**NHẬN DIỆN CẢM XÚC CỦA SINH VIÊN DỰA TRÊN BIỂU CẢM KHUÔN MẶT TRONG GIÁO DỤC TRỰC TUYẾN**

Phạm Quang Huy1, Đặng Thành Trung2

*1,2Khoa Công Nghệ Thông Tin, Trường Đại học Sư Phạm Hà Nội*

**Tóm Tắt:** Với sự phát triển mạnh mẽ của công nghệ thông tin, 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 và ổn định 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, một phương pháp nhận dạng cảm xúc tự động được giới thiệu dựa trên một mô hình mạng tích chập CNN (Convolution Neural Network). Mô hình cho phép nhận dạng bảy loại cảm xúc khác nhau của con người. Phương pháp đề xuất được thực nghiệm dựa trên hai bộ CSDL về nhận dạng cảm xúc là FER2013 và CK2017. Ngoài ra, một khoá học trực tuyến gồm 30 sinh viên khoa CNTT, trườnng ĐHSPHN cũng được sử dụng để đánh giá. Các kết quả thu được cho thấy mô hình đề xuất không chỉ hiệu quả với các bộ dữ liệu chuẩn mà còn hoạt động mạnh mẽ trong các môi trường thực nghiệm khác nhau.

# 1. 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, hạnh phúc, buồn bã và ngạc nhiên. Trong một nghiên cứu mở rộng khác, Ekman và Heider đã bổ sung thêm một loại cảm xúc nữa là khinh bỉ [21].

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, hạnh phúc, buồn bã, ngạc nhiên và bình thường. Trong các cuộc 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 [6]. Hình 1 là 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.

Ghê tởm

Sợ hãi

Vui vẻ

Bình thường

Ngạc nhiên

Bất ngờ

Tức giận

Hình 1. Một số hình ảnh cảm xúc trong CSDL FER2013

Với sự phát triển của công nghệ thông tin, đặc biệt 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 được thể hiện trên khuôn mặt. Các phương pháp sử dụng các mô hình trí tuệ nhân tạo đã cho thấy một hiệu xuất tốt hơn so với các phương pháp phân lớp. Các hình ảnh được sử dụng trong bài toán nhận dạng nói chung được chia ra là hai loại: hình ảnh tĩnh (ảnh đơn lẻ) và hình ảnh động (một chuỗi hình ảnh trong video)[8]. Việc nhận dạng 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.

Các 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 dạng cảm xúc. Trong nghiên cứu này, hai bộ cơ sở dữ liệu được sử dụng bao gồm: bộ dữ liệu CK+ 48 [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, 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 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.

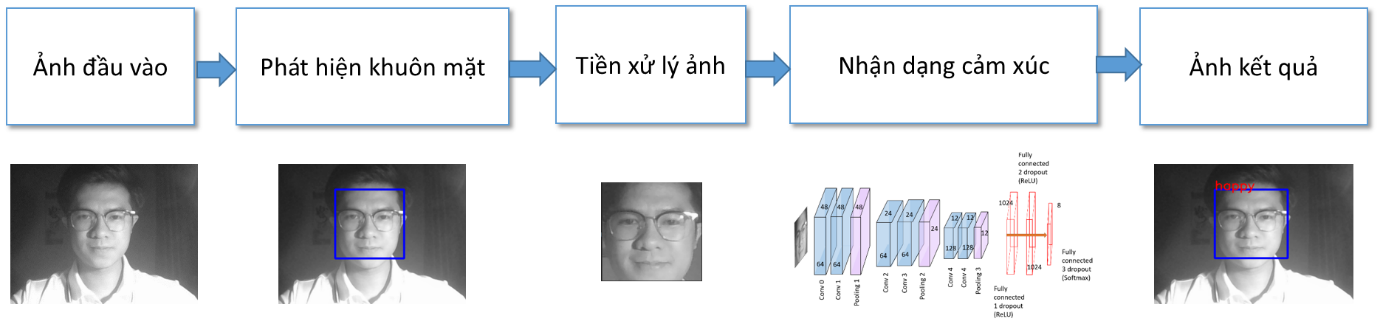
Do hạn chế về khả năng xử lý và phần cứng, hầu hết các phương pháp phân lớp truyền thống sử dụng các đặc trưng thủ công hoặc các thuật toán học nông như: đặc trưng nhị phân cục bộ (LBP)[8] và phân tích nhân tử ma trận không âm (NMF)[11]. Với sự phát triển của khả năng xử lý và mô phỏng máy tính, tất 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 dạng 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 dạng cảm xúc [13]. Nhiều mô hình khác nhau dựa trên cấu trúc CNN đã được đề xuất liên tục 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 liện. 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 kỹ lưỡng rằng các mạng RNN này dễ tối ưu hóa hơn trong khi tăng độ chính xác lên đáng kể. Ngoài ra, một nhóm các nhà nghiên cứu đã chứng minh rằng độ chính xác của nhận dạng 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 dạng cảm xúc. Vấn đề khó khăn lớn giữa lớp học trực tuyến 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, chắc chắn sẽ dẫn đến sự nghi ngờ của giảng viên cũng như sinh viên, 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 dạng nét mặt dựa trên kiến ​​trúc của mạng nơ-ron phức hợp, 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 giáo 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.

