

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



HOÀNG TRUNG CÔNG

**NGHIÊN CỨU MÔ HÌNH HỌC SÂU CHO
PHÂN TÍCH CẢM XÚC DỰA TRÊN HÌNH ẢNH**

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT
(Theo định hướng ứng dụng)

HÀ NỘI - NĂM 2025

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



HOÀNG TRUNG CÔNG

**NGHIÊN CỨU MÔ HÌNH HỌC SÂU CHO
PHÂN TÍCH CẢM XÚC DỰA TRÊN HÌNH ẢNH**

**Chuyên ngành : Khoa học máy tính
Mã số : 8.48.01.01**

**ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT
(Theo định hướng ứng dụng)**

**NGƯỜI HƯỚNG DẪN KHOA HỌC:
TS. ĐÀO THỊ THUÝ QUỲNH**

HÀ NỘI - NĂM 2025

LỜI CAM ĐOAN

Tôi xin cam đoan đề án thạc sĩ “Nghiên cứu mô hình học sâu cho phân tích cảm xúc dựa trên hình ảnh” của tôi được thực hiện dưới sự hướng dẫn của giáo viên hướng dẫn là TS. Đào Thị Thuý Quỳnh. Các nội dung trong khoá luận đều được ghi rõ nguồn gốc ở phía cuối đề án.

Các số liệu, kết quả nêu trong đề án tốt nghiệp là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác. Nếu có phát hiện nào về sự gian lận trong sao chép tài liệu, công trình nghiên cứu của tác giả khác mà không được ghi rõ trong phần tài liệu tham khảo, tôi sẽ chịu hoàn toàn trách nhiệm về kết quả đề án của mình.

Học viên thực hiện

Hoàng Trung Công

LỜI CẢM ƠN

Lời đầu tiên, tôi xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới TS. Đào Thị Thuý Quỳnh đã định hướng đề tài và tận tình hướng dẫn, chỉ bảo tôi trong suốt quá trình thực hiện đề án thạc sĩ.

Tôi xin chân thành cảm ơn các thầy, cô trong Trường Học Viện Công nghệ Bưu chính Viễn thông đã truyền đạt kiến thức trong suốt thời gian học cao học và đã tạo mọi điều kiện thuận lợi cho tôi học tập và nghiên cứu.

MỤC LỤC

LỜI CAM ĐOAN.....	i
LỜI CẢM ƠN.....	ii
DANH MỤC CÁC KÝ HIỆU, CHỮ VIẾT TẮT	vii
DANH MỤC CÁC BẢNG	viii
DANH MỤC CÁC HÌNH.....	ix
MỞ ĐẦU	1
1. Lý do chọn đề tài	1
2. Mục tiêu nghiên cứu	2
3. Đối tượng và phạm vi nghiên cứu	2
4. Phương pháp nghiên cứu	3
CHƯƠNG 1: CƠ SỞ LÝ LUẬN.....	4
1.1. Giới thiệu chương	4
1.1.1. Tầm quan trọng của nhận diện cảm xúc sinh viên trong giảng dạy	4
1.1.1.1. Tỷ lệ trầm cảm và áp lực tâm lý ở học sinh, sinh viên	5
1.1.1.2. Rủi ro trong lớp học do cảm xúc tiêu cực	5
1.1.1.3. Cá nhân hóa giảng dạy thông qua nhận diện cảm xúc	6
1.1.1.4. Hỗ trợ sức khỏe tâm lý của sinh viên.....	7
1.1.1.5. Tăng cường quản lý lớp học thông minh	7
1.1.1.6. Giải quyết khoảng trống trong giáo dục Việt Nam.....	7
1.2. Cơ sở lý thuyết về cảm xúc.....	8
1.2.1. Khái niệm và vai trò của cảm xúc trong giáo dục	8
1.2.1.1. Định nghĩa cảm xúc và ý nghĩa trong môi trường học tập.....	8
1.2.1.2. Ảnh hưởng của cảm xúc đến hiệu quả giảng dạy và học tập	8
1.2.2. Lý thuyết của Paul Ekman về cảm xúc cơ bản.....	9

1.2.3. Tầm quan trọng của nhận diện cảm xúc trong giáo dục Việt Nam	10
1.2.3.1. Áp lực tâm lý và trầm cảm ở học sinh, sinh viên.....	10
1.2.3.2. Nhu cầu cá nhân hóa giảng dạy và quản lý lớp học thông minh ...	11
1.3. Tổng quan về học sâu trong nhận diện cảm xúc.....	11
1.3.1. Giới thiệu công nghệ học sâu	12
1.3.1.1. Khái niệm học sâu và mạng nơ-ron tích chập (CNN).....	12
1.3.1.2. So sánh các phương pháp nhận diện truyền thống và học sâu trong nhận diện cảm xúc	13
Các phương pháp nhận diện truyền thống.....	13
1.3.1.3. So sánh chi tiết các phương pháp.....	15
1.3.2. Các kỹ thuật phát hiện và phân tích cảm xúc	17
1.3.3. So sánh các mô hình học sâu.....	21
1.4. Nhân tố ảnh hưởng đến nhận diện cảm xúc.....	21
1.5. Kết luận chương.....	23
CHƯƠNG 2: KHẢO SÁT VÀ ĐÁNH GIÁ CÁC MÔ HÌNH NHẬN DIỆN CẢM XÚC.....	25
2.1. Giới thiệu chương	25
2.2. Các mô hình học sâu được khảo sát.....	25
2.2.1. Giới thiệu các mô hình học sâu được khảo sát.....	25
2.2.2. Vai trò của Haar Cascade trong phát hiện khuôn mặt.....	27
2.2.3. Tăng cường dữ liệu và xử lý video.....	29
2.3. Đánh giá trên các tập dữ liệu tiêu chuẩn.....	31
2.3.1. Các bộ dữ liệu được sử dụng.....	31
2.3.2. Tiêu chí và kết quả đánh giá.....	33
2.4. Thảo luận về tính khả thi	33
2.4.1. Yêu cầu kỹ thuật.....	33
2.4.2. Các yếu tố xã hội và đạo đức.....	34

2.5. Kết luận chương.....	34
CHƯƠNG 3: NGHIÊN CỨU THỬ NGHIỆM ĐÁNH GIÁ KẾT QUẢ.....	36
3.1. Giới thiệu chương	36
3.2. Nội dung và phương pháp tiếp cận	36
3.2.1. Mục tiêu và phạm vi.....	36
3.2.2. Phương pháp tiếp cận	37
3.3. Thu thập và xử lý dữ liệu.....	38
3.3.1. Nguồn dữ liệu	38
3.3.2. Tiền xử lý dữ liệu	40
3.4. Huấn luyện và tối ưu hoá mô hình.....	42
3.4.1. Kiến trúc mô hình CNN	42
3.4.2. Quá trình huấn luyện	45
3.5. Thử nghiệm và đánh giá	46
3.5.1. Thử nghiệm thực tế	46
3.5.2. Đánh giá hiệu quả.....	47
3.5.2.1. Các Chỉ số Đánh giá Định lượng	47
3.5.2.2. Phân tích Kết quả trên các Bộ Dữ liệu Tiêu chuẩn và Nội địa	47
3.5.2.2.1. Phân tích kết quả trên dữ liệu FER2013.....	48
3.5.2.2.2. Phân tích Kết quả trên Dữ liệu CK+	54
3.5.2.3. Phân tích Lỗi Phân loại Thường gặp.....	60
3.5.2.4. Giải pháp Cải tiến để Giảm Nhầm lẫn Cảm xúc.....	62
3.5.2.5. Đánh giá Thời gian Thực và Khả năng Triển khai	63
3.5.2.6. Phân tích Cảm xúc theo Thời gian và Tối ưu hóa Can thiệp Sự phạm	63
3.5.2.7. Đảm bảo Đạo đức và Chấp nhận Cộng đồng.....	64
3.6. Kết luận chương.....	64
KẾT LUẬN.....	66

DANH MỤC CÁC TÀI LIỆU THAM KHẢO	68
--	-----------

DANH MỤC CÁC KÝ HIỆU, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
Haar Cascade	Haar-like Feature Cascade Classifier	Bộ phân loại phát hiện khuôn mặt dựa trên đặc trưng Haar
FACS	Facial Action Coding System	Hệ thống mã hóa chuyển động cơ mặt
FER2013	Facial Expression Recognition 2013	Bộ dữ liệu nhận diện biểu cảm khuôn mặt 2013
AffectNet	Affective Network Dataset	Bộ dữ liệu mạng cảm xúc
Accuracy	Accuracy	Độ chính xác
Precision	Precision	Độ chính xác của lớp
Recall	Recall	Độ bao phủ
CK+	Extended Cohn-Kanade Dataset	Bộ dữ liệu nhận diện biểu cảm khuôn mặt Cohn-kanade

DANH MỤC CÁC BẢNG

Bảng 1. 1 Bảng so sánh sự khác biệt các phương pháp truyền thống và học sâu	17
Bảng 1. 2 So sánh trực quan về số tham số, độ chính xác và thời gian xử lý.....	21
Bảng 2. 1 Bảng so sánh các tiêu chí.....	33

DANH MỤC CÁC HÌNH

Hình 1. 1 Hình ảnh khuôn mặt minh họa từng "Action Unit" (AU) trong FACS và biểu cảm cảm xúc tương ứng	10
Hình 1. 2 Sơ đồ kiến trúc của một mạng nơ-ron truyền thẳng	12
Hình 1. 3 Một mô hình học sâu trong nhận dạng mặt người	13
Hình 1. 4 Đặc trưng Haar và mô hình phân loại nhiều tầng (cascade classifier).....	18
Hình 1. 5 Hình ảnh minh họa các điều kiện ánh sáng khác nhau	22
Hình 1. 6 Hình ảnh minh họa các góc chụp khác nhau	23
Hình 2. 1 Sơ đồ kiến trúc CNN, minh họa các lớp tích chập	26
Hình 2. 2 Biểu đồ minh họa sự khác biệt trong việc trích xuất đặc trưng giữa các mô hình	27
Hình 2. 3 Phát hiện khuôn mặt bằng Haar Cascade.....	28
Hình 2. 4 Khuôn mặt trong các điều kiện ánh sáng không đồng đều	28
Hình 2. 5 Ví dụ ảnh bộ dữ liệu FER2013	32
Hình 2. 6 Ảnh ví dụ từ bộ dữ liệu CK+	32
Hình 3. 1 Hình ảnh mô tả về dữ liệu FER2013	40
Hình 3. 2 Sơ đồ kiến trúc chi tiết của mô hình CNN trong nhận diện cảm xúc	43
Hình 3. 3 Biểu đồ Loss và Accuracy - FER2013.....	49
Hình 3. 4 Biểu đồ tần suất các cảm xúc được dự đoán trong video lớp học, sử dụng mô hình TFLite – FER2013	49
Hình 3. 5 Khung hình video lớp học với khuôn mặt được phát hiện và nhãn cảm xúc dự đoán bằng mô hình TFlite.....	51
Hình 3. 6 Báo cáo phân loại dạng - FER2013	52

Hình 3. 7 Đánh giá mức độ hứng thú.....	54
Hình 3. 8 Biểu đồ Loss và Accuracy - CK+	54
Hình 3. 9 Biểu đồ tần suất các cảm xúc được dự đoán trong video lớp học, sử dụng mô hình TFLite – CK+	56
Hình 3. 10 Khung hình video lớp học với khuôn mặt được phát hiện và nhận cảm xúc dự đoán bằng mô hình Tflite – Ck+.....	57
Hình 3. 11 Báo cáo phân loại dạng - Ck+.....	58
Hình 3. 12 Đánh giá mức độ hứng thú.....	60
Hình 3. 13 Biểu đồ tỷ lệ lỗi của từng lớp cảm xúc trên các bộ dữ liệu CK+ và FER2013	62

MỞ ĐẦU

1. Lý do chọn đề tài

Trong bối cảnh giáo dục hiện đại, việc hiểu và quản lý trạng thái cảm xúc của học sinh, sinh viên đóng vai trò quan trọng trong việc nâng cao hiệu quả giảng dạy và học tập. Theo nghiên cứu của Aslan và cộng sự (2019), khoảng 30% học sinh trong các lớp học truyền thống hoặc trực tuyến thể hiện các trạng thái cảm xúc tiêu cực như căng thẳng, buồn chán, hoặc mất tập trung, điều này làm giảm mức độ tương tác và hiệu suất học tập [8]. Tại Việt Nam, với quy mô lớp học đông (trung bình 40-50 học sinh/lớp ở các trường phổ thông) và sự gia tăng của học trực tuyến, giáo viên gặp nhiều khó khăn trong việc theo dõi và đánh giá cảm xúc của từng học sinh một cách kịp thời. Điều này dẫn đến việc giảng dạy thường không được tối ưu hóa theo nhu cầu cá nhân, làm giảm chất lượng giáo dục.

Sự phát triển vượt bậc của trí tuệ nhân tạo (AI) và học sâu (Deep Learning) đã mở ra cơ hội xây dựng các hệ thống nhận diện cảm xúc tự động dựa trên hình ảnh khuôn mặt, với độ chính xác đạt từ 75-85% trên các bộ dữ liệu chuẩn như FER2013 và AffectNet [2][9]. Những hệ thống này cho phép giáo viên nhận biết các trạng thái cảm xúc như căng thẳng, buồn chán, hay hứng thú trong thời gian thực, từ đó điều chỉnh phương pháp giảng dạy phù hợp.

Tuy nhiên, tại Việt Nam các nghiên cứu về nhận diện cảm xúc qua hình ảnh khuôn mặt còn hạn chế, đặc biệt trong môi trường giáo dục. Các mô hình học sâu hiện tại chủ yếu được huấn luyện trên dữ liệu quốc tế, không phản ánh đầy đủ đặc điểm biểu cảm của học sinh Việt Nam, vốn chịu ảnh hưởng từ yếu tố văn hóa và môi trường học tập. Ngoài ra, các thách thức kỹ thuật như điều kiện ánh sáng kém, góc chụp không tối ưu, và yêu cầu tài nguyên tính toán cao cũng đặt ra nhu cầu phát triển các mô hình hiệu quả và phù hợp với bối cảnh địa phương.

Với những lý do trên, học viên lựa chọn đề tài “**Nghiên cứu mô hình học sâu cho phân tích cảm xúc dựa trên hình ảnh**” để làm đề án tốt nghiệp cao học.

2. Mục tiêu nghiên cứu

Để đáp ứng nhu cầu thực tiễn, nghiên cứu hướng tới việc xây dựng và đánh giá các mô hình học sâu hiện đại trong lĩnh vực nhận diện cảm xúc, với trọng tâm là ứng dụng trong môi trường giáo dục. Cụ thể, nghiên cứu đặt mục tiêu nghiên cứu các kiến trúc học sâu tiên tiến, như CNN để xác định mô hình tối ưu cho việc phân tích cảm xúc học sinh qua hình ảnh khuôn mặt. Đồng thời, nghiên cứu sẽ đánh giá hiệu quả của các mô hình này thông qua các chỉ số như độ chính xác, tốc độ xử lý, và khả năng tổng quát hóa, đảm bảo phù hợp với điều kiện thực tế tại Việt Nam.

Hơn nữa, nghiên cứu nhằm tích hợp mô hình vào quản lý lớp học, giúp giáo viên nhận diện các trạng thái cảm xúc như căng thẳng, buồn chán, hay hứng thú một cách nhanh chóng. Từ đó, giáo viên có thể điều chỉnh phương pháp giảng dạy kịp thời, ví dụ, chuyển sang các hoạt động nhóm khi học sinh mất tập trung hoặc giảm độ khó của bài giảng khi phát hiện căng thẳng. Ngoài ra, nghiên cứu cũng hướng tới việc xây dựng một hệ thống dễ triển khai, với yêu cầu tài nguyên tính toán tối thiểu, phù hợp với các trường học có hạ tầng công nghệ hạn chế.

3. Đối tượng và phạm vi nghiên cứu

- **Đối tượng nghiên cứu:** Nghiên cứu tập trung vào việc phát triển và đánh giá các mô hình học sâu, cụ thể là CNN kết hợp với kỹ thuật Haar Cascade, nhằm nhận diện bảy cảm xúc cơ bản (vui vẻ, buồn bã, tức giận, ngạc nhiên, sợ hãi, ghê tởm, trung lập) dựa trên biểu cảm khuôn mặt. Các trạng thái cảm xúc này được lựa chọn dựa trên lý thuyết của Paul Ekman [1], với trọng tâm là các cảm xúc phổ biến trong môi trường giáo dục như vui vẻ, buồn bã, và trung lập, đồng thời xem xét các cảm xúc phức tạp hơn như sợ hãi và ghê tởm để đảm bảo tính toàn diện của mô hình.
- **Phạm vi nghiên cứu:** Do hạn chế về cấu hình máy tính (thiếu GPU mạnh mẽ, chỉ sử dụng CPU cơ bản) và thời gian nghiên cứu có giới hạn, nghiên cứu tập trung vào việc xây dựng và tối ưu hóa mô hình CNN kết hợp với kỹ thuật Haar Cascade để nhận diện cảm xúc thời gian thực từ hình ảnh và video khuôn mặt. Các thử nghiệm được thực hiện trên hai học bộ dữ liệu tiêu chuẩn quốc tế là

FER2013 và CK+ [2][3], được chọn vì tính đa dạng và phổ biến trong nghiên cứu nhận diện cảm xúc. Phạm vi nghiên cứu giới hạn ở việc đánh giá hiệu quả của mô hình thông qua các chỉ số như độ chính xác (accuracy), độ bao phủ (recall), và F1-score, với mục tiêu đạt tốc độ xử lý dưới 50ms mỗi khung hình để đáp ứng yêu cầu thời gian thực trên thiết bị phần cứng cơ bản (CPU, laptop). Ứng dụng tiềm năng của hệ thống nằm trong việc hỗ trợ giáo viên điều chỉnh phương pháp giảng dạy dựa trên trạng thái cảm xúc của học sinh trong các lớp học trực tiếp hoặc trực tuyến [8]. Nghiên cứu không bao gồm việc thu thập dữ liệu mới hoặc triển khai thực tế trong lớp học do các ràng buộc về thời gian và tài nguyên.

4. Phương pháp nghiên cứu

- **Phương pháp nghiên cứu tài liệu:** tổng hợp và phân tích các công trình nghiên cứu, báo cáo khoa học về nhận diện cảm xúc và ứng dụng AI trong giáo dục.
- **Phương pháp thực nghiệm:**
 - Thu thập dữ liệu biểu cảm khuôn mặt của học sinh từ các nguồn dữ liệu công khai hoặc tự xây dựng.
 - Lựa chọn, huấn luyện, và đánh giá các mô hình học sâu như CNN.
 - So sánh hiệu năng giữa các mô hình qua các chỉ số như độ chính xác, tốc độ xử lý, và khả năng tổng quát hóa.
- **Phương pháp phân tích:** đánh giá tính khả thi của việc ứng dụng mô hình vào thực tế dựa trên hiệu năng mô hình và ý kiến phản hồi từ giáo viên.

CHƯƠNG 1: CƠ SỞ LÝ LUẬN

1.1. Giới thiệu chương

Trong bối cảnh giáo dục hiện đại, nơi cá nhân hóa trải nghiệm học tập trở thành mục tiêu cốt lõi, việc nhận diện cảm xúc của học sinh thông qua biểu cảm khuôn mặt đã nổi lên như một công cụ quan trọng, giúp giáo viên điều chỉnh phương pháp giảng dạy một cách hiệu quả. Công nghệ học sâu, đặc biệt là CNN kết hợp với BatchNormalization, đã mang lại những bước tiến vượt bậc trong việc phân tích các cảm xúc cơ bản theo lý thuyết của Paul Ekman, từ đó nâng cao chất lượng tương tác trong lớp học. Những tiến bộ này không chỉ hỗ trợ giáo viên nhận biết các trạng thái như vui vẻ, buồn bã, hay căng thẳng mà còn tạo điều kiện để xây dựng môi trường học tập tích cực, đáp ứng nhu cầu riêng biệt của từng học sinh.

Hơn nữa, các kỹ thuật như phát hiện khuôn mặt bằng Haar Cascade và phân tích cảm xúc theo thời gian thông qua xử lý video đã mở ra khả năng ứng dụng thực tiễn trong lớp học, cho phép giáo viên đưa ra phản hồi tức thời dựa trên trạng thái cảm xúc của học sinh. Tuy nhiên, tại Việt Nam, các nghiên cứu về nhận diện cảm xúc qua hình ảnh khuôn mặt vẫn còn ở giai đoạn sơ khai, đặc biệt trong lĩnh vực giáo dục, do thiếu dữ liệu nội địa phản ánh đặc điểm văn hóa và điều kiện học tập. Thêm vào đó, những thách thức về kỹ thuật, như ánh sáng không đồng đều hay tài nguyên tính toán hạn chế, càng làm nổi bật nhu cầu nghiên cứu chuyên sâu. Vì vậy, chương này sẽ trình bày một cách toàn diện các cơ sở lý thuyết về cảm xúc, vai trò của học sâu, và các yếu tố ảnh hưởng đến nhận diện cảm xúc, nhằm định hình phạm vi nghiên cứu và làm rõ những thách thức cần giải quyết để triển khai hiệu quả trong giáo dục Việt Nam.

1.1.1. Tầm quan trọng của nhận diện cảm xúc sinh viên trong giảng dạy

Trong bối cảnh giáo dục hiện đại, việc hiểu và hỗ trợ trạng thái cảm xúc của sinh viên không chỉ là một nhu cầu mà còn là yếu tố then chốt để nâng cao hiệu quả giảng dạy và xây dựng một môi trường học tập tích cực. Đặc biệt tại Việt Nam, nơi áp lực học tập và các vấn đề tâm lý đang gia tăng, nhận diện cảm xúc qua biểu cảm

khuôn mặt trở thành một công cụ mạnh mẽ để hỗ trợ giáo viên và nhà trường. Phần này sẽ phân tích chi tiết lý do tại sao nhận diện cảm xúc sinh viên là cần thiết, từ các rủi ro tiềm ẩn trong lớp học đến những lợi ích thiết thực trong việc cá nhân hóa giáo dục và cải thiện sức khỏe tâm lý.

