $I = \{i_k | k \in [1, n]\}$. Nghiên cứu này xem xét việc xác định phân cực tình cảm thể hiện trên hình ảnh là một nhiệm vụ phân lớp đa phân (multiple classification), trong đó, mỗi hình ảnh được gán một nhãn đơn thể hiện một trong số các loại cảm xúc (anger, disgust, fear, pleasure, sadness, surprise, and neutral). Nói một cách khác, giả sử chúng ta có một hình ảnh i_k và chúng ta đặt $I_{sen} = \{\text{anger, disgust, fear, pleasure, sadness, surprise, and neutral}\} \in R^{m \times 1}$ để biểu diễn các loại nhãn cảm xúc. Cụ thể, nếu i_k biểu hiện sự tức giận, $i_{sen} = "anger"$; nếu i_k biểu hiện sự ghê tởm, $i_{sen} = "disgust"$; nếu i_k biểu hiện sự sợ hãi, $i_{sen} = "fear"$; nếu i_k biểu hiện sự vui vẻ, $i_{sen} = "pleasure"$; nếu i_k biểu hiện sự buồn bã, $i_{sen} = "sadness"$; nếu i_k biểu hiện sự bất ngờ, i_{sen} $"surprise"$; nếu i_k biểu hiện sự trung lập, $i_{sen} = "neutral"$. Sau đây là cách xây dựng vấn đề nghiên cứu cho nghiên cứu này với các ký hiệu đã nói ở trên: Vectơ nhãn cảm xúc I_{sen} và ma trận biểu diễn hình ảnh H_i được trích xuất. Mục tiêu của nghiên cứu này là đề xuất một cách hiệu quả để dự đoán vectơ $\hat{I}_{sen}$ của các hình ảnh không được gắn nhãn cảm xúc bằng cách trích xuất và học các đặc trưng cục bộ cũng như toàn cục từ các hình ảnh có độ phân giải thấp một cách hiệu quả.

4. Phương pháp đề xuất



Hình 1. *Các bước xây dựng phương pháp đề xuất.*

Phần này trình bày chi tiết phương pháp đề xuất. Đầu vào của quá trình huấn luyện là một tập I_{train} bao gồm n_1 hình ảnh về khuôn mặt con người đang biểu hiện các trạng thái cảm xúc khác nhau, mỗi hình ảnh được gán một nhãn cảm xúc tương ứng. Đầu vào của quá trình kiểm tra là một tập I_{test} bao gồm n_2 hình ảnh về khuôn mặt con người đang biểu hiện các trạng thái cảm xúc khác nhau nhưng chưa được gán nhãn cảm xúc. Đầu ra của phương pháp là tỷ lệ các hình ảnh được gán đúng nhãn cảm xúc. Phương pháp đề xuất bao gồm các bước chính sau đây: (i) Biểu diễn các đặc trưng: bước này nhằm chuyển hình ảnh thành ma trận các vectơ đặc trưng bằng cách trích xuất các đặc trưng liên quan đến các trạng thái cảm xúc được thể hiện trên hình ảnh. Trong nghiên cứu này chúng tôi dùng hai mô hình

VGG19 [6] và ResNet50 để thực hiện bước này. (ii) Xây dựng hai khối ResNet50-based CNN và VGG19-based CNN: bước này nhằm huấn luyện và trích xuất các đặc trưng cục bộ và toàn cục từ ma trận các vectơ đặc trưng. (iii) Xây dựng khối kết hợp: bước này nhằm kết hợp các loại đặc trưng đã được trích xuất từ hai mô hình ResNet50-based CNN và VGG19-based CNN. (iv) Xây dựng khối phân loại tình cảm: bước này nhằm xác định phân cực cảm xúc của hình ảnh dựa trên các đặc trưng đã được kết hợp. Quy trình làm việc của phương pháp đề xuất được minh họa ở Hình 1.