Bài báo được cấu trúc gồm 3 phần chính. Sau phần giới thiệu, một mô hình nhận dạng cảm xúc dựa trên mạng tích chập CNN được giới thiệu. Các kết quả thực nghiệm của mô hình đề xuất được thảo luận và đánh giá trong phần 3. Cuối cùng, phần kết luận sẽ tổng hợp các nội dung nghiên cứu đã được trình bày trong bài báo.

# 2. Phương pháp đề xuất

Phương pháp đề xuất chủ yếu bao gồm năm phần: hình ảnh đầu vào, phát hiện khuôn mặt, tiền xử lý hình ảnh, nhận diệm cảm xúc và kết quả hình ảnh đầu ra, trong bài báo này, tôi sẽ sử dụng các lớp học trực tuyến trên nền tảng học trực tuyến zoom và google meet của Khoa công nghệ thông tin trường đại học sư phạm Hà Nội làm ví dụ cho cuộc thử nghiệm và mô hình học sâu dựa trên CNN, trước đó có lưu ý rằng các hình ảnh gốc được thu thập từ các lớp học trực tuyến cần phải được xử lý trước, bao gồm nhận diện khuôn mặt, căn chỉnh, xoay và thay đổi kích thước, tùy theo các yếu tố khác nhau trong hình ảnh gốc. Hình 2 trình bày quá trình của quá trình nhận dạng biểu cảm khuôn mặt(FER), và các bước chi tiết của phương pháp đề xuất như sau: đầu tiên, các máy ảnh tích hợp trong các thiết bị điện tử được sử dụng để chụp ảnh khuôn mặt của các sinh viên đang theo học. Tiếp theo, thuật toán nhận dạng nét mặt được đào tạo bởi cơ sở dữ liệu biểu cảm khuôn mặt được sử dụng để phát hiện các khuôn mặt và phân loại các biểu hiện trên khuôn mặt như tức giận, ghê tởm, sợ hãi, vui vẻ, bình thường, buồn vã và ngạc nhiên. Cuối cùng, biểu đồ thống kê tổng số các cảm xúc và cung cấp cho giáo viên để có thể điều chỉnh kế hoạch dạy học kịp thời.

