

**ĐẠI HỌC SƯ PHẠM HÀ NỘI
KHOA CÔNG NGHỆ THÔNG TIN**



PHẠM QUANG HUY

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

KHÓA LUẬN TỐT NGHIỆP

Ngành: Công nghệ thông tin

HÀ NỘI, 04-2022

**ĐẠI HỌC SƯ PHẠM HÀ NỘI
KHOA CÔNG NGHỆ THÔNG TIN**



PHẠM QUANG HUY

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

KHÓA LUẬN TỐT NGHIỆP

Ngành: Công nghệ thông tin

Cán bộ hướng dẫn: TS. Đặng Thành Trung

HÀ NỘI, 04-2022

LỜI CẢM ƠN

[illegible]

[illegible]

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn “Nhận diện cảm xúc của sinh viên dựa trên biểu cảm khuôn mặt trong giáo dục trực tuyến.” là kết quả nghiên cứu của tôi dưới sự hướng dẫn của giảng viên hướng dẫn, không có bất kỳ hành vi sao chép lại hoặc đạo văn của người khác. Các tài liệu được luận văn tham khảo, kế thừa và trích dẫn đều được liệt kê trong danh mục các tài liệu tham khảo.

Tôi xin chịu hoàn toàn trách nhiệm về lời cam đoan trên.

Hà Nội, ngày ... tháng ... năm 2022

Sinh viên

Phạm Quang Huy

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. Đặc 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 diện cảm xúc tự động được giới thiệu dựa trên một mô hình mạng tích chập CNN (Convolutional 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 bộ CSDL về nhận diện cảm xúc là FER2013. Ngoài ra, ba lớp học trực tuyến gồm ba lớp sinh viên khoa CNTT, trường 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.

MỤC LỤC

LỜI CẢM ƠN	3
NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN	4
NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN	5
LỜI CAM ĐOAN	6
MỤC LỤC	8
DANH MỤC HÌNH ẢNH	11
DANH MỤC BẢNG BIỂU	12
CHƯƠNG 1: GIỚI THIỆU CHUNG	13
1.1 Lý do chọn đề tài	13
1.2 Giới thiệu	13
1.1.1 Bài toán nhận diện cảm xúc dựa trên biểu cảm khuôn mặt	14
1.1.2 Nhận diện cảm xúc dựa trên mạng nơ-ron tích chập	15
1.3 Tiềm năng ứng dụng của bài toán nhận diện cảm xúc dựa trên biểu cảm khuôn mặt trong giáo dục trực tuyến.....	16
1.4 Mục tiêu của luận văn	17
1.5 Cấu trúc luận văn	17
CHƯƠNG 2. TỔNG QUAN VỀ MẠNG NƠ-RON VÀ GIỚI THIỆU VỀ MẠNG NƠ-RON TÍCH CHẬP.....	18
2.1 Học sâu và mạng nơ-ron	18
2.1.1 Trí tuệ nhân tạo, học máy và học sâu.....	18
2.1.2 Sơ lược lịch sử mạng nơ-ron trong học sâu	18
2.1.3 Cấu tạo và quá trình xử lý của một nơ-ron sinh học.....	19
2.2 Mạng nơ-ron trong lĩnh vực học sâu.....	20
2.2.1 Cấu tạo và quá trình xử lý của một nơ-ron trong học sâu.....	20
2.2.2 Các mô hình hàm kích hoạt phổ biến của mạng nơ-ron	21
2.3.1 Mô hình mạng nơ-ron tích chập.....	24
2.3.1.1 Tầng tích chập (convolutional)	25
2.3.1.2 Tầng tổng hợp (pooling layer)	26
2.3.1.3 Tầng kết nối đầy đủ (fully-connected).....	26

2.3.2	Mô hình quá khớp (over-fitting) và mô hình chưa khớp (under-fitting)	27
2.3.3	Chuẩn hóa dữ liệu đầu ra (phương pháp Batch Normalization)	28
2.3.4	Thuật toán tối ưu (optimizers) trong huấn luyện mạng nơ-ron.....	29
2.3.4.1	Gradient Descent	29
2.3.4.2	SGD với động lượng (SGD with momentum).....	29
2.3.4.3	RMSProp (Root Mean Square Propagation)	30
2.3.4.4	Adagrad	30
2.3.4.5	Adadelata.....	30
2.3.4.6	Adam	31
2.3.5	Một số mô hình mạng huấn luyện nổi tiếng	31
2.3.5.1	Mạng huấn luyện AlexNet.....	31
2.3.3.2	Mạng huấn luyện VGG16.....	32
2.3.6	Huấn luyện mô hình.....	33
2.3	Kết luận	33
CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT.....		35
3.1	Nhận diện cảm xúc dựa trên mạng CNN	35
3.2	Lược đồ đề xuất	35
3.3	Hình ảnh đầu vào	35
3.4	Phát hiện khuôn mặt.....	36
3.4.1	Tổng quan về Haar Cascade	36
3.4.2	Cách hoạt động của phương pháp Haar Cascade.....	36
3.5	Tiền xử lý ảnh	38
3.6	Nhận diện cảm xúc.....	39
3.6.1	Bộ dữ liệu đào tạo	39
3.6.2	Xây dựng mô hình nhận diện cảm xúc	39
3.6.3	Môi trường đào tạo	41
3.6.4	Đào tạo mô hình nhận diện cảm xúc.....	42
3.7	Kết luận.....	44
CHƯƠNG 4. KẾT QUẢ THỰC NGHIỆM.....		45

4.1 Cài đặt chương trình nhận diện cảm xúc	45
4.2 So sánh với mô hình mạng nơ-ron VGG16	46
4.2 Kết quả thực nghiệm với bộ dữ liệu FER 2013	47
4.3 Ứng dụng thực tế trên khuôn mặt	48
4.4 Thử nghiệm thực tế tại lớp học trường Đại học sư phạm Hà Nội	51
4.5 Kết quả	54
CHƯƠNG 5. KẾT LUẬN	55
TÀI LIỆU THAM KHẢO	57

DANH MỤC HÌNH ẢNH

Hình 1.1 Một số hình ảnh được gán nhãn cảm xúc trong CSDL FER2013	14
Hình 2.1 Hình ảnh một nơ-ron sinh học [25].....	19
Hình 2.2 Minh họa mạng nơ-ron nhân tạo.....	20
Hình 2.3 Quá trình tính toán của một nơ-ron.....	21
Hình 2.4 Đồ thị hàm Sigmoid.....	22
Hình 2.5 Đồ thị hàm TanH	23
Hình 2.6 Đồ thị hàm ReLU.....	24
Hình 2.7 Mô hình lớp (layer) trong mạng CNN	24
Hình 2.8 Mô hình lớp (layer) trong mạng CNN	25
Hình 2.9 Phương thức tổng hợp Max Pooling và Average Pooling	26
Hình 2.10 Minh họa mô hình kiến trúc mạng CNN	27
Hình 2.11 Các trạng thái over-fitting, under-fitting và cân bằng.....	28
Hình 3.1 Lược đồ phương pháp đề xuất	35
Hình 3.2 Đặc trưng hình chữ nhật trong phương pháp Haar-Cascade.....	37
Hình 3.3 Các đặc trưng hình chữ nhật khác trong phương pháp Haar-Cascade.....	37
Hình 3.4 Phát hiện khuôn mặt bằng phương pháp Haar-Cascade	38
Hình 3.5 Tiền xử lý hình ảnh đầu vào	38
Hình 3.6 Kiến trúc mạng tích chập cho nhận diện cảm xúc	40
Hình 3.7 Sơ đồ khối đào tạo mô hình nhận diện cảm xúc	43
Hình 4.1 Sơ đồ khối nhận diện cảm xúc.....	Error! Bookmark not defined.
Hình 4.2 Hình ảnh lớp học trực tuyến	53
Hình 4.3 Nhận diện cảm xúc khuôn mặt trong lớp học trực tuyến.....	53
Hình 4.4 Biểu đồ đánh giá cảm xúc.....	54

DANH MỤC BẢNG BIỂU

Bảng 1.1 Mô tả đầu vào và đầu ra của bài toán nhận diện cảm xúc dựa trên biểu cảm khuôn mặt.....	14
Bảng 2.1 Các tham số chi tiết cho mô hình đề xuất.....	40
Bảng 3.2 Cấu hình phần cứng GoogleColab	42
Bảng 4.1 So sánh cấu trúc và thời gian đào tạo mô hình đề xuất và VGG16	47
Bảng 4.2 Kết quả thí nghiệm kiểm tra mô hình với bộ dữ liệu kiểm thử	47
Bảng 4.3 Một số kết quả thử nghiệm	48
Bảng 4.4 Kết quả thực nghiệm với khuôn mặt tác giả trong thời gian thực với mô hình đề xuất	49
Bảng 4.5 Kết quả thực nghiệm với khuôn mặt tác giả trong thời gian thực với mô hình VGG16	50
Bảng 4.6 Kết quả thử nghiệm tại lớp học Khoa Công nghệ thông tin, trường ĐH Sư phạm Hà Nội.....	52

CHƯƠNG 1: GIỚI THIỆU CHUNG

1.1 Lý do chọn đề tài

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. Đặc biệt trong hoàn cảnh nghiêm trọng của dịch bệnh COVID-19 như hiện nay. Theo số liệu từ WHO tính đến ngày 30/03/2022 có đến 6.132.461 trường hợp tử vong, trước tình hình nguy hiểm của dịch bệnh, điều này buộc các trường học tại Việt Nam cũng như trên toàn thế giới buộc người dân hạn chế ra khỏi nhà khi không cần thiết. Do đó, hầu hết tất cả các trường học đều đang đóng cửa, giao tiếp trực tuyến qua trong thời gian thực không chỉ đem lại hiệu quả trong hoàn cảnh dịch bệnh này mà phương pháp này còn đem lại nhiều lợi ích khác như: giao tiếp từ xa không quan trọng khoảng cách địa lý, tính tiện lợi và linh hoạt,... Nhờ vào những ưu điểm trên, giáo dục trực tuyến được xem là một trong những giải pháp tối ưu nhất hiện nay.

Tuy nhiên, hiệu quả của các lớp học trực tuyến từ lâu đã bị đặt nhiều dấu hỏi. So với các lớp học trực tiếp truyền thống, các lớp học trực tuyến thiếu sự giao tiếp và phản hồi trực tiếp, kịp thời và hiệu quả giữa giáo viên và học viên.

Vì vậy, nhằm nâng cao chất lượng giáo dục trực tuyến và khắc phục nhược điểm lớn nhất của phương pháp giáo dục này, thì việc tìm ra một giải pháp giúp cải thiện tính tương tác giữa người dạy và người học là vấn đề cấp thiết.

1.2 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 1.1 minh họa một số biểu cảm khuôn mặt cơ bản kèm theo các nhãn cảm xúc tương ứng trong bộ cơ sở dữ liệu FER2013, sẽ được sử dụng để thử nghiệm trong nghiên cứu này.



Hình 1.1 Một số hình ảnh được gán nhãn cảm xúc trong CSDL FER2013

1.1.1 Bài toán nhận diện cảm xúc dựa trên biểu cảm khuôn mặt


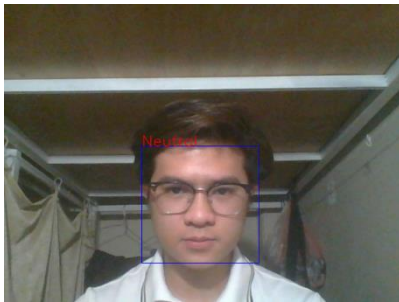
Với sự phát triển của công nghệ thông tin, đặc biệt trong lĩnh vực trí tuệ nhân tạo và học sâu, nhiều thuật toán nhận diện cảm xúc được đề xuất để nhận diện các biểu cảm được thể hiện trên khuôn mặt. Bài toán nhận diện cảm xúc dựa trên biểu cảm khuôn mặt chính là việc xác định loại cảm xúc của một người thông qua đặc trưng của nét mặt.

Đầu vào: là hình ảnh từ các lớp học dựa trên các nền tảng trực tuyến, dựa trên hình ảnh của sinh viên được trích xuất từ thiết bị học trực tuyến hình ảnh đầu và là một bức ảnh chứa các khuôn mặt của sinh viên trong lớp học trực tuyến.

Đầu ra: là kết quả dự đoán cảm xúc của sinh viên, kết quả ở đây có thể là kết quả tổng hợp được biểu diễn theo dạng biểu đồ thông số các cảm xúc hoặc tỉ lệ phần trăm giữa các cảm xúc được đánh giá trong một lớp học trực tuyến tại một thời điểm.

Mô tả cụ thể về bài toán thông qua ví dụ sau:

Bảng 1.1 Mô tả đầu vào và đầu ra của bài toán nhận diện cảm xúc dựa trên biểu cảm khuôn mặt

Ảnh đầu vào	Ảnh đầu ra	Nhãn kết quả (cảm xúc)
		Bình thường

1.1.2 Nhận diện cảm xúc dựa trên mạng nơ-ron tích chập

Các phương pháp sử dụng các mô hình trí tuệ nhân tạo đã cho thấy một 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 ra là 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.

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 diện cảm xúc. Trong nghiên cứu này, bộ cơ sở dữ liệu được sử dụng là: bộ dữ liệu FER2013, là một bộ CSDL không kiểm soát, được thu thập từ các môi trường phức tạp hơn với phong 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 so với những bộ CSDL có sẵn được tạo ra trong phòng thí nghiệm có kích thước dữ liệu nhỏ đem lại hiệu quả không cao như CK Plus[9].

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 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 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 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 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.

1.3 Tiềm năng ứng dụng của bài toán nhận diện cảm xúc dựa trên biểu cảm khuôn mặt trong giáo dục trực tuyến

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 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 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.

1.4 Mục tiêu của luận văn

Luận văn đề xuất phương pháp giải quyết bài toán nhận diện cảm xúc dựa trên biểu cảm khuôn mặt trong giáo dục trực tuyến giúp cải thiện hình thức học trực tuyến trong hoàn cảnh dịch bệnh nghiêm trọng. Mô hình mạng nơ-ron tích chập cho phép trích xuất các đặc trưng của biểu cảm khuôn mặt với tốc độ xử lý nhanh và cho ra độ chính xác cao.

1.5 Cấu trúc luận văn

Dựa trên mục tiêu cụ thể đã trình bày, luận văn được chia ra làm bốn chương với nội dung cụ thể như sau:

Chương 1: Giới thiệu chung

- Giới thiệu chung về bối cảnh và tầm quan trọng của bài toán nhận diện cảm xúc dựa trên biểu cảm khuôn mặt trong giáo dục trực tuyến. Cơ sở khoa học để thực hiện đề tài dựa trên mạng nơ-ron tích chập là tốt nhất ở thời điểm hiện tại so với các phương pháp khác

Chương 2: Tổng quan về mạng nơ-ron và giới thiệu về mạng nơ-ron tích chập

- Giới thiệu về cấu trúc và mô hình của một mạng nơ-ron truyền thông so với mạng nơ-ron tích chập, từ đó làm nổi bật lên ưu và nhược điểm của mạng nơ-ron tích chập.

