Nhận diện cảm xúc của sinh viên dựa trên biểu cảm khuôn mặt trong giáo dục trực tuyến

Student emotion recognition based on facial expressions in online education

# Tóm tắt

Giáo dục trực tuyến dần trở thành một xu hướng mới đầy tiềm năng và thách thức. Đặt biệt trong hoàn cảnh nghiêm trọng của dịch bệnh COVID-19 như hiện nay, hầu hết các trường học đều đang đóng cửa, giáo dục trực tuyến được xem là một trong những giải pháp tối ưu nhất hiện nay. Có nhiều nghiên cứu trước đây đã chỉ ra rằng, có một mối quan hệ chặt chẽ giữa biểu cảm khuôn mặt và cảm xúc của một người nào đó. Do đó, để đánh giá khách quan chất lượng của các lớp học trực tuyến, chúng tôi đề xuất phương pháp nhận diện cảm xúc tự động dựa trên mạng tích chập CNN (Convolution Neural Network). Mô hình cho phép nhận diện bảy loại cảm xúc khác nhau của con người. Phương pháp đề xuất được thực nghiệm dựa trên hai bộ CSDL về nhận diện cảm xúc là FER2013 và CK Plus. Thực nghiệm trên ba lớp học trực tuyến gồm ba lớp sinh viên khoa CNTT, trường Đại học sư phạm Hà Nội. Các kết quả cho thấy mô hình đề xuất không chỉ hiệu quả với các bộ dữ liệu chuẩn mà còn hoạt động tốt trong các môi trường thực nghiệm khác nhau.

**Từ khóa**: Giáo dục trực tuyến; nhận diện cảm xúc; mạng nơ ron tích chập.

# Abstract

Online education is gradually becoming a new trend full of potential and challenges. Especially in the current serious situation of the COVID-19 epidemic, most schools are closed, online education is considered one of the most optimal solutions today. There are many previous studies that have shown that there is a strong relationship between a person's facial expressions and emotions. Therefore, to objectively assess the quality of online classes, we propose an automatic emotion recognition method based on a convolutional network (CNN). The model allows to identify seven different types of human emotions. The proposed method is experimentally based on two databases on emotion recognition, FER2013 and CK Plus. Experiment on three online classes including three classes of students of IT faculty, Hanoi National University of Education. The results show that the proposed model is not only effective with standard data sets, but also works well in different experimental environments.

**Keywords:** Online education; emotion recognition; Convolution Neural Network.

# GIỚI THIỆU

Với hầu hết mọi người thì biểu cảm trên khuôn mặt là một trong những tín hiệu mạnh mẽ, tự nhiên và phổ biến nhất để con người truyền tải trạng thái cảm xúc và ý nghĩ của họ [1], [2], có rất nhiều ứng dụng liên quan đến vấn đề này như: quản lý sức khỏe [3], hỗ trợ lái xe, giao tiếp, … [4].

Ekman và Friesen [5] đã chỉ ra rằng con người nhận thức được một số cảm xúc cơ bản theo cùng một cách bất kể nền tảng văn hóa hay quốc gia nào và họ đã xác định có sáu loại cảm xúc cơ bản bao gồm: giận dữ, ghê tởm, sợ hãi, vui vẻ, buồn bã và ngạc nhiên. Trong một nghiên cứu mở rộng

khác, Ekman và Heider [21] đã bổ sung thêm một loại cảm xúc nữa là khinh bỉ.

Ngoài ra, FER 2013, một bộ cơ sở dữ liệu quy mô lớn được giới thiệu trong IMCL 2013, cũng giới thiệu và phân loại các khuôn mặt với bảy loại trạng thái cảm xúc khác nhau bao gồm: giận dữ, ghê tởm, sợ hãi, vui vẻ, buồn bã, ngạc nhiên và bình thường. Trong các nghiên cứu khác, các nhà khoa học cũng đã giới thiệu nhiều loại mô hình khác nhau để cung cấp nhiều loại cảm xúc hơn do sự phức tạp của nét mặt. Tuy nhiên, các cảm xúc mở rộng này chiếm một phần khá nhỏ trong các biểu

hiện cảm xúc hàng ngày nên chưa được đưa vào trong nghiên cứu này [7]. Hình 1 minh hoạ một số biểu cảm khuôn mặt cơ bản kèm theo các nhãn



cảm xúc tương ứng trong bộ cơ sở dữ liệu FER2013 [23], sẽ được sử dụng để thử nghiệm trong nghiên cứu này.

Sợ hãi Vui vẻ



Giận dữ



Ghê tởm

Buồn bã Ngạc nhiên



Bình thường



Hình 1. *Một số hình ảnh được gán nhãn cảm xúc trong cơ sở dữ liệu FER2013*

Trong lĩnh vực trí tuệ nhân tạo và học sâu, nhiều thuật toán nhận diện cảm xúc được đề xuất để nhận diện các biểu cảm thể hiện trên khuôn mặt. Các phương pháp sử dụng mô hình trí tuệ nhân tạo cho thấy hiệu suất tốt hơn so với các phương pháp phân lớp. Các hình ảnh được sử dụng trong bài toán nhận diện nói chung được chia làm hai loại: hình ảnh tĩnh (ảnh đơn lẻ)[8] và hình ảnh động (một chuỗi hình ảnh trong video). Việc nhận diện các hình ảnh trong video sẽ có nhiều thông tin hơn nhưng mức độ phức tạp sẽ cao hơn. Ngoài ra, các phương pháp dựa trên thị giác và sinh trắc học khác cũng có thể được áp dụng trong việc nhận diện cảm xúc khuôn mặt.