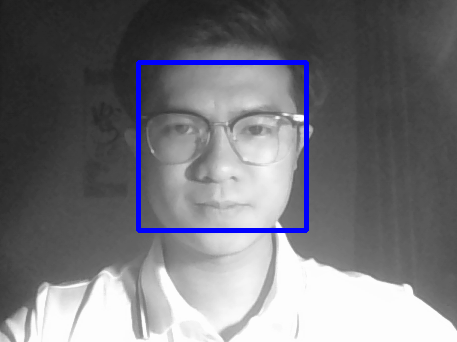
Hình 2. Qúa trình của nhận dạng biểu cảm khuôn mặt

## 2.1 Hình ảnh đầu vào

Những tiến bộ trong công nghệ đã tạo ra một số lượng lớn các nền tảng giáo dục trực tuyến và cải thiện đáng kể tính linh hoạt của giáo dục, cho phép giáo viên áp dụng các phương pháp kỹ thuật đa dạng để hỗ trợ giảng dạy mà không phải lo lắng về giới hạn số lượng sinh viên trong lớp học dựa trên lớp học truyền thống và sinh viên ở các khu vực khác nhau có thể giao tiếp trong thời gian thực mà không cần phải xem xét đến lưu lượng truy cập và các vấn đề khác. Các tài liệu giảng dạy tương tự như các lớp học truyền thống có thể được tải lên các nền tảng này để sinh viên tham khảo. Hiện tại, trong các nền tảng có chức năng giảng dạy trực tuyến, chẳng hạn như Zoom, Google meet, MS Team, giáo viên có thể áp dụng phương pháp họp video và tận dụng camera tích hợp trong thiết bị để chụp và nhận dạng nét mặt của sinh viên trong thời gian thực. Các hình ảnh được chụp sẽ được xử lý trước và sau đó được sử dụng làm đầu vào cho mô hình của CNN.

## 2.2 Phát hiện khuôn mặt

Xử lý hình ảnh trước khi dự đoán có thể làm giảm sự can thiệp của các vật thể giống như khuôn mặt trong nền khi phát hiện khuôn mặt trong một hình ảnh và sau đó chuẩn hóa các hình ảnh về ảnh đa mức xám, điều này sẽ giúp nâng cao hiệu quả của mô hình học sâu. Trong bài nghiên cứu này, tôi sử dụng phương pháp haar-cascade[20], một gói phần mềm có sẵn công khai tích hợp các thuật toán để theo dõi đặc điểm khuôn mặt, ước tính tư thế đầu, phát hiện thuộc tính khuôn mặt, v.v., làm công cụ xử lý trước. Haar-cascade cũng có thể được sử dụng để phát hiện nhiều khuôn mặt cùng một lúc. Các đặc điểm chính của từng khuôn mặt bao gồm lông mày, mắt, đầu mũi và miệng có thể được nhận ra một cách hiệu quả, và biểu hiện trên khuôn mặt có thể được phát hiện bằng các đường viền hình chữ nhật cho phù hợp, những đường viền này được xây dựng bởi các điểm đặc trưng ở cạnh của mọi mặt, bao gồm cả mặt trên và mặt dưới, xác định chiều rộng dọc, ngoài cùng bên phải và ngoài cùng bên trái, xác định chiều ngang của hình ảnh khuôn mặt. Để tránh bỏ sót thông tin trên khuôn mặt đồng thời giảm nhiễu nền, tôi đã chỉnh kích thước đường viền hình chữ nhật lên 3px lần để che được nhiều nội dung trên khuôn mặt hơn. Hơn nữa, xem xét kích thước hình ảnh đầu vào mô hình học tập được đặt trước là 48 × 48, hình ảnh được phát hiện sẽ được xoay với đầu mũi làm trung tâm và được thay đổi kích thước phù hợp để phù hợp với kích thước đầu vào Hình 3 là một ví dụ minh họa về một khuôn mặt đã được phát hiện bằng phương pháp haar-cascade và được tô viền xung quanh khuôn mặt.