4.1. Xây dựng mô hình CNN kết hợp với ResNet50

Mô hình CNN kết hợp với ResNet50 bao gồm hai khối chính như sau:

Khối biểu diễn đặc trưng: Khối này nhằm biểu diễn hình ảnh thành ma trận đặc trưng bao gồm các vectơ. Có nhiều kỹ thuật có thể áp dụng để thực hiện bước này như các mô hình Inception-V3, ResNet50, và VGG-ImageNet. Trong nghiên cứu này, chúng tôi chọn ResNet50 vì mô hình này có khả năng xác định và trích xuất tốt các đặc trưng liên quan đến những cảm xúc cụ thể, đặc biệt là các cảm xúc tiêu cực như ghê tởm, tức giận và buồn bã.

Điều này có nghĩa là hình ảnh i được đưa vào mô hình ResNet50 để tạo ma trận vectơ tương ứng như sau:

$$r_k = ResNet(i_k) \in \mathbb{R}^{d \times d}, với\ k = [1, n] \quad (1)$$

Trong đó, $ResNet(i_k)$ là bộ chuyển đổi của mô hình ResNet50. d là chiều của vectơ đặc trưng. Như vậy, từ hình ảnh i_k kết thúc bước này chúng ta thu được ma trận đặc trưng r_k.

Khối CNN: Khối CNN được xây dựng để giảm kích thước của ma trận đặc trưng v_kbằng cách trích xuất các đặc trưng cục bộ của từng trạng thái cảm xúc cụ thể. Khối này bao gồm các bước sau:

Tạo các ánh xạ đặc trưng (feature map) x_k bằng cách dùng 1 bộ lọc $F \in \mathbb{R}^{f \times d_f}$ để trượt trên ma trận đặc trưng r_k với chiều dài trượt f từ k đến $k + f - 1$ như sau:

$$x_k = ReLU\left(F \ominus r_{k:k+f-1} + b\right) \quad (2)$$

Trong đó, $k = [1, n]$; $\ominus$ là toán tử tích chập; $ReLU$ là hàm kích hoạt; và b chỉ bias của hàm kích hoạt. Như vậy, từ ma trận đặc trưng v kết thúc bước này chúng ta có ma trận ánh xạ đặc trưng $x = \{x_1, x_2, \dots, x_k\}$.

Thực hiện max-pooling để tạo các ma trận đặc trưng cục bộ bằng cách lựa chọn các giá trị cao nhất trong ma trận ánh xạ đặc trưng. Lý do chính là vì ma trận ánh xạ đặc trưng chứa nhiều đặc trưng, trong số đó có những đặc trưng có ý nghĩa và có những đặc trưng không có ý nghĩa. Và để nâng cao hiệu suất của mô hình, chúng ta chỉ nên tập trung vào những đặc trưng có ý nghĩa. Trong khối này, các đặc trưng cục bộ và bối cảnh sẽ được chú ý. Ma trận đặc trưng cục bộ $\hat{x}$ được xác định như sau:

$$\hat{x} = \{Max(x_1), Max(x_2), \dots, Max(x_k)\} \quad (3)$$

Trong đó, $Max(x_k)$ là hàm lấy giá trị lớn nhất từ ma trận ánh xạ đặc trưng x.

4.2. Xây dựng mô hình CNN kết hợp với VGG19

Tương tự như mô hình CNN kết hợp với ResNet50, mô hình CNN kết hợp với VGG19 bao gồm hai khối chính như sau:

Khối biểu diễn đặc trưng: Khối này nhằm biểu diễn hình ảnh thành ma trận đặc trưng bao gồm các vectơ. Trong nghiên cứu này, chúng tôi chọn VGG19 vì mô hình này có khả năng tập trung vào các vùng cơ thể rộng lớn, ví dụ như khuôn mặt, do đó có thể trích xuất được toàn bộ đặc trưng về cảm xúc trên khuôn mặt.