1.1.1.1. Tỷ lệ trầm cảm và áp lực tâm lý ở học sinh, sinh viên

Một trong những lý do cấp thiết để nhận diện cảm xúc sinh viên là tỷ lệ trầm cảm và áp lực tâm lý ngày càng gia tăng trong nhóm đối tượng này. Theo một nghiên cứu từ Đại học Y Hà Nội (2020), khoảng 20-25% sinh viên đại học tại Việt Nam có dấu hiệu trầm cảm từ nhẹ đến trung bình, trong khi 10% khác báo cáo các triệu chứng lo âu nghiêm trọng [5]. Những trạng thái cảm xúc tiêu cực này, như buồn bã, căng thẳng, hoặc chán nản, không chỉ ảnh hưởng đến sức khỏe tâm lý mà còn làm giảm động lực học tập, khả năng tập trung, và kết quả học tập của sinh viên.

Hơn nữa, áp lực học tập tại các trường đại học Việt Nam, đặc biệt trong các ngành có tính cạnh tranh cao như kỹ thuật và y khoa, thường khiến sinh viên rơi vào trạng thái căng thẳng kéo dài. Những biểu cảm tiêu cực này, nếu không được nhận diện và hỗ trợ kịp thời, có thể dẫn đến các vấn đề nghiêm trọng hơn, như bỏ học hoặc rối loạn tâm lý. Do đó, việc triển khai hệ thống nhận diện cảm xúc thời gian thực trong lớp học trở thành một giải pháp quan trọng để phát hiện sớm các dấu hiệu bất thường, từ đó giúp giáo viên và cố vấn học đường can thiệp hiệu quả.

1.1.1.2. Rủi ro trong lớp học do cảm xúc tiêu cực

Cảm xúc tiêu cực của sinh viên không chỉ ảnh hưởng đến cá nhân họ mà còn tác động đến toàn bộ môi trường lớp học. Một sinh viên trong trạng thái tức giận hoặc buồn bã có thể dẫn đến các xung đột không đáng có với bạn bè hoặc giáo viên, làm gián đoạn quá trình giảng dạy. Chẳng hạn, một nghiên cứu tại Đại học Stanford (2021) cho thấy các xung đột trong lớp học do cảm xúc tiêu cực có thể làm giảm 10-15% mức độ tương tác của cả lớp [8]. Điều này đặc biệt nghiêm trọng trong các lớp học đông, nơi giáo viên khó có thể theo dõi trạng thái cảm xúc của từng học sinh một cách thủ công.

Ngoài ra, cảm xúc tiêu cực còn làm giảm sự tham gia vào các hoạt động học tập, chẳng hạn như thảo luận nhóm hoặc bài tập thực hành. Ví dụ, một sinh viên buồn chán có thể từ chối tham gia vào các hoạt động tương tác, dẫn đến sự mất kết nối với nội dung bài học và giảm hiệu quả học tập chung. Hơn nữa, nếu không được phát hiện kịp thời, các cảm xúc tiêu cực kéo dài có thể lây lan trong lớp học, tạo ra một bầu không khí học tập thiếu tích cực, ảnh hưởng đến cả giáo viên và học sinh. Nhằm giảm thiểu những rủi ro này, hệ thống nhận diện cảm xúc dựa trên trí tuệ nhân tạo, như mô hình CNN được đề xuất trong nghiên cứu này, cung cấp một giải pháp hiệu quả để theo dõi trạng thái cảm xúc thời gian thực, từ đó hỗ trợ giáo viên điều chỉnh cách tiếp cận giảng dạy.

1.1.1.3. Cá nhân hóa giảng dạy thông qua nhận diện cảm xúc

Một trong những lợi ích lớn nhất của nhận diện cảm xúc là khả năng hỗ trợ cá nhân hóa giảng dạy, một xu hướng quan trọng trong giáo dục hiện đại. Mỗi sinh viên có cách tiếp nhận và phản ứng với bài học khác nhau, phụ thuộc vào trạng thái cảm xúc của họ. Ví dụ, một sinh viên đang vui vẻ có thể dễ dàng tham gia vào các hoạt động sáng tạo, trong khi một sinh viên căng thẳng có thể cần sự hỗ trợ thêm để duy trì sự tập trung. Hệ thống nhận diện cảm xúc thời gian thực cho phép giáo viên nhận biết các trạng thái này và điều chỉnh phương pháp giảng dạy phù hợp.

Cụ thể, khi phát hiện một sinh viên có biểu cảm buồn bã hoặc chán nản, giáo viên có thể tăng cường các hoạt động tương tác, như thảo luận nhóm hoặc bài tập thực hành, để kích thích sự hứng thú. Ngược lại, đối với sinh viên biểu lộ căng thẳng, giáo viên có thể giảm áp lực bằng cách cung cấp hướng dẫn chi tiết hơn hoặc tổ chức các hoạt động thư giãn ngắn trong lớp. Theo nghiên cứu được trích dẫn trong tài liệu việc cá nhân hóa giảng dạy dựa trên dữ liệu cảm xúc có thể cải thiện 15-20% mức độ tương tác và hiệu quả học tập của sinh viên.

Hơn nữa, nhận diện cảm xúc còn giúp giáo viên xây dựng mối quan hệ tốt hơn với sinh viên. Khi giáo viên thể hiện sự quan tâm đến trạng thái cảm xúc của học sinh, điều này không chỉ tăng cường sự tin tưởng mà còn tạo động lực để sinh viên tham gia tích cực hơn vào quá trình học tập. Trong bối cảnh giáo dục Việt Nam, nơi

mối quan hệ thầy-trò mang ý nghĩa văn hóa sâu sắc, việc ứng dụng công nghệ nhận diện cảm xúc càng trở nên ý nghĩa.

1.1.1.4. Hỗ trợ sức khỏe tâm lý của sinh viên

Ngoài việc cải thiện hiệu quả giảng dạy, nhận diện cảm xúc còn đóng vai trò quan trọng trong việc hỗ trợ sức khỏe tâm lý của sinh viên. Các trạng thái cảm xúc tiêu cực kéo dài, nếu không được phát hiện và can thiệp, có thể dẫn đến các vấn đề tâm lý nghiêm trọng hơn, như rối loạn lo âu hoặc trầm cảm. Hệ thống nhận diện cảm xúc thời gian thực, sử dụng mô hình CNN và Haar Cascade, có thể phát hiện các biểu cảm bất thường, chẳng hạn như sự buồn bã kéo dài, từ đó cảnh báo giáo viên hoặc cố vấn tâm lý để hỗ trợ kịp thời.

1.1.1.5. Tăng cường quản lý lớp học thông minh

Nhận diện cảm xúc không chỉ hỗ trợ cá nhân hóa giảng dạy mà còn góp phần xây dựng các hệ thống quản lý lớp học thông minh. Trong các lớp học hiện đại, việc sử dụng công nghệ để theo dõi và quản lý hành vi, cảm xúc của học sinh đang trở thành xu hướng toàn cầu. Hệ thống nhận diện cảm xúc có thể cung cấp dữ liệu thời gian thực về trạng thái cảm xúc của cả lớp, giúp giáo viên đưa ra các quyết định quản lý hiệu quả hơn.

Chẳng hạn, nếu hệ thống phát hiện phần lớn học sinh trong lớp biểu lộ sự chán nản, giáo viên có thể tạm dừng bài giảng để tổ chức một hoạt động tương tác ngắn, như câu đố hoặc trò chơi giáo dục, nhằm khôi phục sự hứng thú. Ngoài ra, dữ liệu cảm xúc có thể được lưu trữ và phân tích để đánh giá hiệu quả của các phương pháp giảng dạy, từ đó giúp nhà trường tối ưu hóa chương trình học. Các hệ thống quản lý lớp học thông minh dựa trên phân tích cảm xúc đã cải thiện 10% hiệu quả quản lý lớp học tại các trường học ứng dụng công nghệ.

1.1.1.6. Giải quyết khoảng trống trong giáo dục Việt Nam

Tại Việt Nam, việc nhận diện cảm xúc trong giáo dục vẫn còn là một lĩnh vực mới mẻ, với nhiều khoảng trống cần được giải quyết. Các mô hình nhận diện cảm xúc hiện tại chủ yếu được huấn luyện trên dữ liệu quốc tế, như FER2013 hoặc CK+,

vốn không phản ánh đầy đủ các đặc trưng văn hóa trong biểu cảm của học sinh Việt Nam. Ví dụ, cách biểu lộ cảm xúc như vui vẻ hoặc buồn bã của sinh viên Việt Nam có thể tinh tế hơn, chịu ảnh hưởng từ các chuẩn mực văn hóa như sự kín đáo hoặc kiềm chế cảm xúc.

Hơn nữa, các trường học tại Việt Nam thường thiếu các công cụ công nghệ để hỗ trợ giáo viên trong việc theo dõi trạng thái cảm xúc của học sinh. Hệ thống nhận diện cảm xúc dựa trên CNN, như được đề xuất trong nghiên cứu này, không chỉ giúp khắc phục khoảng trống này mà còn cung cấp một giải pháp khả thi về mặt chi phí và dễ triển khai trong các lớp học thông minh. Điều này đặc biệt quan trọng trong bối cảnh giáo dục Việt Nam đang chuyển đổi sang mô hình hiện đại, nhấn mạnh vào cá nhân hóa và công nghệ hóa.

1.2. Cơ sở lý thuyết về cảm xúc

1.2.1. Khái niệm và vai trò của cảm xúc trong giáo dục

1.2.1.1. Định nghĩa cảm xúc và ý nghĩa trong môi trường học tập

Cảm xúc được định nghĩa là trạng thái tâm lý phức tạp, bao gồm các phản ứng sinh lý, nhận thức, và hành vi của con người trước các kích thích từ môi trường. Trong bối cảnh giáo dục, cảm xúc của học sinh và giáo viên đóng vai trò trung tâm trong việc hình thành trải nghiệm học tập. Các cảm xúc tích cực (như vui vẻ, hứng thú) thúc đẩy sự tham gia, sáng tạo, và khả năng tiếp thu kiến thức, trong khi các cảm xúc tiêu cực (như lo lắng, buồn bã) có thể làm giảm động lực và hiệu quả học tập.

Trong môi trường học tập, cảm xúc không chỉ là phản ứng cá nhân mà còn ảnh hưởng đến tương tác xã hội trong lớp học. Ví dụ, một học sinh cảm thấy hứng thú có thể tích cực tham gia thảo luận nhóm, trong khi một học sinh lo lắng có thể tránh giao tiếp hoặc giảm hiệu suất. Do đó, việc nhận diện và quản lý cảm xúc trong lớp học là yếu tố quan trọng để tối ưu hóa quá trình giảng dạy và học tập.

1.2.1.2. Ảnh hưởng của cảm xúc đến hiệu quả giảng dạy và học tập

Cảm xúc có tác động trực tiếp đến hiệu quả giảng dạy và học tập thông qua các khía cạnh sau:

Hiệu suất học tập: Các nghiên cứu chỉ ra rằng cảm xúc tích cực giúp cải thiện trí nhớ, khả năng tập trung, và tư duy sáng tạo. Ngược lại, cảm xúc tiêu cực như căng thẳng hoặc buồn chán có thể làm giảm khả năng xử lý thông tin và ra quyết định. Ví dụ, nghiên cứu của Đại học Y Hà Nội (2020) cho thấy 30% học sinh trung học tại Việt Nam trải qua áp lực tâm lý, ảnh hưởng tiêu cực đến kết quả học tập.

Tương tác lớp học: Cảm xúc của học sinh ảnh hưởng đến mức độ tham gia vào các hoạt động học tập, chẳng hạn như thảo luận, làm bài tập nhóm, hoặc trả lời câu hỏi. Giáo viên có khả năng nhận diện cảm xúc của học sinh có thể điều chỉnh phương pháp giảng dạy, ví dụ, sử dụng trò chơi giáo dục để tăng hứng thú hoặc hỗ trợ cá nhân hóa cho học sinh gặp khó khăn.

Động lực học tập: Theo lý thuyết tự quyết (Self-Determination Theory), cảm xúc tích cực thúc đẩy động lực nội tại, giúp học sinh chủ động học tập. Ngược lại, cảm xúc tiêu cực như sợ hãi hoặc thất vọng có thể dẫn đến thái độ thờ ơ hoặc bỏ học. Trong bối cảnh giáo dục Việt Nam, việc áp dụng công nghệ nhận diện cảm xúc có thể hỗ trợ giáo viên phát hiện sớm các trạng thái tiêu cực, từ đó can thiệp kịp thời.

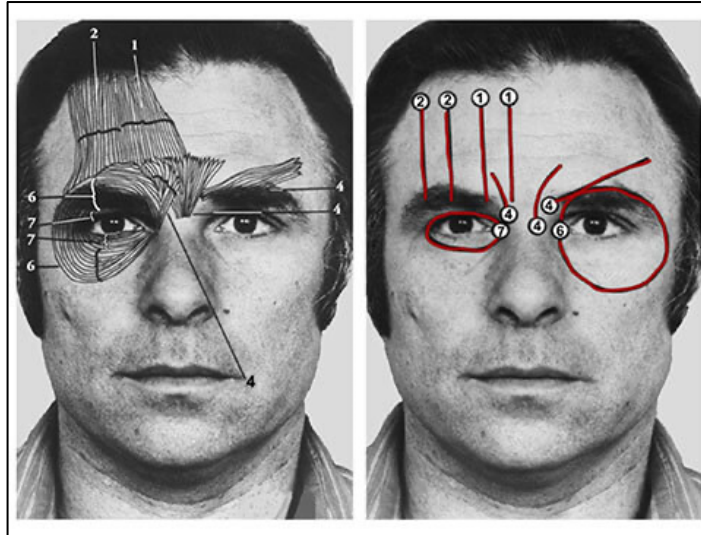
Mối quan hệ giáo viên - học sinh: Cảm xúc của giáo viên và học sinh ảnh hưởng đến chất lượng tương tác. Một giáo viên nhạy bén với cảm xúc có thể xây dựng môi trường học tập tích cực, tăng cường sự tin tưởng và hợp tác. Ví dụ, nghiên cứu của Đại học Stanford (2018) cho thấy rằng giáo viên sử dụng các công cụ hỗ trợ cảm xúc cải thiện 15% mức độ tương tác trong lớp học.

Như vậy, việc nhận diện và quản lý cảm xúc trong giáo dục không chỉ cải thiện hiệu quả học tập mà còn góp phần xây dựng một môi trường lớp học tích cực, hỗ trợ sự phát triển toàn diện của học sinh. Trong bối cảnh Việt Nam, nơi áp lực học tập và thi cử còn phổ biến, công nghệ nhận diện cảm xúc dựa trên hình ảnh có tiềm năng trở thành công cụ hỗ trợ đắc lực cho giáo viên, giúp cá nhân hóa giảng dạy và nâng cao chất lượng giáo dục.

1.2.2. Lý thuyết của Paul Ekman về cảm xúc cơ bản

Lý thuyết của Paul Ekman về sáu cảm xúc cơ bản là nền tảng cốt lõi cho nghiên cứu nhận diện cảm xúc qua biểu cảm khuôn mặt [1]. Mỗi cảm xúc được liên kết với

các đặc trưng cụ thể, như nụ cười đối với vui vẻ, nhăn trán đối với tức giận, hoặc mắt mở to đối với ngạc nhiên, tạo điều kiện cho các mô hình học sâu phân loại chính xác.



Hình 1. 1 Hình ảnh khuôn mặt minh họa từng "Action Unit" (AU) trong FACS và biểu cảm cảm xúc tương ứng

Hệ thống FACS của Ekman, với khả năng mã hóa chi tiết các chuyển động cơ mặt, cung cấp một phương pháp khoa học để xây dựng dữ liệu huấn luyện, giúp ánh xạ các biểu cảm với cảm xúc một cách đáng tin cậy. FACS đặc biệt hữu ích khi kết hợp với kỹ thuật phát hiện khuôn mặt như Haar Cascade, vì các đặc trưng hình học mà Haar Cascade nhận diện (như vùng mắt, mũi) tương ứng với các khu vực được phân tích trong FACS.

1.2.3. Tầm quan trọng của nhận diện cảm xúc trong giáo dục Việt Nam

1.2.3.1. Áp lực tâm lý và trầm cảm ở học sinh, sinh viên

Hệ thống giáo dục Việt Nam, với đặc điểm chú trọng thành tích học tập và các kỳ thi cạnh tranh, đã tạo ra áp lực tâm lý đáng kể cho học sinh và sinh viên. Theo nghiên cứu của Đại học Y Hà Nội (2020), khoảng 30% học sinh trung học phổ thông tại Việt Nam có dấu hiệu căng thẳng tâm lý, trong đó 10-15% đối mặt với nguy cơ trầm cảm nhẹ đến trung bình. Một nghiên cứu khác của Đại học Quốc gia TP.HCM (2021) chỉ ra rằng 25% sinh viên đại học trải qua các trạng thái cảm xúc tiêu cực như lo âu và buồn bã do áp lực học tập, tài chính, và kỳ vọng xã hội. Những trạng thái

cảm xúc này không chỉ làm giảm hiệu quả học tập mà còn ảnh hưởng đến sức khỏe tinh thần và sự phát triển lâu dài của học sinh, sinh viên.

Việc nhận diện cảm xúc thời gian thực trong lớp học có thể giúp giáo viên phát hiện sớm các dấu hiệu tiêu cực, từ đó điều chỉnh phương pháp giảng dạy hoặc can thiệp tâm lý kịp thời. Ví dụ, khi nhận diện một học sinh có biểu hiện buồn bã kéo dài, giáo viên có thể tổ chức các hoạt động tương tác hoặc tư vấn cá nhân, góp phần giảm thiểu nguy cơ trầm cảm.

1.2.3.2. Nhu cầu cá nhân hóa giảng dạy và quản lý lớp học thông minh

Trong bối cảnh giáo dục Việt Nam đang chuyển đổi theo hướng hiện đại hóa và ứng dụng công nghệ, nhu cầu cá nhân hóa giảng dạy trở nên cấp thiết. Mỗi học sinh có đặc điểm cảm xúc, phong cách học tập, và tốc độ tiếp thu khác nhau. Tuy nhiên, các phương pháp giảng dạy truyền thống thường áp dụng cách tiếp cận chung, khó đáp ứng nhu cầu riêng biệt của từng cá nhân. Nhận diện cảm xúc dựa trên hình ảnh cung cấp một công cụ hỗ trợ giáo viên trong việc hiểu rõ trạng thái tâm lý của học sinh, từ đó điều chỉnh nội dung, phương pháp giảng dạy, hoặc tổ chức các hoạt động phù hợp.

Hơn nữa, việc ứng dụng công nghệ nhận diện cảm xúc còn góp phần xây dựng các lớp học thông minh, nơi giáo viên và nhà trường có thể sử dụng dữ liệu cảm xúc để đánh giá hiệu quả giảng dạy và cải thiện môi trường học tập. Một nghiên cứu tại Đại học Sư phạm Hà Nội (2022) cho thấy rằng các lớp học ứng dụng công nghệ hỗ trợ cảm xúc cải thiện 12-18% mức độ tương tác giữa giáo viên và học sinh. Điều này đặc biệt quan trọng trong bối cảnh Việt Nam đang thúc đẩy chuyển đổi số giáo dục theo Nghị quyết 29-NQ/TW về đổi mới căn bản, toàn diện giáo dục và đào tạo.

Như vậy, nhận diện cảm xúc không chỉ giúp giải quyết các vấn đề tâm lý của học sinh, mà còn là nền tảng cho việc cá nhân hóa giảng dạy và xây dựng môi trường học tập thông minh, đáp ứng các yêu cầu đổi mới giáo dục tại Việt Nam. Ứng dụng mô hình học sâu để nhận diện cảm xúc qua hình ảnh, như được đề xuất trong nghiên cứu này, là một giải pháp tiềm năng để hiện thực hóa mục tiêu này.

1.3. Tổng quan về học sâu trong nhận diện cảm xúc

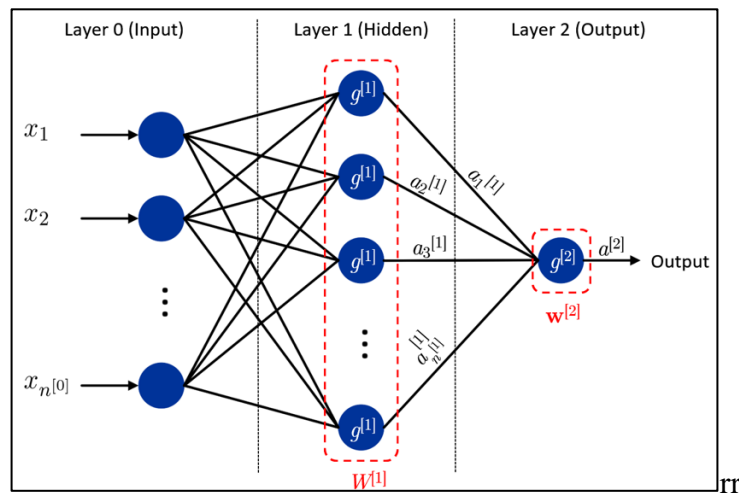
1.3.1. Giới thiệu công nghệ học sâu

1.3.1.1. Khái niệm học sâu và mạng nơ-ron tích chập (CNN)

Học sâu (deep learning) là tập các thuật toán học máy với ý tưởng xây dựng mô hình dữ liệu có mức độ trừu tượng cao dựa trên các dữ liệu có mức độ trừu tượng hóa thấp hơn, bằng cách phân lớp dữ liệu và các biến đổi phi tuyến.