Cơ sở dữ liệu hình ảnh được dán nhãn đầy đủ bao gồm nhiều loại biểu cảm khuôn mặt là yếu tố quan trọng đối với các nhà nghiên cứu để thiết kế và thử nghiệm các mô hình hoặc hệ thống nhận diện cảm xúc. Trong nghiên cứu này, hai bộ cơ sở dữ liệu được sử dụng bao gồm: bộ dữ liệu CK Plus [9], là một bộ CSDL có kiểm soát, được tạo ra trong phòng thí nghiệm với đầy đủ ánh sáng và phông nền và bộ dữ liệu FER2013 [23], là một bộ CSDL không kiểm soát, được thu thập từ các môi trường phức tạp hơn với phông nền, ánh sáng rất khác nhau. Những hình ảnh trong CSDL FER2013 [23] được tạo ra giống với tình huống thực tế hơn nhằm giúp các mô hình có thể hoạt động tốt hơn trong môi trường thực tế.

Do hạn chế về khả năng xử lý và phần cứng, hầu hết các phương pháp phân lớp truyền thống sử dụng các đặc trưng thủ công hoặc các thuật toán học nông như: đặc trưng nhị phân cục bộ (LBP)[8] và phân tích nhân tử ma trận không âm (NMF)[11]. Với sự phát triển của khả năng xử lý và mô phỏng máy tính, tất cả các loại thuật toán học máy, chẳng hạn như mạng nơ ron nhân tạo (ANN), bộ phân lớp SVM và bộ phân loại Bayes, đã được áp dụng cho việc nhận diện cảm xúc với độ chính xác cao

hơn và đã được chứng minh trong môi trường được thí nghiệm (có kiểm soát) để có thể phát hiện khuôn mặt một cách hiệu quả. Tuy nhiên, các phương pháp này hạn chế về khả năng khái quát hóa trong khi đây là chìa khóa để đánh giá tính thực tiễn của một mô hình [12]. Các thuật toán học sâu có thể giải quyết vấn đề này và có hiệu suất khá mạnh mẽ và ổn định cả trong các môi trường thực nghiệm lẫn môi trường thực tế. Có nhiều nghiên cứu đã chỉ ra tính hiệu quả của mạng nơ- ron tích chập (CNN). Đây là một xu hướng mới khá tiềm năng vì tính hiệu quả của chúng trong các bài toán phân lớp và phát hiện đối tượng. Các mô hình này có thể hoạt động tốt trong việc giải quyết các bài toán trong lĩnh vực thị giác máy tính, đặc biệt là đối với bài toán nhận diện cảm xúc [13]. Nhiều mô hình khác nhau dựa trên cấu trúc CNN đã được đề xuất và đã đạt được kết quả tốt hơn các phương pháp trước đây. Simonyan và Zisserman [14] đã thông qua kiến trúc của các bộ lọc tích chập rất nhỏ (3×3) để tiến hành đánh giá toàn diện các mạng với độ sâu ngày càng tăng và hai mô hình ConvNet hoạt động tốt nhất đã được công bố công khai để tạo điều kiện cho các nghiên cứu sâu hơn trong lĩnh vực này. Bằng cách tăng chiều sâu và chiều rộng của mạng trong khi vẫn giữ nguyên cách tính toán, Szegedy và đồng nghiệp [15] đã giới thiệu một kiến trúc mạng nơ- ron phức hợp sâu, gọi là “Inception”, cho phép tăng hiệu suất và giảm đáng kể việc sử dụng tài nguyên tính toán. Jahandad và đồng nghiệp [16] đã giới thiệu hai kiến trúc mạng nơ-ron phức hợp (Inception-v1 và Inception-v3) dựa trên “Inception” và đã chứng minh rằng 2 mô hình này hoạt động tốt hơn các mô hình khác. Inception-v1 với mạng học sâu 22 lớp hoạt động tốt hơn mạng Inception- v3 với 42 lớp sau khi thực nghiệm với hình ảnh đầu vào có độ phân giải thấp và hình ảnh chữ ký hai chiều; tuy nhiên, Inception-v3 hoạt động tốt hơn với bộ dữ liệu ImageNet. Xu hướng chung

của mạng nơ-ron là tăng độ sâu của mạng và độ rộng của lớp. Về lý thuyết, các mô hình mạng nơ- ron càng sâu thì khả năng học càng mạnh nhưng độ phức tạp càng cao và khó huấn luyện. Ông và cộng sự [17] đã đề xuất một mô hình mạng nơ-ron dư thừa (RNN - Residual Neural Network) nhằm làm giảm độ phức tạp trong huấn luyện của các mạng sâu hơn và đã chứng minh rằng các mạng RNN này dễ tối ưu hóa hơn trong khi độ chính xác tăng lên đáng kể. Ngoài ra, một nhóm các nhà nghiên cứu đã chứng minh rằng độ chính xác của nhận diện có thể được cải thiện hơn nữa bằng cách kết hợp CNN với RNN trong đó CNN được sử dụng làm đầu vào cho RNN.