Chương 3: Mô hình đề xuất

- Từ những cơ sở khoa học đã nghiên cứu từ những chương trước đó, ứng dụng mạng nơ-ron tích chập để giải quyết bài toán, từ đó đưa ra mô hình phù hợp cũng như xây dựng kiến trúc mạng nơ-ron sao cho phù hợp nhằm giải quyết bài toán nhận diện cảm xúc dựa trên biểu cảm khuôn mặt với bộ dữ liệu đã sưu tầm được.

Chương 4: Kết quả thực nghiệm

- So sánh kết quả mà mô hình đạt được với mô hình phổ biến khác
- Đưa ra kết quả thực nghiệm thu được với chính bộ dữ liệu đào tạo và kiểm thử mô hình.
- Đưa ra kết quả thực nghiệm thu được với khuôn mặt của tác giả trong thời gian thực
- Cuối cùng, ứng dụng mô hình trên vào trong thực tiễn và đưa ra kết quả với mô hình lớp học khoa Công nghệ thông tin trường Đại học sư phạm Hà Nội

Chương 5: Kết luận

- Tóm lược lại mục tiêu đã đặt ra của đề tài, các cơ sở khoa học, phương pháp thực hiện đề tài và kết quả đạt được trong đề tài
- Mở ra hướng nghiên cứu, phát triển và cải tiến mới trong tương lai

CHƯƠNG 2. TỔNG QUAN VỀ MẠNG NƠ-RON VÀ GIỚI THIỆU VỀ MẠNG NƠ-RON TÍCH CHẬP

2.1 Học sâu và mạng nơ-ron

2.1.1 Trí tuệ nhân tạo, học máy và học sâu

Ngày nay, trong kỷ nguyên số, máy tính là một phần không thể thiếu trong nghiên cứu khoa học cũng như trong đời sống hàng ngày. Tuy nhiên, do hệ thống máy tính dựa trên lý thuyết cổ điển (tập hợp, logic nhị phân), nên dù có khả năng tính toán lớn và độ chính xác cao, thì máy tính cũng chỉ có thể làm việc theo một chương trình gồm các thuật toán được viết sẵn do lập trình viên chứ chưa thể tự lập luận hay sáng tạo.

Trí tuệ nhân tạo (Artificial Intelligence - AI) là đề cập đến trí thông minh do máy móc đạt được, trái ngược với trí thông minh tự nhiên của con người. Trí tuệ nhân tạo được con người tạo ra nhằm giải quyết một hoặc một vài vấn đề cụ thể.

Học máy (Machine Learning) là một tập con các phương thức bên trong AI, đề cập đến các thuật toán mô hình số được thiết lập để phân tích dữ liệu hoặc học khả năng đưa ra quyết định để giải quyết một nhiệm vụ cụ thể. Mục tiêu của học máy là phát hiện ra các mô hình ẩn trong dữ liệu dưới các ràng buộc dữ liệu, ví dụ như kích thước và chất lượng dữ liệu, cho phép giải quyết các vấn đề đang được quan tâm.

Học sâu (Deep Learning) là một tập con của Machine Learning, có khả năng khác biệt ở một số khía cạnh quan trọng so với Machine Learning truyền thống, sử dụng mạng lưới thần kinh với nhiều lớp cho phép máy có thể tự đào tạo chính mình, do đó Deep learning đòi hỏi rất nhiều dữ liệu đầu vào và sức mạnh tính toán hơn là Machine Learning.

2.1.2 Sơ lược lịch sử mạng nơ-ron trong học sâu

Vào năm 1943, nhà thần kinh học Warren McCulloch đã cùng nhà toán học Walter Pitts đã viết một cuốn sách về cách mạng thần kinh hoạt động. Và họ đã thực hiện mô phỏng một mạng thần kinh đơn giản trên một mạch điện. [32]

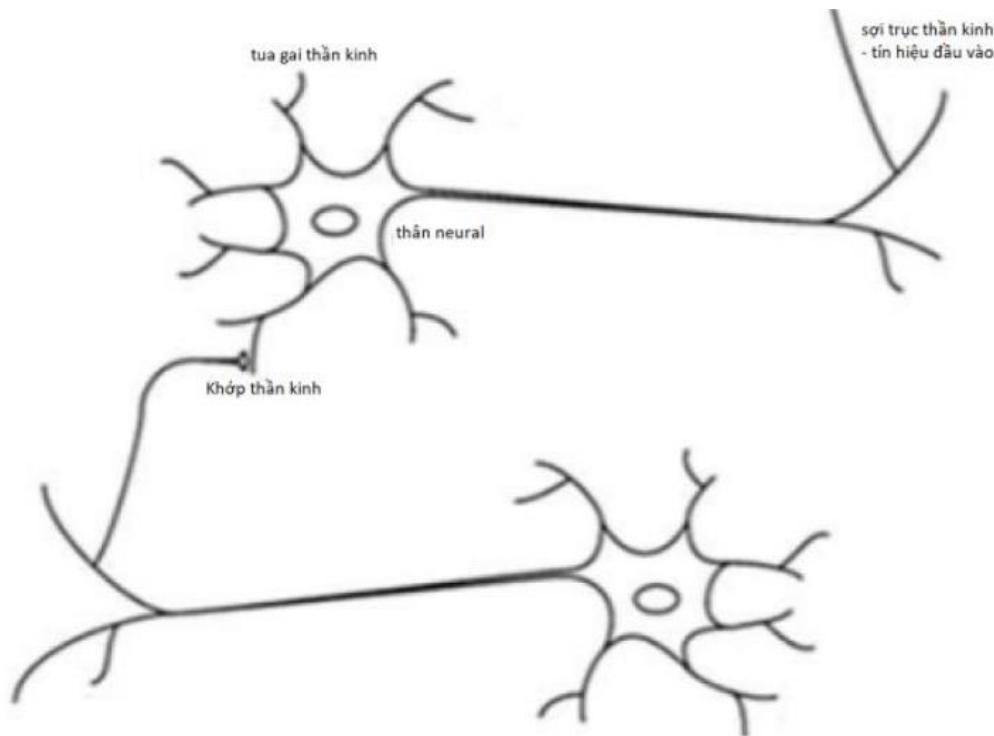
Vào năm 1949, Donald Hebb đã viết cuốn sách Organization of Behavior. Điểm nhấn chính là mạng thần kinh nào được sử dụng nhiều sẽ được tăng cường.

Vào năm 1959, David Hubel và Torsten Wiesel đã xuất bản cuốn sách Receptive fields of single neurons in the cat's striate cortex, miêu tả về phản ứng của các tế bào thần kinh thị giác trên loài mèo, cũng như cách loài mèo ghi nhớ và nhận diện hình dạng trên kiến trúc vỏ não của nó.

Vào năm 1989, Yann LeCun đã áp dụng thuật toán học cho mạng nơ-ron theo kiểu lan truyền ngược vào kiến trúc mạng nơ-ron tích chập của Fukushima. Sau đó vài năm, LeCun đã công bố LeNet-5 [33]. Có thể nói, LeNet-5 là một trong những mạng nơ-ron tích chập sơ khai

nhất, tuy nhiên các dấu ấn của nó vẫn tồn tại tới ngày nay, có thể thấy thông qua một số thành phần thiết yếu mà các mạng nơ-ron tích chập của ngày nay vẫn đang sử dụng

2.1.3 Cấu tạo và quá trình xử lý của một nơ-ron sinh học



Hình 2.1 Hình ảnh một nơ-ron sinh học [25]

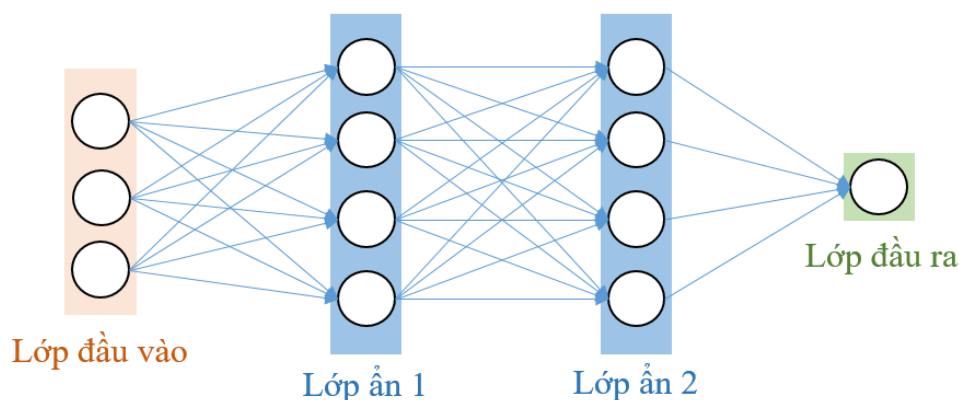
Một nơ-ron gồm có: thân nơ-ron, tua gai thần kinh, sợi trục thần kinh, trong đó:

- *Thân nơ-ron*: là nơi xử lý các tín hiệu được đưa vào;
- *Tua gai thần kinh*: là nơi nhận các xung điện vào trong nơ-ron;
- *Sợi trục thần kinh*: là nơi đưa tín hiệu ra ngoài sau khi được xử lý bởi nơ-ron;
- *Khớp thần kinh*: vị trí nằm giữa tua gai thần kinh và sợi trục thần kinh, đây là điểm liên kết đầu ra của nơ-ron này với đầu vào của nơ-ron khác.

2.2 Mạng nơ-ron trong lĩnh vực học sâu

2.2.1 Cấu tạo và quá trình xử lý của một nơ-ron trong học sâu

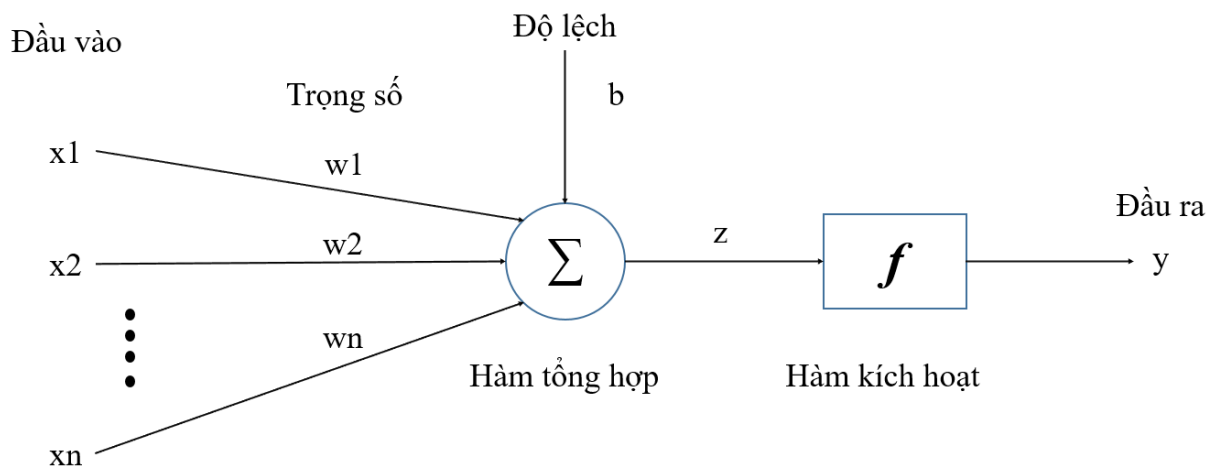
Trong Deep Learning, mạng nơ-ron nhân tạo hay còn gọi là các Multilayer Perceptron là các kiểu kiến trúc đơn giản nhất. Mỗi mạng nơ-ron nhân tạo bao gồm 3 thành phần chính: lớp đầu vào, các lớp ẩn và lớp đầu ra. Hình 2.2 mô tả kiến trúc một mô hình nơ-ron nhân tạo.



Hình 3.2 Minh họa mạng nơ-ron nhân tạo

Trong đó, các lớp ẩn của mô hình nơ-ron nhân tạo được tạo nên từ một hoặc nhiều đơn vị được gọi là các tế bào (perceptron). Các tế bào tại một lớp thực hiện tổ hợp đầu ra của lớp trước đó thành đầu vào của lớp đứng sau hoặc là đầu ra của mạng. Cụ thể, mỗi tế bào sẽ có các liên kết tới lớp đứng trước gọi là trọng số (weights). Quá trình tính toán của mỗi tế bào diễn ra như sau:

- Mỗi đầu ra của lớp phía trước sẽ được nhân với giá trị trọng số tương ứng với chúng và cộng tổng lại
- Các giá trị sau khi được cộng tổng lại sẽ được cộng thêm với độ lệch (bias)
- Cuối cùng giá trị sau khi được tính toán sẽ được đưa qua một hàm kích hoạt (activation function).
- Giá trị đầu vào của lớp đứng sau hoặc lớp hiện tại là lớp cuối cùng trong mạng thì sẽ được sử dụng như là lớp đầu ra của mạng. Hình 2.3 mô tả quá trình tính toán của một tế bào.



Hình 4.3 *Quá trình tính toán của một nơ-ron*

Trong đó:

- **Đầu vào:** Là các thuộc tính đầu vào của một nơ-ron. Số lượng thuộc tính đầu vào thường lớn hơn một, do dữ liệu đầu vào thường là một vector nhiều chiều, hoặc nơ-ron tầng trước kết nối với một nơ-ron sau.
- **Trọng số:** Các liên kết được thể hiện độ mạnh yếu qua một giá trị được gọi là trọng số liên kết. Kết hợp với các đầu truyền
- **Hàm tổng hợp:** Tổng các tích của đầu vào với trọng số liên kết mô phỏng các khớp kết nối. Sau đó đi qua hàm tính tổng để tính ra giá trị trước khi đưa vào hàm truyền.
- **Độ lệch:** Độ lệch được đưa vào sau khi tính toán xong hàm tổng tạo ra giá trị cuối cùng trước khi đưa vào hàm truyền. Mục đích của việc thêm vào độ lệch nhằm dịch chuyển chức năng của hàm kích hoạt sang trạng thái lệch trái hoặc lệch phải, giúp ích trong quá trình huấn luyện.
- **Hàm kích hoạt:** Được sử dụng để tính toán giá trị của đầu vào dựa trên giá trị của hàm tổng hợp.