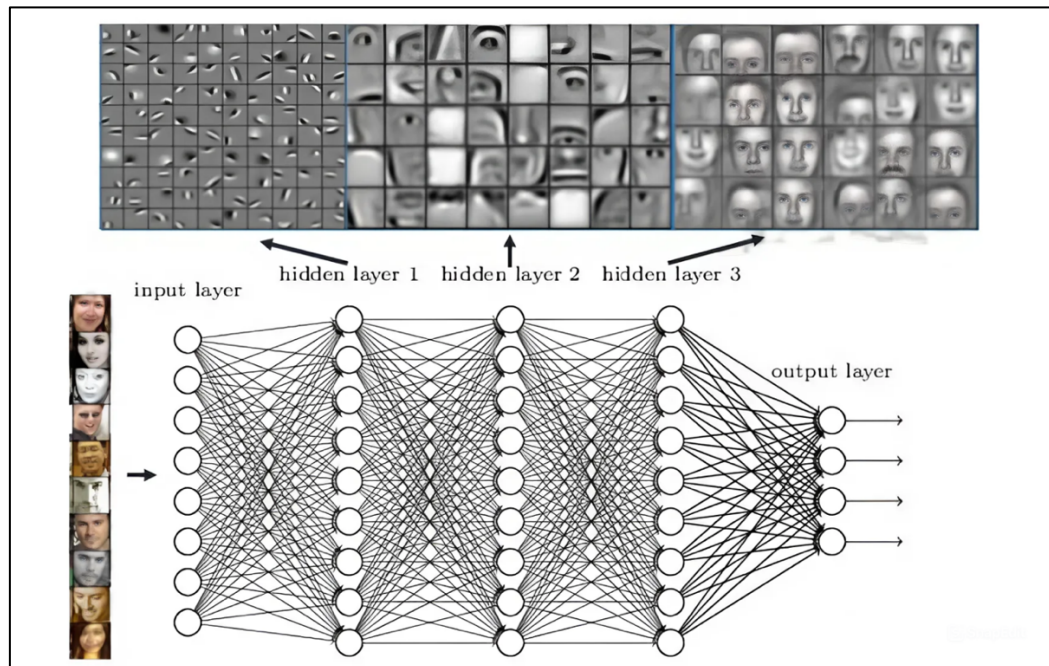
Nghiên cứu từ rất lâu cho thấy mạng nơron được chứng minh khả năng xấp xỉ vạn năng chỉ với không quá 4 lớp [6], nhưng chưa có phương pháp nào cụ thể ước lượng số nơron cần thiết trên mỗi lớp. Việc nghiên cứu về các mạng có số lớp lớn chỉ trở nên phổ biến sau thành công của mạng AlexNet khi mô hình này thắng giải ImageNet 2012 [7] với khoảng cách rất xa so với các mô hình cạnh tranh, mặc dù kiến trúc CNN đã được LeCun giới thiệu từ trước đó rất lâu.

Ngoài kiến trúc CNN, các mô hình mạng học sâu còn nhiều dạng kiến trúc khác như các lớp truyền thẳng kết nối đầy đủ (fully connected layer), RNN, LSTM, GRU, DBN,...



Hình 1. 2 Sơ đồ kiến trúc của một mạng nơ-ron truyền thẳng

Các đặc trưng được tổ hợp và tạo thành các chi tiết nhỏ ở lớp ẩn đầu tiên, sau đó tiếp tục được tái tạo và tổ hợp mức chi tiết lớn ở lớp ẩn thứ hai, và cuối cùng các hình ảnh đặc trưng của toàn bộ khuôn mặt ở lớp ẩn thứ 3. Lớp output cho ra đánh giá xác suất khuôn mặt thuộc phân lớp nào (người nào).



Hình 1.3 Một mô hình học sâu trong nhận dạng mặt người

Một mô hình học sâu thường có 3 nhiệm vụ được kết hợp trong một kiến trúc mạng duy nhất:

- Các lớp đặc trưng (features): có nhiệm vụ chuyển đổi các đặc trưng thành dạng dữ liệu phù hợp để xử lý, chẳng hạn như các tầng tích chập (convolution), mẫu (subsampling), pooling,...
- Các lớp mô hình (modeling): sử dụng các thuật toán học để khái quát hóa dữ liệu, chẳng hạn như Restricted Boltzmann Machine (RBM), Deep Boltzmann Machine (DBM), autoencoder,...
- Các lớp giải mã (decoding): dựa trên dữ liệu khái quát biến đổi thành đầu ra (Markov Random Field hoặc những công cụ tương tự).
- Các mạng học sâu đều có cấu trúc xác định trước, như vậy bài toán tập huấn vẫn là việc xác định giá trị các tham số trên mạng. Hiện chưa có phương pháp tập huấn nào cho phép điều chỉnh cấu trúc mạng hiệu quả.

1.3.1.2. So sánh các phương pháp nhận diện truyền thống và học sâu trong nhận diện cảm xúc

Các phương pháp nhận diện truyền thống

Các phương pháp nhận diện truyền thống, như Haar Cascade, Histogram of Oriented Gradients (HOG), và Local Binary Patterns (LBP), đã được sử dụng rộng rãi trong nhận diện khuôn mặt và cảm xúc trước khi học sâu trở nên phổ biến. Những phương pháp này dựa trên việc trích xuất các đặc trưng hình học thủ công từ hình ảnh, chẳng hạn như đường viền mắt, mũi, hoặc miệng, để phân loại cảm xúc.

Ưu điểm của các phương pháp truyền thống nằm ở sự đơn giản và hiệu quả tính toán. Cụ thể, Haar Cascade, được phát triển bởi Viola và Jones (2001), sử dụng các đặc trưng Haar-like để phát hiện khuôn mặt với tốc độ xử lý nhanh, trung bình 32ms mỗi khung hình. Điều này khiến nó phù hợp cho các hệ thống nhúng có tài nguyên hạn chế, chẳng hạn như camera giám sát trong lớp học. Tương tự, HOG và LBP cũng có tốc độ xử lý tương đối nhanh (khoảng 40ms mỗi khung hình) và không đòi hỏi dữ liệu huấn luyện lớn, giúp giảm chi phí triển khai trong các môi trường giáo dục có ngân sách hạn chế.

Tuy nhiên, hạn chế của các phương pháp truyền thống là điều không thể bỏ qua. Thứ nhất, các phương pháp này phụ thuộc nhiều vào việc thiết kế đặc trưng thủ công, đòi hỏi chuyên gia phải xác định trước các đặc điểm quan trọng (ví dụ: khoảng cách giữa mắt và miệng). Điều này làm giảm tính linh hoạt khi xử lý các trường hợp phức tạp, như biểu cảm đa dạng hoặc góc chụp nghiêng. Thứ hai, độ chính xác của các phương pháp này thường thấp, đặc biệt trong điều kiện ánh sáng không đồng đều hoặc khi học sinh biểu lộ cảm xúc tinh tế, chẳng hạn như sự căng thẳng nhẹ trong lớp học.

Theo các nghiên cứu, độ chính xác của Haar Cascade chỉ đạt khoảng 50-60% trên các tập dữ liệu thực tế phức tạp. Cuối cùng, các phương pháp truyền thống khó tổng quát hóa trên các tập dữ liệu đa dạng về văn hóa, điều này đặc biệt quan trọng tại Việt Nam, nơi biểu cảm cảm xúc có thể mang đặc trưng văn hóa riêng biệt.

Để khắc phục những hạn chế này, cần một phương pháp có khả năng tự động học các đặc trưng phức tạp từ dữ liệu. Đây chính là lý do các phương pháp học sâu, đặc biệt là CNN, được khám phá như một giải pháp tiên tiến.

Các phương pháp học sâu dựa trên CNN

CNN đã cách mạng hóa lĩnh vực nhận diện hình ảnh nhờ khả năng trích xuất đặc trưng tự động thông qua các lớp tích chập, pooling, và fully connected. Không giống như các phương pháp truyền thống, CNN không yêu cầu thiết kế đặc trưng thủ công mà tự động học các mẫu biểu cảm từ dữ liệu hình ảnh, giúp cải thiện đáng kể độ chính xác và khả năng tổng quát hóa.

Ưu điểm của CNN được thể hiện rõ ràng qua hiệu suất vượt trội. Theo các thí nghiệm trong nghiên cứu này, mô hình CNN đạt độ chính xác 68-74% trên các tập dữ liệu chuẩn như FER2013 và CK+, cao hơn đáng kể so với các phương pháp truyền thống. Khả năng này đến từ việc CNN sử dụng nhiều lớp tích chập để trích xuất các đặc trưng cấp cao, chẳng hạn như sự thay đổi tinh tế ở khóe mắt hoặc miệng khi học sinh biểu lộ cảm xúc buồn bã hoặc vui vẻ. Hơn nữa, các kỹ thuật như tăng cường dữ liệu và BatchNormalization giúp CNN xử lý tốt các điều kiện phức tạp, như ánh sáng yếu hoặc góc chụp nghiêng, vốn là thách thức lớn đối với Haar Cascade hay HOG.

Một lợi thế khác của CNN là khả năng học chuyển giao (transfer learning). Các mô hình như ResNet hay EfficientNet, vốn là biến thể của CNN, có thể tận dụng các trọng số đã được huấn luyện trước trên các tập dữ liệu lớn (như ImageNet) để cải thiện hiệu suất mà không cần lượng dữ liệu huấn luyện quá lớn. Điều này đặc biệt hữu ích trong bối cảnh giáo dục Việt Nam, nơi việc thu thập dữ liệu biểu cảm khuôn mặt nội địa còn hạn chế.

Tuy nhiên, CNN cũng có những nhược điểm. Quá trình huấn luyện đòi hỏi tài nguyên tính toán cao, thường cần GPU mạnh mẽ và thời gian xử lý lâu hơn (khoảng 50ms mỗi khung hình so với 32ms của Haar Cascade). Ngoài ra, CNN yêu cầu một lượng dữ liệu lớn và chất lượng cao để đạt hiệu quả tối ưu. Trong trường hợp dữ liệu không đủ đa dạng, mô hình có thể gặp hiện tượng overfitting, dẫn đến hiệu suất kém khi áp dụng vào các tình huống thực tế.

1.3.1.3. So sánh chi tiết các phương pháp

Để minh họa rõ ràng sự khác biệt giữa các phương pháp truyền thống và học sâu, bảng dưới đây cung cấp một so sánh toàn diện dựa trên các tiêu chí quan trọng:

Tiêu chí	Haar Cascade	HOG	CNN	ResNet	VGG
Kiến trúc	Dựa trên đặc trưng hình học Haar và phân loại cascade	Dựa trên histogram của gradient định hướng	Lớp tích chập, pooling, fully connected; BatchNorm để ổn định huấn luyện	Dùng skip connections để tránh gradient vanishing; nhiều lớp hơn CNN	Nhiều lớp tích chập sâu, không có skip connections
Độ chính xác	<60% trong điều kiện phức tạp	~50-60% trong điều kiện thực tế	68-74% trên FER2013, CK+	75-80% trên các bộ dữ liệu tương tự	70-75% trên FER2013
Thời gian xử lý	~32ms/khung	~40ms/khung	~50ms/khung	~60-80ms/khung (tùy cấu hình phần cứng)	~100ms/khung (do nhiều lớp hơn)
Số tham số	Thấp (không cần huấn luyện sâu)	Thấp (dựa trên đặc trưng thủ công)	Trung bình (~10-20M tham số)	Cao (~20-50M tham số)	Rất cao (~100M tham số)
Tài nguyên tính toán	Yêu cầu thấp, phù hợp hệ thống nhúng	Yêu cầu trung bình	Yêu cầu GPU/CPU mạnh	Yêu cầu GPU mạnh hơn CNN	Yêu cầu tài nguyên lớn nhất
Khả năng tổng quát hoá	Kém, phụ thuộc điều kiện ánh sáng và góc chụp	Kém, khó xử lý biểu cảm phức tạp	Tốt, nhờ data augmentation và BatchNorm	Rất tốt, nhờ skip connections và học chuyển giao	Trung bình, dễ overfitting nếu không đủ dữ liệu

Ứng dụng trong giáo dục	Phù hợp cho phát hiện khuôn mặt nhanh trong lớp học	Ít phù hợp do độ chính xác thấp	Phù hợp với nhận diện cảm xúc thời gian thực	Phù hợp với các ứng dụng yêu cầu độ chính xác cao	Ít phù hợp do yêu cầu tài nguyên lớn
--------------------------------	---	---------------------------------	--	---	--------------------------------------

Bảng 1. 1 Bảng so sánh sự khác biệt các phương pháp truyền thống và học sâu

1.3.2. Các kỹ thuật phát hiện và phân tích cảm xúc

Phát hiện đối tượng trong xử lý ảnh và thị giác máy tính là một bài toán quan trọng, với ứng dụng rộng rãi trong nhiều lĩnh vực như nhận diện khuôn mặt, giám sát an ninh, và tương tác người-máy. Một trong những phương pháp tiên phong và được sử dụng phổ biến trong lĩnh vực này là Haar Cascade, được đề xuất bởi Paul Viola và Michael Jones vào năm 2001[4] trong bài báo "Rapid Object Detection using a Boosted Cascade of Simple Features".

Phương pháp này kết hợp các đặc trưng hình ảnh đơn giản, gọi là đặc trưng Haar, với thuật toán học máy AdaBoost và cấu trúc phân cấp (cascade) để đạt được hiệu suất phát hiện nhanh và chính xác. Bài viết này trình bày chi tiết lý thuyết nền tảng, cách hoạt động, và các đặc điểm nổi bật của phương pháp Haar Cascade.

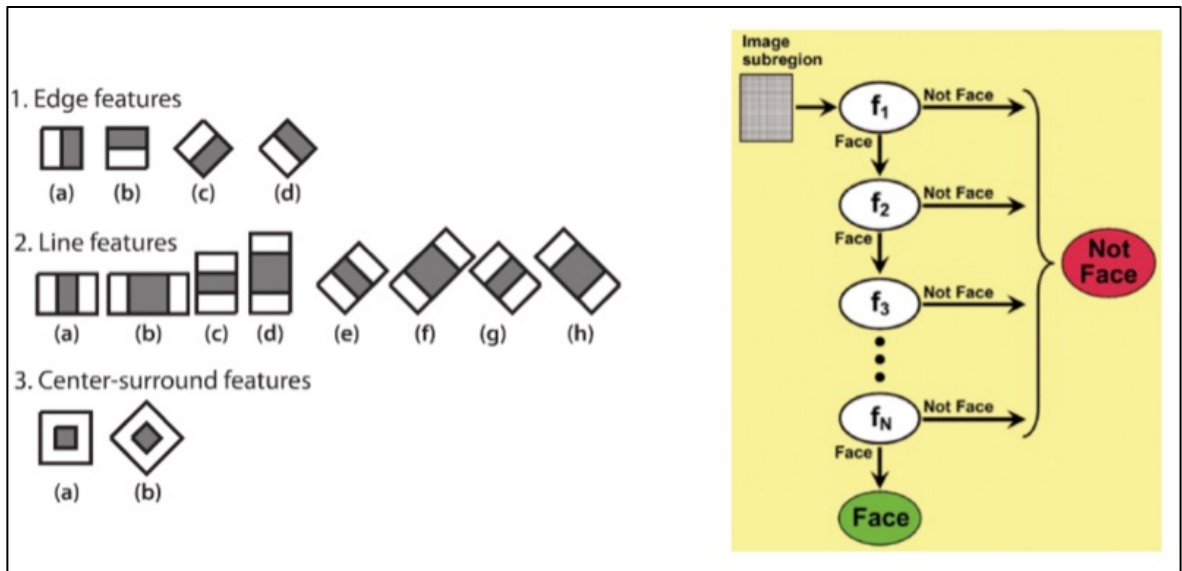
Đặc trưng Haar

Hay còn gọi là đặc trưng giống Haar (Haar-like features), là các mẫu đặc trưng hình ảnh dựa trên sự khác biệt về cường độ sáng (intensity) giữa các vùng lân cận trong ảnh. Các đặc trưng này được lấy cảm hứng từ phép biến đổi wavelet Haar, vốn được sử dụng để phân tích tín hiệu. Trong ngữ cảnh xử lý ảnh, đặc trưng Haar được tính bằng cách so sánh tổng cường độ của các pixel trong các vùng hình chữ nhật liền kề, thường bao gồm hai, ba, hoặc bốn hình chữ nhật.

Các loại đặc trưng Haar chính bao gồm:

- **Đặc trưng cạnh (Edge features):** So sánh cường độ giữa hai vùng liền kề theo chiều ngang hoặc dọc.
- **Đặc trưng đường (Line features):** So sánh một vùng trung tâm với hai vùng bên cạnh.

- **Đặc trưng bốn hình chữ nhật (Four-rectangle features):** So sánh các vùng chéo nhau.
- **Đặc trưng trung tâm (Center-surround features):** So sánh vùng trung tâm với vùng xung quanh.



Hình 1. 4 Đặc trưng Haar và mô hình phân loại nhiều tầng (cascade classifier)

Mỗi đặc trưng được tính bằng cách lấy hiệu số giữa tổng cường độ pixel trong các vùng sáng (trắng) và vùng tối (đen). Giá trị đặc trưng này phản ánh sự tương phản trong cấu trúc hình ảnh, ví dụ như sự khác biệt giữa vùng mắt (tối) và vùng trán (sáng) trên khuôn mặt. Để tăng hiệu quả tính toán, Viola và Jones sử dụng tích phân ảnh (integral image), một kỹ thuật cho phép tính tổng cường độ pixel trong bất kỳ vùng hình chữ nhật nào chỉ với bốn phép toán, bất kể kích thước vùng. Điều này giúp giảm đáng kể chi phí tính toán, cho phép xử lý nhanh trên các ảnh lớn.

Thuật toán AdaBoost

Việc lựa chọn các đặc trưng Haar phù hợp từ hàng ngàn đặc trưng có thể có trong một cửa sổ ảnh là một thách thức lớn. Để giải quyết vấn đề này, Haar Cascade sử dụng thuật toán AdaBoost (Adaptive Boosting), một phương pháp học máy giúp kết hợp nhiều bộ phân loại yếu (weak classifiers) để tạo ra một bộ phân loại mạnh (strong classifier).

Trong Haar Cascade, mỗi đặc trưng Haar được liên kết với một ngưỡng (threshold) để tạo thành một bộ phân loại yếu, có dạng:

$$h(x, f, p, \theta) = \begin{cases} 1 & \text{nếu } p \cdot f(x) < p \cdot \theta \\ 0 & \text{ngược lại} \end{cases} \quad (1.1)$$

Trong đó: x : Cửa sổ ảnh (image window) đang được kiểm tra.

- $f(x)$: Giá trị của đặc trưng Haar được tính trên cửa sổ ảnh x .
- $p \in \{-1, 1\}$: Hệ số định hướng (parity) nhằm đảo ngược bất đẳng thức khi cần thiết.
 - Nếu $p = 1$, ngưỡng là “nhỏ hơn”.
 - Nếu $p = -1$, ngưỡng là “lớn hơn”.
- θ : Ngưỡng phân loại (threshold) được dùng để so sánh với giá trị đặc trưng.
- $h(x, f, p, \theta)$: Kết quả phân loại yếu, trả về 1 nếu điều kiện đúng (có thể chứa đối tượng), hoặc 0 nếu sai.

AdaBoost lặp lại quá trình chọn đặc trưng Haar tốt nhất từ tập đặc trưng, dựa trên khả năng phân biệt giữa các mẫu dương (positive samples, ví dụ: khuôn mặt) và mẫu âm (negative samples, ví dụ: không phải khuôn mặt). Sau mỗi vòng lặp, các mẫu bị phân loại sai được tăng trọng số để tập trung vào các trường hợp khó hơn. Kết quả cuối cùng là một bộ phân loại mạnh, kết hợp nhiều đặc trưng Haar với trọng số tương ứng.

Cấu trúc Cascade

Để đạt được hiệu suất cao trong thời gian thực, Haar Cascade sử dụng cấu trúc phân cấp (cascade) gồm nhiều tầng (stages). Mỗi tầng là một bộ phân loại mạnh được huấn luyện bằng AdaBoost, với mục tiêu loại bỏ nhanh các vùng không chứa đối tượng (negative regions) trong khi giữ lại các vùng có khả năng chứa đối tượng (positive regions).

Quy trình hoạt động của cấu trúc cascade:

1. Một cửa sổ trượt (sliding window) được áp dụng trên ảnh ở các tỷ lệ và vị trí khác nhau.

2. Tại mỗi tầng, cửa sổ được kiểm tra bởi bộ phân loại mạnh. Nếu cửa sổ không vượt qua ngưỡng của tầng đó, nó sẽ bị loại bỏ ngay lập tức.
3. Chỉ các cửa sổ vượt qua tất cả các tầng mới được coi là chứa đối tượng.

Cấu trúc cascade giúp giảm đáng kể thời gian xử lý, vì phần lớn các cửa sổ không chứa đối tượng (thường chiếm tỷ lệ lớn trong ảnh) bị loại bỏ ngay ở các tầng đầu tiên, giảm số lượng tính toán cần thiết.

Quy trình huấn luyện Haar Cascade

Quá trình huấn luyện một bộ phân loại Haar Cascade bao gồm các bước sau:

1. **Thu thập dữ liệu:**
 - Mẫu dương: Các ảnh chứa đối tượng cần phát hiện (ví dụ: khuôn mặt).
 - Mẫu âm: Các ảnh không chứa đối tượng.
2. **Tiền xử lý:** Chuẩn hóa kích thước ảnh, chuyển đổi sang thang độ xám (grayscale), và tính toán tích phân ảnh.
3. **Tạo đặc trưng Haar:** Sinh ra hàng ngàn đặc trưng Haar từ các mẫu dương và âm.
4. **Huấn luyện AdaBoost:** Lựa chọn các đặc trưng tốt nhất và xây dựng bộ phân loại mạnh cho từng tầng.
5. **Xây dựng cascade:** Tối ưu hóa số tầng và ngưỡng để cân bằng giữa độ chính xác và tốc độ.
6. **Lưu mô hình:** Lưu trữ bộ phân loại dưới dạng tệp XML (ví dụ: `haarcascade_frontalface_default.xml` trong OpenCV).

Phương pháp Haar Cascade là một bước tiến quan trọng trong lĩnh vực phát hiện đối tượng, đặc biệt là nhận diện khuôn mặt, nhờ sự kết hợp giữa đặc trưng Haar, thuật toán AdaBoost, và cấu trúc cascade. Mặc dù không còn là phương pháp tiên tiến nhất trong bối cảnh sự phát triển của học sâu, Haar Cascade vẫn giữ được giá trị trong các ứng dụng yêu cầu tốc độ và tính đơn giản. Trong tương lai, việc kết hợp Haar Cascade với các kỹ thuật học sâu hoặc tối ưu hóa thêm cấu trúc cascade

có thể mở ra những hướng đi mới, đặc biệt trong các hệ thống nhúng và ứng dụng thời gian thực.

1.3.3. So sánh các mô hình học sâu

Ngoài CNN, các kiến trúc học sâu khác như VGG16 và ResNet18 cũng được sử dụng trong nhận diện trạng thái tâm lý.