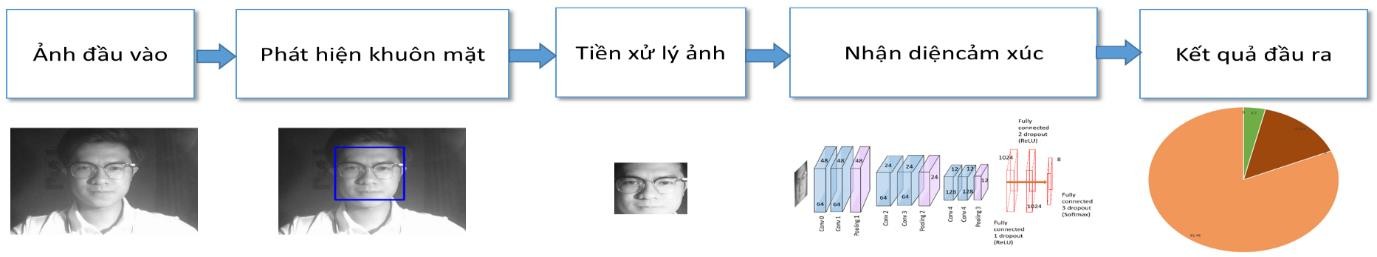
Trong suốt những thập kỷ qua, giáo dục trực tuyến đã phát triển nhanh chóng dù là tại các trường đại học hay cơ sở đào tạo [18], điều này mang lại cơ hội ứng dụng tiềm năng cho các hệ thống nhận diện cảm xúc. Vấn đề khó khăn lớn giữa lớp học trực tuyến và học trực tiếp truyền thống đó là các lớp học trực tuyến thường được coi là ít ràng buộc hơn và giao tiếp kém hiệu quả. Điều này chắc chắn sẽ dẫn đến sự nghi ngờ của giảng viên cũng như sinh viên đối với phương pháp giáo dục mới lạ này. Trong khi có một số nghiên cứu cho rằng kết quả học tập của sinh viên đạt được bằng giáo dục trực tuyến có thể tương đương với các lớp học truyền thống, ngoại trừ các kỹ năng đòi hỏi độ chính xác tối ưu và mức độ nhận thức xúc giác cao hơn [19]. Không thể phủ nhận rằng tốc độ phát triển nhanh chóng của giáo dục trực tuyến có thể mang lại sự thuận tiện và linh hoạt cho nhiều sinh viên hơn, vì vậy nó cũng có không gian phát triển rộng rãi trong tương lai. Do đó, làm thế nào để đảm bảo rằng sinh viên giữ được mức độ tập

trung và hiệu quả học tập như các lớp học truyền thống trong quá trình giáo dục trực tuyến là rất quan trọng để thúc đẩy sự phát triển hơn nữa của giáo dục trực tuyến. Để giải quyết vấn đề này, cần phải có những công cụ đánh giá chủ quan và khách quan làm cơ sở cho những sự thay đổi, cải tiến nhằm nâng cao chất lượng đào tạo.

Bằng cách kết hợp các nền tảng giáo dục trực tuyến hiện có với mô hình nhận diện nét mặt dựa trên kiến trúc của mạng nơ-ron tích chập, chúng tôi đã đề xuất một phương pháp cho phép theo dõi thời gian thực cảm xúc của sinh viên trong các khóa học trực tuyến và đảm bảo rằng phản hồi được thể hiện bằng nét mặt có thể cung cấp cho giảng viên một công cụ đánh giá khách quan, kịp thời. Giúp các nhà quản lý, giảng viên có thêm một công cụ để họ có thể linh hoạt điều chỉnh chương trình dạy học một cách phù hợp hơn và cuối cùng là nâng cao chất lượng và hiệu quả của giáo dục trực tuyến.

# PHƯƠNG PHÁP ĐỀ XUẤT

Trong phần này, giới thiệu lược đồ nhận diện cảm xúc dựa trên nền tảng học trực tuyến. Hiện tại, có hai nền tảng học trực tuyến được sử dụng phổ biến tại trường ĐH Sư phạm Hà Nội là Zoom và Google meet. Do đó, các ảnh đầu vào sẽ được thu thập chủ yếu dựa trên hai nền tảng này. Lược đồ nhận diện đề xuất bao gồm năm bước chính: thu thập ảnh đầu vào, phát hiện khuôn mặt, tiền xử lý ảnh đầu vào, nhận diện cảm xúc và hiển thị kết quả. Hình 2 minh hoạ một cách trực quan các bước của lược đồ. Một biểu đồ thống kê tổng số các cảm xúc hiện có trong lớp được tổng hợp và cung cấp cho các giảng viên.