Hình 3. Khuôn mặt được phát hiện và tô viền minh họa bằng phương pháp haar-cascade

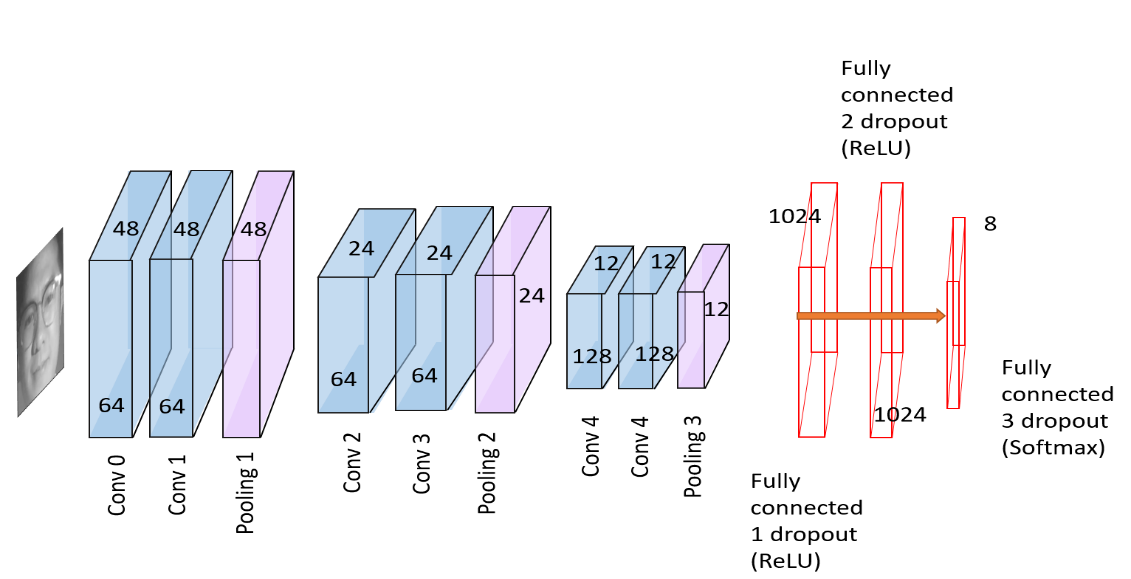
## 2.3 Tiền xử lý hình ảnh

Sau phát hiện khuôn mặt bằng phương pháp haar-cascade thì việc thực hiện nhận diện cảm xúc là hoàn toàn có thể nhưng chưa thực sự khả thi bởi hình ảnh sau khi được phát hiện khuôn mặt chưa loại bỏ đi các chi tiết dư thừa, đồng thời do cơ chế nhận diện của phương pháp haar-cascade thì hình ảnh đầu ra của phương pháp này là một hình ảnh mới có kích thước thay đổi liên tục theo thời gian thực. Trong bài nghiên cứu này, việc tiền xử lý hình ảnh được chưa ra làm hai nhiệm vụ chính: Đầu tiên, tôi tạo ra một mảng dữ liệu mới được tạo ra bằng cách cắt vừa vặn khuôn mặt từ ảnh đa mức xám được phát hiện theo các tọa độ, chiều cao, chiều rộng có được từ dữ liệu đầu ra của phương pháp haar-cascade. Tiếp theo, do cơ chế phát hiện khuôn mặt từ phương pháp haar-cascade cho ra là một hình ảnh khuôn mặt có kích thước thay đổi liên tục theo thời gian thực và không đúng với kích cỡ ảnh mà mô hình chúng ta đã đào tạo để nhận diện cảm xúc trước đó, như vậy việc thực hiện thay đổi kích thước ảnh về theo một kích thước cố định là điều tất yếu. Hình 4 mô phỏng khuôn mặt sau khi được tiền xử lý