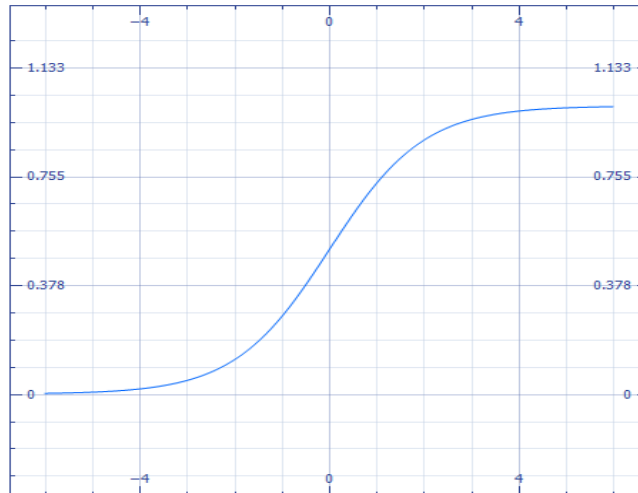
2.2.2 Các mô hình hàm kích hoạt phổ biến của mạng nơ-ron

Các hàm kích hoạt trong mạng nơ-ron nhân tạo thường là các hàm kích hoạt phi tuyến tính nhằm đem lại các mối quan hệ phức tạp giữa đầu ra và đầu vào. Một số hàm kích hoạt thường được sử dụng và phổ biến có thể kể đến như Sigmoid, tanh, rectified linear unit (ReLU). Việc chọn các hàm kích hoạt sao cho phù hợp phụ thuộc vào quá trình thực nhiệm vụ cụ thể. Các mô hình hàm kích hoạt phổ biến bao gồm:

- **Hàm Sigmoid:** thường được sử dụng vì ngưỡng của hàm nằm trong khoảng (0, 1). Do đó hàm này được sử dụng nhiều cho các mô hình dự đoán xác suất đầu ra, tức kết quả chỉ tồn tại trong khoảng từ 0 đến 1: khi đầu vào là số dương lớn, đầu ra của hàm sigmoid gần bằng 1. Khi nhỏ hơn 0, đầu ra gần bằng 0. Tuy nhiên, việc tối ưu của hàm này khó

khăn, nguyên nhân vì nếu giá trị đầu vào của hàm là 1 số rất lớn, thì đầu ra của hàm càng về 2 đầu xấp xỉ 1 hoặc 0, nên tốc độ hội tụ sẽ rất chậm[26]

- Biểu diễn hàm: $f(x) = \frac{1}{1 + e^{-x}}$
- Đạo hàm của hàm: $f'(x) = f(x)(1 - f(x))$
- Hình 2.4 mô tả đồ thị hàm Sigmoid

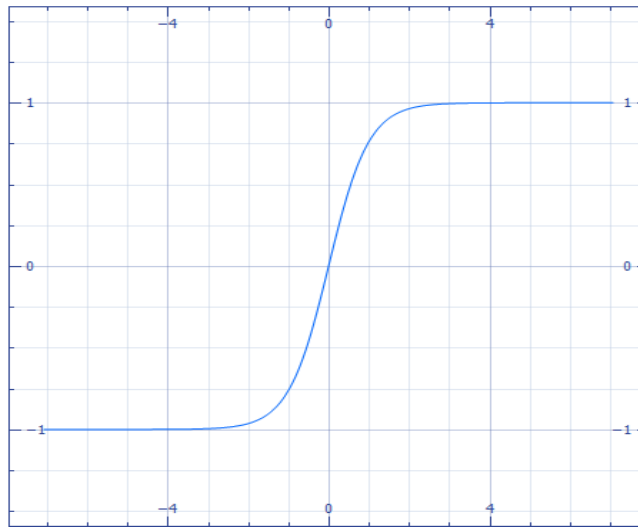


Hình 5.4 Đồ thị hàm Sigmoid

- Hàm TanH: được sử dụng vì đầu ra của hàm nằm trong khoảng , thích hợp với các mô hình đầu ra có ba giá trị: âm, trung tính (0) và dương. Hàm nhận đầu vào là một số thực và chuyển thành một giá trị trong khoảng (-1; 1). Cũng như Sigmoid, hàm tanH bị bão hoà ở 2 đầu (gradient thay đổi rất ít ở 2 đầu). Tuy nhiên hàm tanH lại đối xứng qua 0 nên khắc phục được một nhược điểm của Sigmoid. Hàm tanH và đồ thị được biểu diễn như sau[26]:

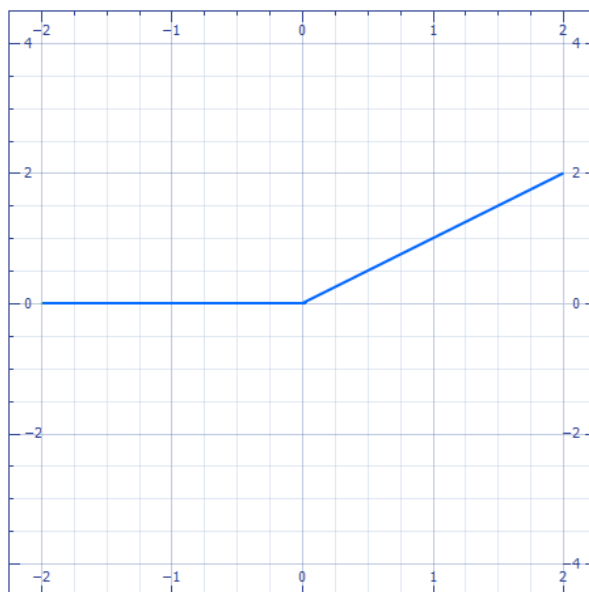
- Biểu diễn hàm: $f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
- Đạo hàm của hàm: $f'(x) = 1 - f(x)^2$

- Hình 2.5 mô tả đồ thị hàm TanH



Hình 6.5 Đồ thị hàm TanH

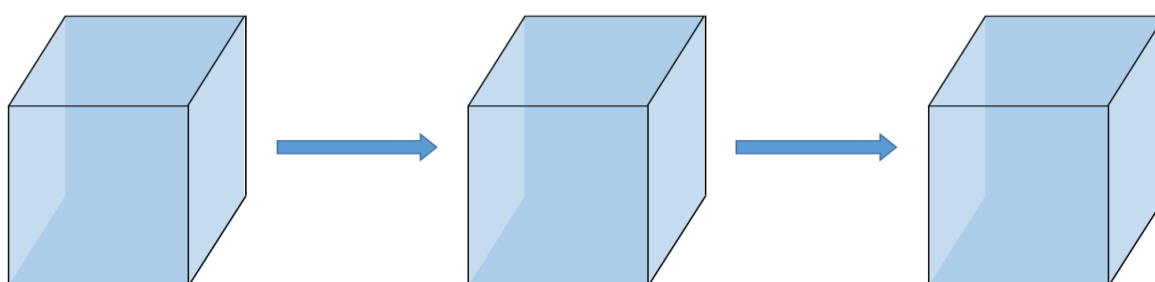
- Hàm ReLU: đây là hàm được sử dụng phổ biến trong những năm gần đây khi huấn luyện các mạng nơ-ron. Được áp dụng với những trường hợp cần đầu ra nằm trong khoảng $(0, +\infty)$ [26].
 - Đồ thị hàm ReLU: $f(x) = \begin{cases} 0 & \text{khi } x < 0 \\ x & \text{khi } x \geq 0 \end{cases}$
 - Đồ thị đạo hàm ReLU: $f'(x) = \begin{cases} 0 & \text{khi } x < 0 \\ 1 & \text{khi } x \geq 0 \end{cases}$
 - Hình 2.6 mô tả đồ thị hàm ReLU
 - So sánh ReLU với Sigmoid và tanH: Hàm ReLU ưu điểm có thể kể đến như: có tốc độ tính toán rất nhanh, gán các giá trị âm trở thành 0 ngay lập tức, phù hợp cho việc huấn luyện từ dữ liệu chuẩn. Tốc độ hội tụ nhanh hơn hẳn vì ReLU có tốc độ hội tụ nhanh gấp 6 lần tanH, điều này có thể do ReLU không bị bão hòa ở 2 đầu như Sigmoid và tanH. ReLU tính toán nhanh hơn vì tanH và Sigmoid sử dụng hàm exp và công thức phức tạp hơn ReLU rất nhiều do vậy sẽ tốn nhiều chi phí hơn để tính toán. Tuy nhiên, ReLU cũng có những nhược điểm sau: ReLU không ánh xạ các giá trị âm một cách thích hợp vì với các node có giá trị nhỏ hơn 0, qua ReLU activation sẽ thành 0, hiện tượng này gọi là “Dying ReLU”. Nếu các node bị chuyển thành 0 thì sẽ không có ý nghĩa với bước linear activation ở lớp tiếp theo và các hệ số tương ứng từ node này cũng không được cập nhật với gradient descent và khi learning rate lớn, các trọng số (weights) có thể thay đổi theo cách làm tắt cả neuron dừng việc cập nhật.



Hình 7.6 Đồ thị hàm ReLU

2.3 Mạng nơ-ron tích chập (CNN)

Mạng nơ-ron tích chập là một trong những mạng truyền thẳng đặc biệt. Đây cũng là mô hình học sâu kinh điển và phổ biến nhất hiện nay có ảnh hưởng nhiều nhất trong lĩnh vực thị giác máy tính (computer vision). Mạng nơ-ron tích chập thường được dùng để giải quyết vấn đề xử lý ảnh (nhận dạng ảnh, phân tích video, ...). Với tốc độ xử lý nhanh và độ chính xác cao, mạng nơ-ron tích chập được ứng dụng hầu hết trong các hệ thống nhận diện và xử lý ảnh hiện nay. So với mạng nơ-ron truyền thống, các tầng được coi là một chiều, thì trong mạng nơ-ron tích chập, các tầng được coi là 3 chiều bao gồm: chiều cao, chiều rộng và chiều sâu. Hình 2.7 mô tả các tầng (layer) trong CNN là 3 chiều.



Hình 8.7 Mô hình lớp (layer) trong mạng CNN

2.3.1 Mô hình mạng nơ-ron tích chập

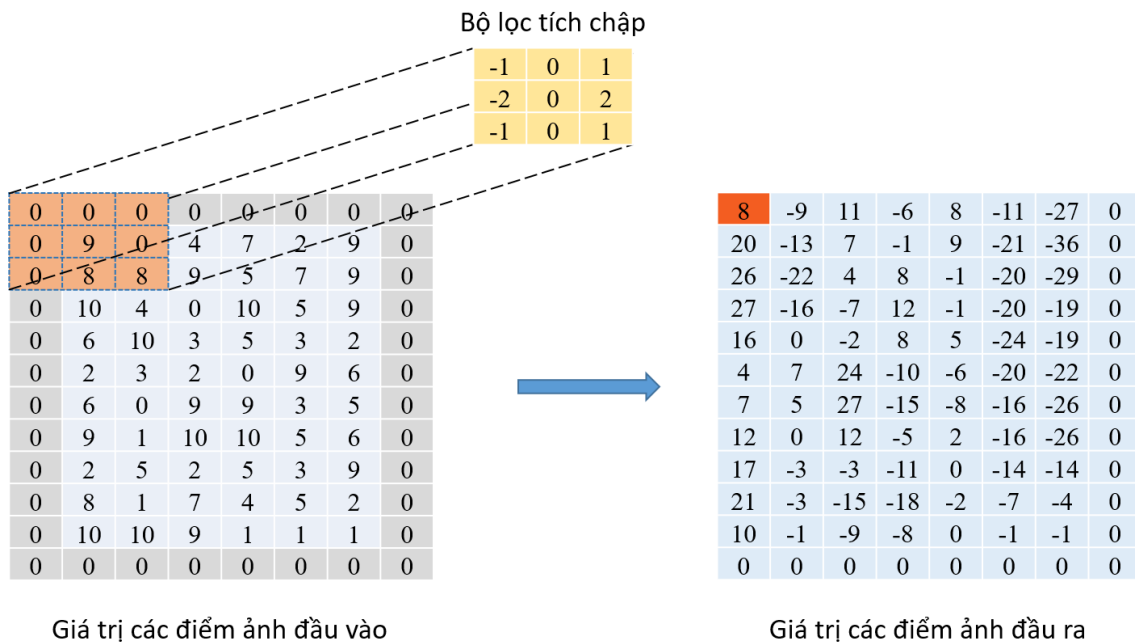
Kiến trúc của mô hình mạng nơ-ron tích chập bao gồm ba phần chính:

1. Tầng tích chập (convolutional)
2. Tầng tổng hợp (pooling layer)

3. Tầng kết nối đầy đủ (fully-connected)

2.3.1.1 Tầng tích chập (convolutional)

Có nhiệm vụ trích xuất các đặc trưng của ảnh, đầu ra của tầng tích chập này là đầu vào của tầng tích chập tiếp theo, tầng tích chập bao gồm: dữ liệu đầu vào và bộ lọc tích chập. Phép tích chập là phép toán tuyến tính thực hiện trên 2 đồ thị hàm số để đo lường sự chồng chéo của chúng. Đây cũng là thành phần quan trọng nhất trong mạng CNN, cũng là nơi thể hiện tư tưởng xây dựng liên kết cục bộ thay vì kết nối toàn bộ các điểm ảnh. Các liên kết cục bộ này được tính toán bằng phép tích chập giữa các giá trị điểm ảnh trong một vùng ảnh cục bộ với các bộ lọc (filters) có kích thước nhỏ. Hình 2.8 minh họa bộ lọc tích chập được sử dụng trên ma trận điểm ảnh.



Hình 9.8 Mô hình lớp (layer) trong mạng CNN

Trong ví dụ ở Hình 2.8, ta thấy bộ lọc được sử dụng là một ma trận có kích thước 3x3. Bộ lọc này được dịch chuyển lần lượt qua từng vùng ảnh đến khi hoàn thành quét toàn bộ bức ảnh, tạo ra một bức ảnh mới có kích thước nhỏ hơn hoặc bằng với kích thước ảnh đầu vào. Kích thước này được quyết định tùy theo kích thước các khoảng trắng được thêm ở viền bức ảnh gốc và được tính theo công thức (2.1) [23]:

$$o = \frac{i + 2 * p - k}{s} + 1 \quad (2.1)$$

Trong đó:

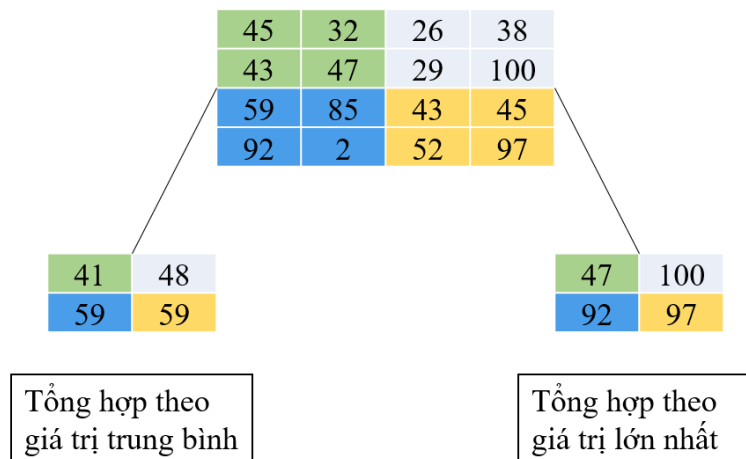
- o: kích thước ma trận ảnh đầu ra
- i: kích thước ma trận ảnh đầu vào

- p : kích thước khoảng trắng phía ngoài viền của ảnh gốc
- k : kích thước ma trận bộ lọc
- s : bước nhảy của bộ lọc

Như vậy, sau khi đưa một bức ảnh đầu vào cho lớp Tích chập ta nhận được kết quả đầu ra là một loạt ảnh hưởng tương ứng với các bộ lọc đã được sử dụng để thực hiện phép tích chập. Các trọng số của các bộ lọc này được khởi tạo ngẫu nhiên trong lần đầu tiên và sẽ được cải thiện dần xuyên suốt quá trình huấn luyện.