# Hình ảnh đầu vào

Hình 2. *Lược đồ phương pháp đề xuất*

# Phát hiện khuôn mặt

Những tiến bộ trong công nghệ đã tạo ra một số lượng lớn các nền tảng giáo dục trực tuyến và tăng tính linh hoạt trong đào tạo. Hiện tại, hầu hết các nền tảng này đều tích hợp chức năng dạy trực tuyến như Zoom, Google meet, MS Team... Khi đó, giảng viên có thể dễ dàng tương tác với sinh viên thời gian thực và cũng dễ dàng thu được hình ảnh khuôn mặt của sinh viên dựa trên các camera tích hợp. Các hình ảnh khuôn mặt này có thể được sử dụng như là tập các dữ liệu đầu vào cho hệ thống đề xuất để có thể đánh giá và nhận diện cảm xúc của người học theo thời gian thực.

Các hình ảnh khuôn mặt đầu vào có thể chứa nhiều thông tin khác nhau ngoài hình ảnh khuôn mặt cần nhận diện (nhiều chi tiết khác trên ảnh nền, …) do đó, cần phải xác định chính xác vị trí khuôn mặt trong ảnh trước khi tiến hành nhận diện. Trong nhiều trường hợp, người học có thể sử dụng các loại background khác nhau, sẽ khiến cho việc phát hiện khuôn mặt khó khăn hơn. Trong nghiên cứu này, để có thể phát hiện và cắt được chính xác vị trí khuôn mặt trong ảnh, sử dụng phương pháp Haar-Cascade[20]. Các video lớp học trực tuyến sẽ được phát hiện khuôn mặt một cách tự động dựa trên phương pháp này bằng cách cắt từng khung hình của video thành từng ảnh riêng lẻ và thực hiện phát hiện khuôn mặt trên từng ảnh trong video lớp học trực tuyến.

# Tiền xử lý hình ảnh

Sau phát hiện khuôn mặt trong ảnh đầu vào dựa trên phương pháp Haar-Cascade thì việc thực hiện nhận diện cảm xúc là hoàn toàn khả thi. Một ảnh mới (chỉ có khuôn mặt) sẽ được cắt ra một cách thủ công theo tọa độ, chiều rộng, chiều cao của khuôn mặt dựa trên những khuôn mặt được phát hiện nhờ phương pháp Haar-Cascade, từ đó thu được một mảng các khuôn mặt có trong ảnh để làm hình ảnh đầu vào cho bước nhận diện tiếp theo. Việc cắt hình ảnh khuôn mặt sẽ làm giảm bớt các chi tiết dư thừa trong ảnh, nâng cao hiệu suất nhận diện. Tuy nhiên, trong quá trình thực nghiệm, các kết quả cho thấy việc nhận diện cảm xúc vẫn chưa thực sự hiệu quả một phần là do chất lượng ảnh đầu vào chưa tốt (quá tối, hoặc nhiễu, …), một phần là do kích thước hình ảnh đầu vào khác nhau, nên kích thước ảnh khuôn mặt sau khi được phát hiện cũng sẽ khác nhau. Do đó, cần phải tiến hành thêm bước tiền xử lý để chuẩn hoá các ảnh khuôn mặt đầu vào trước khi tiến hành nhận diện. Một số thao tác tiền xử lý được thực hiện trong lược đồ đề xuất bao gồm: nâng cấp hình ảnh (dựa trên việc cân bằng histogram) việc cân bằng histogram giúp cải thiện chất lượng ảnh và làm tăng độ tương phản của ảnh bằng phương cân bằng histogram cơ bản: thông kê số lượng các mức xám, tính mật độ xác suất theo số lượng mức xám và tổng số điểm ảnh cuối cùng mức xác mới sau khi cân bằng sẽ được tính theo hàm phân bố tích lũy cho mỗi mức xám (CDF), giảm nhiễu với bộ lọc Gaussian giúp làm mịn ảnh giúp các chi tiết trong ảnh được nổi bật và rõ ràng hơn, xoay ảnh dựa trên việc xác định mũi là trung tâm khuôn mặt, thay đổi kích thước ảnh cho phù hợp với kích thước đầu vào của bộ nhận diện (ảnh được chuẩn hoá về kích thước 48x48), …

# Nhận diện cảm xúc

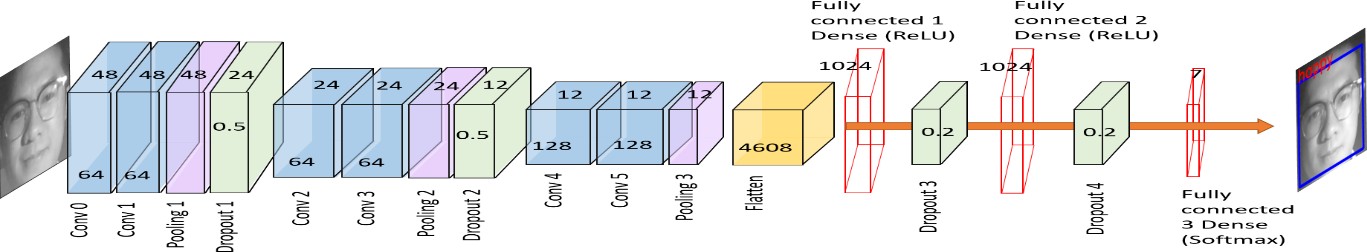
Sau khi hình ảnh khuôn mặt đã được tiền xử lý và chuẩn hoá, giai đoạn tiếp theo trong lược đồ đề xuất sẽ là việc nhận diện cảm xúc từ thông tin hình ảnh đầu vào. Trong nghiên cứu này, chúng tôi đề xuất một mô hình học sâu mạng tích chập CNN dựa trên mô hình gốc của Kuo [22] do sự vượt trội về hiệu suất và độ chính xác của nó so với các cách tiếp cận khác. Hình 3 minh hoạ các lớp của mô hình nhận diện, bao gồm các khối chính sau:

* Khối thứ nhất chứa 2 lớp tích chập mỗi lớp gồm 64 bộ lọc (channel); mỗi bộ lọc có kích thước cỡ 3 x 3 và kích thước ảnh đầu vào của bộ lọc có kích thước 48x48x1. Sau đó là hai lớp tổng hợp (pooling) có kích cỡ 2x2, bước nhảy là 2x2 và lớp dropout có tỷ lệ là 0.5 nhằm loại bỏ một vài trường hợp trong quá trình huấn luyện mạng. Việc bỏ các điểm đầu vào được thực hiện bằng cách lấy ngẫu nhiên nhưng đảm bảo một ngưỡng xác suất nào đó. Việc bổ sung thêm lớp dropout nhằm trách trường hợp overfiting trong quá trình huấn luyện.
* Khối thứ hai có cấu trúc tương tự như khối thứ nhất bao gồm 2 lớp tích chập gồm 64 bộ lọc cỡ 3x3, một lớp tổng hợp pooling cỡ 2x2 với bước nhảy 2x2 và cuối cùng là một lớp dropout với tỷ lệ

0.5. Tuy nhiên, khác với khối thứ nhất, kích thước ảnh đầu vào bộ lọc khối thứ 2 sẽ giảm một nửa còn 24x24 để giảm độ phức tạp của thuật toán và tăng độ chính xác về việc trích chọn đặc trưng của ảnh.

* Khối thứ ba cũng có cấu trúc tương tự như hai khối trước. Kích thước ảnh đầu vào cũng được tiếp tục giảm đi một nửa còn 12x12. Ngoài ra, hai lớp tích chập trong khối này được tăng cường số lượng kênh lên là 128 đồng thời bổ sung thêm lớp flatten nhằm làm phẳng dữ liệu và kết hợp các đặc trưng của ảnh đề có được đầu ra cho mô hình.
* Khối cuối cùng bao gồm các lớp kết nối đầy đủ (fully connected layer) gồm 3 lớp. Lớp đầu tiên có 1024 nơ-ron, trong đó sử dụng hàm kích hoạt ReLUs. Lớp kết nối đầu đủ sau cùng gồm 7 nơ- ron và sử dụng hàm softmax làm hàm kích hoạt để phân loại các biểu cảm bao gồm: Tức giận, ghê tởm, sợ hãi, vui vẻ, buồn, ngạc nhiên, bình thường.

Thông tin chi tiết về các lớp trong các khối của mô hình mạng nơ-ron tích chập đề xuất được mô tả trong Bảng 1.



Hình 3*. Kiến trúc mạng tích chập cho nhận diện cảm xúc*

Bảng 1. *Các tham số chi tiết cho mô hình đề xuất*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Lớp** | **Số kernel** | **Kích thước mỗi kernel** | **Bước nhảy** | **Kích thước ảnh** |
| Input | 0 | 0 | None | 48 x 48 x 1 |
| Conv2D-0 | 64 | 3 x 3 | 1 | 48 x 48 x 64 |
| Conv2D-1 | 64 | 3 x 3 | 1 | 48 x 48 x 64 |
| Pooling 1 | 0 | 2 x 2 | 2 | 48 x 48 x 64 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Lớp** | **Số kernel** | **Kích thước mỗi kernel** | **Bước nhảy** | **Kích thước ảnh** |
| Dropout 1 |  | Dropout=0.5 |  | 24 x 24 x 64 |
| Conv2D-2 | 64 | 3 x 3 | 1 | 24 x 24 x 64 |
| Conv2D-3 | 64 | 3 x 3 | 1 | 24 x 24 x 64 |
| Pooling 2 | 0 | 2 x 2 | 2 | 24 x 24 x 64 |
| Dropout 2 |  | Dropout=0.5 |  | 12 x 12 x 128 |
| Conv2D-4 | 128 | 3 x 3 | 1 | 12 x 12 x 128 |
| Conv2D-5 | 128 | 3 x 3 | 1 | 12 x 12 x 128 |
| Pooling 3 | 0 | 2 x 2 | 2 | 12 x 12 x 128 |
| Flatten |  |  |  | 1 x 1 x 4608 |
| Dense | 1024 | activation='relu' |  | 1 x 1 x 1024 |
| Dropout 3 |  | Dropout=0.2 |  | 1 x 1 x 1024 |
| Dense | 1024 | activation='relu' |  | 1 x 1 x 1024 |
| Dropout 3 |  | Dropout=0.2 |  | 1 x 1 x 1024 |
| Dense | 7 | activation='softmax' |  | 1 x 1 x 7 |
| Output | 0 | 0 | None | 1 x 1 x 7 |

# Kết quả thực nghiệm

* 1. **Bộ dữ liệu huấn luyện**

Bộ dữ liệu FER2013 [23 ] và CK Plus [24] được sử dụng để huấn luyện mô hình nhận diện cảm xúc, hai bộ dữ liệu bao gồm các ảnh đa mức xám có kích thước 48x48. Trong cơ sở dữ liệu này, hình ảnh khuôn mặt đã được cắt bỏ phần ảnh nền dư thừa xung quanh và khuôn mặt được căn giữa hình ảnh. Các hình ảnh được gán nhãn với bảy loại cảm xúc khác

nhau: giận dữ, ghê tởm, sợ hãi, vui vẻ, buồn, ngạc nhiên, bình thường. Thông tin chi tiết về cơ sở dữ liệu được mô tả trong Bảng 2.