Điều này có nghĩa là hình ảnh i được đưa vào mô hình VGG19 để tạo ma trận vectơ tương ứng như sau:

$$v_k = VGG(i_k) \in \mathbb{R}^{d \times d}, với\ k = [1, n] \quad (4)$$

Trong đó, $VGG(i_k)$ là bộ chuyển đổi của mô hình VGG19. d là chiều của vectơ đặc trưng. Như vậy, từ hình ảnh i_k kết thúc bước này chúng ta thu được ma trận đặc trưng v_k.

Khối CNN: Khối CNN được xây dựng để giảm kích thước của ma trân đặc trưng v_kbằng cách trích xuất các đặc trưng về cảm xúc được thể hiện trên hình ảnh. Khối này bao gồm các bước sau:

JTE
HCMUTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Tạo các ánh xạ đặc trưng (feature map) y_k bằng cách dùng 1 bộ lọc $F \in \mathbb{R}^{f \times d_f}$ để trượt trên ma trận đặc trưng v_k với chiều dài trượt f từ k đến $k + f - 1$ như sau:

$$y_k = ReLU\left(F \ominus v_{k:k+f-1} + b\right) \tag{5}$$

Trong đó, $k = [1, n]$; $\ominus$ là toán tử tích chập; $ReLU$ là hàm kích hoạt; và b chỉ bias của hàm kích hoạt. Như vậy, từ ma trận đặc trưng v kết thúc bước này chúng ta có ma trận ánh xạ đặc trưng $y = \{y_1, y_2, \dots, y_k\}$.

Thực hiện max-pooling để tạo các ma trận đặc trưng cục bộ bằng cách lựa chọn các giá trị cao nhất trong ma trận ánh xạ đặc trưng. Lý do chính là vì ma trận ánh xạ đặc trưng chứa nhiều đặc trưng, trong số đó có những đặc trưng có ý nghĩa và có những đặc trưng không có ý nghĩa. Và để nâng cao hiệu suất của mô hình, chúng ta chỉ nên tập trung vào những đặc trưng có ý nghĩa. Trong khối này, các đặc trưng cục bộ và bối cảnh sẽ được chú ý. Ma trận đặc trưng cục bộ $\hat{x}$ được xác định như sau:

$$\hat{y} = \{Max(y_1), Max(y_2), \dots, Max(y_k)\} \tag{6}$$

Trong đó, $Max(y_k)$ là hàm lấy giá trị lớn nhất từ ma trận ánh xạ đặc trưng y.

4.3. Xây dựng khối kết hợp

Khối kết hợp được thiết kế để kết hợp các ma trận đặc trưng đã được tạo ra từ khối CNN kết hợp với ResNet50 và khối CNN kết hợp với VGG19. Khối này được xây dựng dựa trên cơ chế Early feature fusion [10] có hiệu quả hơn vì nó hợp nhất các nguồn dữ liệu khi bắt đầu quá trình xử lý. Trong nghiên cứu này, khối kết hợp dựa trên cơ chế early fusion được xác định như sau:

$$\hat{e} = fusion(\hat{x} \oplus \hat{y}) \tag{7}$$

Trong đó $fusion$ là khối kết hợp các đặc trưng dựa trên cơ chế tổng hợp sớm, $\oplus$ là một toán tử kết hợp, $\hat{e}$ là ma trận kết hợp đặc trưng.

4.4. Xây dựng bộ phân loại tình cảm

Ma trận kết hợp đặc trưng được đưa vào khối được kết nối đầy đủ (fully connected layer) và hàm softmax tính toán đầu ra của khối được kết nối đầy đủ để xác định giá trị phân phối của nhãn cảm xúc như sau:

$$p = softmax(W \cdot \hat{e} + b) \tag{8}$$

Trong đó W và b là ma trận trọng số và độ lệch của hàm softmax.

Huấn luyện mô hình: Phương pháp đề xuất được huấn luyện bằng cách giảm thiểu sai số entropy chéo của phân phối nhãn đúng và nhãn dự đoán theo phương trình sau:

$$\hat{l} = -[ylogp + (1 - y)\log(1 - p)] \tag{9}$$

Trong đó, y biểu thị giá trị phân phối thực của nhãn tin tức và p là giá trị phân phối dự đoán của nhãn tin tức.