Hình 4. Hình ảnh khuôn mặt sau khi được tiền xử lý trước khi đưa vào mô hình CNN

## 2.4 Nhận diện cảm xúc

Kiến trúc ứng dụng của mô hình học sâu dựa trên CNN được minh họa trong Hình 5, dựa trên kết quả nghiên cứu do Kuo đề xuất[22] và hiệu suất của mô hình này trong FER so với mô hình tương tự khác cũng đã được chứng minh. Khởi tạo một mô hình dự đoán với các khối chứa các lớp như 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 định nghĩa kích thước đầu vào của bộ lọc với kích thước 48 x 48 x 1, theo sau đó là hai lớp tổng hợp (pooling) có kích cỡ 2 x 2, bước nhảy là 2 x 2 và lớp dropout có tỷ lệ là 0.5 nhằm loại bỏ một vài unit trong suốt quá trình training theo xác suất, đây là một kỹ thuật nhằm tránh trường hợp over-fitting trong suốt quá trình training. Khối thứ hai vẫn tương tự như khối thứ nhất bao gồm 2 lớp tích chập gồm 64 bộ lọc cỡ 3 x 3, một lớp tổng hợp cỡ 2 x 2 với bước nhảy 2 x 2 và cuối cùng là một lớp dropout với tỷ lệ 0.5. Khối thứ ba cũng tương tự như khối thứ nhất và khối thứ hai nhưng ở khối này hai lớp tích chập được tăng cường số lượng channel lên là 128 channel đồng thời thêm một lớp flatten nhằm làm phẳng dữ liệu và kết hợp các đặc điểm của ảnh đề có được đầu ra cho model. Khối cuối cùng bao gồm các lớp kết nối đầy đủ (fully connected layer) gồm 1024 nơ-ron, trong đó sử dụng hàm kích hoạt ReLUs[23-24]. Lớp kết nối đầu đủ sau cùng gồm 7 nơ-ron và sử dụng hàm softmax[25] làm hàm kích hoạt để phân loại các biểu cảm bao gồm tức giận, ghê tởm, sợ hãi, vui vẻ, bình thường, buồn, ngạc nhiên. Chi tiết hơn về mô hình mạng nơ-ron tích chập được mô tả trong Bảng 1.



Hình 5. Kiến trúc được áp dụng của mô hình CNN

Bảng 1. Các thông số chi tiết được áp dụng trong mô hình mạng nơ-ron tích chập

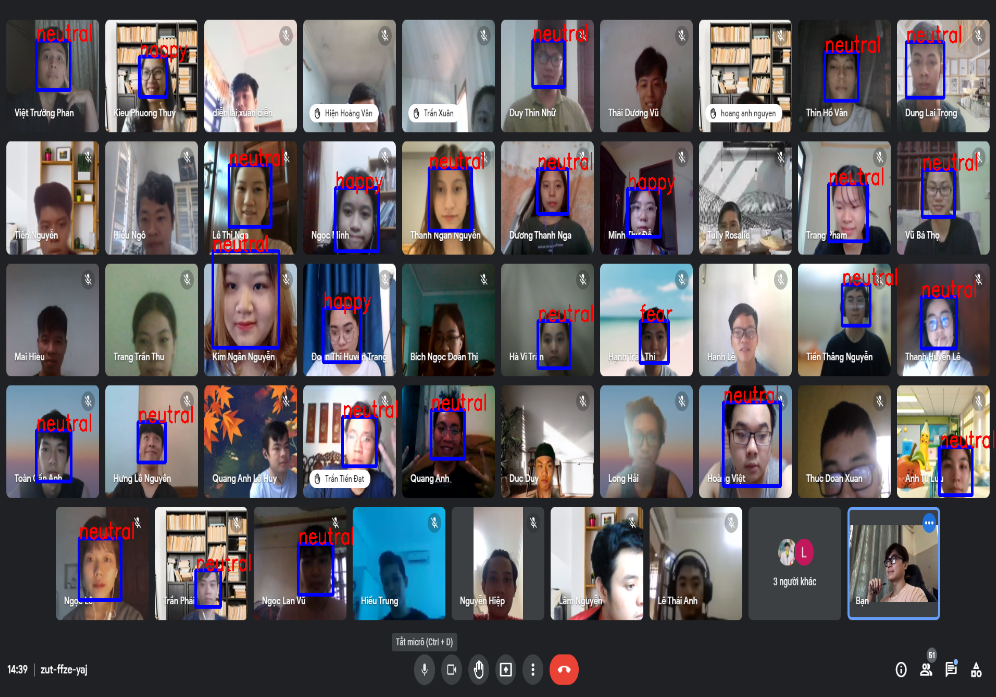
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Lớp | Số kernel | Kích thước mỗi kernel | Bước nhảy | Kích thước ảnh |
| Input | 0 | 0 | None | 48 x 48 x 1 |
| Convolutional-0 | 64 | 3 x 3 | 1 | 48 x 48 x 64 |
| Convolutional-1 | 64 | 3 x 3 | 1 | 48 x 48 x 64 |
| Pooling 1 | 0 | 2 x 2 | 2 | 48 x 48 x 64 |
| Convolutional-2 | 64 | 3 x 3 | 1 | 24 x 24 x 64 |
| Convolutional-3 | 64 | 3 x 3 | 1 | 24 x 24 x 64 |
| Pooling 2 | 0 | 2 x 2 | 2 | 24 x 24 x 64 |
| Convolutional-4 | 128 | 3 x 3 | 1 | 12 x 12 x 128 |
| Convolutional-5 | 128 | 3 x 3 | 1 | 12 x 12 x 128 |
| Pooling 3 | 0 | 2 x 2 | 2 | 12 x 12 x 128 |
| Fully connected 1 |  | Dropout=0.2 |  | 1 x 1 x 1024 |
| Fully connected 2 |  | Dropout=0.2 |  | 1 x 1 x 1024 |
| Output | 0 | 0 | None | 1 x 1 x 7 |