2.3.1.2 Tầng tổng hợp (pooling layer)

Có chức năng tổng hợp các bộ lọc lại với nhau, bằng cách gom một tập các điểm ảnh và cho ra kích thước nhỏ hơn, làm giảm dần kích thước không gian để giảm số lượng tham số và tính toán, tầng tổng hợp pooling thường được đặt sau lớp tích chập để làm giảm kích thước ảnh đầu ra trong khi vẫn giữ được các thông tin quan trọng của ảnh đầu vào. Việc giảm kích thước dữ liệu có tác dụng làm giảm số lượng tham số cũng như tăng hiệu quả khi tính toán và đào tạo mô hình. Tầng tổng hợp cũng sử dụng một cửa sổ trượt để quét toàn bộ các vùng trong ảnh tương tự tầng tích chập và thực hiện phép lấy mẫu thay vì tích chập, có nghĩa là ta sẽ chọn một giá trị duy nhất để đại diện cho toàn bộ thông tin trong vùng ảnh đó. Hình 2.9 mô tả các phương thức lấy mẫu thường được sử dụng hiện nay, đó là Max Pooling (lấy ra giá trị điểm ảnh lớn nhất) và Average Pooling (lấy giá trị trung bình của các điểm ảnh trong vùng ảnh cục bộ)[24]



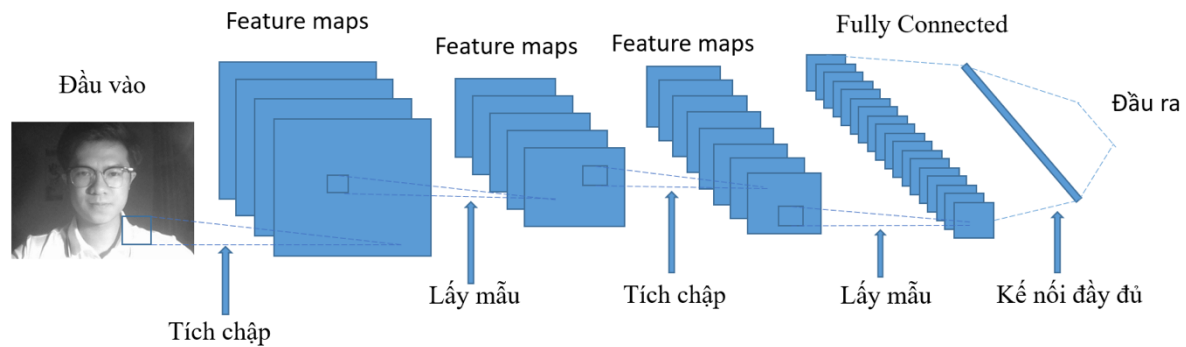
Hình 10.9 Phương thức tổng hợp Max Pooling và Average Pooling

Như vậy, với mỗi ảnh đầu vào được đưa qua lấy mẫu ta thu được một ảnh đầu ra tương ứng, có kích thước giảm xuống đáng kể nhưng vẫn giữ được các đặc trưng cần thiết cho quá trình tính toán sau này cũng tăng hiệu quả khi đào tạo.

2.3.1.3 Tầng kết nối đầy đủ (fully-connected)

Giống như các mạng nơ-ron thông thường, tầng kết nối đầy đủ được tích chập nhiều lần từ các tầng trước đó. Tầng gộp có thể làm giảm kích thước mẫu trên từng khối của tầng trước.

Trong mô hình mạng nơ-ron tích chập, kiến trúc mạng nơ-ron tích chập thường chồng cả ba tầng này để tạo thành một kiến trúc đầy đủ. Hình 2.10 minh họa về cấu trúc mạng của CNN. So với mạng nơ-ron truyền thống thì đầu vào của tầng kết nối đầy đủ có kích thước đã được giảm bớt rất nhiều. Do vậy, việc tính toán nhận dạng sử dụng mô hình truyền thẳng đã không còn phức tạp và tốn thời gian như trong mạng nơ-ron truyền thống.

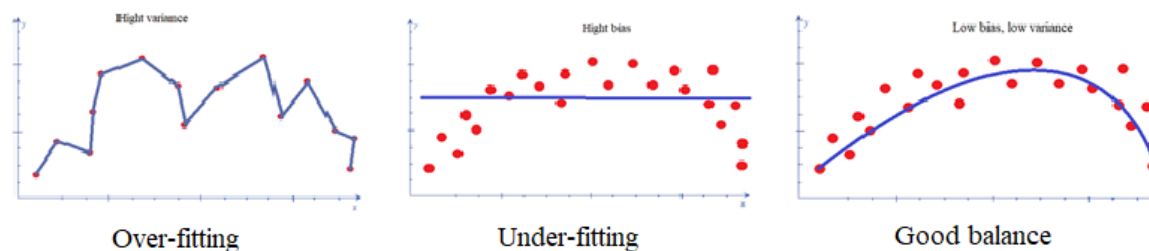


Hình 11.10 Minh họa mô hình kiến trúc mạng CNN

2.3.2 Mô hình quá khớp (over-fitting) và mô hình chưa khớp (under-fitting)

Cấu trúc của một mạng CNN rất lớn, trong mạng có rất nhiều nơ-ron, kết nối, cũng như có rất nhiều trọng số cần thiết để huấn luyện. Nhưng với lượng dữ liệu huấn luyện thường không đủ lớn để huấn luyện hoàn thiện cho một mạng nơ-ron lớn. Mô hình sau khi huấn luyện có thể đạt hiệu quả không tốt khi dự đoán với một bộ dữ liệu mới. Do đó có thể dẫn đến một số vấn đề về quá khớp và chưa khớp, khiến cho kết quả thực tế gây ra sai số lớn. Có một số kỹ thuật để cải thiện điều này.

- Mô hình chưa khớp là mô hình mà bias (độ sai lệch giữa giá trị model dự đoán với giá trị thật) có độ lớn cao và variance (độ phân tán dữ liệu – phương sai) thấp. Hiện tượng này xảy ra khi lượng dữ liệu quá ít hoặc cố gắng mô tả các dữ liệu phức tạp bằng các mô hình tuyến tính đơn giản. Khi gặp hiện tượng này chúng ta phải khắc phục bằng cách tìm kiếm thêm dữ liệu đầu vào và tăng độ phức tạp của model.
- Mô hình quá khớp: là mô hình quá khớp với dữ liệu, nó sẽ đúng trên tập training (đào tạo) nhưng trên tập test (kiểm thử) thì kết quả rất tệ. Mô hình này thường có bias nhỏ và variance lớn. Bởi vì khi đào tạo model trên nhiều dữ liệu dẫn đến model bị quá phức tạp so với mức cần thiết dẫn đến model không tổng quát hóa được khi gặp các dữ liệu mới và dẫn đến dự đoán sai.



Hình 12.11 Các trạng thái over-fitting, under-fitting và cân bằng

Một trong những phương pháp đó là giảm trọng số trong lúc huấn luyện. Dropout là một trong những kỹ thuật nổi tiếng và khá phổ biến để khắc phục vấn đề này. Dropout đặt đầu ra của mỗi nơ-ron ẩn thành 0 với xác suất 0,5. Vì vậy, các nơ-ron này sẽ không đóng góp vào lan truyền tiến, do đó và sẽ không tham gia vào lan truyền ngược. Thông thường, đối với các đầu vào khác nhau, mạng nơ-ron xử lý dropout theo một cấu trúc khác nhau. Một cách khác để cải thiện việc quá khớp là tăng lượng dữ liệu. Chúng ta có thể phản chiếu hình ảnh, lộn ngược hình ảnh, lấy mẫu hình ảnh, v.v. Những cách này sẽ tăng số lượng dữ liệu huấn luyện. Vì vậy, nó có khả năng ngăn chặn quá khớp. Hình 2.11 mô tả ba trạng thái over-fitting, under-fitting và cân bằng dưới góc nhìn của variance và bias.

2.3.3 Chuẩn hóa dữ liệu đầu ra (phương pháp Batch Normalization)

Chuẩn hóa dữ liệu (Normalization) luôn là kỹ thuật được nghiên cứu tích cực trong Deep Learning. Các phương pháp Normalization có thể giúp mô hình huấn luyện nhanh và kết quả tốt [31].

- Chuẩn hóa dữ liệu mỗi đặc trưng sẽ giữ được đầy đủ thông tin quan trọng của mọi đặc trưng trong quá trình huấn luyện.
- Làm giảm Internal Covariate Shift (ICS). Việc mô hình càng sâu sẽ có nhiều layer cùng với đó là có nhiều hàm kích hoạt, nó sẽ làm biến đổi đi phân phối của dữ liệu. Do đó chúng ta cần chuẩn hóa lại nó để có được sự đồng bộ phân phối của dữ liệu trong quá trình huấn luyện.

Batch Normalization (BN) là một kỹ thuật để chuẩn hóa các kích hoạt trong các lớp trung gian của mạng nơ-ron. Xu hướng cải thiện độ chính xác và tăng tốc độ đào tạo đã khiến BN trở thành một kỹ thuật phổ biến trong học sâu. BN chủ yếu cho phép đào tạo với tỷ lệ học tập lớn (learning rates), đây là nguyên nhân giúp mô hình khi đào tạo có thể hội tụ nhanh hơn và tổng quát hóa tốt hơn [31].

Batch Normalization còn có vai trò như một dạng của chính quy hóa (regularization) giúp cho việc giảm thiểu overfitting. Sử dụng batch normalization, chúng ta sẽ không cần phải sử dụng quá nhiều dropout và điều này rất có ý nghĩa vì chúng ta sẽ không cần phải lo lắng vì bị mất quá nhiều thông tin khi sử dụng dropout của mạng. Những lợi ích của phương pháp chuẩn hóa dữ liệu Batch Normalization có thể kể đến bao gồm:

- Làm giảm internal covariate shift (ICS) và tăng tốc độ huấn luyện cho mô hình deep learning.
- Cách tiếp cận này làm giảm sự phụ thuộc của gradients vào tỉ lệ của các tham số hoặc giá trị ban đầu của chúng, dẫn đến learning rate cao hơn mà không có nguy cơ phân kỳ.
- Batch normalization giúp sử dụng các chế độ phi tuyến bão hòa bằng cách ngăn mạng khỏi bị kẹt trong các chế độ bão hòa.

2.3.4 Thuật toán tối ưu (optimizers) trong huấn luyện mạng nơ-ron

Thuật toán tối ưu là cơ sở để xây dựng mô hình neural network với mục đích "học" được các đặc trưng (hay pattern) của dữ liệu đầu vào, từ đó có thể tìm 1 cặp weights và bias phù hợp để tối ưu hóa model. Tuy nhiên, chúng ta phải tìm 1 thuật toán để cải thiện weight và bias theo từng bước, và đó là lý do các thuật toán optimizer ra đời.

2.3.4.1 Gradient Descent

Gradient Descent (GD) là thuật toán tìm tối ưu chung cho các hàm số. Ý tưởng chung của GD là điều chỉnh các tham số để lặp đi lặp lại thông qua mỗi dữ liệu huấn luyện để giảm thiểu hàm chi phí.

$$w^{(k+1)} = w^{(k)} - \eta \nabla_w J(w^{(k)}) \quad (2.2)$$

Với $w^{(k)}$ là tham số tại bước cập nhật tại lớp k, η là tỉ lệ học, $J(w)$ là hàm lỗi, $\nabla_w J(w^{(k)})$: đạo hàm của hàm lỗi tại điểm $w^{(k)}$.

2.3.4.2 SGD với động lượng (SGD with momentum)

SGD với momentum là phương pháp giúp tăng tốc các vector độ dốc theo đúng hướng, và giúp hệ thống hội tụ nhanh hơn. Đây là một trong những thuật toán tối ưu hóa phổ biến nhất và nhiều mô hình hiện đại sử dụng nó để đào tạo. Mô tả như sau:

$$v_j \leftarrow \alpha * v_j - \eta * \nabla_w \sum_1^m L_m(w) \quad (2.3)$$

$$w_j \leftarrow v_j + w_j$$

Phương trình (2.3) có hai phần. Thuật ngữ đầu tiên là độ dốc v_j được giữ lại từ các lần lặp trước. Hệ số động lượng α là tỉ lệ phần trăm của độ dốc được giữ lại mỗi lần lặp. L là hàm mất mát, η là tỉ lệ học.

2.3.4.3 RMSProp (Root Mean Square Propagation)

RMSProp sử dụng trung bình bình phương của gradient để chuẩn hóa gradient. Có tác dụng cân bằng kích thước bước - giảm bước cho độ dốc lớn để tránh hiện tượng phát nổ độ dốc (Exploding Gradient), và tăng bước cho độ dốc nhỏ để tránh biến mất độ dốc (Vanishing Gradient). RMSProp tự động điều chỉnh tốc độ học tập, và chọn một tỉ lệ học tập khác nhau cho mỗi tham số. Phương pháp cập nhật các trọng số được thực hiện như mô tả:

$$\begin{aligned} s_t &= \rho s_{t-1} + (1 - \rho) * g_t^2 \\ \Delta x_t &= -\frac{\eta}{\sqrt{s_t + \epsilon}} * g_t \\ x_{t+1} &= x_{t+1} + \Delta x_t \end{aligned} \quad (2.4)$$

Với s_t : tích lũy phương sai của các gradient trong quá khứ, ρ : tham số suy giảm, Δx_t : sự thay đổi các tham số trong mô hình, g_t : gradient của các tham số tại vòng lặp t , ϵ : tham số đảm bảo kết quả xấp xỉ có ý nghĩa.

2.3.4.4 Adagrad

Adagrad là một kỹ thuật học máy tiên tiến, thực hiện giảm dần độ dốc bằng cách thay đổi tốc độ học tập. Adagrad được cải thiện hơn bằng cách cho trọng số học tập chính xác dựa vào đầu vào trước nó để tự điều chỉnh tỉ lệ học theo hướng tối ưu nhất thay vì với một tỉ lệ học duy nhất cho tất cả các nút.

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{G_t + \epsilon}} * g_t \quad (2.5)$$

Trong công thức (2.5), G_t là ma trận đường chéo chứa bình phương của đạo hàm vector tham số tại vòng lặp t ; g_t là vector của độ dốc cho vị trí hiện tại và η là tỉ lệ học.

2.3.4.5 Adadelta

Adadelta là một biến thể khác của AdaGrad. Adadelta không có tham số tỉ lệ học. Thay vào đó, nó sử dụng tốc độ thay đổi của chính các tham số để điều chỉnh tỉ lệ học nghĩa là bằng cách giới hạn cửa sổ của gradient tích lũy trong quá khứ ở một số kích thước cố định của trọng số w .

$$\begin{aligned} g'_t &= \sqrt{\frac{\Delta x_{t-1} + \epsilon}{s_t + \epsilon}} * g_t \\ x_t &= x_{t-1} - g'_t \\ \Delta x_t &= \rho \Delta x_{t-1} + (1 - \rho) x_t^2 \end{aligned} \quad (2.6)$$

Từ công thức (2.6), Adadelta sử dụng 2 biến trạng thái: st để lưu trữ trung bình của khoảng thời gian thứ hai của gradient và Δx_t để lưu trữ trung bình của khoảng thời gian thứ 2 của sự thay đổi các tham số trong mô hình. g'_t : căn bậc hai thương của trung bình tốc độ thay đổi bình phương và trung bình mô-men bậc hai của gradient.