5. Thực nghiệm

5.1. Dữ liệu

Để chứng minh hiệu suất của phương pháp đề xuất, chúng tôi tiến hành thực nghiệm phương pháp đề xuất trên tập dữ liệu chuẩn, là tập dữ liệu đã được công bố và được các phương pháp khác thực nghiệm. Trong nghiên cứu này, chúng tôi chọn tập dữ liệu FER2013 [9] để huấn luyện phương pháp đề xuất. Tập dữ liệu FER2013 là tập dữ liệu được công bố công khai, được phát triển để tham gia cuộc thi Kaggle ICML 2013. FER2013 bao gồm 35887 hình ảnh, mỗi hình ảnh được gán một trong bảy nhãn cảm xúc đó là anger, disgust, fear, pleasure, sadness, surprise, and neutral. Tập dữ liệu này được chia thành tập huấn luyện với 28709 hình ảnh và tập kiểm tra với 3589 hình ảnh. Tập dữ liệu này chứa một thách thức lớn mà các phương pháp phải vượt qua đó là các hình ảnh tương ứng với các nhãn cảm xúc được phân bổ không đồng đều. Nhiều hình ảnh bị lỗi như không có khuôn mặt, chỉ có một phần khuôn mặt, một số bị gán nhãn cảm xúc sai.

5.2. Thiết lập thực nghiệm

Phần này trình bày cách thiết lập các tham số khi huấn luyện và kiểm tra phương pháp được đề xuất. Chi tiết các thiết lập cho các tham số được trình bày ở Bảng 3.

Bảng 3. *Các tham số được thiết lập cho phương pháp đề xuất.*

Các thiết lập	Giá trị
Tập huấn luyện	n× 80%
Tập xác thực	Tập huấn luyện × 10%
Tập kiểm tra	n× 20%
Cách chia các tập dữ liệu	Random
ResNet50 optimizer	SGD với learning rate là 0,001
VGG19 optimizer	SGD với learning rate là 0,001
CNN optimizer	Adam với learning rate là 0,0001
Batch size của ResNet50 and VGG19	128
Batch size của CNN	64
Loss	Cross entropy
Số lượng Epoch	50

n là số lượng các hình ảnh có trong tập dữ liệu.

5.3. Phương pháp cơ sở

Trong nghiên cứu này, đối với các phương pháp cơ sở, chúng tôi chỉ thừa nhận lại kết quả đã được các nghiên cứu trước đây công bố chứ không tiến hành tái xây dựng lại các phương pháp đó. Vì vậy, để đảm bảo sự công bằng khi so sánh và đánh giá, chúng tôi đã chọn các tập dữ liệu thực nghiệm và phương pháp đánh giá trùng với các nghiên cứu được chọn làm phương pháp cơ sở. Cụ thể, các phương pháp cơ sở và phương pháp lược bỏ được dùng trong nghiên cứu này như sau:

Phương pháp cơ sở: Đối với phương pháp cơ sở, chúng tôi chọn các phương pháp được thực nghiệm trên tập dữ liệu FER2013 và dùng độ chính xác để đánh giá, bao gồm các phương pháp: Inception-v3-based CNN [1], VGG19-based CNN [7], FER [11], Jia và cộng sự [12], Shengtao và cộng sự [13], Jha và cộng sự [14], và Liu và cộng sự [8]. Phương pháp cơ sở được dùng để khẳng định tính hiệu quả của phương pháp đề xuất.

Phương pháp lược bỏ: Đối với phương pháp lược bỏ, chúng tôi tạo phương pháp lược bỏ bằng cách bỏ bớt các khối trong phương pháp đề xuất và tiến hành thực nghiệm chúng trên tập dữ liệu FER2013. Phương pháp lược bỏ được dùng để khẳng định lại vai trò và sự cần thiết phải có của các thành phần trong phương pháp đề xuất. Trong nghiên cứu này, chúng tôi thiết kế ba phương pháp lược bỏ như sau:

- NoVGG là phương pháp đề xuất đã bỏ đi khối CNN kết hợp với VGG19. Nghĩa là phương pháp đề xuất chỉ còn lại mô hình CNN kết hợp với ResNet50 bao gồm các khối là ResNet50, CNN, và bộ phân loại tình cảm.