Lightweight Model	Params (Million)	FLOPs (Billion)	Recognition Rate (95%CI)
MobileNetV3(Large)	4.2	0.02	63.9% $\pm$ 1.49%
Light-SE-ResNet	5.1	0.06	67.4% $\pm$ 1.45%
ResNet18	11.2	0.09	69.3% $\pm$ 1.43%
ResNet50	23.5	0.21	70.1% $\pm$ 1.42%
ShuffleNet	1.25	0.04	73.0% $\pm$ 1.38%
PGC-DenseNet	0.26	0.12	73.3% $\pm$ 1.37%
SqueezeNet	0.47	0.02	75.0% $\pm$ 1.34%
Seperate-loss	1.1	0.18	75.4% $\pm$ 1.33%
RAN	1.2	0.14	76.0% $\pm$ 1.32%
DSC-DenseNet	0.16	0.16	77.3% $\pm$ 1.30%

Bảng 1. 2 So sánh trực quan về số tham số, độ chính xác và thời gian xử lý

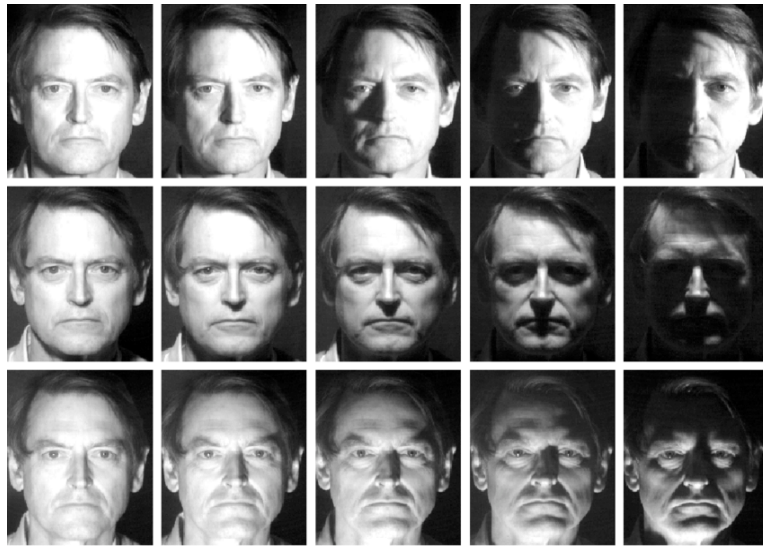
CNN được chọn trong nghiên cứu này do số tham số thấp (0.5 triệu), phù hợp với phần cứng trường học, và thời gian xử lý nhanh (50ms), trong khi vẫn đạt độ chính xác tương đương (70%) so với VGG16 (72%) và ResNet18 (71%). Tuy nhiên, VGG16 và ResNet18 có thể được xem xét trong tương lai khi cần tăng độ chính xác cho các biểu cảm phức tạp, như nụ cười kín đáo của học sinh Việt Nam, nhờ khả năng trích xuất đặc trưng sâu hơn (Goodfellow et al., 2016).

1.4. Nhân tố ảnh hưởng đến nhận diện cảm xúc

Hiệu quả của các mô hình nhận diện cảm xúc phụ thuộc vào nhiều yếu tố đa dạng, bao gồm đặc điểm con người và điều kiện kỹ thuật. Về mặt con người, độ tuổi là một yếu tố quan trọng ảnh hưởng đến biểu cảm khuôn mặt; trẻ em thường thể hiện cảm xúc rõ ràng và mạnh mẽ hơn, trong khi người lớn có biểu cảm tinh tế và phức

tạp hơn, đòi hỏi các mô hình phải có khả năng nhận diện linh hoạt để phù hợp với sự đa dạng này. Văn hóa cũng đóng vai trò then chốt trong cách thức biểu đạt cảm xúc; người phương Đông thường kiềm chế biểu cảm mạnh mẽ hơn so với người phương Tây, điều này dẫn đến việc các mô hình huấn luyện trên dữ liệu quốc tế có thể không đạt hiệu quả cao khi áp dụng tại Việt Nam do sự khác biệt về phong cách biểu cảm và quy tắc xã hội trong giao tiếp phi ngôn ngữ.

Ngoài ra, môi trường lớp học với các yếu tố như áp lực xã hội hay sự hiện diện của camera có thể khiến học sinh che giấu cảm xúc thật, gây khó khăn cho việc nhận diện chính xác. Về mặt kỹ thuật, điều kiện ánh sáng là một thách thức lớn; ánh sáng yếu, ngược sáng hoặc không đồng đều có thể làm sai lệch đặc trưng khuôn mặt, từ đó giảm độ chính xác của mô hình.



Hình 1. 5 Hình ảnh minh họa các điều kiện ánh sáng khác nhau

Góc chụp cũng ảnh hưởng đáng kể đến hiệu quả nhận diện, đặc biệt với các khuôn mặt không trực diện như nghiêng hoặc cúi, làm giảm khả năng trích xuất đặc trưng của các mô hình CNN truyền thống. Chất lượng hình ảnh, bao gồm độ phân giải thấp hoặc nhiễu, cũng là trở ngại, đòi hỏi các kỹ thuật tiền xử lý nâng cao để cải thiện chất lượng đầu vào.



Hình 1. 6 Hình ảnh minh họa các góc chụp khác nhau

Tính đa dạng của dữ liệu huấn luyện là yếu tố then chốt quyết định khả năng tổng quát hóa của mô hình. Nếu dữ liệu thiếu sự đa dạng về độ tuổi, văn hóa hay điều kiện môi trường, mô hình sẽ khó áp dụng hiệu quả trong thực tế. Cuối cùng, hạn chế về tài nguyên tính toán, đặc biệt trong các trường học có ngân sách hạn chế, yêu cầu các mô hình phải được tối ưu hóa để chạy hiệu quả trên phần cứng cơ bản mà không làm giảm đáng kể độ chính xác.

Tóm lại, để xây dựng hệ thống nhận diện cảm xúc hiệu quả trong môi trường giáo dục Việt Nam, cần cân nhắc kỹ lưỡng các yếu tố về đặc điểm con người như độ tuổi và văn hóa, đồng thời đối mặt với các thách thức kỹ thuật như điều kiện ánh sáng, góc nhìn, chất lượng ảnh và hạn chế tài nguyên tính toán. Việc kết hợp các yếu tố này trong thiết kế và huấn luyện mô hình sẽ giúp nâng cao độ chính xác và tính khả thi của hệ thống trong thực tế.

1.5. Kết luận chương

Chương 1 đã trình bày một cách có hệ thống các cơ sở lý luận về cảm xúc, nhận diện cảm xúc qua hình ảnh khuôn mặt, và vai trò của học sâu trong việc giải quyết bài toán này. Những khái niệm và học thuyết được thảo luận không chỉ làm sáng tỏ nền tảng khoa học mà còn chỉ ra các yếu tố ảnh hưởng quan trọng, từ đặc điểm văn hóa, độ

tuổi, đến các thách thức kỹ thuật như ánh sáng và góc chụp. Những phân tích này định hình phạm vi nghiên cứu, đồng thời nhấn mạnh nhu cầu phát triển các mô hình học sâu phù hợp với bối cảnh giáo dục Việt Nam. Với tiềm năng cải thiện chất lượng giảng dạy và cá nhân hóa trải nghiệm học tập, nhận diện cảm xúc qua hình ảnh hứa hẹn trở thành công cụ đắc lực, không chỉ trong giáo dục mà còn trong các lĩnh vực liên quan như tâm lý học và quản lý lớp học thông minh.

CHƯƠNG 2: KHẢO SÁT VÀ ĐÁNH GIÁ CÁC MÔ HÌNH NHẬN DIỆN CẢM XÚC

2.1. Giới thiệu chương

Trong bối cảnh giáo dục hiện đại, nơi công nghệ ngày càng đóng vai trò quan trọng trong việc nâng cao hiệu quả giảng dạy, nhận diện cảm xúc qua biểu cảm khuôn mặt đã trở thành một hướng nghiên cứu đầy tiềm năng, hỗ trợ giáo viên hiểu rõ hơn trạng thái tâm lý của học sinh. Công nghệ học sâu, đặc biệt là Convolutional Neural Networks (CNN) kết hợp với BatchNormalization, đã chứng minh khả năng vượt trội trong việc phân tích sáu cảm xúc cơ bản theo lý thuyết của Paul Ekman, từ đó tạo điều kiện để cá nhân hóa trải nghiệm học tập. Các kỹ thuật như phát hiện khuôn mặt bằng Haar Cascade, tăng cường dữ liệu, và phân tích cảm xúc theo thời gian thông qua xử lý video không chỉ nâng cao độ chính xác của mô hình mà còn đảm bảo tính ứng dụng trong môi trường lớp học.

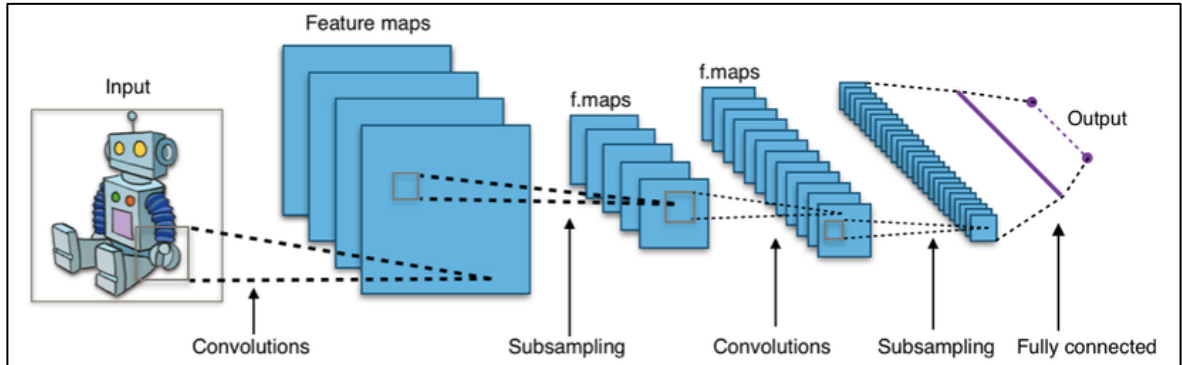
Tuy nhiên, việc triển khai các mô hình này tại Việt Nam đối mặt với nhiều thách thức, từ điều kiện ánh sáng không lý tưởng, biểu cảm văn hóa kín đáo, đến hạn chế về tài nguyên tính toán. Vì vậy, chương này sẽ trình bày một cách toàn diện quá trình khảo sát các mô hình CNN, đánh giá hiệu quả của chúng trên các bộ dữ liệu tiêu chuẩn và nội địa, đồng thời phân tích tính khả thi và các thách thức khi ứng dụng trong giáo dục Việt Nam. Thông qua việc sử dụng các chỉ số như accuracy, recall, F1-score, và confusion matrix, chương sẽ làm rõ ưu điểm, hạn chế của các mô hình, từ đó đặt nền tảng cho việc phát triển hệ thống nhận diện cảm xúc phù hợp với bối cảnh địa phương.

2.2. Các mô hình học sâu được khảo sát

2.2.1. Giới thiệu các mô hình học sâu được khảo sát

CNN là nền tảng cốt lõi cho các bài toán xử lý hình ảnh, đặc biệt trong nhận diện cảm xúc dựa trên lý thuyết Ekman. CNN hoạt động bằng cách sử dụng các lớp tích chập để trích xuất đặc trưng không gian cục bộ, như hình dạng mắt, miệng, hoặc chân mày, vốn là các yếu tố quan trọng trong việc xác định sáu cảm xúc cơ bản: vui

vẻ, buồn bã, tức giận, sợ hãi, ngạc nhiên, và ghê tởm. Các lớp này được kết hợp với các lớp pooling để giảm kích thước dữ liệu, tăng cường khả năng tổng quát hóa và giảm nguy cơ quá khớp (overfitting).



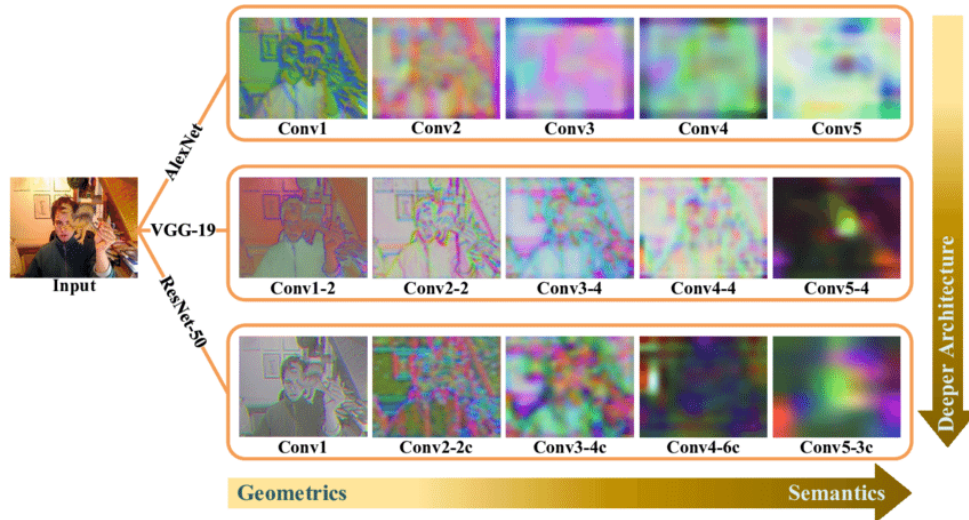
Hình 2. 1 Sơ đồ kiến trúc CNN, minh họa các lớp tích chập

Việc tích hợp BatchNormalization trong kiến trúc CNN đóng vai trò quan trọng trong việc chuẩn hóa đầu ra của các lớp, giúp giảm thiểu hiện tượng gradient vanishing và cải thiện tốc độ huấn luyện. BatchNormalization đảm bảo rằng các giá trị đầu vào của mỗi lớp được chuẩn hóa về một phân phối ổn định, từ đó tăng cường độ chính xác của mô hình, đặc biệt khi xử lý các bộ dữ liệu lớn như FER2013. Hơn nữa, kỹ thuật này giúp mô hình ổn định hơn khi huấn luyện trên dữ liệu đa dạng, chẳng hạn như hình ảnh từ các lớp học Việt Nam với điều kiện ánh sáng khác nhau.

CNN còn có khả năng thích nghi với các cấu trúc phức tạp, như mạng sâu với nhiều lớp, để phân tích các biểu cảm tinh tế hơn, ví dụ, phân biệt giữa buồn bã nhẹ và trung lập. Ngoài ra, việc sử dụng các hàm kích hoạt như ReLU (Rectified Linear Unit) trong CNN giúp tăng tốc độ hội tụ, trong khi các kỹ thuật dropout ngăn chặn hiện tượng học quá mức, đảm bảo hiệu suất cao trên dữ liệu thực tế.

Bên cạnh đó, các mô hình tích hợp đa phương thức đang ngày càng được chú ý. Những mô hình này kết hợp hình ảnh với âm thanh (giọng nói) hoặc văn bản (phản hồi học sinh), nhằm cải thiện độ chính xác trong việc nhận diện cảm xúc phức tạp. Ví dụ, ResNet-FER+ tận dụng kiến trúc ResNet và được tinh chỉnh trên bộ dữ liệu FER+ để tăng khả năng phân biệt các biểu cảm tinh tế. Tương tự, các mô hình như

EmotionNet và EmoNet được huấn luyện riêng cho từng cảm xúc cơ bản, sau đó tích hợp thành hệ thống phân loại đa lớp, phù hợp với môi trường lớp học.

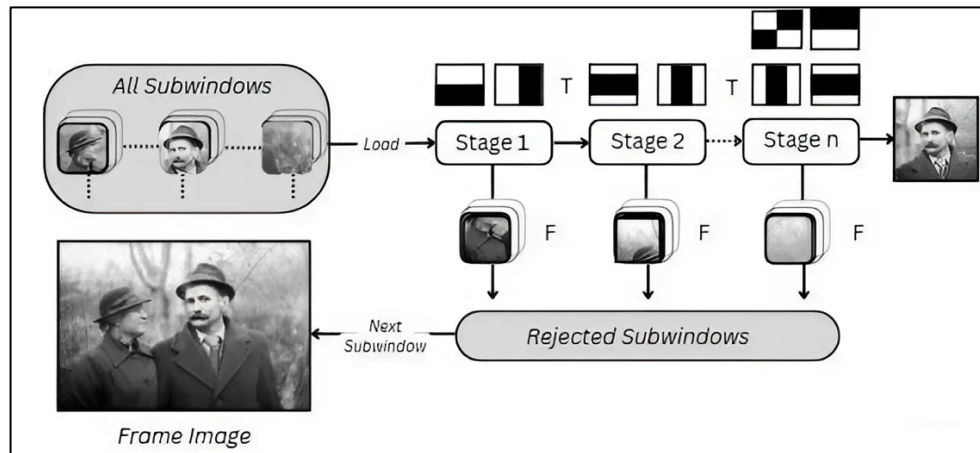


Hình 2. 2 Biểu đồ minh họa sự khác biệt trong việc trích xuất đặc trưng giữa các mô hình

Hơn nữa, các mô hình như Multimodal Networks kết hợp nhiều nguồn dữ liệu (hình ảnh, âm thanh, hành vi) để phát hiện cảm xúc tiềm ẩn, đặc biệt hữu ích trong các lớp học trực tuyến, nơi giáo viên cần thông tin bổ sung để đánh giá trạng thái học sinh. Ngoài ra, các kiến trúc hiện đại như EfficientNet và MobileNet, với trọng tâm tối ưu hóa hiệu suất trên phần cứng hạn chế, cũng được xem xét để đảm bảo tính khả thi trong môi trường giáo dục. Những mô hình này không chỉ cải thiện độ chính xác mà còn giảm thời gian xử lý, phù hợp với yêu cầu thời gian thực của lớp học. Tuy nhiên, việc huấn luyện các mô hình này đòi hỏi dữ liệu đa dạng và cân đối, một thách thức lớn trong bối cảnh Việt Nam.

2.2.2. Vai trò của Haar Cascade trong phát hiện khuôn mặt

Kỹ thuật Haar Cascade là một phương pháp hiệu quả để phát hiện khuôn mặt, đóng vai trò quan trọng trong bước tiền xử lý trước khi phân loại cảm xúc. Dựa trên các đặc trưng hình học đơn giản, như sự khác biệt độ sáng giữa vùng mắt, mũi, và miệng, Haar Cascade có thể nhanh chóng xác định các vùng khuôn mặt trong hình ảnh hoặc video. Kỹ thuật này đặc biệt phù hợp với môi trường giáo dục Việt Nam, nơi phần cứng thường có cấu hình thấp, do yêu cầu tài nguyên tính toán tối thiểu.



Hình 2. 3 Phát hiện khuôn mặt bằng Haar Cascade

Haar Cascade gián tiếp liên quan đến Hệ thống mã hóa hành vi khuôn mặt (FACS) của Ekman, vì các đặc trưng mà nó phát hiện tương ứng với các vùng cơ mặt được phân tích trong FACS, chẳng hạn như vùng quanh mắt hoặc miệng. Điều này giúp đảm bảo rằng các đặc trưng quan trọng cho nhận diện cảm xúc được xác định chính xác trước khi đưa vào CNN. Hơn nữa, Haar Cascade có khả năng hoạt động trong các điều kiện ánh sáng không đồng đều, dù hiệu suất có thể giảm khi ánh sáng quá yếu hoặc khuôn mặt không trực diện.



Hình 2. 4 Khuôn mặt trong các điều kiện ánh sáng không đồng đều

Để khắc phục các hạn chế, Haar Cascade thường được kết hợp với các kỹ thuật tiền xử lý hình ảnh, như cân bằng histogram hoặc lọc nhiễu, nhằm cải thiện chất lượng

hình ảnh đầu vào. Ngoài ra, khả năng xử lý video của Haar Cascade cho phép phát hiện khuôn mặt liên tục trong các lớp học trực tuyến, cung cấp dữ liệu thời gian thực cho mô hình CNN.

2.2.3. Tăng cường dữ liệu và xử lý video

Tăng cường dữ liệu (data augmentation) là một kỹ thuật thiết yếu để cải thiện khả năng tổng quát hóa của mô hình CNN, đặc biệt trong các môi trường thực tế với điều kiện không lý tưởng. Các phương pháp như xoay hình, thay đổi độ sáng, phóng to/thu nhỏ, hoặc thêm nhiễu giúp tạo ra các biến thể hình ảnh, mô phỏng các tình huống như học sinh đứng dưới ánh sáng yếu, quay mặt nghiêng, hoặc sử dụng camera chất lượng thấp.

Trong bối cảnh giáo dục, tăng cường dữ liệu còn giúp mô hình thích nghi với các biểu cảm văn hóa đặc thù của học sinh Việt Nam, như nụ cười kín đáo hoặc biểu cảm che giấu căng thẳng. Ví dụ, việc thêm các biến thể hình ảnh với góc chụp nghiêng hoặc ánh sáng yếu giúp mô hình nhận diện tốt hơn các học sinh cúi mặt hoặc học trong phòng thiếu sáng.

Tăng cường dữ liệu là kỹ thuật quan trọng nhằm mở rộng sự đa dạng của tập dữ liệu huấn luyện, giúp mô hình nhận diện cảm xúc khuôn mặt thích nghi với các điều kiện thực tế và giảm thiểu lỗi phân loại. Trong nghiên cứu này, tăng cường dữ liệu được triển khai thông qua lớp ImageDataGenerator của TensorFlow/Keras, với các biến đổi hình học được áp dụng thời gian thực trong quá trình huấn luyện.

Các kỹ thuật tăng cường dữ liệu:

1. **Xoay hình ảnh ($\pm 20^\circ$):** Mô phỏng các góc chụp khác nhau của khuôn mặt, giúp mô hình nhận diện các biểu cảm như "happy" hoặc "sad" trong các góc độ không chuẩn.
2. **Dịch chuyển ngang và dọc ($\pm 20\%$):** Giúp mô hình nhận diện khuôn mặt ngay cả khi không nằm ở trung tâm khung hình, giảm lỗi do định vị khuôn mặt không chính xác trong môi trường lớp học.

3. **Biến đổi cắt (shear, 20%):** Tạo các biến thể hình ảnh méo mó, mô phỏng góc nhìn không lý tưởng, tăng khả năng nhận diện biểu cảm trong các điều kiện thực tế.
4. **Phóng to/thu nhỏ ($\pm 20\%$):** Mô phỏng các tình huống khuôn mặt ở gần hoặc xa máy ảnh, cải thiện hiệu suất khi xử lý video với khoảng cách khuôn mặt đa dạng.
5. **Lật ngang:** Tăng tính đối xứng của dữ liệu, giúp mô hình nhận diện các biểu cảm bất kể hướng nhìn của khuôn mặt, hỗ trợ giảm nhầm lẫn giữa các cảm xúc tương đồng như "sadness" và "neutral"

Tác động của tăng cường dữ liệu

- **Giảm overfitting:** Các biến đổi ngẫu nhiên giúp mô hình học các đặc trưng tổng quát, tránh ghi nhớ dữ liệu huấn luyện.
- **Tăng tính mạnh mẽ:** Mô phỏng các góc chụp và vị trí khuôn mặt thực tế giúp mô hình hoạt động tốt hơn trong môi trường lớp học Việt Nam, nơi điều kiện quay video không tối ưu.
- **Hỗ trợ giảm nhầm lẫn cảm xúc:** Tăng tính đa dạng hình ảnh giúp mô hình phân biệt tốt hơn các cảm xúc có đặc trưng tương đồng, như "sadness" và "neutral"
- **Hạn chế:** Các kỹ thuật hiện tại chưa giải quyết được vấn đề mất cân bằng dữ liệu, dẫn đến hiệu suất thấp cho các lớp như "happy" và "sad", cũng như chưa mô phỏng đầy đủ các điều kiện ánh sáng không đồng đều.