2.3.4.6 Adam

Adam được xem như là sự kết hợp của RMSprop và Stochastic Gradient Descent với động lượng. Adam là một phương pháp tỉ lệ học thích ứng, nó tính toán tỉ lệ học tập cá nhân cho các tham số khác nhau. Adam sử dụng ước tính của khoảng thời gian thứ nhất và thứ hai của độ dốc để điều chỉnh tỉ lệ học cho từng trọng số của mạng nơ-ron. Tuy nhiên, qua nghiên cứu thực nghiệm, trong một số trường hợp, Adam vẫn còn gặp phải nhiều thiếu sót so với thuật toán SGD. Thuật toán Adam được mô tả:

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t \\ v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \end{aligned} \quad (2.7)$$

Trong công thức (2.7), v_t là trung bình động của bình phương và m_t là trung bình động của gradient; β_1 và β_2 là tốc độ của di chuyển.

2.3.5 Một số mô hình mạng huấn luyện nổi tiếng

2.3.5.1 Mạng huấn luyện AlexNet

Mạng huấn luyện AlexNet là công trình đầu tiên phổ biến mạng CNN trong lĩnh vực Thị giác máy tính, cũng là một trong những mạng huấn luyện CNN nổi tiếng nhất nhờ thành tích ấn tượng mà nó đạt được trong cuộc thi nhận dạng ảnh quy mô lớn tổ chức vào năm 2012. Cuộc thi này có tên chính thức là ILSVRC – ImageNet Large Scale Visual Recognition Challenge [29], được ImageNet - một hãng CSDL ảnh - tổ chức thường niên và được coi là cuộc thi Olympics quy mô thế giới trong lĩnh vực Thị giác máy tính. Mục đích của cuộc thi là nhằm thử nghiệm các công nghệ mới giúp cho máy tính có thể hiểu, phân tích, phát hiện và nhận dạng các vật thể trong một bức ảnh.

Kiến trúc tổng thể của mạng AlexNet bao gồm:

- Lớp 1 (Tích chập):
 - Số bộ lọc: 96
 - Kích thước bộ lọc: 11 x 11
 - Bước trượt (Stride): 4
- Lớp lấy mẫu 1:
- Lớp 2 (Tích chập):
 - Kích thước: 5 x 5
- Lớp 2 (Tích chập):
 - Số bộ lọc: 256/27
 - Kích thước bộ lọc: 5 x 5

- Lớp 3, 4, 5: Tương tự như với lớp 1 và lớp 2 với các kích thước bộ lọc là 3×3 . Đầu ra cuối cùng qua lớp Tích chập thứ 5 là dữ liệu với kích thước 13×13 , dữ liệu này sau khi đi qua một lớp lấy mẫu tối đa cuối cùng sẽ được dùng làm đầu vào cho các lớp sau đó là các lớp Kết nối đầy đủ.
- Lớp 6 (Kết nối đầy đủ):
 - Số nơ-ron: 4096
- Lớp 7 (Kết nối đầy đủ): Tương tự lớp 6.
- Lớp 8 (Kết nối đầy đủ): Tương tự lớp 7 và có số nơ-ron tương ứng với bài toán.

2.3.3.2 Mạng huấn luyện VGG16

VGG16 là mạng convolutional neural network được đề xuất bởi K. Simonyan and A. Zisserman, University of Oxford. Model sau khi train bởi mạng VGG16 đạt độ chính xác 92.7% top-5 test trong dữ liệu ImageNet gồm 14 triệu hình ảnh thuộc 1000 lớp khác nhau [30].

Kiến trúc tổng thể của mạng VGG16 bao gồm:

- Lớp 1 và lớp 2 (Tích chập):
 - Số bộ lọc: 64
 - Kích thước bộ lọc: 3×3
 - Bước trượt (Stride): 4
- Lớp lấy mẫu 1:
 - Kích thước: 2×2
- Lớp 3 và lớp 4 (Tích chập):
 - Số bộ lọc: 128
 - Kích thước bộ lọc: 3×3
- Lớp lấy mẫu 2:
 - Kích thước: 2×2
- Lớp 5, 6, 7: Tương tự như với lớp 3, 4 có số bộ lọc là 256
- Lớp lấy mẫu 3:
 - Kích thước: 2×2
- Lớp 8, 9, 10: Tương tự như với lớp 5, 6, 7 có số bộ lọc là 512
- Lớp lấy mẫu 4:
 - Kích thước: 2×2
- Lớp 11, 12, 13: Tương tự như với lớp 8, 9, 10 có số bộ lọc là 512
- Lớp lấy mẫu 5:
 - Kích thước: 2×2
- Lớp 6 (Kết nối đầy đủ):
 - Số nơ-ron: 4096
- Lớp 7 (Kết nối đầy đủ): Tương tự lớp 6.

- Lớp 8 (Kết nối đầy đủ): Lớp cuối cùng trong mạng VGG16 này có số nơ-ron tương ứng với bài toán.

2.3.6 Huấn luyện mô hình

Để hoàn tất quá trình huấn luyện và đưa ra kết quả cuối cùng là một mô hình có giải quyết được bài toán đề ra thì ngoài việc thu thập và xử lý dữ liệu đầu vào, xây dựng mô hình sao cho phù hợp với bộ dữ liệu thì bước quan trọng cuối cùng đó là thực hiện luồng đào tạo mô hình.

Trong bước thực thi huấn luyện mô hình này, dữ liệu cần được đi từ đầu vào của mô hình đến đầu ra của mô hình qua các lớp tích chập, các lớp tổng hợp và các hàm biến đổi kích hoạt. Tuy nhiên, để hoạt động đào tạo diễn ra một cách nhanh chóng và có hiệu quả hơn thì cần chú ý đến thuật toán và những tham số trong quá trình huấn luyện, ví dụ như: Epoch, Batch size, Iterations.

- Epoch: được tính khi đưa tất cả dữ liệu vào mô hình mạng huấn luyện trong một lần. Ví dụ: có 10 triệu bức ảnh trong tập huấn luyện, tiến hành cho toàn bộ số ảnh làm đầu vào của mô hình 3 lần, như vậy kết luận rằng đã đào tạo mô hình được 3 epoch.
- Batch size: Trong trường hợp dữ liệu quá lớn, không thể đưa hết tất cả các tập dữ liệu vào để huấn luyện, khi đó sẽ cần một siêu máy tính có dung lượng RAM và GPU RAM cực lớn để lưu trữ toàn bộ số lượng hình ảnh trên, điều này là bất khả thi đối với những phòng lab nhỏ và các nghiên cứu sinh. Do đó, việc chia nhỏ tập dữ liệu ra để huấn luyện là cần thiết, từ đó hình thành nên khái niệm batch. Batch size là số lượng mẫu dữ liệu trong một lần huấn luyện, Ví dụ: batch size là 32 thì mỗi một lần lặp sẽ có ngẫu nhiên 32 bức hình lan truyền trong mạng, quá trình diễn ra liên tục và không lặp lại các hình trước đó cho đến khi hoàn thành một epoch.
- Iterations: là số lượng batches cần để có thể hoàn thành một epoch. Ví dụ: dữ liệu có 20,000 mẫu và batch size là 500, như vậy cần có 40 lần lặp (iterations) để hoàn thành một epoch.
- Dữ liệu đánh giá (validation data): thông thường để đào tạo một mô hình, dữ liệu sẽ được chia ra làm ba bộ (dữ liệu đào tạo, dữ liệu đánh giá, dữ liệu thử nghiệm), trong đó bộ dữ liệu đánh giá đóng vai trò quan trọng trong việc đánh giá mức độ tổng quát của mô hình

2.3 Kết luận

Mạng nơ-ron trong Deep Learning là một chuỗi các thuật toán được sử dụng để tìm ra mối quan hệ của một tập dữ liệu thông qua cơ chế vận hành của bộ não sinh học. Mạng nơ-ron thường được huấn luyện qua một tập dữ liệu chuẩn cho trước, từ đó có thể đúc rút được kiến thức từ tập dữ liệu huấn luyện, và áp dụng với các tập dữ liệu khác với độ chính xác cao. Các phương pháp sử dụng để huấn luyện mạng nơ-ron ngày càng tối ưu hơn về mặt tính toán và phục vụ cho nhiều mục đích khác nhau.

Hiện nay, kiến trúc mạng nơ-ron ngày càng được hoàn thiện cho nhiều nhiệm vụ, trong đó mạng nơ-ron tích chập được chú ý rất nhiều vì tính hiệu quả trong thị giác máy tính đặc biệt là bài toán phân loại ảnh. Mạng nơ-ron tích chập có khả năng ghi lại sự phụ thuộc không gian của hình ảnh kể từ khi nó xử lý chúng dưới dạng ma trận và phân tích toàn bộ các phần của một hình ảnh tại một thời điểm, tùy thuộc vào kích thước của bộ lọc, mỗi phần của hình ảnh được cung cấp một tập hợp các tham số (chiều rộng và độ lệch) sẽ tham chiếu mức độ liên quan của tập hợp pixel đó với toàn bộ hình ảnh, tùy thuộc vào bộ lọc. Theo điều này, bằng cách giảm số lượng các tham số và bằng cách phân tích hình ảnh theo từng phần, CNN có thể hiển thị đại diện tốt hơn của hình ảnh.

Với nhiều mô hình nổi tiếng như AlexNet và VGG16 đạt tỷ lệ chính xác cao, trong đó VGG16 sâu hơn so với AlexNet và số lượng tham số nhiều hơn. Bắt đầu từ VGG-16, một hình mẫu chung cho các mạng CNN trong các tác vụ học có giám sát trong xử lý ảnh đã bắt đầu hình thành đó là các mạng trở nên sâu hơn. Mạng nơ-ron tích chập với các cải tiến góp phần giảm thời gian tính toán và tăng độ chính xác hứa hẹn sẽ là một trong những phương pháp được áp dụng rất nhiều vào thực tế trong tương lai

CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT

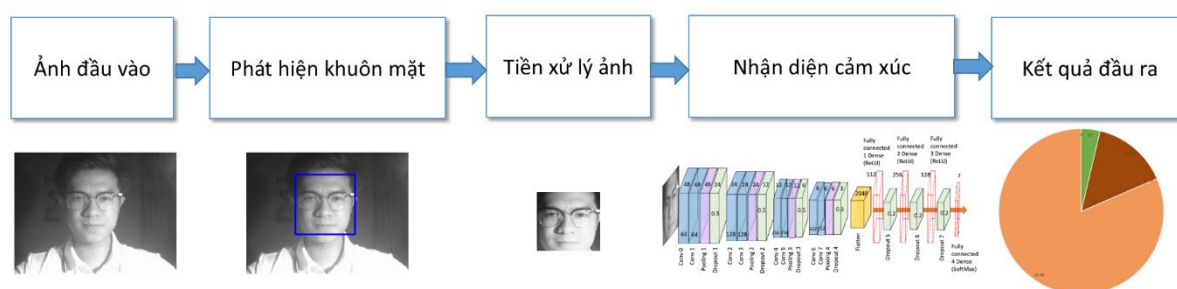
3.1 Nhận diện cảm xúc dựa trên mạng CNN

Từ những ưu điểm của mạng CNN trong những bài toán xử lý ảnh nói chung như đã trình bày ở phần trước, thì việc ứng dụng mạng CNN để xây dựng một mô hình nhận diện cảm xúc và cụ thể là mô hình dựa trên ý tưởng của VGG16 để có thể giải quyết bài toán nhận diện cảm xúc của sinh viên dựa trên biểu cảm khuôn mặt là hoàn toàn khả thi, tuy nhiên việc xây dựng một mô hình mạng dựa trên ý tưởng của mô hình VGG16 để cho ra kết quả nhận diện tốt thì cần có những bước tiền xử lý ảnh đầu vào sao cho hiệu quả, nhằm đảm bảo độ chính xác cũng như hiệu năng của chương trình. Thêm vào đó, dữ liệu đầu ra là các nhãn cảm xúc cần được thống kê một cách cụ thể và khoa học nhằm giúp cho người dạy học có thể nắm bắt số liệu một cách nhanh chóng để điều chỉnh chất lượng dạy học.

Do đó, một lược đồ đề xuất bao gồm năm bước chính là: thu thập ảnh đầu vào, phát hiện khuôn mặt, tiền xử lý ảnh, nhận diện cảm xúc và cuối cùng là thống kê lại kết quả sẽ được đề xuất trong phần này.

3.2 Lược đồ đề xuất

Trong phần này, một lược đồ nhận diện cảm xúc dựa trên các nền tảng học trực tuyến được giới thiệu. 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 bao gồm: 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 3.1 minh họa một cách trực quan các bước của lược đồ. Một biểu đồ thống kê tổng số các cảm xúc hiện có trong lớp sẽ được tổng hợp và cung cấp cho các giảng viên. Dựa trên biểu đồ thống kê này, giảng viên và các nhà quản lý đào tạo sẽ có thêm một kênh đánh giá khách quan để có thể điều chỉnh kế hoạch giảng dạy nhằm nâng cao chất lượng đào tạo.



Hình 1.1 Lược đồ phương pháp đề xuất

3.3 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 tăng tính linh hoạt trong đào tạo. Những nền tảng công nghệ này cho phép giáo viên áp dụng các phương tiện công nghệ cao và đa dạng để hỗ trợ giảng dạy mà không phải lo lắng về

giới hạn số lượng sinh viên trong lớp như các lớp học truyền thống và sinh viên ở các vị trí địa lí khác nhau hoàn toàn có thể giao tiếp trong thời gian thực mà không cần phải đến lớp. 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 thêm. 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.

3.4 Phát hiện khuôn mặt

3.4.1 Tổng quan về Haar Cascade

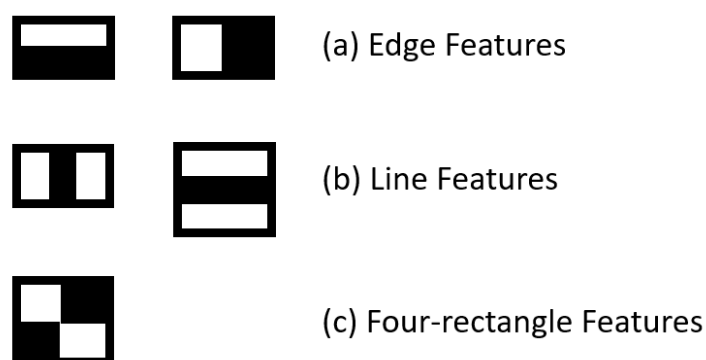
Các hình ảnh khuôn mặt đầu vào có thể chứa nhiều thông tin khác nhau ngoài hình ảnh khuôn mặt cần nhận diện (nhiều chi tiết khác trên ảnh nền, ...) do đó, cần phải xác định chính xác vị trí khuôn mặt trong ảnh trước khi tiến hành nhận diện. Trong nhiều trường hợp, người học có thể sử dụng các loại background 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, phương pháp Haar-Cascade[20] được ứng dụng dựa trên các đặc trưng Haar.

Haar Cascade là một thuật toán được tạo ra để phát hiện đối tượng (có thể là khuôn mặt, mắt, tay, đồ vật,...) được đề xuất vào năm 2001 bởi Paul Viola và Michael Jones trong bài báo của họ với khẳng định “Phát hiện đối tượng một cách nhanh chóng bằng cách sử dụng tầng (Cascade) tăng cường các tính năng đơn giản”.

Triển khai ban đầu được sử dụng để phát hiện khuôn mặt chính diện và các đặc điểm như Mắt, Mũi và Miệng. Tuy nhiên, có nhiều đặc trưng Haar được đào tạo trước cho các đối tượng khác cũng như cho toàn bộ cơ thể, thân trên, thân dưới, nụ cười và nhiều đồ vật khác.