- NoResNet là phương pháp đề xuất đã bỏ đi khối CNN kết hợp với ResNet50. Nghĩa là phương pháp đề xuất chỉ còn lại mô hình CNN kết hợp với VGG19 bao gồm các khối là VGG19, CNN, và bộ phân loại tình cảm.

- NoCNN là phương pháp đề xuất đã bỏ đi khối CNN trong hai mô hình CNN kết hợp với ResNet50 và khối CNN kết hợp với VGG19. Nghĩa là phương pháp đề xuất là phương pháp kết hợp giữa hai mô hình ResNet50 và VGG19. Phương pháp NoCNN bao gồm các khối chính là khối ResNet50, khối VGG, khối kết hợp, và bộ phân loại tình cảm.

Phương pháp đánh giá: Để đảm bảo tính công bằng khi so sánh với các phương pháp cơ sở và các phương pháp lược bỏ, chúng tôi chọn phương pháp đánh giá chung giữa các phương pháp cơ sở đó là độ chính xác.

5.4. Kết quả và Thảo luận

Để đánh giá hiệu suất của phương pháp đề xuất, đầu tiên chúng tôi tiến hành so sánh hiệu suất của nó trên từng loại nhãn tình cảm như trình bày ở Bảng 4.

Bảng 4. *Ma trận nhầm lẫn của phương pháp đề xuất (%).*

Trang thái tình cảm	anger	disgust	fear	pleasure	sadness	surprise	neutral	Total	Accuracy (%)
anger	348	7	65	15	20	15	21	491	70,88
disgust	1	39	6	1	8	0	0	55	70,91
fear	27	8	301	0	115	3	74	528	57,01
pleasure	13	0	23	806	5	17	15	879	91,70
sadness	30	1	66	12	423	12	50	594	71,21
surprise	5	0	9	11	7	359	25	416	86,30
neutral	67	0	58	34	16	10	415	626	70,45

Dựa vào Bảng 4, có thể thấy phương pháp đề xuất đạt hiệu suất tốt nhất khi phân loại các tình cảm tích cực (pleasure và surprise). Phương pháp đề xuất có khả năng phân loại tốt một số tình cảm tiêu cực (anger, disgust và sadness) và tình cảm trung tính (neutral) do tận dụng được ưu điểm của mô hình ResNet50. Tuy nhiên, với tình cảm tiêu cực còn lại (fear), phương pháp đạt hiệu suất chưa tốt. Một lý do có thể dẫn đến nhược điểm này là cả hai mô hình ResNet50 và VGG19 không có mô hình nào có khả năng nắm bắt tốt các đặc trưng liên quan đến biểu hiện ghê tởm do đó dù có kết hợp lại vẫn không có khả năng nắm bắt tốt các đặc trưng đó. Mặc dù vậy, kết quả đạt được cũng cho thấy rằng việc kết hợp nhiều mô hình phân loại có thể cải thiện tỷ lệ dự đoán sai vì các mô hình thành viên có thể bổ sung các phân loại chính xác của nhau.

Tiếp theo, chúng tôi tiến hành so sánh hiệu suất của phương pháp đề xuất với các phương pháp lược bỏ của nó như trình bày ở Bảng 5.