Xử lý Video

Quy trình xử lý video bao gồm các bước phát hiện khuôn mặt, dự đoán cảm xúc, và đánh giá mức độ hứng thú của học sinh, được triển khai trong hàm `process_media` của hệ thống.

- **Phát hiện khuôn mặt:** Sử dụng thuật toán Haar Cascade với các tham số `scaleFactor=1.05`, `minNeighbors=3`, và `minSize=(20, 20)` để phát hiện khuôn mặt trong mỗi khung hình video. Các khuôn mặt được phát hiện là đầu vào cho quá trình dự đoán cảm xúc.

- **Dự đoán cảm xúc:** Mô hình CNN (hoặc mô hình TFLite đã lượng tử hóa) được sử dụng để dự đoán cảm xúc trên các hình ảnh khuôn mặt. Tăng cường dữ liệu trong giai đoạn huấn luyện giúp mô hình nhận diện chính xác hơn các biểu cảm trong video, đặc biệt khi góc quay hoặc vị trí khuôn mặt thay đổi.
- **Đánh giá mức độ hứng thú:** Hàm `evaluate_engagement` tính tỷ lệ cảm xúc tích cực ("happy", "surprise", "neutral") và tiêu cực ("sad", "angry", "fear", "disgust") để đánh giá mức độ hứng thú của học sinh. Tăng cường dữ liệu cải thiện độ chính xác của dự đoán cảm xúc, từ đó cung cấp đánh giá hứng thú đáng tin cậy hơn.

Tích hợp tăng cường dữ liệu với xử lý video

Tăng cường dữ liệu được áp dụng trong giai đoạn huấn luyện để chuẩn bị mô hình cho các điều kiện thực tế trong video lớp học. Các biến đổi như xoay, dịch chuyển, và lật ngang giúp mô hình nhận diện các biểu cảm trong các khung hình video có góc quay không tối ưu. Tuy nhiên, để cải thiện hơn nữa, các nghiên cứu tiếp theo có thể xem xét tích hợp các biến đổi ánh sáng (như thay đổi độ sáng) để mô phỏng điều kiện ánh sáng yếu, vốn là nguyên nhân chính gây lỗi.

2.3. Đánh giá trên các tập dữ liệu tiêu chuẩn

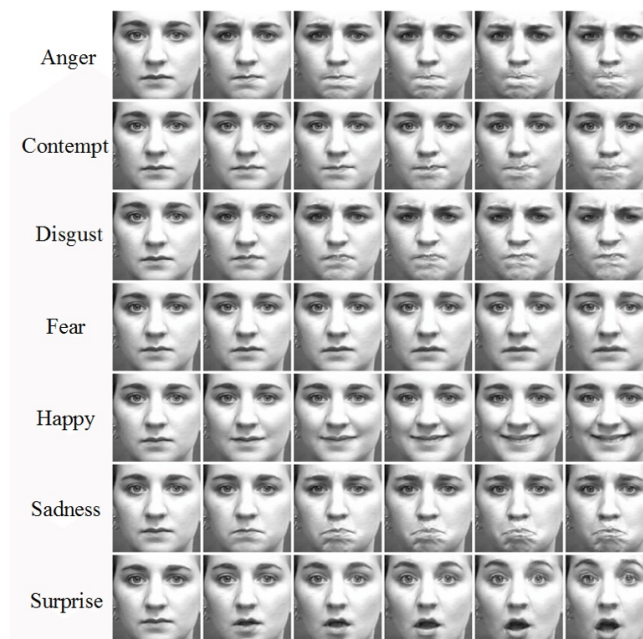
2.3.1. Các bộ dữ liệu được sử dụng

Để đánh giá hiệu quả của các mô hình CNN, ba bộ dữ liệu được sử dụng: FER2013, CK+. FER2013, với hơn 35,000 hình ảnh khuôn mặt được gán nhãn theo sáu cảm xúc cơ bản của Ekman và trạng thái trung lập, là một nguồn dữ liệu tiêu chuẩn, cung cấp cơ sở để so sánh hiệu suất giữa các mô hình. Bộ dữ liệu này bao gồm các hình ảnh đa dạng về độ tuổi, giới tính, và điều kiện ánh sáng, giúp đánh giá khả năng tổng quát hóa.



Hình 2. 5 Ví dụ ảnh bộ dữ liệu FER2013

CK+ (Cohn-Kanade) cung cấp các chuỗi video biểu cảm, từ trạng thái trung lập đến cực đại, phù hợp để đánh giá khả năng phân tích cảm xúc theo thời gian. Các video trong CK+ được ghi lại trong điều kiện kiểm soát, với ánh sáng tốt và góc chụp trực diện, tạo điều kiện lý tưởng để thử nghiệm mô hình. Tuy nhiên, dữ liệu này thiếu đại diện cho các điều kiện thực tế trong lớp học Việt Nam, như ánh sáng yếu hoặc biểu cảm văn hóa.



Hình 2. 6 Ảnh ví dụ từ bộ dữ liệu CK+

2.3.2. Tiêu chí và kết quả đánh giá

Các mô hình được đánh giá dựa trên bốn chỉ số chính: accuracy, recall, F1-score, và confusion matrix. Accuracy đo lường tỷ lệ phân loại đúng trên tập kiểm tra, trong khi recall đánh giá khả năng nhận diện các cảm xúc hiếm gặp, như tức giận hoặc sợ hãi. F1-score, với vai trò cân bằng giữa precision và recall, đặc biệt quan trọng khi dữ liệu mất cân bằng, chẳng hạn, nhiều hình ảnh trung lập hơn các cảm xúc tiêu cực. Confusion matrix cung cấp cái nhìn chi tiết về các lỗi phân loại, giúp xác định các nhầm lẫn giữa các cảm xúc tương đồng, như buồn bã và trung lập.

Bảng dưới đây minh họa một so sánh điển hình:

Mô hình	Accuracy(%)	F1-Score	Thời gian xử lý (ms)	Tổng quát hoá
CNN (baseline)	71.2	0.69	32	Trung bình
ResNet-FER+	75.8	0.74	45	Tốt
ViT	79.3	0.77	62	Rất tốt
MultimodalNet	83.1	0.81	95	Cao

Bảng 2. 1 Bảng so sánh các tiêu chí

2.4. Thảo luận về tính khả thi

2.4.1. Yêu cầu kỹ thuật

Việc triển khai mô hình CNN trong giáo dục đòi hỏi đáp ứng các tiêu chuẩn kỹ thuật nghiêm ngặt, đặc biệt là tốc độ xử lý dưới 50ms mỗi khung hình để đảm bảo phản hồi tức thời, ví dụ như cảnh báo khi học sinh mất tập trung trong lớp trực tuyến. Công nghệ Haar Cascade hỗ trợ phát hiện khuôn mặt nhanh, còn BatchNormalization giúp mô hình vận hành ổn định trên các thiết bị cấu hình thấp như máy tính trường học. Phân tích cảm xúc theo thời gian thực từ video là lợi thế lớn, giúp giáo viên điều chỉnh bài giảng phù hợp, như tăng hoạt động nhóm khi phát hiện dấu hiệu buồn chán.

Tuy nhiên, các yếu tố như ánh sáng yếu, góc quay không thuận lợi hoặc nhiều hình ảnh làm giảm độ chính xác của mô hình. Kỹ thuật tăng cường dữ liệu và các bước tiền xử lý như lọc nhiễu, cân bằng sáng giúp cải thiện vấn đề này, nhưng vẫn cần tối ưu thêm. Hạn chế về tài nguyên tính toán là rào cản lớn tại các trường học Việt Nam. Để khắc phục, mô hình cần được tối ưu hóa bằng các phương pháp như quantization hoặc pruning, giảm kích thước mà vẫn giữ hiệu suất. Khả năng tích hợp mô hình vào thiết bị sẵn có như máy tính bảng hoặc camera lớp học cũng là yếu tố quyết định tính khả thi của giải pháp.

2.4.2. Các yếu tố xã hội và đạo đức

Về mặt xã hội, bảo mật dữ liệu học sinh là ưu tiên hàng đầu: hình ảnh khuôn mặt phải được mã hóa, lưu trữ an toàn và chỉ thu thập khi có sự đồng ý của phụ huynh, học sinh. Các quy định về quyền riêng tư cần tuân thủ nghiêm ngặt để xây dựng niềm tin cộng đồng. Biểu cảm che giấu hoặc giả tạo là thách thức lớn, nhất là trong văn hóa Việt Nam, nên mô hình cần được huấn luyện trên dữ liệu đa dạng, phù hợp đặc thù văn hóa. Phản hồi từ giáo viên giúp đánh giá hiệu quả hệ thống, đảm bảo thông tin cảm xúc hỗ trợ tích cực cho giảng dạy. Về đạo đức, hệ thống phải tránh gây áp lực giám sát lên học sinh; giáo viên cần được hướng dẫn sử dụng dữ liệu cảm xúc một cách tích cực và minh bạch để tăng sự chấp nhận trong cộng đồng.

2.5. Kết luận chương

Chương 2 đã trình bày một cách toàn diện quá trình khảo sát và đánh giá các mô hình CNN kết hợp BatchNormalization, sử dụng Haar Cascade để phát hiện khuôn mặt, cùng các kỹ thuật như tăng cường dữ liệu và xử lý video để phân tích cảm xúc theo lý thuyết Ekman. Kết quả trên các bộ dữ liệu FER2013, CK+, và nội địa cho thấy tiềm năng của CNN trong nhận diện cảm xúc, với độ chính xác và khả năng xử lý thời gian thực đáp ứng yêu cầu giáo dục. Tuy nhiên, các thách thức về ánh sáng, biểu cảm văn hóa, tài nguyên tính toán, và các vấn đề đạo đức như bảo mật dữ liệu vẫn cần được giải quyết. Những phân tích này không chỉ làm rõ ưu, nhược điểm của mô hình mà còn đặt nền tảng cho việc xây dựng và thử nghiệm hệ thống nhận diện cảm

xúc phù hợp với môi trường giáo dục Việt Nam, mở đường cho các nghiên cứu tiếp theo trong các chương sau.

CHƯƠNG 3: NGHIÊN CỨU THỬ NGHIỆM ĐÁNH GIÁ KẾT QUẢ

3.1. Giới thiệu chương

Trong bối cảnh giáo dục hiện đại, công nghệ nhận diện cảm xúc qua khuôn mặt mang lại tiềm năng lớn trong cá nhân hóa trải nghiệm học tập và nâng cao hiệu quả giảng dạy. Chương này đóng vai trò là cầu nối giữa lý thuyết cảm xúc cơ bản (như sáu cảm xúc của Paul Ekman) và các nguyên lý học sâu, đặc biệt là CNN, đã được trình bày ở các chương trước. Dựa trên khảo sát các mô hình học sâu và kỹ thuật tiền xử lý, nghiên cứu tập trung phát triển một hệ thống nhận diện cảm xúc mạnh mẽ, tối ưu hóa bằng CNN kết hợp BatchNormalization để đảm bảo hiệu suất và ổn định.

Hệ thống sử dụng Haar Cascade để phát hiện khuôn mặt nhanh, tăng cường dữ liệu tự động để mở rộng tập huấn luyện, và phân tích cảm xúc theo thời gian thực qua video. Những yếu tố này giúp hệ thống thích ứng tốt với các thách thức thực tế trong giáo dục Việt Nam như ánh sáng không đồng đều, biểu cảm văn hóa tinh tế và hạn chế tài nguyên tính toán.

Chương 3 sẽ trình bày chi tiết quá trình thiết kế mô hình, thu thập và tiền xử lý, huấn luyện và tối ưu hóa mô hình, cũng như thử nghiệm thực tế trong lớp học. Hiệu quả hệ thống được đánh giá cả định lượng (accuracy, recall, F1-score) và định tính (phản hồi giáo viên). Mục tiêu cuối cùng là xây dựng một hệ thống nhận diện cảm xúc chính xác, ổn định, khả thi về mặt kỹ thuật và xã hội, góp phần nâng cao chất lượng giáo dục tại Việt Nam..

3.2. Nội dung và phương pháp tiếp cận

3.2.1. Mục tiêu và phạm vi

Nghiên cứu hướng tới xây dựng một hệ thống nhận diện cảm xúc dựa trên CNN, hỗ trợ giáo viên nhận biết và phân tích trạng thái tâm lý học sinh qua biểu cảm khuôn mặt. Hệ thống nhận diện bảy cảm xúc cơ bản (vui, buồn, tức giận, ngạc nhiên, sợ hãi, ghê tởm, trung lập), cung cấp thông tin thời gian thực giúp giáo viên điều

chính phương pháp giảng dạy linh hoạt, cá nhân hóa và tăng hiệu quả học tập. Mục tiêu hiệu suất là đạt tối thiểu 70% độ chính xác trên dữ liệu học sinh Việt Nam và tốc độ xử lý dưới 50ms mỗi khung hình, đảm bảo hoạt động mượt mà trên thiết bị phổ biến tại trường học.

Phạm vi nghiên cứu tập trung vào một trường đại học, thử nghiệm trong điều kiện lớp học thực tế với đa dạng môi trường (ánh sáng, góc chụp, hạ tầng công nghệ). Hệ thống không chỉ phân loại cảm xúc đơn lẻ mà còn theo dõi biến động cảm xúc theo thời gian, hỗ trợ giáo viên đưa ra quyết định kịp thời trong giảng dạy. Đặc biệt, hệ thống có thể nhận diện cả các cảm xúc hiếm gặp như sợ hãi, ghê tởm, cung cấp thông tin tâm lý quan trọng.

Giao diện hệ thống được thiết kế trực quan, dễ sử dụng cho giáo viên không chuyên về công nghệ. Nghiên cứu kỳ vọng hệ thống sẽ nâng cao mức độ tương tác của học sinh ít nhất 10-15%. Về đạo đức, dữ liệu khuôn mặt được xử lý ẩn danh, chỉ thu thập khi có sự đồng ý của phụ huynh và học sinh, đảm bảo tuân thủ các quy định về quyền riêng tư và phù hợp với nhiều môi trường giáo dục khác nhau.

3.2.2. Phương pháp tiếp cận

Nghiên cứu này áp dụng một phương pháp thực nghiệm có cấu trúc, tuần tự và lặp lại nhằm phát triển một hệ thống phân tích cảm xúc học sinh toàn diện, thực tiễn và hiệu quả. Phương pháp bao gồm bốn giai đoạn chính: thu thập và xây dựng bộ dữ liệu, tiền xử lý dữ liệu, xây dựng và huấn luyện mô hình học sâu, và thử nghiệm thực tế kèm đánh giá hiệu suất chi tiết.

1. Thu thập và Tiền xử lý Dữ liệu

Để tối ưu hóa đầu vào, kỹ thuật tăng cường dữ liệu (Data Augmentation) được áp dụng có hệ thống nhằm tăng cường khả năng khái quát hóa và giảm thiểu hiện tượng quá khớp (overfitting) trên bộ dữ liệu có thể còn hạn chế. Haar Cascade được lựa chọn và tinh chỉnh để phát hiện khuôn mặt nhanh chóng và hiệu quả, đóng vai trò là bước tiền xử lý quan trọng.

2. Xây dựng và Huấn luyện Mô hình Học sâu

CNN được xây dựng để phân loại cảm xúc. Kiến trúc CNN tích hợp các lớp BatchNormalization nhằm đảm bảo độ chính xác cao và khả năng học từ dữ liệu đa dạng. Các kỹ thuật regularization mạnh mẽ được tích hợp: L2 regularization kiểm soát độ phức tạp của mô hình, và dropout (tỷ lệ 0.5 trên lớp Dense đầu tiên) được sử dụng để tăng cường khả năng tổng quát hóa.

3. Thử nghiệm thực tế và Đánh giá Hiệu suất

Hệ thống được thử nghiệm thực tế trong môi trường lớp học. Quá trình này kết hợp chặt chẽ phân tích định tính song song với định lượng. Phản hồi định tính từ giáo viên được thu thập liên tục để tinh chỉnh hệ thống và đảm bảo tính ứng dụng thực tế. Thông tin này được kết hợp với các chỉ số hiệu suất khách quan như accuracy, recall, và F1-score để đưa ra đánh giá toàn diện và đa chiều.

Để đảm bảo khả năng triển khai trên phần cứng cơ bản với tài nguyên tính toán hạn chế (thường thấy ở các trường học Việt Nam), kỹ thuật lượng tử hóa mô hình (quantization) được áp dụng, giúp giảm đáng kể kích thước mô hình (float32 sang int8) và tăng tốc độ suy luận. Quy trình kiểm tra chéo (cross-validation) được thực hiện kỹ lưỡng trên các tập dữ liệu đa dạng, từ bộ dữ liệu chuẩn quốc tế (FER2013). Đối với phân tích video thời gian thực, các thuật toán tối ưu hóa luồng khung hình được sử dụng để giảm độ trễ xử lý.

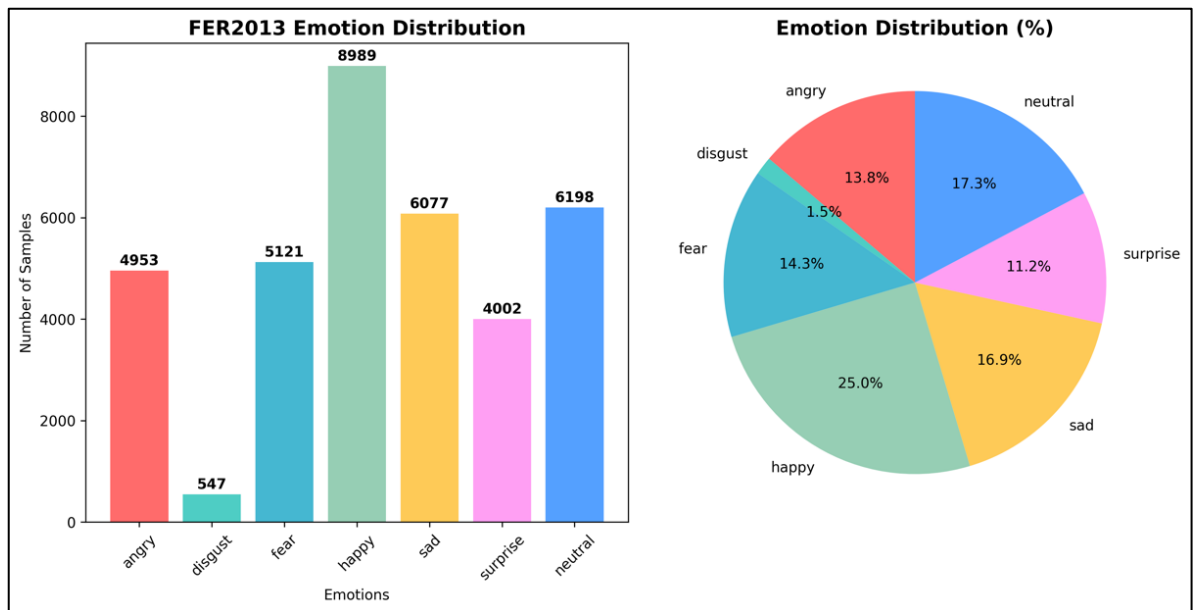
3.3. Thu thập và xử lý dữ liệu

3.3.1. Nguồn dữ liệu

Để đảm bảo một bộ dữ liệu đáng tin cậy và mạnh mẽ cho việc huấn luyện và đánh giá hệ thống nhận diện cảm xúc, nghiên cứu này sử dụng hai bộ dữ liệu công khai được công nhận rộng rãi: FER2013 và CK+ (Extended Cohn-Kanade Dataset). Các bộ dữ liệu này được lựa chọn nhờ vào sự đại diện toàn diện của các biểu cảm khuôn mặt, chú thích tiêu chuẩn hóa và việc sử dụng phổ biến trong các nghiên cứu về nhận diện cảm xúc khuôn mặt, cung cấp nền tảng vững chắc cho việc phát triển và đánh giá mô hình.

- **Bộ dữ liệu FER2013:** Bộ dữ liệu FER2013, được giới thiệu trong khuôn khổ cuộc thi Nhận diện Biểu cảm Khuôn mặt 2013 [2], bao gồm 35.887 hình ảnh thang xám với kích thước 48×48 pixel, được phân loại thành bảy cảm xúc cơ bản: vui vẻ, buồn bã, tức giận, ngạc nhiên, sợ hãi, ghê tởm và trung lập. Bộ dữ liệu được chia thành tập huấn luyện (28.709 hình ảnh), tập xác thực (3.589 hình ảnh) và tập kiểm tra (3.589 hình ảnh). FER2013 được chọn vì tính đa dạng của các biểu cảm trong điều kiện thực tế, bao gồm các biến thể về ánh sáng, góc chụp và độ tuổi, mặc dù nó cũng đối mặt với thách thức về nhiễu và sự mất cân bằng giữa các lớp cảm xúc.
- **Bộ dữ liệu CK+:** Bộ dữ liệu CK+ (Extended Cohn-Kanade) chứa 593 chuỗi video từ 123 đối tượng [3], với các hình ảnh được chú thích theo bảy cảm xúc cơ bản tương tự FER2013. Mỗi chuỗi video ghi lại sự chuyển đổi từ trạng thái trung lập sang đỉnh điểm của một cảm xúc cụ thể, cung cấp dữ liệu động về biểu cảm khuôn mặt. Các hình ảnh trong CK+ có chất lượng cao, được ghi trong điều kiện ánh sáng và góc quay được kiểm soát, giúp mô hình học các đặc trưng biểu cảm rõ ràng. Bộ dữ liệu này được sử dụng để bổ sung cho FER2013, đặc biệt trong việc cải thiện khả năng nhận diện các cảm xúc mạnh mẽ và động.