Với bộ dữ liệu ảnh FER2013, chúng tôi sử dụng 28709 ảnh cho việc huấn luyện mô hình mạng tích chập, và 3589 ảnh được sử dụng để làm dữ liệu kiểm thử. Tương tự, với bộ dữ liệu CK Plus, 900 hình ảnh được bổ sung để huấn luyện mô hình mạng và 81 ảnh được sử dụng để kiểm thử.

Bảng 2. *Thông tin chi tiết số lượng ảnh và cảm xúc trong bộ CSDL FER2013 và CK Plus*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CSDL** | **Tổng số ảnh** | **Bộ ảnh huấn luyện** | **Bộ ảnh kiểm thử** | **Kích thước** | **Số lượng trạng thái** |
| FER2013 | 32298 | 28709 | 3589 | 48x48 | 7 |
| CK Plus | 981 | 900 | 81 | 48x48 | 7 |

# Kết quả thử nghiệm và đánh giá

Mô hình đề xuất được huấn luyện với 28709 ảnh (CSDL FER 2013) và 900 ảnh (CSDL CKPlus).

Trong quá trình thực nghiệm, mô hình đã được triển khai với ngôn ngữ lập trình Python, quá trình huấn luyện được thực hiện trên Google Colab. CPU **(**Intel(R) Xeon(R) CPU @ 2.30 GHz và 13GB RAM), GPU **(**Tesla K80 12GB, GDDR5

VRAM,Intel(R) Xeon(R) CPU @ 2.20 GHz và 13GB RAM), TPU **(**TPU Cloud, Intel(R) Xeon(R)

CPU @ 2.30 GHz và 13GB RAM). Để đánh giá

mô hình đề xuất, chúng tôi sử dụng bộ ảnh kiểm thử từ bộ dữ liệu FER2013 như đã trình bày ở trên và bộ ảnh kiểm thử từ bộ dữ liệu CK Plus với 81 ảnh, các kết quả thực nghiệm thu được được mô tả trong Bảng 3. Kết quả của quá trình huấn luyện đạt hiệu suất cao với độ chính xác (accuracy) là 0.95 sau 100 chu kỳ đào tạo. Kết quả đầu ra cho thấy có đến 3443 trên tổng số 3589 ảnh có kết quả dự đoán đúng, tỷ lệ chính xác là 95,9% đối với bộ dữ liệu ảnh kiểm thử FER2013 và 75 trên tổng số 81 ảnh có kết quả dự đoán đúng, tỷ lệ chính xác là 96,3% đối với bộ dữ liệu ảnh kiểm thử CK Plus.

Bảng 3. *Kết quả thí nghiệm kiểm tra mô hình với bộ dữ liệu kiểm thử*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CSDL** | **Số lượng ảnh tập huấn** | **Số lượng ảnh kiểm thử** | **Số lượng kết quả đúng** | **Tỷ lệ chính xác** | **Thời gian trung bình (ms)** |
| FER2013 | 28709 | 3589 | 3443 | 95,9% | 56,76 |
| CK Plus | 900 | 81 | 78 | 96,3% | 48,6 |

Bảng 4. *Kết quả thử nghiệm trên 2 bộ dữ liệu*

|  |  |  |  |
| --- | --- | --- | --- |
| **Bộ CSDL** | **Ảnh** | **Nhãn CSDL** | **Nhãn kết quả** |
|  |  | Vui vẻ | Vui vẻ |
|  | Bình thường | Buồn |

|  |  |  |  |
| --- | --- | --- | --- |
| **Bộ CSDL** | **Ảnh** | **Nhãn CSDL** | **Nhãn kết quả** |
| FER2013 |  | Vui vẻ | Vui vẻ |
|  | Sợ hãi | Sợ hãi |
|  | Tức giận | Tức giận |
|  | Buồn | Buồn |
|  | Bình thường | Bình thường |
| CK Plus |  | Tức giận | Tức giận |
|  | Ghê tởm | Ghê tởm |
|  | Bất ngờ | Bất ngờ |

# Ứng dụng thực tế

Để kiểm tra hiệu quả của phương pháp được đề xuất trong các ứng dụng thực tế, chúng tôi đã sử dụng hình ảnh học trực từ một số lớp học trên ứng dụng Zoom và đưa mô hình mạng nơ-ron tích chập vào nhận dạng cảm xúc trong ảnh, đây là hình ảnh được chụp trước khi kết thúc lớp học người giáo viên đã có vài phát biểu trước khi kết thúc lớp học trong một bầu không khí vui vẻ. Chúng tôi đã tiến hành thực nghiệm thu thập thông tin hình ảnh trong một số môn của Khoa Công nghệ thông tin, trường ĐH Sư phạm Hà Nội. Các môn học được thực nghiệm bao gồm cả ngành Sư phạm Tin và Công nghệ thông tin. Các lớp học bao gồm chủ yếu là các bạn sinh viên năm thứ 2 và năm thứ 3. Trong một nghiên cứu của Toguc và Ozkara [25] có chỉ ra rằng, mức độ cảm xúc vui vẻ của sinh viên sẽ được cải thiện đáng kể trong vòng vài phút trước khi kết thúc bài giảng, do đó, các thực nghiệm của chúng tôi được thực hiện tại một thời điểm ngẫu nhiên giữa tiết học (từ phút 30