3.4.2 Cách hoạt động của phương pháp Haar Cascade

Bằng cách sử dụng các điểm hình chữ nhật như một bộ lọc để phát hiện các đặc điểm khác nhau của khuôn mặt như mắt và các nốt như trong hình hình 3.2. Các cửa sổ trượt hình chữ nhật được chạy lần lượt trên hình ảnh và tổng số pixel nằm trong phần màu trắng được trừ cho tổng số pixel nằm trong phần màu đen.

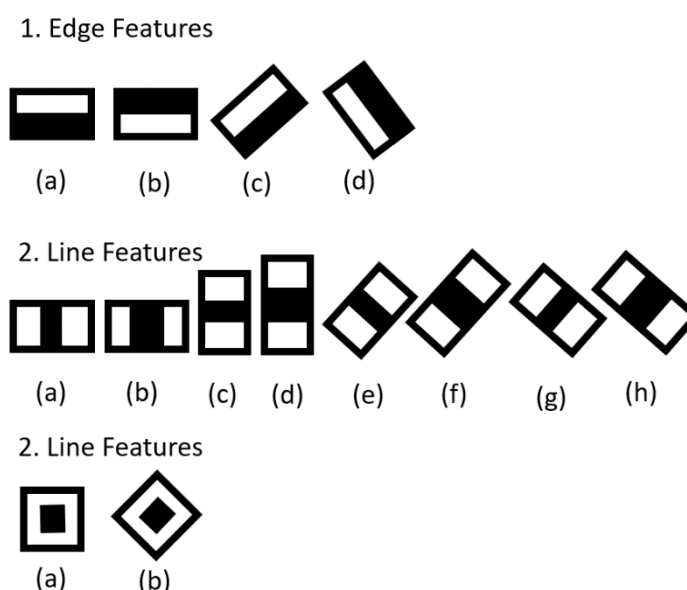


Hình 2.2 Đặc trưng hình chữ nhật trong phương pháp Haar-Cascade

Trong đó:

- a) Là các bộ lọc bắt các cạnh trong ảnh
- b) Bắt các đường thẳng trong ảnh
- c) Về đặc trưng 4 hình vuông

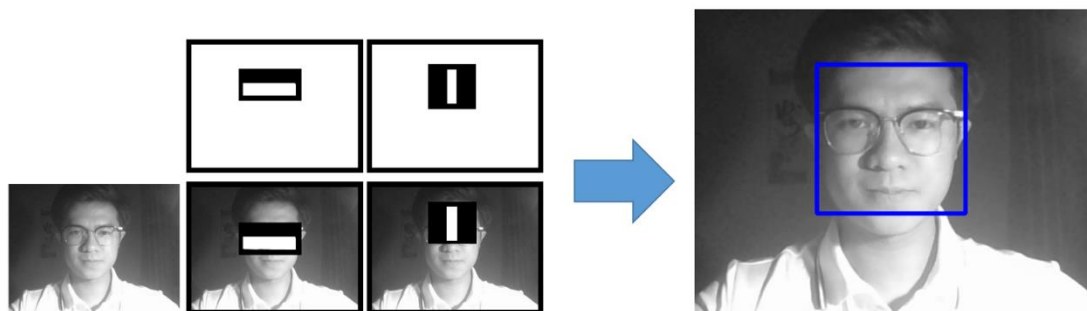
Hoặc các đặc trưng được nằm gọn trong trung tâm một vùng như trong hình 3.3



Hình 3.3 Các đặc trưng hình chữ nhật khác trong phương pháp Haar-Cascade

Các đặc trưng Haar cho phép phát hiện các khuôn mặt trong ảnh một cách nhanh chóng, thời gian thực và không phụ thuộc vào vị trí hoặc tỉ lệ ảnh. Haar-cascade cũng có thể được sử dụng để phát hiện nhiều khuôn mặt trong ảnh 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, các đường viền của hình chữ nhật định vị khuôn mặt sẽ được đặt là 3px.

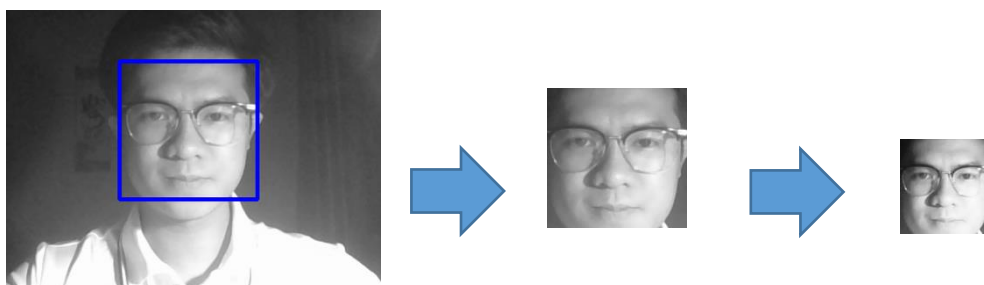
Tuy nhiên, cách áp dụng các bộ lọc này khác một chút so với các cửa sổ bộ lọc trong mạng nơ-ron tích chập, trong mạng nơ-ron tích chập bộ lọc chiếm toàn bộ cửa sổ trượt, trong khi ở đặc trưng Haar, bộ lọc chỉ chiếm một phần trong cửa sổ trượt thôi. Hình 3.3 minh họa một ví dụ về một khuôn mặt đã được phát hiện dựa trên phương pháp Haar-Cascade với các cửa sổ trượt và được tô viền xung quanh khuôn mặt sau khi phát hiện khuôn mặt.



Hình 4.4 Phát hiện khuôn mặt bằng phương pháp Haar-Cascade

3.5 Tiền xử lý 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 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.



Hình 5.5 Tiền xử lý hình ảnh đầu vào

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 nhiều, ...), 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 hóa 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): Nhằm giảm sự ảnh hưởng do chiếu sáng (chói), thiếu ánh sáng (ảnh tối), ..., các giải thuật xử lý ảnh thường nhạy cảm với ánh sáng, cùng nội dung ảnh nhưng với các điều kiện ánh sáng khác nhau có thể làm sai lệch kết quả xử lý. Do đó, cân bằng sáng ở bước tiền xử lý là một trong những cách giúp làm giảm các ảnh hưởng này.

- Giảm nhiễu với bộ lọc Gaussian: nhiễu gauss có được do bản chất rời rạc của bức xạ (hệ thống ghi ảnh bằng cách đếm các photon-lượng tử ánh sáng). Mỗi pixel trong ảnh nhiễu là tổng giá trị pixel đúng và pixel ngẫu nhiên. Lọc gaussian được thực hiện bằng cách nhân chập ảnh đầu vào với một ma trận lọc Gauss sau đó cộng chúng lại để tạo thành ảnh đầu ra.
- 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 hóa về kích thước 48x48), ... Hình 3.4 mô phỏng hình ảnh khuôn mặt sau khi được tiền xử lý.

3.6 Nhận diện cảm xúc

3.6.1 Bộ dữ liệu đào tạo

Bộ cơ sở dữ liệu ảnh là một trong các thành phần quan trọng hàng đầu trong các phương pháp học máy nói chung, được sử dụng để phục vụ cho quá trình tính toán tham số và huấn luyện, tinh chỉnh các mô hình. Thông thường, bộ dữ liệu càng lớn và càng được chọn lọc tỉ mỉ cẩn thận thì độ chính xác của mô hình càng được cải thiện.

Bộ dữ liệu đào tạo FER 2013 (Facial Emotion Recognition) Tập dữ liệu nguồn mở được tạo ra cho một dự án bởi PierreLuc Carrier và Aaron Courville, được chia sẻ công khai trong cuộc thi Kaggle (2013). Bộ dữ liệu FER 2013 bao gồm 35.887 ảnh xám: hình ảnh khuôn mặt kích thước 48x48 pixel từ nhiều góc độ khác nhau. Hình ảnh được phân loại thành một trong bảy lớp thể hiện cảm xúc khuôn mặt khác nhau, tất cả được gán nhãn từ 0 – 7 (0 = Giận dữ, 1 = Ghê tởm, 2 = Sợ hãi, 3 = Vui vẻ, 4 = Buồn, 5 = Ngạc nhiên, 6 = Bình thường) [28].

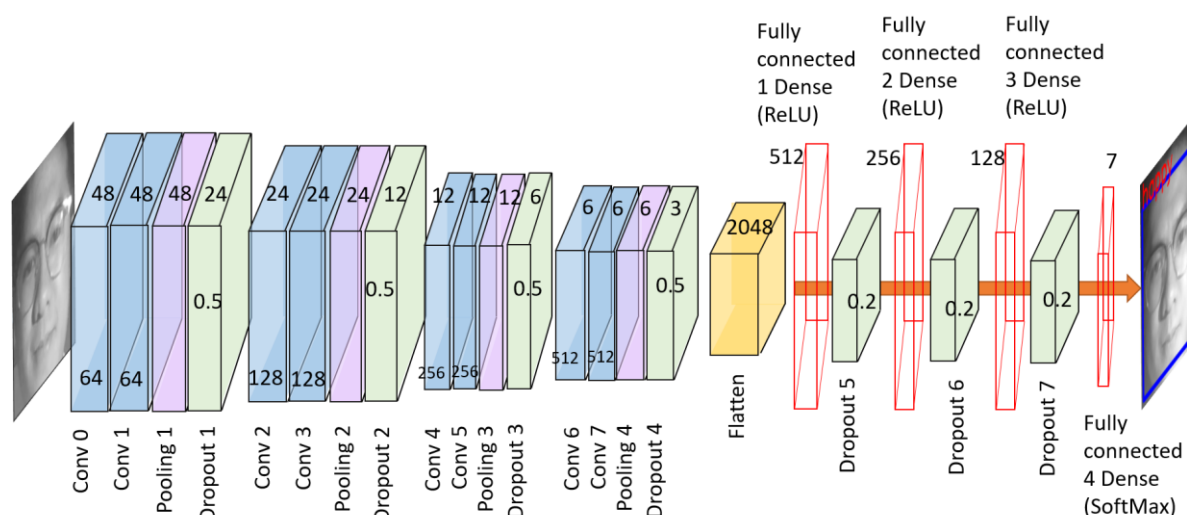
3.6.2 Xây dựng mô hình nhận diện cảm xúc

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. Mô hình nhận diện đề xuất dưới đây được xây dựng dựa trên ý tưởng của mạng VGG16 như đã trình bày ở phần trước. Hình 3.5 minh họa một cách chi tiết các lớp của mô hình nhận diện, bao gồm ba khối chính 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 kích thước ảnh đầu vào của bộ lọc có kích thước 48x48x1. Theo 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 tránh trường hợp overfitting trong quá trình huấn luyện.
- Khối thứ hai có cấu trúc tương tự như khối thứ nhất bao gồm 2 lớp tích chập gồm 128 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 có cấu trúc tương tự hai khối trước với kích thước đầu vào được giảm còn 12x12. Trong đó, hai lớp tích chập trong khối được tăng lên 256 bộ lọc nhằm tăng cường độ phức tạp cho mô hình cho phù hợp với số lượng dữ liệu đầu vào điều này giúp hạn chế tình trạng mô hình chưa khớp (under-fitting) cho mô hình.
 - Khối thứ tư về cơ bản cũng có cấu trúc tương tự như ba 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 512 đồng thời bổ sung thêm lớp flatten nhằm làm phẳng dữ liệu và kết hợp các đặc trưng của ảnh để có được đầu ra cho mô hình.
 - Khối cuối cùng bao gồm các lớp kết nối đầy đủ (fully connected layer) gồm 4 lớp. Lớp đầu tiên có 512 nơ-ron, trong đó sử dụng hàm kích hoạt ReLUs, các lớp kết nối đầy đủ phía sau lần lượt là 256 và 128 nơ-ron. 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 3.1.



Hình 6.6 Kiến trúc mạng tích chập cho nhận diện cảm xúc

Bảng 2.1 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 0	0	2 x 2	2	24 x 24 x 64
Dropout 0		Dropout=0.5		24 x 24 x 64
Conv2D-2	128	3 x 3	1	24 x 24 x 128
Conv2D-3	128	3 x 3	1	24 x 24 x 128

Pooling 1	0	2 x 2	2	12 x 12 x 128
Dropout 1		Dropout=0.5		12 x 12 x 128
Conv2D-4	256	3 x 3	1	12 x 12 x 256
Conv2D-5	256	3 x 3	1	12 x 12 x 256
Pooling 2	0	2 x 2	2	6 x 6 x 256
Dropout 2		Dropout=0.5		6 x 6 x 256
Conv2D-6	512	3 x 3	1	6 x 6 x 512
Conv2D-7	512	3 x 3	1	6 x 6 x 512
Pooling 3	0	2 x 2	2	3 x 3 x 512
Dropout 3		Dropout=0.5		3 x 3 x 512
Flatten				1 x 1 x 2048
Dense-0	512	activation='relu'		1 x 1 x 512
Dropout 4		Dropout=0.2		1 x 1 x 512
Dense-1	256	activation='relu'		1 x 1 x 256
Dropout 5		Dropout=0.2		1 x 1 x 256
Dense-2	128	activation='relu'		1 x 1 x 128
Dropout 6		Dropout=0.2		1 x 1 x 128
Dense-3	7	activation='softmax'		1 x 1 x 7
Output	0	0	None	1 x 1 x 7