# 3. Kết quả thực nghiệm

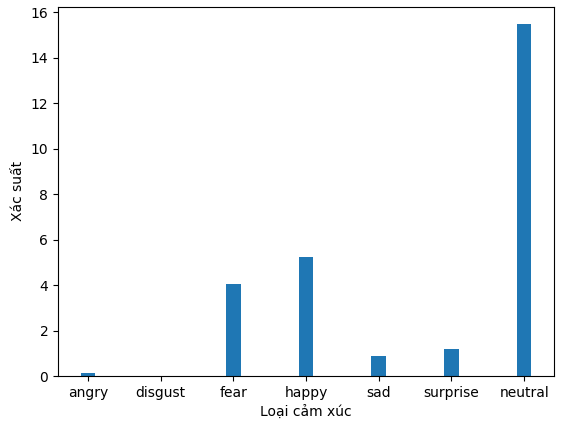
Để kiểm tra hiệu quả của phương pháp được đề xuất trong các ứng dụng thực tế, tôi đã chụp một hình ảnh bao gồm 48 người từ một lớp học trực tuyến được tổ chức trên Zoom và đưa mô hình mạng nơ-ron tích chập vào nhận dảng cảm xúc trong ảnh, đây là hình ảnh được chụp trước khi kết thúc lớp học người giáo viên đã có vài phát biểu trước khi kết thúc lớp học trong một bầu không khí vui vẻ. Ngoài ra, mọi người đều được thông báo rằng lớp học sắp kết thúc, một thí nghiệm được thực hiện bởi Toguc và Ozkara [25], mức độ hạnh phúc của sinh viên sẽ được cải thiện đáng kể trong vòng vài phút trước khi kết thúc bài giảng, do đó, tương tự trong hình ảnh được đưa vào thực nghiệm như hình số 4, có thể suy ra rằng cảm xúc của hầu hết các khuôn mặt trong hình ảnh này là vui vẻ hoặc bình thường. Hình 6 cho thấy đầu vào và Hình 7 cho thấy đầu ra hình ảnh của mô hình CNN. Có thể thấy rõ ràng từ kết quả rằng tất cả các khuôn mặt đã được nhận dạng và đánh dấu bằng các đường viền hình chữ nhật, và các biểu cảm của các khuôn mặt được chụp một cách rõ nét cũng được gắn nhãn. 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”. Tuy nhiên, có tới 21 khuôn mặt chưa được tô viền và đánh nhãn, nguyên nhân là do các khuôn mặt này thiếu đi các chi tiết nét mặt cơ bản hay các đặc điểm của một khuôn mặt cho việc nhận diện khuôn mặt hoặc do ánh sáng chưa đủ từ các hình ảnh ghi được từ thiết bị ghi hình của các bạn sinh viên. Hình 8 cho thấy biểu đồ phân bố xác suất cảm xúc của 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 “giận dữ”, “buồn” và “ngạc nhiên” đều xuất hiện trong biểu đồ xác suất nhưng trong hình ảnh đầu ra khuôn mặt được dán nhãn cảm xúc “giận dữ”, “buồn” và “ngạc nhiên” lạ không hề xuất hiện. Điều này có thể giải thích như sau: trên cùng một khuôn mặt sau khi đi qua mô hình CNN thì kết quả thu được sẽ là một mảng gồm xác suất của các loại cảm xúc có trong khuôn mặt, nhưng cảm xúc tổng thể của khuôn mặt có thể được đánh giá bằng nhiều phương pháp khác nhau, trong bài nghiên cứu này 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 “hạnh phúc”, trong khi ở một số khuôn mặt được đánh nhãn là “hạnh phúc” thì xác suất cảm xúc “bình thường” có thể chỉ thấp hơn một chút so với cảm xúc “hạnh phúc”. Nhìn chung, kết quả của thí nghiệm này có thể hỗ trợ thuật lợi cho hoạt động của mô hình khi áp dụng vào môi trường thực tế.