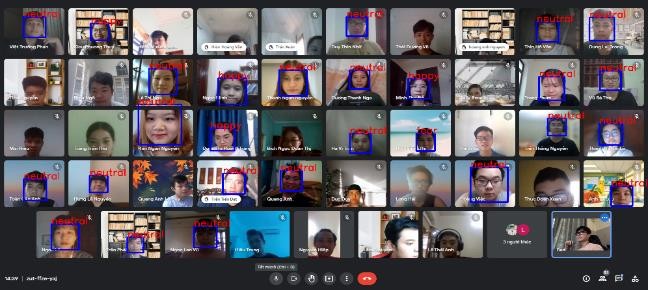
– 40, với tiết học có thời lượng 50 phút).

Bảng 5. *Một số kết quả thử nghiệm*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tên môn** | **Số lượng sinh viên** | **Số khuôn mặt phát hiện được** | **Số khuôn mặt được gán nhãn** | **Tỷ lệ nhận diện** | **Thời gian trung bình (ms)** |
| Một số vấn đề xã hội của CNTT | 48 | 27 | 27 | 56,2% | 1817.491 |
| Phần mềm nhúng  và di động | 47 | 15 | 15 | 32% | 1413.18 |
| Phát triển phần mềm cho thiết bị di động K69 | 28 | 17 | 17 | 60,7% | 1332.91 |

Hình 4 và Hình 5 minh hoạ một ví dụ về việc đánh giá cảm xúc của lớp học. Hầu hết các khuôn mặt đã được phát hiện và đánh dấu bằng các đường viền hình chữ nhật; các biểu cảm của các khuôn mặt được tiền xử lý một cách rõ nét và đã được nhận diện với các nhãn tương ứng. Trong tổng số 48 khuôn mặt, có 4 khuôn mặt được gắn nhãn “vui vẻ”, 22 khuôn mặt được gắn nhãn “bình thường” và 1 khuôn mặt được gắn nhãn “sợ hãi”. Các khuôn mặt chưa được tô viền và đánh nhãn, nguyên nhân là do các hình ảnh khuôn mặt này thiếu đi các chi tiết nét đặc trưng của khuôn mặt cơ bản hoặc do ánh sáng chưa đủ từ các thiết bị ghi hình của sinh viên.

Hình 6 minh hoạ thống kê về số lượng cảm xúc và tỷ lệ % cảm xúc nhận diện được tại một lớp học, từ đó chúng ta có thể quan sát tổng thể các cảm xúc một cách trực quan và phán đoán trạng thái cảm xúc của lớp cho phù hợp. Tuy nhiên, cần lưu ý rằng cảm xúc tổng thể của khuôn mặt có thể được đánh giá bằng nhiều phương pháp khác nhau, trong nghiên cứu này chúng tôi sử dụng phương pháp tìm ra giá trị lớn nhất của cảm xúc có trong kết quả dự đoán. Ở một số khuôn mặt được đánh dấu là “bình thường” có xác suất cao hơn nhiều so với “vui vẻ”, trong khi ở một số khuôn mặt được đánh nhãn là “vui vẻ” thì xác suất cảm xúc “bình thường” có thể chỉ thấp hơn một chút so với cảm xúc “vui vẻ”.

Hình 4. *Hình ảnh lớp học trực tuyến* Hình 5. *Nhận diện cảm xúc khuôn mặt*

# KẾT LUẬN

Bằng cách kết hợp các nền tảng lớp học trực tuyến và mô hình học sâu dựa trên kiến trúc của mô hình mạng tích chập CNN, chúng tôi đã đề xuất phương pháp phân tích cảm xúc của sinh viên dựa trên nét mặt. Các kết quả thu được được trình bày dưới dạng biểu đồ một cách trực quan giúp giảng viên, người quản lý giáo dục có thể điều chỉnh phương pháp giảng dạy, kế hoạch giảng dạy sao cho phù hợp và nâng cao hiệu quả

của việc giảng dạy trực tuyến. Để đánh giá mô hình đề xuất, chúng tôi đã sử dụng hai bộ cơ sở dữ liệu hình ảnh chuẩn FER 2013 và CK Plus để thực nghiệm. Các kết quả thực nghiệm cho thấy, mức độ nhận diện cảm xúc với độ chính xác 95,9% và 96,3% với hai bộ CSDL FER2013 và CK Plus. Các kết quả thu được cho thấy mức độ tin cậy của mô hình đề xuất là chấp nhận được và hoàn toàn có thể đáp ứng được các ứng dụng thực tế.