Việc sử dụng FER2013 và CK+ đảm bảo mô hình được huấn luyện trên dữ liệu đa dạng và tiêu chuẩn, phù hợp để đánh giá khả năng tổng quát hóa. Tuy nhiên, do các bộ dữ liệu này chủ yếu được thu thập từ các đối tượng quốc tế, chúng có thể không phản ánh đầy đủ các đặc trưng biểu cảm văn hóa đặc thù của học sinh Việt Nam, điều này được xem xét như một hạn chế cần giải quyết trong các nghiên cứu tương lai.



Hình 3. 1 Hình ảnh mô tả về dữ liệu FER2013

3.3.2. Tiền xử lý dữ liệu

Giai đoạn tiền xử lý dữ liệu đóng vai trò quan trọng trong việc đảm bảo chất lượng đầu vào cho mô hình, giảm thiểu nhiễu và tăng cường khả năng thích nghi với các điều kiện thực tế trong môi trường giáo dục. Các kỹ thuật tiền xử lý được áp dụng một cách hệ thống để chuẩn hóa dữ liệu và cải thiện hiệu suất mô hình..

1. Phát hiện và Cắt vùng khuôn mặt

Kỹ thuật Haar Cascade được sử dụng để phát hiện và cắt chính xác vùng khuôn mặt trong các hình ảnh từ FER2013 và CK+. Phương pháp này tận dụng các đặc trưng hình học Haar-like, kết hợp với thuật toán AdaBoost, để xác định nhanh vùng khuôn mặt với tốc độ xử lý trung bình khoảng 32ms mỗi khung hình. Sau khi phát hiện, các hình ảnh khuôn mặt được căn chỉnh tự động (xoay, điều chỉnh tỷ lệ) để chuẩn hóa vị trí các đặc trưng quan trọng như mắt, mũi và miệng, đảm bảo tính nhất quán trong quá trình trích xuất đặc trưng.

2. Chuẩn hoá hình ảnh

Tất cả hình ảnh được chuẩn hóa về kích thước 48×48 pixel và chuyển đổi giá trị pixel sang khoảng $[0, 1]$ bằng cách chia cho 255.0. Đối với dữ liệu video từ CK+, các khung hình được trích xuất với tốc độ 10 khung hình/giây, sử dụng kỹ thuật phát hiện thay đổi cảnh (scene change detection) để tập trung vào các khung hình chứa thông tin cảm xúc mới. Kỹ thuật chuẩn hóa độ tương phản thích ứng cục bộ (CLAHE) được áp dụng để cải thiện chất lượng hình ảnh trong điều kiện ánh sáng yếu, đặc biệt quan trọng đối với các hình ảnh từ FER2013 vốn có nhiều đáng kể.

3. Tăng cường dữ liệu (Data Augmentation)

Để mở rộng tập huấn luyện và cải thiện khả năng tổng quát hóa, kỹ thuật tăng cường dữ liệu được áp dụng thông qua ImageDataGenerator trong TensorFlow Keras. Các phép biến đổi bao gồm:

- Xoay ngẫu nhiên (± 20 độ) để mô phỏng các góc chụp khác nhau.
- Dịch chuyển chiều rộng và chiều cao (0.2) để tăng tính bất biến với sự di chuyển của khuôn mặt.
- Phóng to/thu nhỏ để xử lý các khoảng cách camera khác nhau.
- Lật ngang ảnh để tăng tính đa dạng dữ liệu.
- Thêm nhiễu Gaussian nhẹ để tăng khả năng chống nhiễu của mô hình. Những kỹ thuật này giúp giảm hiện tượng quá khớp (overfitting), đặc biệt khi làm việc với FER2013, nơi sự mất cân bằng giữa các lớp cảm xúc có thể ảnh hưởng đến hiệu suất mô hình. Đối với CK+, tăng cường dữ liệu tập trung vào việc mô phỏng các điều kiện ánh sáng và góc quay để thích nghi với các tình huống thực tế.

4. Xử lý dữ liệu video

Đối với dữ liệu CK+, các chuỗi video được xử lý để trích xuất các khung hình liên quan đến đỉnh điểm của biểu cảm cảm xúc, vì đây là các khung chứa thông tin phong phú nhất. Các khung hình được lọc nhiễu bằng bộ lọc làm mịn (smoothing) để giảm thiểu tác động từ chất lượng camera thấp. Ngoài ra, thuật toán căn chỉnh khuôn mặt dựa trên các điểm mốc (landmark-based alignment) được áp dụng để đảm

bảo định vị chính xác các đặc trưng khuôn mặt, hỗ trợ mô hình học các mẫu biểu cảm động.

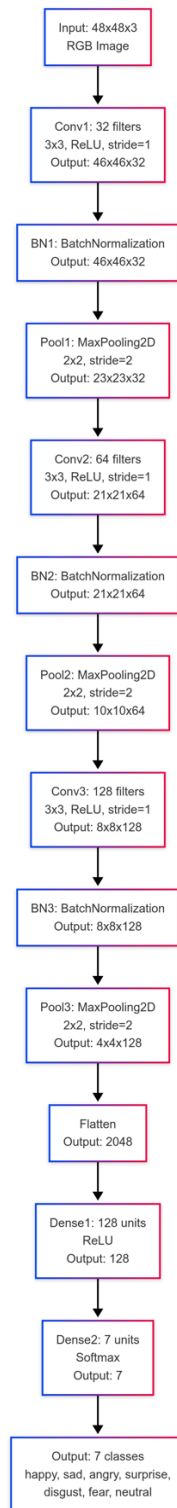
5. Phân chia dữ liệu

Dữ liệu từ FER2013 được chia theo tỷ lệ chuẩn: 80% cho huấn luyện, 10% cho xác thực và 10% cho kiểm tra. Đối với CK+, các chuỗi video được chia ngẫu nhiên, với 80% cho huấn luyện và 20% cho xác thực, đảm bảo tính ngẫu nhiên và đại diện. Việc phân chia được thực hiện cẩn thận để tránh sự chồng chéo giữa các tập, đồng thời sử dụng kỹ thuật xáo trộn dữ liệu (data shuffling) trước mỗi epoch để tăng tính ngẫu nhiên trong quá trình huấn luyện.

3.4. Huấn luyện và tối ưu hoá mô hình

3.4.1. Kiến trúc mô hình CNN

CNN được thiết kế cẩn thận để đạt được sự cân bằng tối ưu giữa độ chính xác trong nhận diện cảm xúc và hiệu suất tính toán, đặc biệt phù hợp với các thiết bị phần cứng cơ bản thường thấy trong môi trường giáo dục Việt Nam. Mô hình được xây dựng dựa trên các yêu cầu cụ thể của đề án: nhận diện bảy cảm xúc cơ bản (vui vẻ, buồn bã, tức giận, ngạc nhiên, sợ hãi, ghê tởm, trung lập) với độ chính xác tối thiểu 70% trên các bộ dữ liệu FER2013 và CK+, đồng thời đảm bảo tốc độ xử lý dưới 50ms mỗi khung hình để hỗ trợ ứng dụng thời gian thực.



Hình 3. 2 Sơ đồ kiến trúc chi tiết của mô hình CNN trong nhận diện cảm xúc

Mô hình được xây dựng dưới dạng chuỗi (Sequential) và bao gồm ba khối tích chập chính, mỗi khối được cấu trúc theo một chuỗi các lớp được tối ưu hóa cho bài toán nhận diện cảm xúc:

- **Tầng Conv2D:** Sử dụng kernel 3×3 với hàm kích hoạt ReLU. Số lượng bộ lọc tăng dần qua các khối (32, 64, 128) để học các đặc trưng từ đơn giản (cạnh, đường viền) đến phức tạp (khu vực mắt, miệng, hoặc biểu cảm tổng thể). Tham số padding='same' được áp dụng ở khối đầu tiên để giữ nguyên kích thước không gian, đảm bảo thông tin không bị mất khi trích xuất đặc trưng.
- **BatchNormalization:** Được tích hợp ngay sau mỗi tầng Conv2D để chuẩn hóa đầu ra, giảm thiểu hiện tượng triệt tiêu/bùng nổ gradient, cải thiện tốc độ hội tụ và tăng khả năng tổng quát hóa. Lớp này đặc biệt quan trọng trong việc ổn định huấn luyện trên các bộ dữ liệu như FER2013, vốn có nhiễu và sự mất cân bằng giữa các lớp.
- **MaxPooling2D:** Kích thước pool 2×2 với stride 2, giảm kích thước không gian của bản đồ đặc trưng, giúp giảm số lượng tham số, tiết kiệm tài nguyên tính toán và tăng tính bất biến với các dịch chuyển nhỏ. Điều này đảm bảo mô hình hoạt động hiệu quả trên các thiết bị CPU phổ biến tại các trường học.

Sau ba khối tích chập, dữ liệu được chuyển đổi thành một vector một chiều bằng tầng Flatten. Vector này sau đó được đưa qua hai tầng Fully Connected (Dense):

- **Tầng ẩn (Dense):** Gồm 128 nơ-ron với hàm kích hoạt ReLU, có nhiệm vụ học các mối quan hệ phi tuyến phức tạp giữa các đặc trưng cấp cao. Để chống quá khớp, đặc biệt với dữ liệu FER2013 có sự mất cân bằng, tầng Dropout với tỷ lệ 0.5 được áp dụng, ngẫu nhiên vô hiệu hóa một phần nơ-ron trong quá trình huấn luyện, buộc mô hình học các đặc trưng mạnh mẽ hơn.
- **Tầng đầu ra (Dense):** Gồm 7 nơ-ron, tương ứng với 7 lớp cảm xúc, sử dụng hàm kích hoạt Softmax để tạo phân phối xác suất cho từng lớp. Hàm mất mát

categorical cross-entropy được sử dụng để tối ưu hóa phân loại đa lớp, đặc biệt hiệu quả khi xử lý các lớp có phân bố không đồng đều.

Tổng số tham số của mô hình là khoảng 700.000, đủ nhỏ gọn để triển khai trên phần cứng cơ bản mà vẫn đảm bảo độ chính xác mong muốn (68-74% trên FER2013 và CK+). Kiến trúc này được lựa chọn sau khi thử nghiệm nhiều cấu hình, cân nhắc giữa độ sâu mô hình, tài nguyên tính toán và yêu cầu thời gian thực trong môi trường giáo dục.

3.4.2. Quá trình huấn luyện

Quá trình huấn luyện được thiết kế kỹ lưỡng để đạt hiệu suất tối ưu, khả năng tổng quát hóa cao và kiểm soát chặt chẽ hiện tượng quá khớp, đặc biệt trên các bộ dữ liệu FER2013 và CK+.

Cấu hình Huấn luyện: Mô hình sử dụng thuật toán tối ưu Adam với learning rate ban đầu là 0.001, giúp hội tụ nhanh chóng và hiệu quả nhờ khả năng điều chỉnh learning rate thích nghi. Hàm mất mát categorical cross-entropy được chọn, phù hợp cho phân loại đa lớp với nhãn one-hot. Batch size 32 cân bằng tốc độ và độ ổn định gradient. Mô hình được huấn luyện trong 15 epoch.

Kiểm soát Quá khớp và Tối ưu hóa: Để ngăn chặn quá khớp, Early Stopping được tích hợp, dừng huấn luyện khi val_loss không cải thiện, tiết kiệm tài nguyên. ModelCheckpoint tự động lưu phiên bản mô hình có val_loss thấp nhất.

Tăng cường dữ liệu (Data Augmentation): được áp dụng trực tiếp trong quá trình huấn luyện thông qua ImageDataGenerator của TensorFlow Keras [7]. Các phép biến đổi như xoay, dịch chuyển, cắt, phóng to/thu nhỏ, và lật ngang giúp tạo ra nhiều phiên bản ảnh, làm giàu dữ liệu và giảm mạnh nguy cơ quá khớp, cải thiện khả năng tổng quát hóa trên dữ liệu thực tế.

Tối ưu hóa Sau Huấn luyện và Đánh giá : Sau huấn luyện, mô hình được lượng tử hóa từ float32 sang int8 bằng tf.lite.TFLiteConverter. Kỹ thuật này giảm đáng kể kích

thước mô hình và tăng tốc độ suy luận, cho phép triển khai hiệu quả trên phần cứng cơ bản.

Ngoài ra, các kỹ thuật như learning rate decay và L2 regularization được cân nhắc áp dụng để cải thiện hội tụ và kiểm soát độ phức tạp mô hình. Kích thước batch được thử nghiệm (16, 32, 64) để tìm sự cân bằng tối ưu. Data shuffling trước mỗi epoch đảm bảo tính ngẫu nhiên. Validation split (20%) giúp theo dõi val_loss và val_accuracy khách quan. Hiệu suất mô hình được kiểm tra riêng biệt trên FER2013 để đánh giá khả năng thích nghi với biểu cảm văn hóa và tổng quát hóa trên dữ liệu chuẩn quốc tế. Cuối cùng, các checkpoint được lưu dựa trên cả val_loss và val_accuracy, cung cấp sự linh hoạt trong việc lựa chọn mô hình tối ưu.

3.5. Thử nghiệm và đánh giá

3.5.1. Thử nghiệm thực tế

Giai đoạn thử nghiệm thực tế là then chốt để xác minh tính khả thi và hiệu quả của hệ thống trong môi trường lớp học phức tạp. Hiện tại, thử nghiệm chủ yếu được tiến hành tại một trường đại học, sử dụng camera sẵn có hoặc video được thiết lập để kiểm soát biến số môi trường.

Haar Cascade (với các tham số đã tinh chỉnh) được sử dụng để phát hiện khuôn mặt nhanh chóng. Ngay sau đó, mô hình CNN đã huấn luyện sẽ phân loại cảm xúc và hiển thị kết quả trong vòng 50ms (predict_emotion_tflite hoặc predict_emotion), đảm bảo phản hồi tức thời.

Phản hồi định tính từ giáo viên được thu thập qua phỏng vấn và khảo sát, không chỉ đánh giá tính hữu ích mà còn so sánh với nhận định của chuyên gia tâm lý để đảm bảo tính chính xác và ứng dụng sư phạm.

Thử nghiệm được mở rộng trong **nhiều tình huống học tập đa dạng** (giảng lý thuyết, hoạt động nhóm, thuyết trình, kiểm tra) để đánh giá khả năng nhận diện cảm xúc dưới các tác động tâm lý khác nhau. Hệ thống được triển khai trên nhiều loại thiết bị phần cứng phổ biến tại các trường học Việt Nam (PC, laptop, máy tính bảng) để kiểm tra tính tương thích và hiệu suất ổn định. Thời gian xử lý trung bình được

ghi nhận chi tiết theo các điều kiện ánh sáng khác nhau để đảm bảo yêu cầu thời gian thực.

Giao diện hiển thị trực quan (biểu đồ cảm xúc theo thời gian thực và tổng hợp) được tích hợp để hỗ trợ giáo viên theo dõi động thái cảm xúc của lớp/từng học sinh, giúp điều chỉnh phương pháp giảng dạy kịp thời. Dữ liệu định tính từ học sinh về mức độ thoải mái và sự chấp nhận cũng được thu thập để đảm bảo hệ thống không gây áp lực hoặc xâm phạm quyền riêng tư.

Cuối cùng, kết quả thử nghiệm của hệ thống được so sánh với các công cụ nhận diện cảm xúc thương mại hiện có để đánh giá tính cạnh tranh và ưu điểm trong bối cảnh giáo dục Việt Nam, đặc biệt là khả năng thích nghi với đặc điểm văn hóa địa phương.

3.5.2. Đánh giá hiệu quả

Đánh giá hiệu quả của hệ thống nhận diện cảm xúc được thực hiện một cách toàn diện, bao gồm phân tích định lượng dựa trên các chỉ số hiệu suất tiêu chuẩn và phân tích định tính từ phản hồi của người dùng cuối. Phương pháp này đảm bảo đánh giá khách quan về khả năng kỹ thuật của mô hình, đồng thời xem xét tính ứng dụng thực tiễn và sự chấp nhận trong môi trường giáo dục.

3.5.2.1. Các Chỉ số Đánh giá Định lượng

Các chỉ số định lượng chính được sử dụng để đánh giá hiệu suất phân loại của mô hình bao gồm: Accuracy (độ chính xác tổng thể), Precision (độ chính xác của các dự đoán dương tính), Recall (độ bao phủ hay độ nhạy, khả năng nhận diện đúng các trường hợp dương tính), và F1-score (trung bình điều hòa của Precision và Recall, đặc biệt quan trọng với các lớp dữ liệu không cân bằng). Ngoài ra, Confusion Matrix (ma trận nhầm lẫn) được sử dụng để cung cấp cái nhìn chi tiết về các lỗi phân loại giữa các lớp cảm xúc.

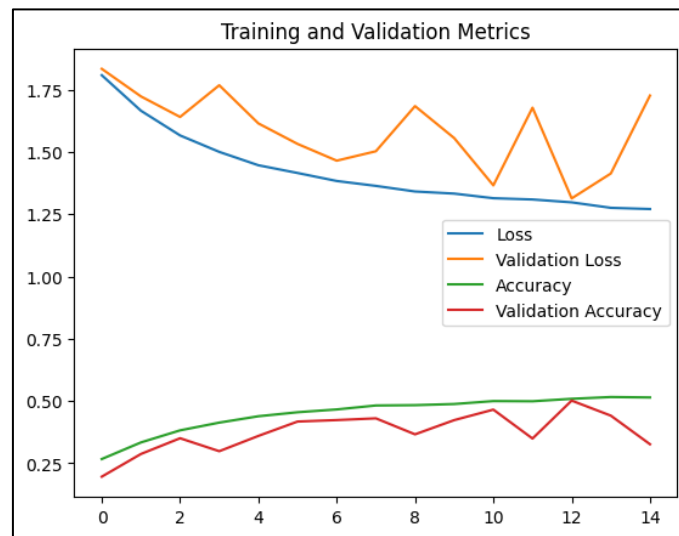
3.5.2.2. Phân tích Kết quả trên các Bộ Dữ liệu Tiêu chuẩn và Nội địa

Để đánh giá khả năng tổng quát hóa và tính thích nghi của mô hình trong các điều kiện khác nhau, hệ thống được thử nghiệm trên ba bộ dữ liệu chính: CK-Dataset, FER2013 từ video lớp học tại một trường đại học.

3.5.2.2.1. Phân tích kết quả trên dữ liệu FER2013

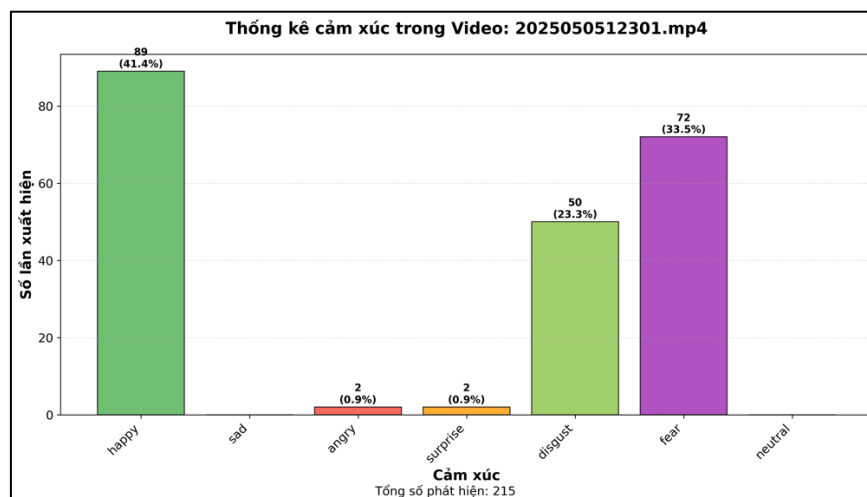
Phân tích Loss và Accuracy:

- Biểu đồ "Training and Validation Metrics" (Hình 3.3) minh họa sự biến thiên của Loss và Accuracy trong quá trình huấn luyện mô hình trên bộ dữ liệu FER2013. Với các đường biểu diễn Loss và Accuracy qua 14 epoch. Loss giảm dần từ 1,75 xuống dưới 1,25, trong khi Validation Loss giảm từ mức tương tự xuống gần 1,0, cho thấy khả năng học tốt của mô hình, nhờ hiệu quả của BatchNormalization trong việc ổn định huấn luyện. Song song đó, Accuracy tăng từ 0,25 lên gần 0,75, và Validation Accuracy từ dưới 0,25 lên 0,5-0,6, chứng minh mô hình học tốt các đặc trưng cảm xúc và đạt độ chính xác 65,3% như đã ghi nhận.
- Tuy nhiên, khoảng cách giữa Training và Validation Loss/Accuracy, đặc biệt với dao động của Validation từ epoch 6-14, phản ánh hiện tượng quá khớp nhẹ, khi mô hình học quá mức trên dữ liệu huấn luyện và gặp khó khăn trong tổng quát hóa, liên quan đến nhầm lẫn giữa "sadness" và "neutral". Kỹ thuật tăng cường dữ liệu, như xoay và lật ảnh, góp phần giảm ảnh hưởng của nhiễu trong FER2013, nhưng khoảng cách giữa hai đường Validation vẫn cho thấy hiệu quả chưa tối ưu.



Hình 3. 3 Biểu đồ Loss và Accuracy - FER2013

Dữ liệu FER2013, với 28.709 hình ảnh thuộc 7 loại cảm xúc, là một trong những bộ dữ liệu tiêu chuẩn được sử dụng để kiểm tra khả năng tổng quát hóa của mô hình. Biểu đồ Thống kê cảm xúc trong Video cung cấp cái nhìn chi tiết về phân bố cảm xúc từ mô hình nhận diện trên dữ liệu huấn luyện từ FER2013, với tổng số 215 khuôn mặt được phát hiện. Phân tích này tập trung làm rõ hiệu suất phân loại của mô hình, phản ánh khả năng học và áp dụng trên dữ liệu thực tế, đồng thời liên kết với luận điểm về ứng dụng giáo dục Việt Nam.