Hình 6. Hình ảnh đầu vào của mô hình CNN



Hình 7. Hình ảnh đầu ra của mô hình CNN



Hình 8. Biểu đồ phân bổ xác suất cảm xúc

# 4. Kết luận

Trong nghiên cứu này, bằng cách kết hợp các nền tảng lớp học trực tuyến và mô hình học sâu dựa trên kiến ​​trúc của CNN, tôi xây dựng một phương pháp phân tích cảm xúc của sinh viên dựa trên nét mặt của họ từ góc độ mô phỏng máy tính. Kết quả tổng thể có thể được trình bày dưới dạng biểu đồ một cách trực quan và giáo viên có thể điều chỉnh phương pháp giảng dạy của sao mình cho phù hợp để nâng cao hiệu quả của việc giảng dạy trực tuyến. Với việc tham khảo các nghiên cứu của Ekman và cộng sự của ông ấy vào năm 2013, các cảm xúc được phân loại thành tức giận, ghê tởm, sợ hãi, hạnh phúc, buồn bã, ngạc nhiên và bình thường giống như trong phương pháp đề xuất. Để chứng minh tính khả thi của phương pháp này trong môi trường thực tế, tôi đã chụp một hình ảnh bao gồm những hình ảnh chứa khuôn mặt của tất cả những sinh viên tham gia cùng một lúc trong một lớp học trực tuyến thực sự của Khoa công nghệ thông tin Trường đại học sư phạm Hà Nội, trong hình ảnh ví dụ minh họa có 48 người tham gia lớp học này và thời gian chụp lại bức hình được xác định vào lúc chuẩn bị kết thúc lớp học. Tổng cộng 48 khuôn mặt đã được chụp, trong đó 27 khuôn mặt chứa đầy đủ các đặc điểm đặc trung của khuôn mặt do đó có thể phát hiện một cách hiệu quả. Bằng cách đưa hình ảnh này áp dụng vào mô hình CNN, tôi đã thu được các nhãn cảm xúc cho từng khuôn mặt hợp lệ và có được cảm xúc tổng thể tại thời điểm đó. Người ta đã chứng minh rằng phương pháp có khả năng ứng dụng tốt trong hoạt động thực tiễn và đóng vai trò tích cực trong việc giải quyết các vấn đề như thiếu tính ràng buộc đối với sinh viên, giáo viên không thể phản hồi kịp thời. Cuối cùng, phương pháp đề xuất sẽ góp phần nâng cao chất lượng giáo dục trực tuyến.