50

22

0 0

1 4

0

Tức giận Ghê tởm Sợ hãi Vui vẻ Buồn Ngạc Bình

nhiên thường

Số cảm xúc

30.7

14.814

0

81.48

Tức giận Ghê tởm Sợ hãi

Vui vẻ

Buồn Ngạc nhiên Bình thường

*Hình 6.* Biểu đồ đánh giá cảm xúc

Dựa trên các kết quả thực nghiệm, chúng tôi cũng đã tiến hành áp dụng mô hình vào môi trường thực tế. Một số môn học của Khoa Công nghệ thông tin, trường Đại học Sư phạm Hà nội được sử dụng làm môi trường thu thập và đánh giá. Các hình ảnh được thu thập từ 3 môn của 3 lớp. Tổng số 123 sinh viên tham gia 3 lớp học được thu thập trong đó 59 khuôn mặt chứa đầy đủ các đặc điểm đặc trưng của khuôn mặt nên có thể phát hiện một cách hiệu quả. Một số kết quả thực nghiệm cũng đã thu được và đã thể hiện được trên các lược đồ

tương ứng. Các kết quả thực nghiệm cho thấy kết quả tiềm năng của giải pháp đề xuất.

# TÀI LIỆU THAM KHẢO

1. C. Darwin and P. Prodger. *The Expression of the Emotions in Man and Animals*. John Murray, 1998.
2. Y. Tian, T. Kanade, and J. F. Cohn. *Recognizing action units for facial expression analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 2, 2001.
3. M. Bani, S. Russo, S. Ardenghi, G. Rampoldi, V. Wickline, S. Nowicki Jr, M. G. Strepparava *Behind the Mask: Emotion Recognition in Healthcare Students*. Med.Sci.Educ. 2021.
4. M. Jeong, B. C. Ko. *Driver’s Facial*

*Expression Recognition in Real-Time for Safe Driving*. Department of Computer Engineering, Keimyung University, Daegu 42601, Korea, 4

December 2018.

1. P. Ekman and W. V. Friesen. *Constants across cultures in the face and emotion*. Journal of Personality and Social Psychology*, vol. 17, no. 2, 124–129, 1971.*
2. Z. Zeng, M. Pantic, G. I. Roisman, and T.

S. Huang. *A survey of affect recognition methods: audio, visual, and spontaneous expressions.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pp. 39–58, 2009.

1. S. Li and W. *Deng. Deep facial expression*

*recognition: a survey*. IEEE Transactions on Affective Computing, In press.

1. C. Shan, S. Gong, and P. W. McOwan. *Facial expression recognition based on local binary patterns: a comprehensive study*. Image and Vision Computing, vol. 27, no. 6, pp. 803–816, 2009.
2. 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*. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern RecognitionWorkshops, pp. 94–101, San Francisco, CA, USA, July 2010
3. D. Matsumoto. *More evidence for the universality of a contempt expression*. Motivation and Emotion, vol. 16, no. 4, pp. 363–368, 1992.
4. R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn. *Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition*. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 41, no. 1, pp. 38–52, 2011.
5. A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. *Emotion recognition in the wild challenge 2014: baseline, data and protocol*. In Proceedings of the 16th International Conference

on Multimodal Interaction, pp. 461–466, ACM, Istanbul Turkey, November 2014.

1. J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju. *Attention mechanism-based CNN for facial expression recognition*. Neurocomputing, vol. 411, pp. 340–350, 2020.
2. K. Simonyan and A. Zisserman. *Very*

*deep convolutional networks for large-scale image recognition*. 2014, https:// arxiv.org/abs/1409.1556.

1. C. Szegedy, W. Liu, Y. Jia et al. *Going*

*deeper with convolutions*. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9, Boston, MA, USA, June 2015.

1. A. Jahandad, S. M. Sam, K. Kamardin, N.

N. Amir Sjarif, and N. Mohamed. *Offline signature verification using deep learning convolutional neural network (CNN) architectures GoogLeNet inception-v1 and inception-v3*. Procedia Computer Science, vol. 161, pp. 475–483, 2019.

1. K. He, X. Zhang, S. Ren, and J. Sun.

*Deep residual learning for image recognition*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, Las Vegas, NV, USA, June 2016.

1. I. Allen and J. Seaman. *Digital compass*

*learning: distance education enrollment report 2017*. Babson Survey Research Group, Babson Park, MA, USA, 2017.

1. E. Dolan, E. Hancock, and A. Wareing. *An*

*evaluation of online learning to teach practical competencies in undergraduate health science students*. The Internet and Higher Education, vol. 24, pp. 21–25, 2015.

1. A.B.Shetty , Bhoomika , Deeksha ,

J.Rebeiro , Ramyashree. *Facial Recognition using Haar Cascade and LBP Classifiers*. Journal Pre- proof, 28 July 2021.

1. P. Ekman and W. V. Friesen. A new pan cultural facial expression of emotion . Motivation and Emotion, vol. 10, no. 2, pp. 159–168, 1986.
2. C. M. Kuo, S. H. Lai, and M. Sarkis. *A compact deep learning model for robust facial expression recognition*. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.
3. P. Carrier and A. Courville. *The Facial Expression Recognition 2013 (FER-2013) Dataset.* <https://www.kaggle.com/msambare/fer2013>
4. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I.Matthews. *The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression.* <https://www.kaggle.com/shawon10/ckplus>