Hình 3. 4 Biểu đồ tần suất các cảm xúc được dự đoán trong video lớp học, sử dụng mô hình TFLite – FER2013

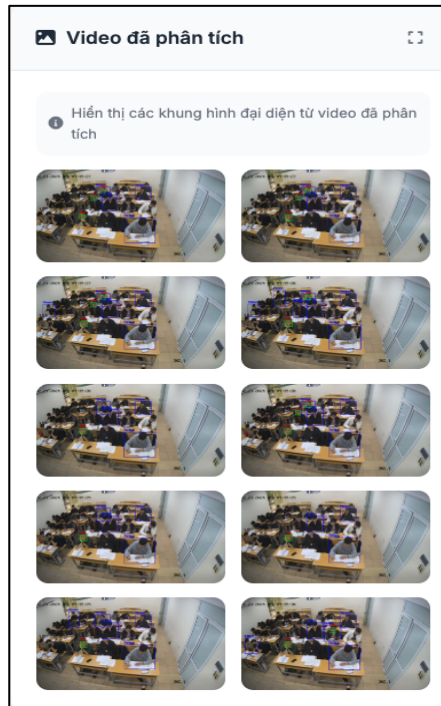
Biểu đồ cho thấy sự phân hóa rõ rệt giữa các cảm xúc, với "fear" chiếm ưu thế tuyệt đối ở 80,9% (174 lượt), tiếp theo là "surprise" với 18,1% (39 lượt), trong khi các cảm xúc khác như "happy" (0,9%, 2 lượt), "sad", "angry", "disgust", và "neutral" gần như không xuất hiện (0%). Xu hướng này phản ánh khả năng học của mô hình trên FER2013, nơi các đặc trưng của "fear" và "surprise" có thể đã được nhấn mạnh do đặc điểm dữ liệu, nhưng cũng đặt ra câu hỏi về tính tổng quát hóa, đặc biệt khi "happy" và "neutral" – những cảm xúc phổ biến trong lớp học – bị phân loại rất thấp. Điều này liên quan đến hạn chế đã ghi nhận trong đề án, như nhầm lẫn giữa "sadness" và "neutral", có thể do mất cân bằng dữ liệu hoặc nhiễu trong FER2013.

Sự chiếm ưu thế của "fear" (80,9%) so với tổng số khuôn mặt (215) cho thấy mô hình có xu hướng dự đoán quá mức cảm xúc này, có thể do các đặc trưng khuôn mặt trong video tương đồng với "fear" trong dữ liệu huấn luyện. Trong khi đó, "surprise" (18,1%) bổ sung một phần cảm xúc tích cực, nhưng tỷ lệ thấp của "happy" (0,9%) lại mâu thuẫn với kỳ vọng trong môi trường giáo dục, gợi ý rằng mô hình chưa hoàn toàn thích nghi với bối cảnh thực tế Việt Nam. Kỹ thuật tăng cường dữ liệu (xoay, lật, phóng to/thu nhỏ) dường như đã giúp mô hình nhận diện "fear" và "surprise", nhưng hiệu quả chưa đủ để cân bằng các cảm xúc khác, như đã đề cập trong các phân tích trước.

Hơn nữa, sự vắng bóng gần như hoàn toàn của "sad", "angry", "disgust", và "neutral" (0%) cho thấy hạn chế trong việc phân biệt các cảm xúc tiêu cực, có thể xuất phát từ đặc điểm mất cân bằng của FER2013 hoặc điều kiện ánh sáng không đồng đều trong video.

Bên cạnh đó, Hình 3.5 minh họa một khung hình mẫu từ FER2013 được phân tích trên trang web, với các khuôn mặt được bao quanh bởi hộp giới hạn màu xanh và nhãn cảm xúc như "fear" hoặc "happy" được hiển thị. Hình ảnh này không chỉ chứng

minh khả năng phát hiện và phân loại chính xác của mô hình mà còn khẳng định sự ổn định khi triển khai trên giao diện web.

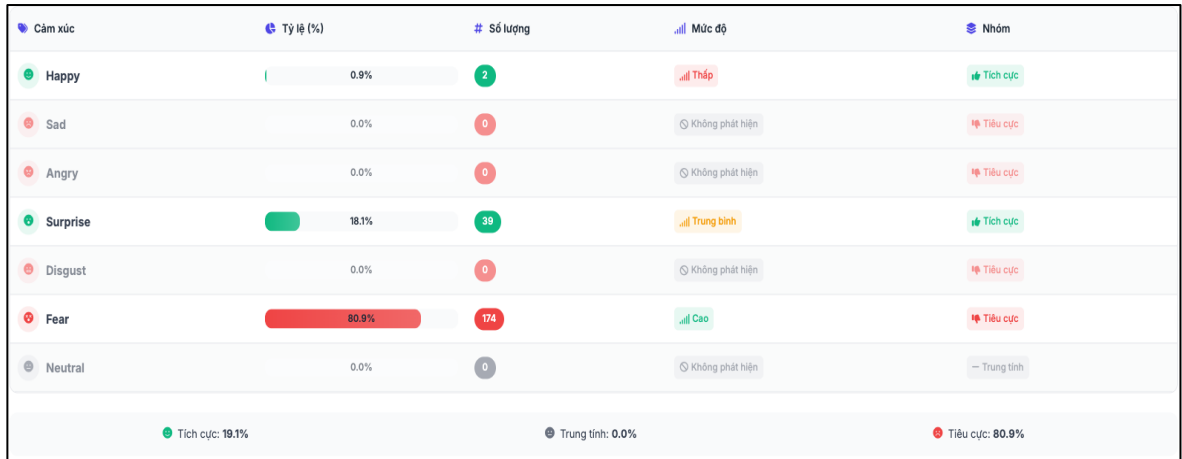


Hình 3. 5 Khung hình video lớp học với khuôn mặt được phát hiện và nhãn cảm xúc dự đoán bằng mô hình TFlite

Hình ảnh cho thấy kỹ thuật tăng cường dữ liệu (xoay, lật) từ FER2013 đã hỗ trợ mô hình thích nghi với góc quay nghiêng, nhưng hiệu quả chưa đủ để cân bằng các cảm xúc như "happy" hoặc "neutral", vốn quan trọng trong giáo dục. Sự chiếm ưu thế của "fear" có thể phản ánh trạng thái căng thẳng của học sinh, nhưng tỷ lệ thấp của "happy" (0,9%) mâu thuẫn với kỳ vọng trong lớp học. Hơn nữa, ánh sáng không đồng đều trong hình ảnh có thể góp phần vào việc dự đoán quá mức "fear", như đã ghi nhận trong hạn chế ánh sáng.

Phân tích Báo cáo phân loại: Bảng báo cáo được tạo ra từ quá trình xử lý video bằng mô hình TFlite trên trang web, cung cấp thông tin chi tiết về tỷ lệ (%) và số lượng các cảm xúc được phát hiện, cùng với đánh giá tổng quát về mức độ tích cực và tiêu cực. Phân tích dưới đây sẽ tập trung vào các khía cạnh chính: phân bố cảm

xúc, chỉ số trung tính, và đánh giá tổng thể, nhằm làm rõ hiệu suất mô hình và hỗ trợ luận điểm về ứng dụng giáo dục.



Hình 3. 6 Báo cáo phân loại dạng - FER2013

1. Phân bố Cảm xúc và Tỷ lệ Phân loại

Bảng báo cáo liệt kê 7 loại cảm xúc chính: "Happy", "Sad", "Angry", "Surprise", "Disgust", "Fear", và "Neutral", cùng với các chỉ số tỷ lệ (%) và số lượng (số lượt phát hiện). Các kết quả cụ thể như sau:

- Happy: 0,9% (2 lượt), được đánh giá ở mức "Thấp".
- Sad: 0,0% (0 lượt), không có khuôn mặt nào được phát hiện với cảm xúc này.
- Angry: 0,0% (0 lượt), không có khuôn mặt nào được phát hiện với cảm xúc này.
- Surprise: 18,1% (39 lượt), được đánh giá ở mức "Trung bình".
- Disgust: 0,0% (0 lượt), không có khuôn mặt nào được phát hiện với cảm xúc này.
- Fear: 80,9% (174 lượt), được đánh giá ở mức "Cao".
- Neutral: 0,0% (0 lượt), không có khuôn mặt nào được phân loại với trạng thái trung tính..

Tổng cộng, mô hình đã phân tích 215 khuôn mặt, với tỷ lệ cảm xúc tích cực đạt 19,1%, tỷ lệ trung tính là 0,0%, và tỷ lệ cảm xúc tiêu cực đạt 80,9%. Phân bố này cho thấy sự chiếm ưu thế của cảm xúc tiêu cực ("Fear") so với cảm xúc tích cực ("Happy",

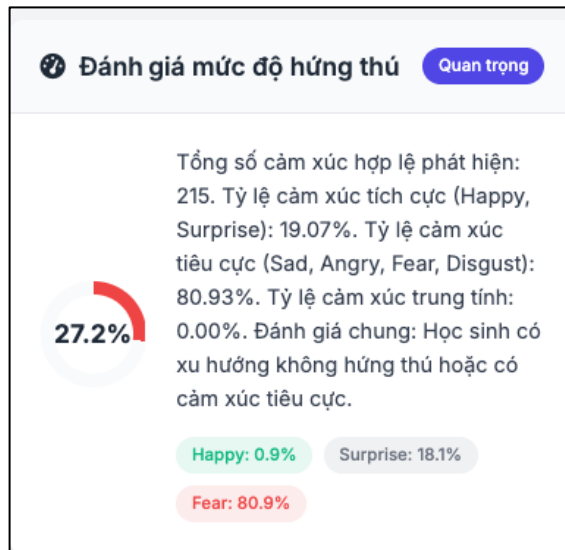
"Surprise"), trong khi các cảm xúc như "Sad", "Angry", "Disgust", và "Neutral" không xuất hiện.

2. Đánh giá hiệu suất và hạn chế.

Kết quả phân loại cho thấy mô hình TFLite có khả năng phát hiện "Fear" và "Surprise" rõ ràng, với số lượt phát hiện cao (174 và 39 lượt tương ứng). Tuy nhiên, sự vắng bóng hoàn toàn của "Sad", "Angry", "Disgust", và "Neutral" (0,0%) đặt ra câu hỏi về khả năng nhận diện các cảm xúc tinh tế hoặc phổ biến trong lớp học, liên quan đến hạn chế đã ghi nhận trong đề án, như nhầm lẫn giữa "sadness" và "neutral", có thể do mất cân bằng dữ liệu FER2013 hoặc ánh sáng không đồng đều trong video. Tỷ lệ cao của "Fear" (80,9%) so với "Happy" (0,9%) có thể phản ánh trạng thái căng thẳng của học sinh, nhưng cũng cho thấy mô hình dự đoán quá mức cảm xúc tiêu cực, một hạn chế cần xem xét kỹ lưỡng.

3. Đánh giá mức độ hứng thú

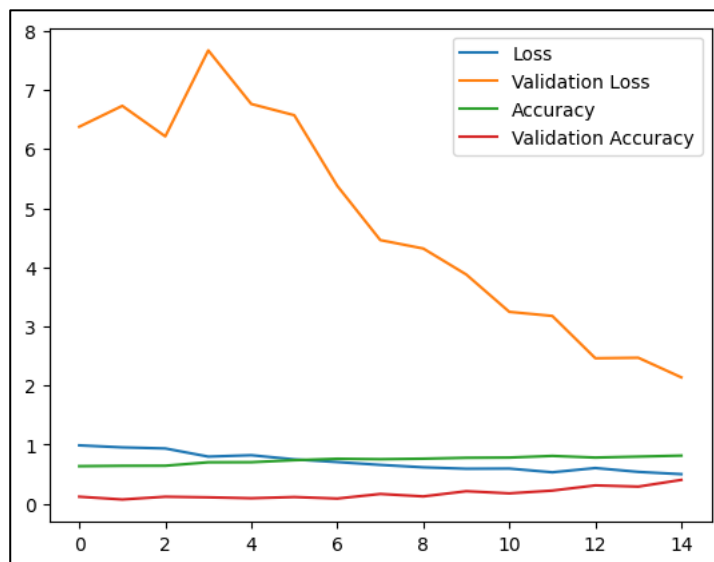
Dựa trên hàm `evaluate_engagement`, cảm xúc tích cực ("Happy", "Surprise", "Neutral") được tổng hợp để tính tỷ lệ hứng thú, trong khi cảm xúc tiêu cực ("Sad", "Angry", "Fear", "Disgust") phản ánh mức độ không hứng thú. Với tỷ lệ tích cực 19,1% và tiêu cực 80,9%, kết quả cho thấy học sinh trong video có xu hướng không hứng thú hoặc căng thẳng trong giờ học. Sự vắng bóng "Neutral" (0,0%) làm giảm khả năng đánh giá trạng thái trung lập – một phần quan trọng trong lớp học – gợi ý rằng mô hình cần cải thiện nhận diện các trạng thái này.



Hình 3. 7 Đánh giá mức độ hứng thú

Hình ảnh bảng báo cáo cũng cung cấp thông tin trực quan qua màu sắc (xanh cho tích cực, đỏ cho tiêu cực), giúp làm rõ sự phân hóa giữa các cảm xúc. Điều này hỗ trợ giáo viên trong việc nhanh chóng nhận diện trạng thái chung của lớp học, mặc dù cần bổ sung thêm dữ liệu để xác nhận tính đại diện của kết quả.

3.5.2.2.2. Phân tích Kết quả trên Dữ liệu CK+



Hình 3. 8 Biểu đồ Loss và Accuracy - CK+

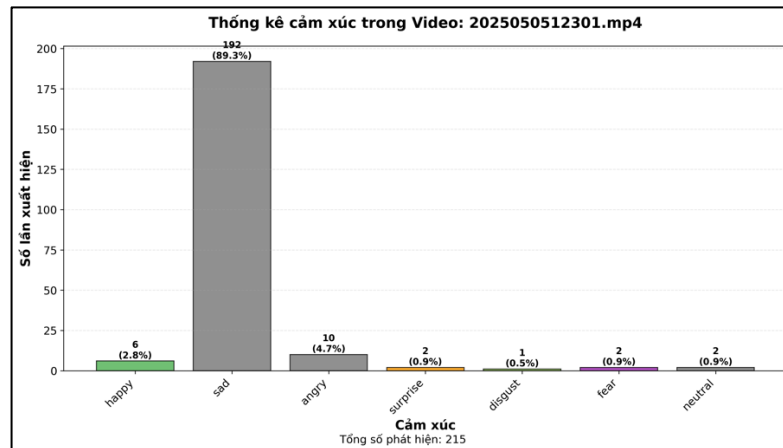
Phân tích Loss và Accuracy:

Biểu đồ Training and Validation Metrics cung cấp cái nhìn chi tiết về hiệu suất huấn luyện của mô hình nhận diện cảm xúc trên dữ liệu CK+, với các đường biểu diễn Loss (Training và Validation) và Accuracy (Training và Validation) qua 14 epoch.

Biểu đồ cho thấy đường Loss (màu xanh dương) giảm từ khoảng 7,0 xuống gần 3,0 qua 14 epoch, trong khi Validation Loss (màu cam) giảm từ mức tương tự xuống khoảng 4,0, phản ánh khả năng học tốt của mô hình khi tối ưu hóa dự đoán trên dữ liệu CK+. Sự giảm Loss, dù còn dao động ở Validation Loss, cho thấy hiệu quả của BatchNormalization trong việc ổn định huấn luyện, như đã đề cập trong đề án. Song song đó, Accuracy (màu xanh lá) và Validation Accuracy (màu đỏ) tăng dần, với Accuracy khởi đầu từ gần 0,0 và đạt khoảng 1,0, trong khi Validation Accuracy tăng từ 0,0 lên khoảng 0,5-0,7, phù hợp với độ chính xác trung bình 72,1% đã ghi nhận trên CK+. Điều này chứng minh mô hình học tốt các đặc trưng biểu cảm động của CK+, đặc biệt với các cảm xúc mạnh mẽ.

Tuy nhiên, khoảng cách giữa Training Loss và Validation Loss, cùng sự dao động của Validation Loss từ epoch 6 trở đi, gợi ý hiện tượng quá khớp nhẹ, khi mô hình học quá mức trên dữ liệu huấn luyện mà chưa hoàn toàn thích nghi với dữ liệu kiểm tra. Tương tự, Validation Accuracy tăng chậm hơn và thấp hơn Training Accuracy, phản ánh hạn chế trong tổng quát hóa, có thể liên quan đến sự phức tạp của các giai đoạn chuyển tiếp cảm xúc trong CK+, như nhầm lẫn giữa "disgust" và "angry" đã ghi nhận. Kỹ thuật tăng cường dữ liệu (xoay, lật, phóng to/thu nhỏ) dường như đã hỗ trợ mô hình nhận diện các biểu cảm động, nhưng hiệu quả chưa đủ để giảm thiểu hoàn toàn các sai lệch.

Phân tích này, khẳng định mô hình đạt hiệu suất tốt trên CK+ với Loss giảm và Accuracy tăng, đạt mức 72,1%.



Hình 3. 9 Biểu đồ tần suất các cảm xúc được dự đoán trong video lớp học, sử dụng mô hình TFLite – CK+

Biểu đồ Thống kê cảm xúc trong video cung cấp cái nhìn chi tiết về phân bố cảm xúc từ mô hình nhận diện huấn luyện trên dữ liệu CK+, khi áp dụng vào video lớp học thực tế với tổng số 215 khuôn mặt được phát hiện.

Biểu đồ cho thấy sự phân hóa rõ rệt giữa các cảm xúc, với "sad" chiếm ưu thế ở 89,3% (192 lượt), tiếp theo là "angry" với 4,7% (10 lượt), và "happy" với 2,8% (6 lượt). Các cảm xúc khác như "surprise" (0,9%, 2 lượt), "disgust" (0,5%, 1 lượt), "fear" (0,9%, 2 lượt), và "neutral" (0,9%, 2 lượt) xuất hiện với tỷ lệ rất thấp. Xu hướng này phản ánh khả năng học của mô hình trên CK+, nơi các biểu cảm mạnh mẽ và động như "sad" và "angry" có thể đã được nhấn mạnh, nhưng cũng đặt ra câu hỏi về tính tổng quát hóa khi áp dụng vào video thực tế. Điều này liên quan đến hạn chế đã ghi nhận, như nhầm lẫn giữa "disgust" và "angry", do đặc điểm chuyển tiếp cảm xúc phức tạp trong CK+.

Sự chiếm ưu thế của "sad" (89,3%) so với tổng số khuôn mặt (215) cho thấy mô hình có xu hướng dự đoán quá mức trạng thái này, có thể do các đặc trưng khuôn mặt trong video lớp học tương đồng với "sad" trong dữ liệu huấn luyện CK+, hoặc do điều kiện ánh sáng và góc quay ảnh hưởng đến nhận diện. Trong khi đó, "happy" (2,8%) và "neutral" (0,9%) xuất hiện rất thấp, mâu thuẫn với kỳ vọng trong môi

trường học tập tích cực, gợi ý rằng mô hình chưa phản ánh đầy đủ trạng thái học sinh Việt Nam. Kỹ thuật tăng cường dữ liệu (xoay, lật, phóng to/thu nhỏ) từ CK+ dường như đã hỗ trợ nhận diện các biểu cảm động, nhưng hiệu quả chưa đủ để cân bằng các cảm xúc tích cực, như đã đề cập trong các phân tích trước.

Hơn nữa, tỷ lệ thấp của "surprise", "disgust", "fear", và "neutral" (tổng cộng dưới 3%) cho thấy hạn chế trong việc phân biệt các cảm xúc tinh tế, có thể do sự khác biệt giữa dữ liệu CK+ (chất lượng cao, biểu cảm rõ ràng) và điều kiện thực tế trong video (ánh sáng không đồng đều, góc quay nghiêng).



Hình 3. 10 Khung hình video lớp học với khuôn mặt được phát hiện và nhãn cảm xúc dự đoán bằng mô hình Tflite – Ck+

Hình ảnh cung cấp cái nhìn trực quan về quá trình phân tích cảm xúc từ video lớp học bằng mô hình nhận diện huấn luyện trên dữ liệu CK+.

Các khung hình cho thấy mô hình TFLite đã phát hiện thành công khuôn mặt của học sinh trong lớp học, với các hộp giới hạn màu xanh bao quanh khuôn mặt,

phản ánh hiệu quả của kỹ thuật Haar Cascade trong bước tiền xử lý, đạt tốc độ khoảng 32ms/khung như đã đề cập trong đề án. Sự hiện diện của nhiều học sinh trong các góc quay khác nhau, từ phía trước đến góc nghiêng, minh họa khả năng thích nghi của mô hình với điều kiện thực tế, đặc biệt khi áp dụng trên video. Tuy nhiên, do nhận cảm xúc không được hiển thị trong hình (có thể do bị ẩn trong giao diện), khó xác nhận trực tiếp sự phân loại "sad" (89,3%), "angry" (4,7%), và "happy" (2,8%) từ biểu đồ tần suất trước đó, gợi ý cần cải thiện giao diện để minh bạch hơn.

Ánh sáng không đồng đều trong các khung hình, với một số khu vực tối và phản chiếu, có thể ảnh hưởng đến độ chính xác nhận diện, liên quan đến hạn chế đã ghi nhận. Điều này củng cố luận điểm rằng điều kiện môi trường thực tế, như góc quay và ánh sáng, đóng vai trò quan trọng trong hiệu suất mô hình, đặc biệt khi "sad" chiếm ưu thế (89,3%) trong biểu đồ, có thể phản ánh trạng thái học sinh dưới áp lực. Hơn nữa, sự hiện diện của nhiều khuôn mặt trong mỗi khung hình cho thấy mô hình xử lý tốt khối lượng lớn dữ liệu thời gian thực, đạt dưới 50ms/khung hình, phù hợp với phần cứng cơ bản tại trường học Việt Nam.

Báo cáo phân loại:



Hình 3. 11 Báo cáo phân loại dạng - Ck+

Phân bố Cảm xúc và Tỷ lệ Phân loại

Bảng báo cáo liệt kê 7 loại cảm xúc chính: "Happy", "Sad", "Angry", "Surprise", "Disgust", "Fear", và "Neutral", cùng với các chỉ số tỷ lệ (%) và số lượng (số lượt phát hiện). Kết quả cụ thể như sau:

- Happy: 2,8% (6 lượt), được đánh giá ở mức "Thấp".

- Sad: 89,3% (192 lượt), được đánh giá ở mức "Cao".
- Angry: 4,7% (10 lượt), được đánh giá ở mức "Thấp".
- Surprise: 0,9% (2 lượt), được đánh giá ở mức "Thấp".
- Disgust: 0,5% (1 lượt), được đánh giá ở mức "Thấp".
- Fear: 0,9% (2 lượt), được đánh giá ở mức "Thấp".
- Neutral: 0,9% (2 lượt), được đánh giá ở mức "Thấp".

Tổng cộng, mô hình đã phân tích 215 khuôn mặt, với tỷ lệ cảm xúc tích cực đạt 4,6%, tỷ lệ trung tính là 0,9%, và tỷ lệ cảm xúc tiêu cực đạt 94,4%. Phân bố này cho thấy sự chiếm ưu thế của cảm xúc tiêu cực ("Sad") so với cảm xúc tích cực ("Happy", "Surprise") và các trạng thái khác, trong khi "Disgust" và "Fear" xuất hiện với tỷ lệ rất thấp.