Cùng với những lợi ích nêu trên, vẫn còn nhiều vấn đề cần được cải thiện trong phương pháp này và các ứng dụng của nó. Từ góc độ công nghệ, với sự phát triển của mô phỏng máy tính, các thuật toán có hiệu suất tốt hơn và thời gian hoạt động ngắn hơn, bao gồm các mô hình tiền xử lý và học sâu, sẽ liên tục được phát triển theo thời gian. Ví dụ: tiền xử lý hình ảnh sau khi phát hiện khuôn mặt bao gồm căn chỉnh, cắt ảnh về đúng kích thước, nhưng khi gặp các vấn đề, chẳng hạn như ngược sáng và sự không hoàn chỉnh của khuôn mặt do môi trường phức tạp gây ra, các phương pháp hiện tại này luôn bất lực và những thiếu sót này có thể được giải quyết trong tương lai. Hơn nữa, mặc dù mô hình CNN trong phương pháp đề xuất hiện đang hoạt động tốt, nó sẽ được thay thế bằng các mô hình có khả năng học tập cao hơn và độ chính xác phân loại cao hơn trong tương lai. Do đó phương cần được điều chỉnh và duy trì thường xuyên, đồng thời áp dụng các thuật toán và công nghệ tiên tiến hơn để cập nhật nó.

Ngoài ra, với số lượng lớn người tham gia các lớp học trực tuyến, tôi không có cách nào để đảm bảo rằng tất cả mọi người đều giữ được mức độ tập trung cao và khi đó biểu hiện cảm xúc của sinh viên có thể không thể hiện đầy đủ cảm xúc của họ do những yếu tố chủ quan. Thực hiện các biện pháp như đặt ngưỡng có thể lọc ra một số thông tin không hợp lệ và làm nổi bật cảm xúc chính trong ảnh. Cuối cùng, kết quả là hiệu quả giảng dạy có thể được cải thiện.

# Tính sẵn có của dữ liệu

Dữ liệu được sử dụng trong bài nghiên cứu này có thể được truy cập bởi độc giả thông qua website của công ty Kaggle (một công ty con của Google LLC) là một cộng đồng trực tuyến bao gồm các nhà khoa học dữ liệu và những nhà nghiên cứu về thực hành học máy: <https://www.kaggle.com/msambare/fer2013> và <https://www.kaggle.com/shawon10/ckplus>

# 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 onPattern 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.
5. 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.*
6. 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.
7. S. Li and W. *Deng. Deep facial expression recognition: a survey*. IEEE Transactions on Affective Computing, In press.
8. 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.
9. 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
10. D. Matsumoto. *More evidence for the universality of a contempt expression*. Motivation and Emotion, vol. 16, no. 4, pp. 363–368, 1992.
11. 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.
12. 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.
13. 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.
14. K. Simonyan and A. Zisserman. *Very deep convolutional networks for large-scale image recognition*. 2014, https:// arxiv.org/abs/1409.1556.
15. 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.
16. 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.
17. 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.
18. I. Allen and J. Seaman. *Digital compass learning: distance education enrollment report 2017*. Babson Survey Research Group, Babson Park, MA, USA, 2017.
19. 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.
20. A.B.Shetty , Bhoomika , Deeksha , J.Rebeiro , Ramyashree. *Facial Recognition using Haar Cascade and LBP Classifiers*. Journal Pre-proof, 28 July 2021.
21. P. Ekman and K. G. Heider, *The universality of a contempt expression: a replication.* Motivation and Emotion, vol. 12, no. 3, pp. 303–308, 1988.
22. 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
23. A. Krizhevsky, I. Sutskever, and G. E. Hinton. *ImageNet classification with deep convolutional neural networks*. In Proceedings of the Advances in Neural Information Processing Systems, pp. 1097–1105, Lake Tahoe, Nevada, December 2012.
24. L. Zhang, P. Shen, X. Peng et al. *Simultaneous enhancement and noise reduction of a single low-light image*. IET Image Processing, vol. 10, no. 11, pp. 840–847, 2016.
25. G. Tonguç and B. O. Ozkara. *Automatic recognition of student emotions from facial expressions during a lecture*. Computers & Education, vol. 148, Article ID 103797, 2020