Bảng 5. *Hiệu suất của phương pháp đề xuất và phương pháp lược bỏ (%).*

Phương pháp lược bỏ	anger	disgust	fear	pleasure	sadness	surprise	neutral
NoVGG	65,01	64,96	49,03	90,02	60,15	80,01	75,02
NoResNet	61,79	66,84	45,10	85,16	54,49	76,26	70,69
NoCNN	68,17	69,06	52,65	91,03	63,74	80,08	72,19
Phương pháp đề xuất	70,88	70,91	57,01	91,70	71,21	86,30	70,45

Nhìn vào các Bảng 5 ta có thể thấy mỗi một khối trong mô hình đề xuất đều góp phần vào hiệu suất có thể đạt được của phương pháp. Cụ thể, hiệu suất đạt được của phương pháp NoResNet thấp hơn của phương pháp NoVGG, nghĩa là khối CNN kết hợp với ResNet50 có nhiều ảnh hưởng đến hiệu suất của toàn bộ phương pháp hơn khối CNN kết hợp với VGG19. Và phương pháp đề xuất đạt được hiệu suất tốt nhất cho thấy sự kết hợp giữa ResNet50, VGG19, và CNN cho phép kết hợp và tận dụng ưu điểm của từng phương pháp một cách khá tốt.

Cuối cùng, chúng tôi so sánh hiệu suất của phương pháp đề xuất và các phương pháp cơ sở như trình bày ở Bảng 6. Nhìn vào Bảng 6 ta thấy hầu hết các phương pháp sử dụng mô hình transfer learning để biểu diễn đặc trưng trước khi đưa vào mô hình phân loại tình cảm dựa trên các mô hình học sâu đều có kết quả tốt hơn so với các mô hình không sử dụng. Các phương pháp sử dụng mô hình học sâu đơn lẻ đạt được hiệu suất thấp nhất. Phương pháp đề xuất đã cải thiện được độ chính xác so với các phương pháp trước đây thấp nhất là 0,07% và cao nhất là 9,03%. Lý do chính cho kết quả này là phương pháp được đề xuất bao gồm hai mô hình trích xuất và học cách biểu diễn các đặc trưng có khả năng thực hiện đồng thời, mỗi mô hình chú ý đến một nhóm các đặc trưng nhất định. Ví dụ, mô hình CNN kết hợp với VGG19 tập trung vào các đặc trưng thể hiện cảm xúc trên khuôn mặt, trong khi mô hình ResNet tập trung vào các đặc trưng thể hiện từng loại cảm xúc.

Bảng 6. *So sánh hiệu suất giữa các phương pháp (%).*

Phương pháp	Accuracy
Inception-v3-based CNN [1]	73,09
VGG19-based CNN [7]	65,41
FER [11]	72,30
Jia và cộng sự [12]	71,27
Shengtao và cộng sự [13]	70,74
Jha và cộng sự [14]	69,90
Liu và cộng sự [8]	65,03
Phương pháp đề xuất	74,06

6. Tổng kết

Nghiên cứu này đề xuất một phương pháp mới để phân tích tình cảm trên các hình ảnh. Ưu điểm của phương pháp đề xuất là kết hợp ưu điểm của các phương pháp trích xuất và biểu diễn đặc trưng dựa trên các mô hình transfer learning như ResNet50 và VGG19 cho những hình ảnh có chất lượng thấp. Phương pháp đề xuất đã trả lời được câu hỏi nghiên cứu đã được đặt ra thể hiện ở việc cải thiện được hiệu suất của các phương pháp cơ sở. Tuy nhiên, nhược điểm chính của nghiên cứu này là chúng tôi chỉ thừa nhận lại kết quả đã được bảo cáo bởi các nghiên cứu trước mà không tái tạo lại chúng trong quá trình đánh giá, so sánh để xác nhận chắc chắn hơn về tính hiệu quả của mô hình đề xuất. Ngoài ra, phương pháp đề xuất chỉ được thực nghiệm trên một tập dữ liệu FER2013, và chỉ dùng một phương pháp đánh giá là độ chính xác để so sánh, điều này dẫn đến việc khó đưa ra đánh giá khách quan về hiệu suất đạt được của phương pháp đề xuất. Nghĩa là khó đảm bảo được khả năng tổng quát hóa của phương pháp. Trong tương lai, chúng tôi dự định xây dựng các mô hình kết hợp cho việc phân tích tình cảm đa phương thức hoặc kết hợp giữa fuzzy logic và các mô hình học sâu để giải quyết các vấn đề biểu diễn các đặc trưng không chắc chắn khi xây dựng các phương pháp phân tích tình cảm trên hình ảnh.