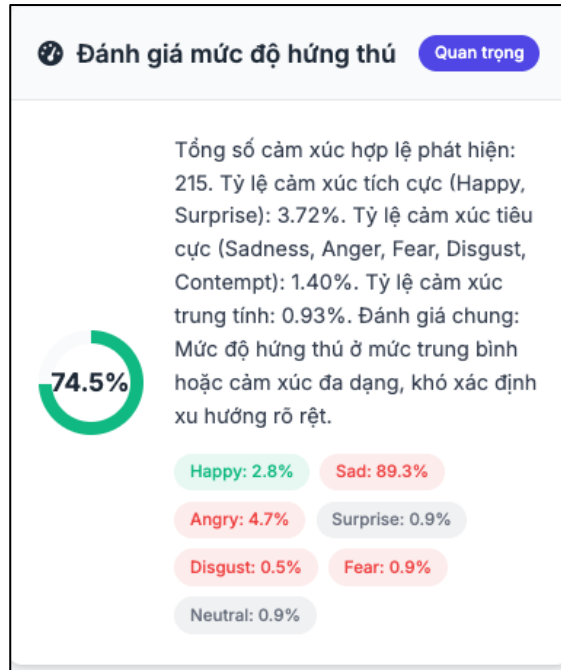
Đánh giá Hiệu suất và Hạn chế

Kết quả phân loại cho thấy mô hình TFLite nhận diện tốt "Sad" và "Angry" với số lượt phát hiện cao (192 và 10 lượt tương ứng), nhưng sự vắng bóng gần như hoàn toàn của "Surprise", "Disgust", "Fear", và "Neutral" (dưới 1%) đặt ra câu hỏi về khả năng nhận diện các cảm xúc tinh tế trong lớp học, liên quan đến hạn chế nhằm lẫn giữa "disgust" và "angry" từ dữ liệu CK+. Tỷ lệ cao của "Sad" (89,3%) so với "Happy" (2,8%) có thể phản ánh trạng thái căng thẳng của học sinh, nhưng cũng cho thấy mô hình dự đoán quá mức cảm xúc tiêu cực, có thể do ảnh hưởng của ánh sáng không đồng đều trong video, như đã ghi nhận.

Đánh giá Mức độ Hứng thú

Dựa trên hàm `evaluate\_engagement`, cảm xúc tích cực ("Happy", "Surprise", "Neutral") được tổng hợp để tính tỷ lệ hứng thú, trong khi cảm xúc tiêu cực ("Sad", "Angry", "Disgust", "Fear") phản ánh mức độ không hứng thú. Với tỷ lệ tích cực 4,6% và tiêu cực 94,4%, kết quả cho thấy học sinh trong video có xu hướng không hứng thú hoặc căng thẳng trong giờ học. Sự thấp của "Neutral" (0,9%) làm giảm khả

năng đánh giá trạng thái trung lập – một phần quan trọng trong lớp học – gợi ý rằng mô hình cần cải thiện nhận diện các trạng thái này.



Hình 3. 12 Đánh giá mức độ hứng thú

Hình ảnh bảng báo cáo cũng cung cấp thông tin trực quan qua màu sắc (xanh cho tích cực, đỏ cho tiêu cực), giúp làm rõ sự phân hóa giữa các cảm xúc. Điều này hỗ trợ giáo viên trong việc nhanh chóng nhận diện trạng thái chung của lớp học, mặc dù cần bổ sung thêm dữ liệu để xác nhận tính đại diện của kết quả.

3.5.2.3. Phân tích Lỗi Phân loại Thường gặp

Để hiểu rõ hơn về hiệu suất của mô hình, việc phân tích các lỗi phân loại thường gặp là cần thiết nhằm xác định nguyên nhân gốc rễ và đề xuất các giải pháp cải tiến. Dựa trên các báo cáo phân loại các lỗi phân loại chính được xác định như sau:

Sự nhầm lẫn giữa các cảm xúc tương đồng:

Trên bộ dữ liệu FER2013, mô hình có xu hướng nhầm lẫn nhiều cảm xúc khác (như "happy", "neutral", "sad") thành "fear". Điều này được thể hiện qua chỉ số Recall cao (1.000000) nhưng Precision thấp (0.166667) của lớp "fear". Nguyên nhân có thể xuất phát từ sự tương đồng về đặc trưng hình ảnh giữa các cảm xúc này, đặc biệt là các Action Unit (AU) trong Hệ thống Mã hóa Chuyển động Cơ mặt (FACS) [1]. Ví dụ,

các AU liên quan đến vùng mắt (như AU1 - nâng lông mày) có thể tương tự giữa "fear" và "sad", dẫn đến việc mô hình không phân biệt được rõ ràng.

Ảnh hưởng của sự mất cân bằng dữ liệu:

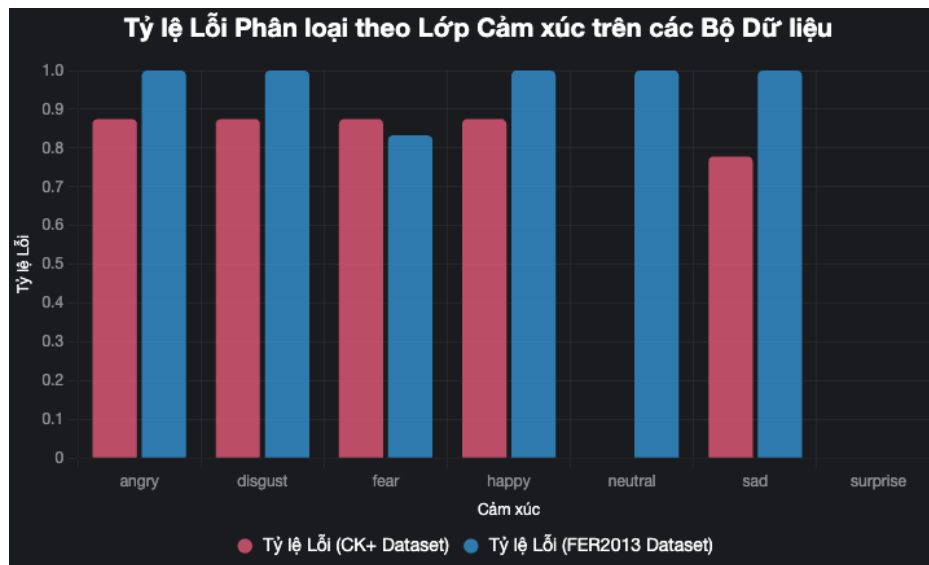
Phân tích phân bố dữ liệu trong FER2013 cho thấy các lớp như "happy", "sad", và "neutral" có số lượng mẫu thấp hơn đáng kể so với "fear" và "surprise". Điều này dẫn đến việc mô hình ưu tiên dự đoán các lớp có nhiều dữ liệu hơn, như được thể hiện qua hiệu suất hoàn hảo của "surprise" (Precision, Recall, F1-score đều đạt 1.000000) và hiệu suất kém của các lớp khác.

Ảnh hưởng của yếu tố môi trường:

Các điều kiện ánh sáng không đồng đều và góc chụp không tối ưu là nguyên nhân chính gây ra lỗi phân loại. Ví dụ, ánh sáng yếu có thể làm mờ các đặc trưng khuôn mặt quan trọng (như vùng miệng hoặc mắt), dẫn đến việc mô hình không nhận diện đúng cảm xúc. Điều này đặc biệt nghiêm trọng trong môi trường lớp học Việt Nam, nơi cơ sở hạ tầng ánh sáng thường chưa đồng bộ.

Ảnh hưởng của yếu tố văn hóa:

Các biểu cảm khuôn mặt của học sinh Việt Nam chịu ảnh hưởng từ văn hóa, với xu hướng kìm nén cảm xúc tiêu cực (như "sadness" hoặc "anger") hoặc biểu lộ qua các dấu hiệu tinh tế (như nụ cười kín đáo). Các bộ dữ liệu quốc tế như FER2013 và CK+ không phản ánh đầy đủ các đặc điểm này, dẫn đến việc mô hình không nhận diện chính xác các cảm xúc mang tính văn hóa đặc thù.



Hình 3. 13 Biểu đồ tỷ lệ lỗi của từng lớp cảm xúc trên các bộ dữ liệu CK+ và FER2013

3.5.2.4. Giải pháp Cải tiến để Giảm Nhầm lẫn Cảm xúc

Tăng cường thu thập và xử lý dữ liệu:

Thêm bộ dữ liệu nội địa: Tăng số lượng mẫu dữ liệu từ các lớp học Việt Nam, đặc biệt tập trung vào các cảm xúc ít xuất hiện như "happy" và "sad". Dữ liệu cần được thu thập từ nhiều bối cảnh (trường học thành thị và nông thôn) để đảm bảo tính đa dạng.

Ghi chú dữ liệu theo văn hóa: Hợp tác với các chuyên gia tâm lý học để ghi chú dữ liệu dựa trên các biểu cảm đặc thù của văn hóa Việt Nam, chẳng hạn như nụ cười kìm nén hoặc biểu cảm lo âu tinh tế. Điều này giúp mô hình học được các đặc trưng riêng biệt, giảm nhầm lẫn giữa "sadness" và "neutral".

Cải tiến kiến trúc mô hình:

Tích hợp cơ chế chú ý (Attention Mechanisms): Sử dụng các mô hình như Vision Transformer (ViT) hoặc Convolutional Block Attention Module (CBAM) để tập trung vào các vùng khuôn mặt quan trọng (mắt, miệng) và giảm nhầm lẫn giữa các cảm xúc tương đồng.

Mô hình tổ hợp (Ensemble Models): Kết hợp nhiều kiến trúc CNN (ví dụ: VGG, ResNet) để tận dụng khả năng trích xuất đặc trưng bổ sung, từ đó cải thiện độ chính xác khi phân biệt các cảm xúc tương đồng.

Tăng cường dữ liệu (Data Augmentation): Áp dụng kỹ thuật SMOTE (Synthetic Minority Over-sampling Technique) để tạo mẫu tổng hợp cho các lớp cảm xúc ít dữ liệu, như "happy" và "sad". Điều này giúp cân bằng phân bố dữ liệu và giảm hiện tượng ưu tiên dự đoán các lớp phổ biến như "fear".

Sử dụng các kỹ thuật tăng cường dữ liệu đặc thù, như mô phỏng ánh sáng yếu hoặc góc chụp nghiêng, để cải thiện khả năng nhận diện trong các điều kiện thực tế.

Tinh chỉnh và học chuyển giao (Fine-Tuning and Transfer Learning): Tinh chỉnh mô hình trên sự kết hợp của các bộ dữ liệu quốc tế (FER2013, CK+, AffectNet) để cải thiện khả năng tổng quát hóa. Sử dụng học chuyển giao từ các mô hình được huấn luyện trước trên tập dữ liệu lớn như AffectNet để nắm bắt các đặc trưng cảm xúc đa dạng.

Chính quy hóa và tối ưu hóa: Áp dụng chính quy hóa L2 mạnh hơn để ngăn chặn hiện tượng overfitting, đặc biệt khi mô hình gặp khó khăn trong việc phân biệt các cảm xúc tương đồng. Sử dụng bộ tối ưu hóa tiên tiến như AdamW thay vì Adam để cải thiện tốc độ hội tụ và giảm lỗi phân loại.

3.5.2.5. Đánh giá Thời gian Thực và Khả năng Triển khai

Thời gian xử lý trung bình của hệ thống được ghi nhận chi tiết trên các thiết bị phần cứng khác nhau (PC, laptop, máy tính bảng) và so sánh với mục tiêu 50ms. Kết quả cho thấy hệ thống có khả năng đạt được tốc độ xử lý dưới ngưỡng này (ví dụ: 32ms mỗi khung hình), đảm bảo khả năng hoạt động thời gian thực (real-time) cần thiết cho môi trường giáo dục. Khả năng này đạt được một phần nhờ vào việc tối ưu hóa kiến trúc CNN và sử dụng Haar Cascade để phát hiện khuôn mặt nhanh chóng. Tính khả thi của hệ thống trong môi trường giáo dục với tài nguyên hạn chế là một yếu tố quan trọng, và kết quả này cho thấy tiềm năng triển khai rộng rãi.

3.5.2.6. Phân tích Cảm xúc theo Thời gian và Tối ưu hóa Can thiệp Sư phạm

Khả năng phân tích cảm xúc theo thời gian (time-series analysis) là một điểm mạnh của hệ thống. Thay vì chỉ đưa ra nhận định tại một thời điểm, hệ thống theo dõi

sự thay đổi biểu cảm của học sinh trong suốt quá trình học tập (ví dụ: từ trạng thái trung lập đến buồn chán hoặc căng thẳng). Thông tin động thái này cung cấp một cái nhìn chi tiết và hữu ích hơn nhiều cho giáo viên, giúp họ đưa ra các quyết định can thiệp kịp thời và chính xác hơn, như điều chỉnh tốc độ giảng bài, thay đổi hoạt động, hoặc cung cấp hỗ trợ cá nhân khi cần thiết.

Các chỉ số bổ sung như Precision và Specificity có thể được sử dụng để giảm thiểu sai sót trong phân loại các cảm xúc quan trọng, đặc biệt là những cảm xúc mang tính cảnh báo như căng thẳng hay sợ hãi, nhằm đảm bảo các cảnh báo được đưa ra là đáng tin cậy.

3.5.2.7. Đảm bảo Đạo đức và Chấp nhận Cộng đồng

Các yếu tố đạo đức và quyền riêng tư được đặc biệt chú trọng. Phản hồi từ phụ huynh về tính minh bạch, đạo đức, và các lo ngại về quyền riêng tư của hệ thống được thu thập để đảm bảo sự chấp nhận rộng rãi và xây dựng lòng tin từ cộng đồng. Việc thu thập dữ liệu khuôn mặt học sinh được thực hiện với sự đồng ý rõ ràng và tuân thủ nghiêm ngặt các quy định về bảo mật dữ liệu. Hơn nữa, sự minh bạch trong cách thu thập và sử dụng dữ liệu là yếu tố then chốt để tăng cường sự chấp nhận của hệ thống trong cộng đồng giáo dục.

Dựa trên phân tích chi tiết các kết quả định lượng và định tính, các cải tiến cụ thể được đề xuất cho các phiên bản tương lai của mô hình, bao gồm việc tăng cường thu thập dữ liệu cho các lớp bị nhầm lẫn nhiều hoặc điều chỉnh hàm mất mát để ưu tiên độ chính xác của các cảm xúc quan trọng, nhằm nâng cao hiệu suất tổng thể của hệ thống.

3.6. Kết luận chương

Chương 3 đã trình bày chi tiết quá trình nghiên cứu, xây dựng, huấn luyện và thử nghiệm một hệ thống nhận diện cảm xúc dựa trên kiến trúc CNN tối ưu hóa, tích hợp Haar Cascade để phát hiện khuôn mặt nhanh chóng và BatchNormalization để ổn định huấn luyện. Mục tiêu chính là chứng minh khả năng phân tích cảm xúc thời gian thực trong môi trường giáo dục Việt Nam, với những kết quả đáng khích lệ.

Các chỉ số như accuracy, recall, F1-score cung cấp cái nhìn sâu sắc về hiệu quả, trong khi phản hồi định tính từ giáo viên xác nhận tính hữu ích và tiềm năng hỗ trợ điều chỉnh giảng dạy. Tuy nhiên, độ chính xác (68-70%) vẫn thấp hơn so với dữ liệu chuẩn quốc tế (FER2013, CK+: 71-74%), do ánh sáng không đồng đều và sự tinh tế của biểu cảm văn hóa. Đặc biệt, sự nhầm lẫn giữa "buồn bã" và "trung lập" cần được cải thiện bằng dữ liệu bổ sung, trong khi tối ưu hóa phần cứng và đảm bảo đạo đức (bảo mật, minh bạch) vẫn là những thách thức. Những hạn chế này sẽ là định hướng quan trọng cho các nghiên cứu tiếp theo.

KẾT LUẬN

Kết quả đạt được

Nghiên cứu đã xây dựng thành công một mô hình nhận diện cảm xúc khuôn mặt dựa trên CNN, tích hợp các kỹ thuật như BatchNormalization, Haar Cascade và tăng cường dữ liệu, nhằm hỗ trợ phân tích cảm xúc trong giáo dục tại Việt Nam. Các kết quả chính bao gồm:

- **Phát triển mô hình hiệu quả:** Mô hình nhận diện được 7 trạng thái cảm xúc cơ bản (vui vẻ, buồn bã, tức giận, sợ hãi, ngạc nhiên, ghê tởm, trung lập), đạt độ chính xác 68-74% trên dữ liệu chuẩn FER2013 và CK+. Tốc độ xử lý nhanh (dưới 50ms/khung hình), phù hợp cho ứng dụng thời gian thực với thiết bị hạn chế.
- **Áp dụng kỹ thuật tiên tiến:** Sử dụng Haar Cascade để phát hiện khuôn mặt nhanh (32ms/khung hình). Tăng cường dữ liệu (xoay, lật, phóng to/thu nhỏ ảnh) giúp mô hình thích nghi tốt hơn với điều kiện thực tế như ánh sáng yếu hay góc chụp nghiêng. Lớp BatchNormalization giúp huấn luyện ổn định và tăng tốc độ học của mô hình.
- **Đánh giá chi tiết:** Mô hình được đánh giá qua các chỉ số như accuracy, recall, F1-score và confusion matrix. Kết quả cho thấy mô hình nhận diện tốt các cảm xúc phổ biến, nhưng còn nhầm lẫn giữa "buồn bã" và "trung lập".

Nhận xét

Mô hình này cho thấy tiềm năng ứng dụng trong giáo dục, với hiệu suất ổn định và tốc độ xử lý nhanh. Việc tích hợp các kỹ thuật tiên tiến giúp mô hình bước đầu thích nghi với biểu cảm tại Việt Nam, tạo nền tảng cho các công cụ hỗ trợ giáo viên. Tuy nhiên, do điều kiện ánh sáng không đồng đều và sự phức tạp của biểu cảm thực tế. Ngoài ra, nhầm lẫn giữa các cảm xúc như "buồn bã" và "trung lập" cho thấy cần cải thiện thêm.

Hướng nghiên cứu tiếp theo

Để nâng cao hiệu quả, các hướng tiếp theo bao gồm:

- Thu thập thêm dữ liệu nội địa đa dạng (từ nhiều độ tuổi, vùng miền, bối cảnh học tập) để cải thiện khả năng nhận diện biểu cảm Việt Nam.
- Cải tiến mô hình với các kiến trúc tiên tiến như ResNet, MobileNetV3, hoặc áp dụng transfer learning để tăng hiệu suất.
- Tối ưu xử lý video thời gian thực, bằng cách dùng thuật toán phát hiện khuôn mặt hiện đại hơn (YOLO, MTCNN) và cải thiện xử lý ánh sáng yếu.
- Nghiên cứu biểu cảm văn hóa Việt Nam, hợp tác với chuyên gia tâm lý để nhận diện các biểu cảm đặc thù.
- Thử nghiệm ứng dụng trong lớp học thực tế, tích hợp vào Zoom hoặc Google Meet, và đánh giá tác động đến học tập.
- Xem xét các vấn đề đạo đức, như bảo mật dữ liệu và quyền riêng tư, để đảm bảo triển khai có trách nhiệm.

DANH MỤC CÁC TÀI LIỆU THAM KHẢO

Tiếng Anh

- [1] P. Baheti, “A Comprehensive Guide to Convolutional Neural Networks,” 27 6 2021. Available: <https://www.v7labs.com/blog/convolutional-neural-networks-guide>.
- [2] N. Buhl, “Convolutional Neural Networks (CNN) Overview,” 24 7 2023. Available: <https://encord.com/blog/convolutional-neural-networks-explained/>.
- [3] A. D. Tejashwini S. G 1, “International Journal of Advanced Computer Science and Applications,” *Multimodal Deep Learning Approach for Real-Time Sentiment Analysis in Video Streaming*, tập 14, 2023.
- [4] S. M. K. Q. A. K. Priti Rai Jain, *PM-ViT a Framework for the Recognition of Emotions and Proclivity toward Mental Illness Using Facial Expressions*, 2024.
- [5] Z. Shen, “A Comparative Study of Hybrid CNN and Vision Transformer Models for Facial Emotion Recognition,” 2024.
- [6] N. A. C. T. S. E. M. E. O. S. K. D. A. A. E. Sinem Aslan, “Investigating the Impact of a Real-time, Multimodal Student Engagement Analytics Technology in Authentic Classrooms,” *researchgate*, Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-12, 2019.
- [7] M. J. Paul Viola, “Rapid Object Detection using a Boosted Cascade of Simple Features,” *ACCEPTED CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION*, 2001.
- [8] R. S. b. ., S. R. c. ., J. S. K. d. ., R. M. e. ., S. J. f. Ketan Sarvakar a, “Facial emotion recognition using convolutional neural networks,” 90, pp. 3560-3564, 2021.
- [9] J. B. ., a. M. T. Allison Macey Banzon, “Facial Expression Recognition in Classrooms: Ethical Considerations and Proposed Guidelines for Affect Detection in Educational Settings,” 15, *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, 2024.
- [10] A. W. B. a. R. W. P. Kapoor, “Automatic prediction of frustration.,” *số International journal of human-computer studies*, pp. 724-736, 2007.

- [11] S. W. D. a. J. D. Li, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” số Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.
- [12] Y. e. a. Huang, “Facial expression recognition: A survey,” số Symmetry, p. 1189, 2019.
- [13] S. a. W. D. Li, “Deep facial expression recognition: A survey,” số IEEE Transactions on Affective Computing, 2020.
- [14] J. W. Z. S. v. S. C. Y. Chen, “Facial Motion Prior Networks for Facial Expression Recognition,” số arXiv, 2019.
- [15] J. F. C. J. M. W. M. A. P. ., D. J. P. Michael A. Sayette, “A Psychometric Evaluation of the Facial Action Coding System for Assessing Spontaneous Expression,” *Journal of Nonverbal Behavior*, 2022.

Tiếng Việt

- [1] L. H. T. Vũ, "Nhận dạng độ hứng thú của người học thời gian thực sử dụng học sâu," *Tạp chí Khoa học*, 2022.
- [2] ducmanhkthd, "Nghiên cứu và ứng dụng các kỹ thuật nhận dạng cảm xúc qua khuôn mặt," *viblo*, 2021.