3.6.3 Môi trường đào tạo

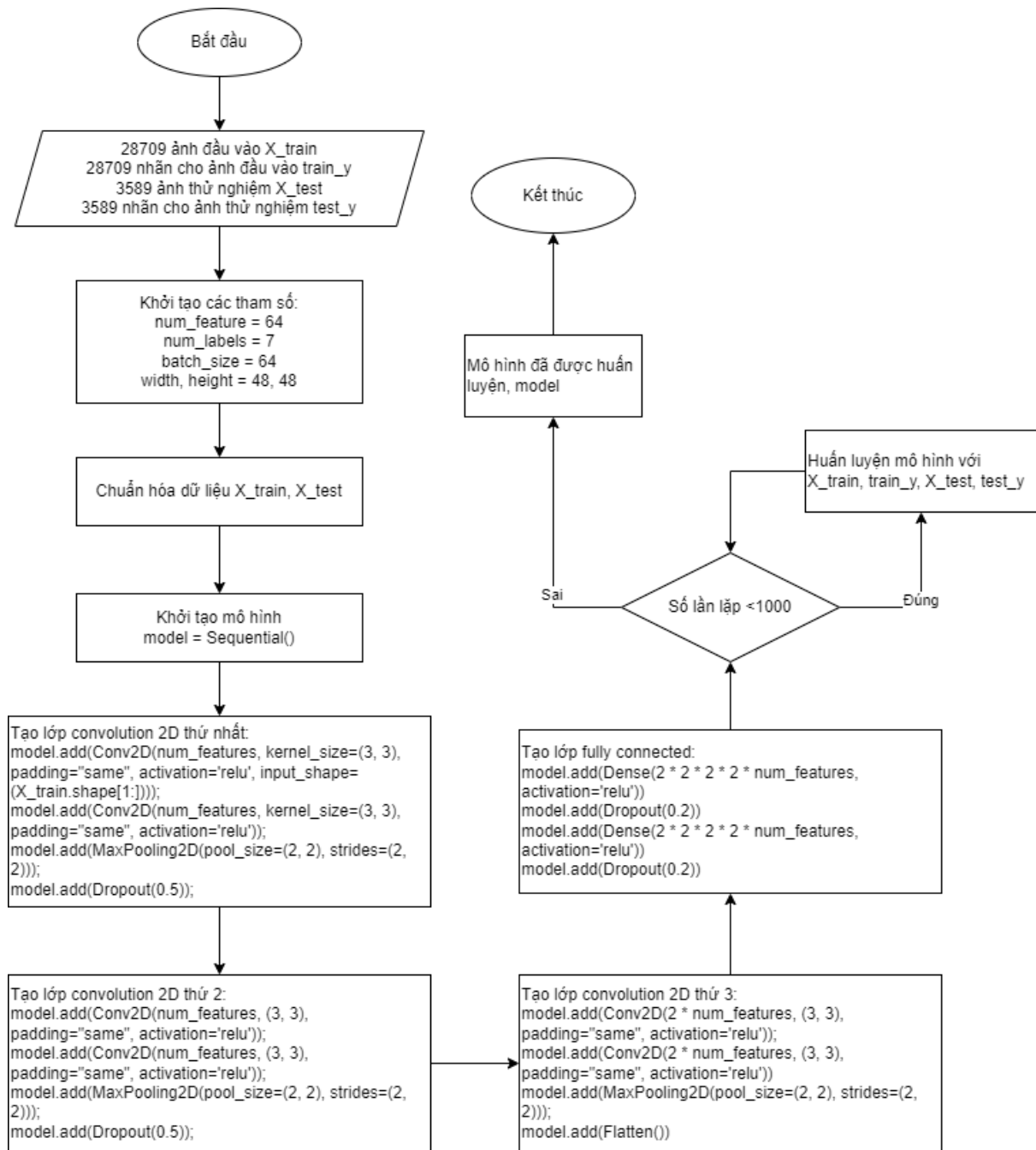
Mô hình đề xuất được huấn luyện với 28709 ảnh trong bộ CSDL FER 2013. Trong quá trình thực nghiệm, mô hình đã được triển khai với ngôn ngữ lập trình Python, quá trình huấn luyện được thực hiện trên Colaboratory hay còn gọi là Google Colab, một dịch vụ máy chủ điện toán đám mây của Google dành cho mục đích nghiên cứu. Dịch vụ này cho phép chạy các dòng code python thông qua trình duyệt, đặc biệt phù hợp với các lĩnh vực nghiên cứu: phân tích dữ liệu, học máy, trí tuệ nhân tạo... Colab cung cấp nhiều loại GPU, thường là Nvidia K80s, T4s, P4s và P100s, tuy nhiên người dùng không thể chọn loại GPU trong Colab, GPU trong Colab thay đổi theo thời gian. Vì là dịch vụ miễn phí, nên Colab sẽ có những thứ tự ưu tiên trong việc sử dụng tài nguyên hệ thống, cũng như giới hạn thời gian sử dụng, thời gian sử dụng tối đa lên tới 12 giờ, Bảng 3 mô tả cấu hình phần cứng Google Colab được sử dụng trong nghiên cứu này.

Bảng 3.2 Cấu hình phần cứng GoogleColab

CPU	GPU	TPU
Intel(R) Xeon(R) CPU @ 2.30 GHz và 13GB RAM	Tesla K80 12GB, GDDR5 VRAM, Intel(R) Xeon(R) CPU @ 2.20 GHz và 13GB RAM	TPU Cloud, Intel(R) Xeon(R) CPU @ 2.30 GHz và 13GB RAM

3.6.4 Đào tạo mô hình nhận diện cảm xúc

- Bước 1: Đọc dữ liệu đầu vào từ bộ dữ liệu FER 2013 dưới dạng file .csv và khởi tạo các tham số cần thiết
- Bước 2: Duyệt lần lượt dữ liệu đầu vào và tiến hành chuẩn hóa dữ liệu
- Bước 4: Khởi tạo mô hình đào tạo
- Bước 5: Tiến hành đào tạo mô hình và lưu lại kết quả



Hình 7.7 Sơ đồ khối đào tạo mô hình nhận diện cảm xúc

3.7 Kết luận

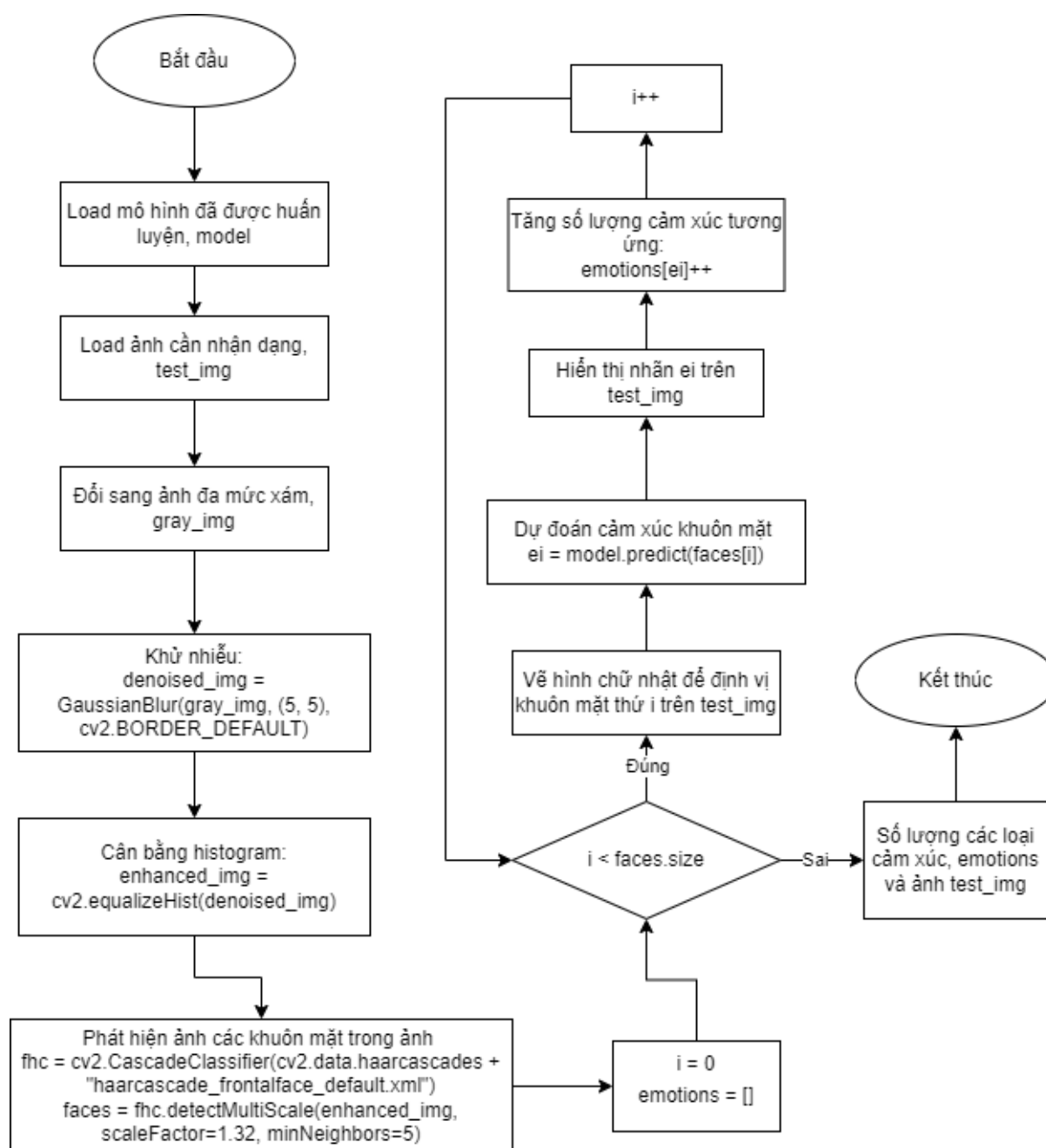
Lược đồ đề xuất bao gồm năm bước chính bao gồm: Thu thập hình ảnh đầu vào từ lớp học, phát hiện khuôn mặt dựa trên phương pháp Haar-Cascade, tiền xử lý ảnh bao gồm: cắt ảnh, cân bằng histogram và sử dụng bộ lọc gaussian, nhận diện cảm xúc, từ đó có thể đưa ra kết quả theo mong muốn. Với mô hình mạng nơ-ron tích chập được xây dựng dựa trên ý tưởng của VGG16, mô hình được đào tạo với bộ CSDL FER 2013 dựa trên môi trường do Colaboratory (dịch vụ điện toán đám mây do Google cung cấp) và được cải tiến phù hợp với bộ dữ liệu đào tạo, hứa hẹn sẽ đem lại hiệu quả cao khi đem vào ứng dụng thực tế. Đây sẽ là bước tiến quan trọng trong việc góp phần cải thiện chất lượng dạy học trực tuyến tận dụng lợi ích của dạy học trực tuyến dựa trên mô hình nhận diện cảm xúc trong thời gian thực.

CHƯƠNG 4. KẾT QUẢ THỰC NGHIỆM

4.1 Cài đặt chương trình nhận diện cảm xúc

Sau khi đào tạo mô hình nhận diện cảm xúc từ bộ dữ liệu FER 2013 kết quả thu được của mô hình sẽ được lưu lại dưới dạng file h5, việc cài đặt chương trình nhận dạng đơn giản sẽ bao gồm ba bước chính như sau:

- Bước 1: Đọc mô hình, khởi tạo tham số mặc định và đọc dữ liệu nhận diện đầu vào
- Bước 2: Tiền xử lý ảnh, tiến hành phát hiện khuôn mặt
- Bước 3: Tiến hành dự đoán cảm xúc khuôn mặt đã được phát hiện và vẽ nhãn lên ảnh cho hình ảnh thêm trực quan, cuối cùng chúng ta hiển thị bức ảnh trước và sau khi xử lý lên màn hình và cài đặt phím tắt để dừng chương trình.



Hình 1.1 Sơ đồ khối nhận diện cảm xúc

4.2 So sánh với mô hình mạng nơ-ron VGG16

Trước khi tiến hành đánh giá thực nghiệm với bộ dữ liệu FER 2013 và đưa vào ứng dụng thực tế, việc so sánh mô hình đã xây dựng và đào tạo với mô hình theo kiến trúc nổi tiếng là VGG16 giúp đánh giá mô hình đã xây dựng có tính khách quan hơn so với những mô hình kiến trúc nổi tiếng.

Bảng 1.1 So sánh cấu trúc và thời gian đào tạo mô hình đề xuất và VGG16

Mô hình	Tổng số lớp tích chập (convolutional)	Tổng số lớp lấy mẫu (Pooling)	Tổng số lớp kết nối đầy đủ (Fully-connect)	Thời gian đào tạo (h)
VGG16	13	5	4	45
Mô hình đề xuất	8	4	4	30











4.2 Kết quả thực nghiệm với bộ dữ liệu FER 2013

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, các kết quả thực nghiệm thu được được mô tả trong Bảng 4.2. 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, Bảng 4.3 minh họa một số ảnh cụ thể trong quá trình kiểm thử đối với bộ dữ liệu trên với mô hình đề xuất và mô hình theo kiến trúc VGG16.

Bảng 2.2 Kết quả thí nghiệm kiểm tra mô hình với bộ dữ liệu kiểm thử

Mô hình	Số lượng ảnh tập huấn	Số lượng ảnh kiểm thử	Số lượng kết quả đúng	Tỷ lệ chính xác	Thời gian trung bình (ms)
VGG16	28709	3589	1536	42,7%	68,33
Mô hình đề xuất	28709	3589	3443	95,9%	56,76

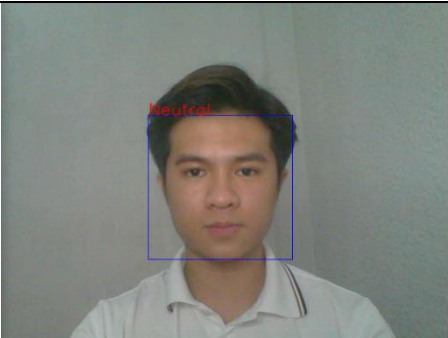

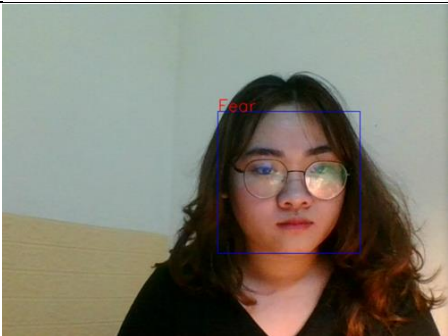
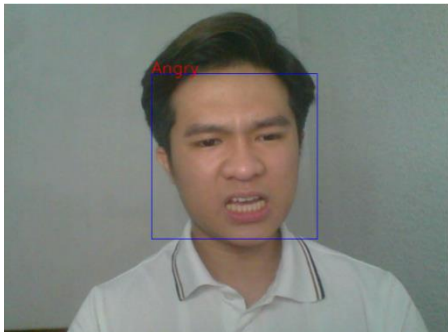
Bảng 3.3 Một số kết quả thử nghiệm



Bộ CSDL	Ảnh	Nhãn CSDL	Nhãn kết quả (VGG16)	Nhãn kết quả (Mô hình đề xuất)
FER2013		Vui vẻ	Vui vẻ	Vui vẻ
		Sợ hãi	Bình thường	Sợ hãi
		Tức giận	Bình thường	Tức giận
		Buồn	Bình thường	Buồn
		Bình thường	Buồn	Bình thường
		Tức giận	Bình thường	Tức giận
		Ghê tởm	Bình thường	Sợ hãi
		Ngạc nhiên	Bình thường	Ngạc nhiên
		Vui vẻ	Vui vẻ	Vui vẻ
		Bình thường	Sợ hãi	Buồn

4.3 Ứng dụng thực tế trên khuôn mặt

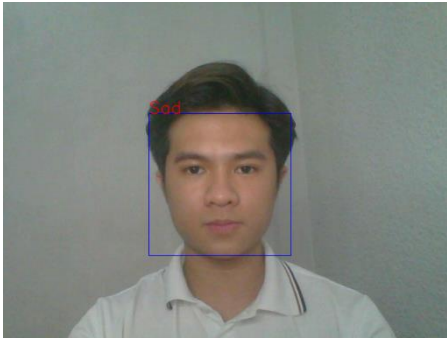
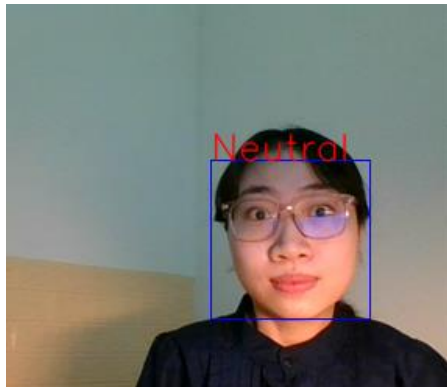
Để kiểm tra hiệu quả của phương pháp được đề xuất trong các ứng dụng thực tế, chúng tôi đã ứng dụng thực tiễn với khuôn mặt của tác giả và một vài khuôn mặt khác trong thực tế làm hình ảnh đầu vào trong thời gian thực và đưa mô hình mạng nơ-ron tích chập vào nhận dạng cảm xúc trong ảnh. Kết quả thực nghiệm cho thấy mô hình nhận dạng cảm xúc trong thời gian thực đạt hiệu quả tốt với độ chính xác trong thời gian thực. Điều này cho thấy mô hình đã có đủ khả năng để có thể áp dụng được vào lớp học thực tế. Bảng 4.5 mô phỏng lại kết quả thực nghiệm trên khuôn mặt riêng lẻ trong thực tế với mô hình được đề xuất, từ đó ta có thể so sánh kết quả này với mô hình VGG16 với bức ảnh tương tự.