Xung đột lợi ích

Các tác giả tuyên bố không có xung đột lợi ích trong bài báo này.

Tuyên bố dữ liệu sẵn có

Dữ liệu hỗ trợ cho các khám phá của nghiên cứu này khi độc giả yêu cầu một cách hợp lý sẽ được tác giả liên hệ cung cấp.

TÀI LIỆU THAM KHẢO

[1] G. Meena, K. K. Mohbey and S. Kumar, "Sentiment analysis on images using convolutional neural networks based Inception-V3 transfer learning approach," *International journal of information management data insights*, p. 100174, 2023.
[2] E. Pranav, S. Kamal, C. S. Chandran and M. Supriya, "Facial emotion recognition using deep convolutional neural network," in *6th International conference on advanced computing and communication Systems (ICACCS)*, IEEE, 2020, pp. 317-320.
[3] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," *SN Applied Sciences*, vol. 2, no. 3, p. 446, 2020.
[4] A. Jaiswal, A. K. Raju and S. Deb, "Facial emotion detection using deep learning," in *International conference for emerging technology (INCET)*, IEEE, 2020, pp. 1-5.
[5] S. Modi and M. H. Bohara, "Facial emotion recognition using convolution neural network," in *5th international conference on intelligent computing and control systems (ICICCS)*, IEEE, 2021, pp. 1339-1344.
[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[7] A. Agrawal and N. Mittal, "Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy," *The Visual Computer*, vol. 36, no. 2, pp. 405-412, 2020.
[8] K. Liu, M. Zhang, and Z. Pan, "Facial expression recognition with CNN ensemble," in *International conference on cyberworlds (CW)*, IEEE, 2016, pp. 163-166.
[9] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing: 20th International Conference, ICONIP*, Daegu, Korea, Springer, 2013, pp. 117-124.
[10] K. Gadzicki, R. Khamsehashari and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *IEEE 23rd international conference on information fusion (FUSION)*, IEEE, 2020, pp. 1-6.
[11] E. G. Moung, C. C. Wooi, M. M. Sufian, C. K. On, and J. A. Dargham, "Ensemble-based face expression recognition approach for image sentiment analysis," *Int. J. Electr. Comput. Eng*, vol. 12, no. 3, pp. 2588-2600, 2022.

[12] C. Jia, C. L. Li, and Z. Ying, "Facial expression recognition based on the ensemble learning of CNNs," in *IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, IEEE, 2020, pp. 1-5.
[13] G. Shengtao, X. Chao, and F. Bo, "Facial expression recognition based on global and local feature fusion with CNNs," in *IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, IEEE, 2019, pp. 1-5.
[14] V. Jha, P. D. Shenoy, and K. Venugopal, "Development of facial expression classifier using neural networks," in *IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*, IEEE, 2019, pp. 1-4.

Do Hoang Nam received a master's degree in information systems from Graduate University of Sciences and Technology, Vietnam. He is currently a lecture in the Faculty of Information Technology, Nguyen Tat Thanh University. His current research interests include natural language processing, multimodal sentiment analysis, machine learning, and deep learning. Email: namdh@ntt.edu.vn

Phan Thi Huyen Trang received an M.S. degree in computer science from the University of Science and Technology, The University of Da Nang, Vietnam, in 2015, and a Ph.D. degree in computer science from Yeungnam University, South Korea, in 2020. She was an assistant professor in the Department of Computer Engineering, Yeungnam University, South Korea, from 2021 to 2024. She is currently a lecture in the Faculty of Information Technology, HCMC University of Technology and Education. She has authored ten journal articles and fifteen conference papers as the first author. Her research interests include sentiment analysis, fake news detection, text summarization, and decision support systems.
Email: trangpth@hcmute.edu.vn. ORCID: https://orcid.org/0000-0002-7466-9562