Bảng 4.4 Kết quả thực nghiệm với khuôn mặt tác giả trong thời gian thực với mô hình đề xuất

Ảnh	Thời gian nhận diện(s)	Nhãn kết quả
	0.15	Bình thường
	0.14	Ngạc nhiên
	0.12	Sợ hãi
	0.19	Tức giận

	0.14	Buồn
	0.17	Vui vẻ

Bảng 5.5 Kết quả thực nghiệm với khuôn mặt tác giả trong thời gian thực với mô hình VGG16

Ảnh	Thời gian nhận diện(s)	Nhãn kết quả
	0.21	Buồn
	0.22	Bình thường

	0.18	Sợ hãi
	0.2	Bình thường
	0.19	Bình thường
	0.21	Vui vẻ

4.4 Thử nghiệm thực tế tại lớp học trường Đại học sư phạm Hà Nội

Trong phần này, 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).

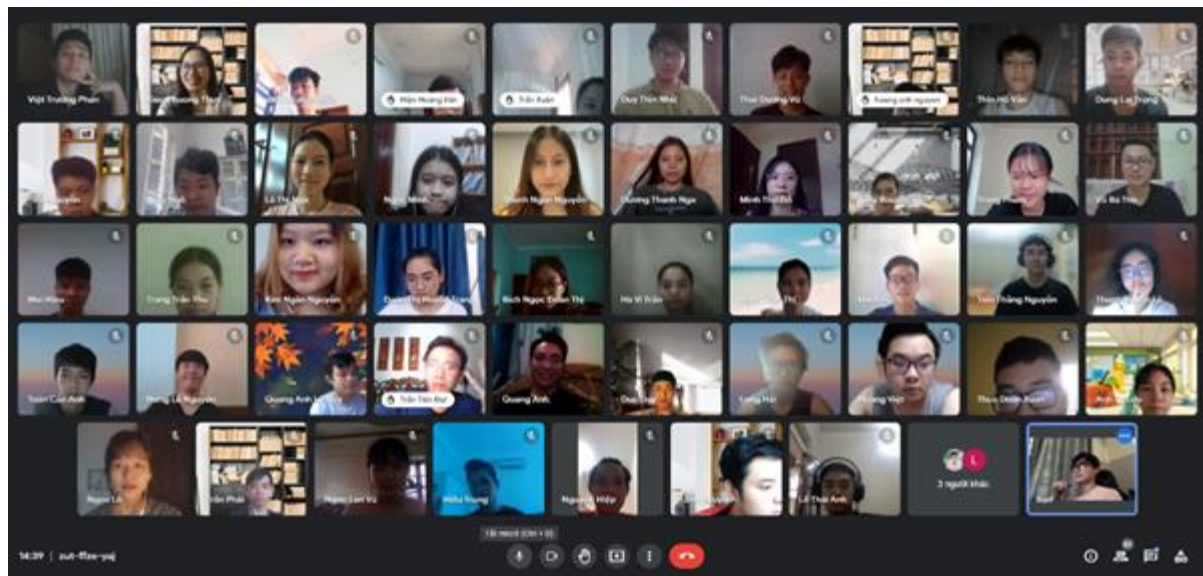
Bảng 6.6 Kết quả thử nghiệm tại lớp học Khoa Công nghệ thông tin, trường ĐH Sư phạm Hà Nội

Tên môn	Số lượng sinh viên	Số khuôn mặt phát hiện được	Số khuôn mặt được gắn 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

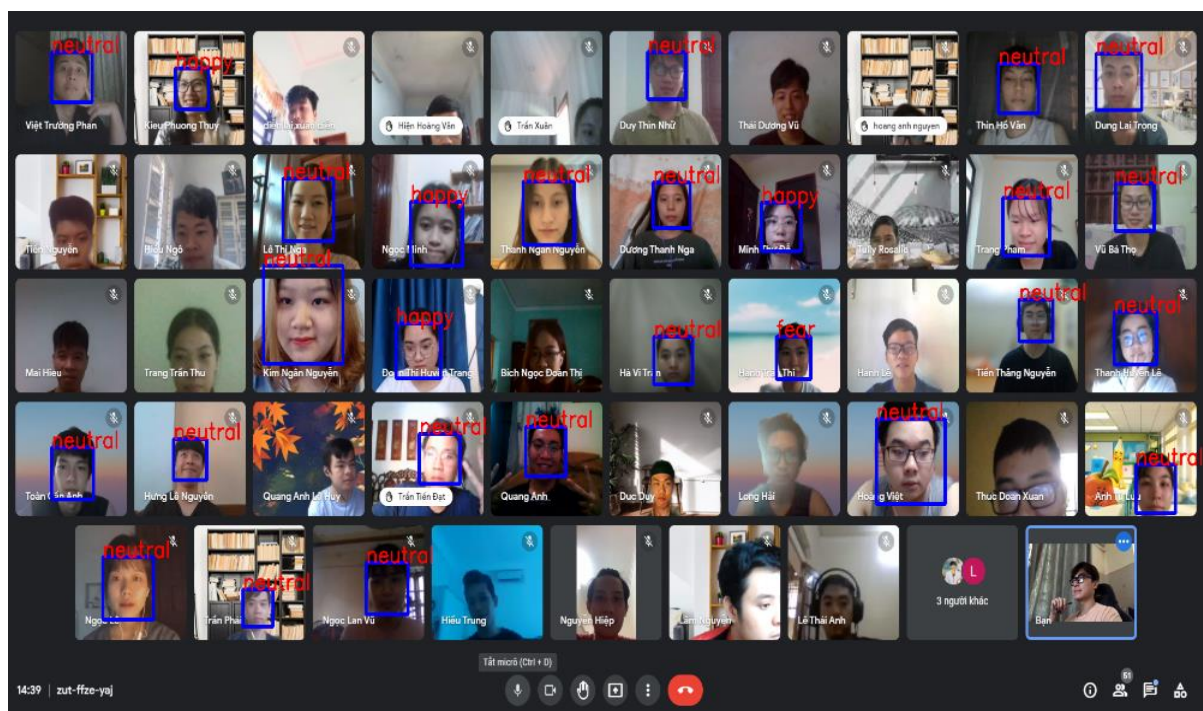
Hình 4.2 và Hình 4.3 minh hoạ một ví dụ về việc đánh giá cảm xúc của lớp học. Hầu hết các khuôn mặt đã được phát hiện và đánh dấu bằng các đường viền hình chữ nhật; các biểu cảm của các khuôn mặt được tiền xử lý một cách rõ nét và đã được nhận diện với các nhãn tương ứng. Trong tổng số 48 khuôn mặt, 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 4.6 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 bài 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”. 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.



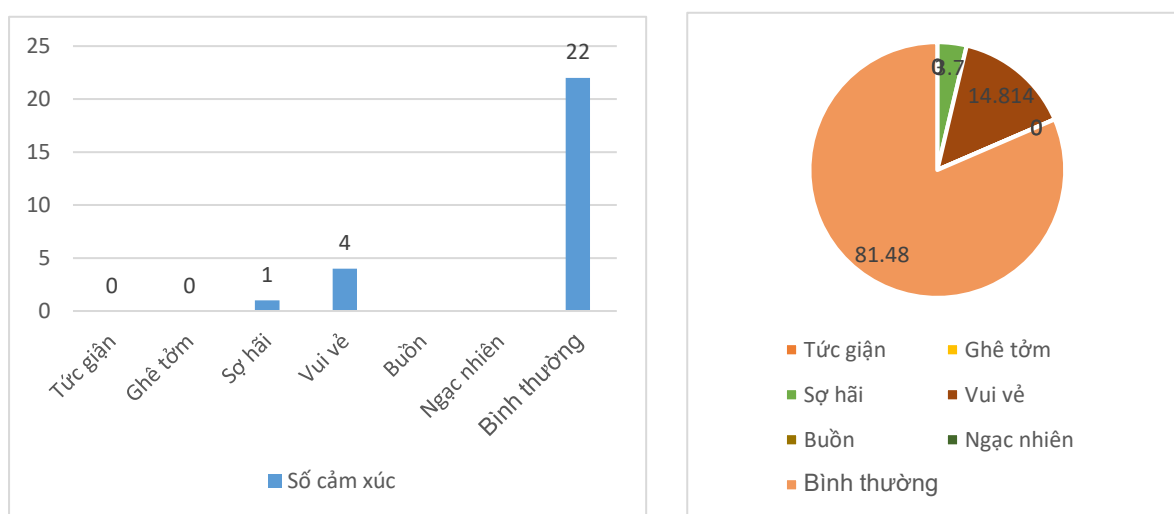
Hình 2.2 Hình ảnh lớp học trực tuyến



Hình 3.3 Nhận diện cảm xúc khuôn mặt trong lớp học trực tuyến

4.5 Kết quả

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 bộ cơ sở dữ liệu hình ảnh chuẩn FER 2013 để thực nghiệm. Các kết quả thực nghiệm cho thấy, mức độ nhận diện cảm xúc với độ chính xác 95,9% đối với bộ CSDL FER2013. 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.



Hình 4.4 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 ĐH 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 một kết quả tiềm năng và thú vị.

CHƯƠNG 5. KẾT LUẬN

Kết quả đã thực hiện được của luận văn

Luận văn đã nghiên cứu và tìm ra phương pháp giúp cải thiện chất lượng giáo dục trực tuyến trong hoàn cảnh dịch bệnh nghiêm trọng. 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 mô hình mạng tích chập CNN, một phương pháp phân tích cảm xúc của sinh viên dựa trên nét mặt đã được giới thiệu. Để giải quyết bài toán có tính cấp thiết cao với kết quả tốt, luận văn đã đạt được một số mục tiêu như sau:

- Đề xuất một phương pháp nhằm cải thiện những vấn đề đặt ra trong giáo dục trực tuyến như đã trình bày ở phần trước, nhằm cải thiện chất lượng dạy học trực tuyến, điều này đã đóng góp một phần trong việc cải thiện tạo dựng niềm tin giữa người dạy và học cũng như cải thiện chất lượng dạy học trong giáo dục trực tuyến.
- Tìm hiểu và nghiên cứu về Deep Learning trong lĩnh vực nhận diện và xử lý ảnh cụ thể là mô hình mạng nơ-ron tích chập, đây là một lĩnh vực còn mới mẻ chứa đựng nhiều tiềm năng và thách thức tại Việt Nam.
- Đề xuất một lược đồ nhằm giải quyết bài toán mà luận văn đã đặt ra bằng những kiến thức đã tìm hiểu. Một mô hình đề xuất bao gồm bốn bước chính là: thu thập ảnh đầu vào, phát hiện khuôn mặt (sử dụng phương pháp haar-cascade), tiền xử lý ảnh, nhận diện cảm xúc, thống kê lại kết quả. Đặc biệt nhất trong đó là bước nhận diện cảm xúc, để có thể xây dựng một mô hình đạt hiệu quả cao, thì ý tưởng dựa trên VGG16 và đã được điều chỉnh lại sao cho phù hợp đã mang lại thành công cho luận văn này.
- Thử nghiệm lược đồ với dữ liệu đầu vào là bộ dữ liệu FER2013 và dữ liệu ảnh lớp học tại khoa công nghệ thông tin trường Đại học sư phạm Hà Nội đã cho thấy lược đồ đề xuất hoạt động rất tốt trong thời gian thực

Phương hướng phát triển luận văn trong tương lai

Hiện tại, kết quả nhận diện vẫn còn hạn chế và còn nhiều nhược điểm trong quá trình đào tạo mô hình cũng như thuật toán nhận diện khuôn mặt, do chất lượng hình ảnh chụp còn chưa đủ tốt, việc phát hiện hình ảnh khuôn mặt có tỷ lệ chưa cao và thuật toán nhận diện khuôn mặt cũng như nhận diện cảm xúc và đánh nhãn cho khuôn mặt còn có thể cải thiện hơn. Do đó, trong tương lai luận văn sẽ khắc phục và cải thiện như sau:

- Nâng cấp khả năng phát hiện khuôn mặt trong điều kiện hạn chế bằng cách điều chỉnh độ phân giải thấp, môi trường chụp hạn chế, ảnh bị mờ chữ lên mặt.
- Đào tạo mô hình cũng như cải thiện thuật toán nhằm tăng khả năng ứng dụng thực tế của mô hình đề xuất.

Ngoài ra, với số lượng lớn người tham gia các lớp học trực tuyến lớn, nhưng màn hình học trực tuyến tại mỗi thời điểm là hạn chế, do đó, không thể đảm bảo việc đánh giá được toàn bộ

người đang học cùng một lúc. Một số giải pháp cũng đã được đề xuất và sẽ được giới thiệu trong các nghiên cứu tiếp theo.

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.
- [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 Recognition Workshops*, 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 W. V. Friesen. A new pan cultural facial expression of emotion . Motivation and Emotion, vol. 10, no. 2, pp. 159–168, 1986.
- [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] Dumoulin, V., & Visin, F. (2016). *A guide to convolution arithmetic for deep learning*.
- [24] Samer, C. H., Rishi, K., & Rowen. *Image Recognition Using Convolutional Neural Networks*. Cadence Whitepaper, 1–12, 2015.
- [25] Reinhard Klette. *Concise Computer Vision*. Springer, 2014.
- [26] T. Szandala *Review and Comparison of Commonly Used Activation*. Wroclaw University of Science and Technology Wroclaw, Poland.
- [27] H.K. Jabbar, R.Z. Khan. *Method To Avoid Over-Fitting And Under-Fitting In Supervised Machine Learning (Comparative Study)*. Computer Science, Communication & Instrumentation Devices, 2015.
- [28] S. Turabzadeh, H. Meng, R. Swash, M. Pleva, and J. Juhar, *Facial Expression Emotion Detection for Real-Time Embedded Systems*, Technologies, vol. 6, no. 1, p. 17, Jan. 2018.
- [29] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Fei-Fei, L. (2015). *ImageNet Large Scale Visual Recognition Challenge*. International Journal of Computer Vision, 115(3), 211–252.
- [30] Q. Guan, Y. Wang , B. Ping , D. Li , J. Du, Y. Qin, H. Lu, X. Wan, J. Xiang . *Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study*. Ivyspring International Publisher, 2019.
- [31] J. Bjorck, C. Gomes, B. Selman, K. Q. Weinberger. *Understanding Batch Normalization*. Cornell University, 2018.
- [32] D. Anderson, G. McNeill. *Artificial Neural Networks Technology*. Kaman Sciences Corporation. DACS State-of-the-Art Report, 1992.
- [33] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner. *Gradient-based learning applied to document recognition*, 1998.

