ĐẠI HỌC BÁCH KHOA HÀ NỘI
Trường Điện – Điện Tử

BÁO CÁO BÀI TẬP LỚN

Nhận diện khuôn mặt và phân loại cảm xúc người thời gian thực qua video

Môn: Trí tuệ nhân tạo và ứng dụng
Nhóm thực hiện: Nhóm 23

SV thực hiện:	Nguyễn Phương Linh	20206203
	Trần Thành Lâm	20200339
	Nguyễn Ngọc Dương	20200122
Mã lớp:	144117	
GV hướng dẫn:	TS. Võ Lê Cường	
Học kỳ:	20231	

Hà Nội, tháng 12 năm 2023

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

DANH MỤC CÁC TỪ VIẾT TẮT

Từ đầy đủ tiếng Việt (nếu có)	Từ đầy đủ tiếng Anh (nếu có)	Từ viết tắt
Trí tuệ nhân tạo	Artificial Intelligence	AI
Bản đặc tả	Specification	SPEC
	Facial Expression Recognition 2013	FER-2013
Mạng nơ-ron tích chập	Convolutional Neural Network	CNN

BẢNG PHÂN CÔNG NHIỆM VỤ

Tuần	Nhiệm vụ	Chi tiết công việc	Sinh viên
Tuần 1	Khảo sát và nghiên cứu đề tài	- Đặt vấn đề: Lý do chọn đề tài, tính ứng dụng - Khảo sát các phương án từ các bài báo, bài nghiên cứu về bài toán trước đó - Lập bảng Related work so sánh các nghiên cứu bao gồm: tên bài, năm công bố, hướng giải quyết, ưu điểm, đặc trưng của phương pháp, điểm hạn chế	Trần Thành Lâm Nguyễn Ngọc Dương
	Viết specification	- Liệt kê ra các khối cần thiết để giải quyết bài toán - Vẽ sơ đồ liên kết các khối, đầu vào, ra của chúng	Nguyễn Phương Linh
Tuần 2 + Tuần 3	Tìm hiểu cơ sở lý thuyết, lựa chọn các model	- Chọn một vài model loại model từ các phương án, bài báo và từ đề xuất của thành viên trong nhóm, thầy hướng dẫn - Phân tích cụ thể sự phù hợp của các model ứng với hai vấn đề classify và detect, ưu, nhược từng loại model - Chốt lại 1 vài model có tính khả thi cao và phù hợp với mục tiêu của nhóm	Trần Thành Lâm Nguyễn Ngọc Dương Nguyễn Phương Linh
	Lựa chọn tệp dataset	- Tìm hiểu các tệp dataset thông dụng trong bài toán đề ra - Chọn tệp dataset (phù hợp với đầu vào/ra của các khối đã định nghĩa trong SPEC)	Nguyễn Phương Linh
Tuần 4	Chạy thử code model	- Tìm kiếm, chạy thử code mẫu đã có ứng với từng model đã chọn (có thể với các bài toán khác) - Quan sát kết quả, hiểu code và ánh xạ phần lý thuyết đã học vào các phần trong code - Chạy test để xem kết quả chạy của các model mẫu	Nguyễn Phương Linh Trần Thành Lâm Nguyễn Ngọc Dương
Tuần 5 + Tuần 6	Viết code cụ thể hóa cho đề tài của nhóm	- Dựa trên các code mẫu đã tìm hiểu và chạy thử từ tuần trước code lại phù hợp để ứng dụng cho đề tài của nhóm - Chạy thử, lựa chọn các bộ tham số phù hợp cho các model	Nguyễn Phương Linh
	Chạy code với tập dữ liệu đã chọn, thống kê performance	- Train các model với các bộ tham số, tệp dữ liệu đã chọn - Test các model đã train - Lựa chọn performance metric phù hợp và thống kê kết quả [performance, số phép tính, tốc độ,... --> Định lượng]	Trần Thành Lâm

	Đánh giá và nhận xét kết quả	- Đánh giá kết quả thu được (so sánh các model nhóm đã code với nhau, so sánh với các model khác có cùng tệp dữ liệu từ báo,...) - Nhận xét kết quả, đề xuất các cải tiến tiềm năng	Nguyễn Ngọc Dương Nguyễn Phương Linh
Tuần 7	Cài đặt hệ thống thực tế và thu thập đánh giá từ người dùng	- Cài đặt và triển khai hệ thống trên thực tế - Gửi hệ thống cho một vài người dùng thử và thu thập các đánh giá từ người dùng (--> đánh giá có thể mang tính định tính)	Nguyễn Ngọc Dương Trần Thành Lâm Nguyễn Phương Linh
Tuần 8	Đề xuất phương án cải thiện	- Từ các nhận xét và đánh giá ở tuần 5, 6, 7 (cả định lượng và định tính), đề xuất một vài phương án cải tiến khả thi	Nguyễn Ngọc Dương Trần Thành Lâm
		- Đọc lý thuyết và thử triển khai một vài phương án - Thu thập và thống kê kết quả sau khi triển khai các phương án cải thiện, đánh giá và đưa ra lựa chọn (có lấy các cải tiến đang đề xuất không)	Nguyễn Phương Linh
Tuần 9	Tổng hợp và viết báo cáo	- Đặt vấn đề - Khảo sát bài báo, khảo sát các giải pháp mô hình nhận diện và phân loại	Trần Thành Lâm
		- Phân tích lựa chọn tệp dữ liệu - Đề xuất cải tiến	Nguyễn Ngọc Dương
		- Viết phần đặt vấn đề, đặc tả - Cụ thể phương pháp - Kết quả, đánh giá và đề xuất cải thiện - Tổng hợp và trình bày báo cáo	Nguyễn Phương Linh
	Chuẩn bị slide	- Chuẩn bị slide thuyết trình	Nguyễn Ngọc Dương Trần Thành Lâm

CHƯƠNG 1
ĐẶT VẤN ĐỀ

1.1 Thực trạng, tính cấp thiết và lý do chọn đề tài

Cảm xúc không chỉ đóng vai trò quan trọng, ảnh hưởng đến các mối quan hệ cá nhân mà còn chi phối nhiều khía cạnh của cuộc sống, từ hiệu quả truyền đạt thông tin, giáo dục, mua sắm, đến tiếp thị và an ninh. Trong môi trường giáo dục, việc đánh giá độ hứng thú của sinh viên có thể cung cấp thông tin quý báu về hiệu suất học tập và phản ánh chất lượng của quá trình giảng dạy. Trong lĩnh vực mua sắm và tiếp thị, phân tích cảm xúc của khách hàng giúp hiểu rõ phản hồi, đánh giá sự ưa thích sản phẩm và mức độ hài lòng đối với nhân viên, mang lại cơ hội tối ưu hóa chiến lược kinh doanh. Chính vì vậy việc có thể phân loại được cảm xúc con người đang là lĩnh vực đầy thách thức và triển vọng, nhất là trong bối cảnh bùng nổ về công nghệ cũng như trí tuệ nhân tạo (AI) như hiện nay.

Có hai phương pháp chính để có thể phát hiện cảm xúc đó chính là dựa vào các thông tin hình ảnh (như ảnh khuôn mặt, cử chỉ) và dựa vào các sinh trắc học (như nhịp tim, nồng độ máu và các dữ liệu khác). Tuy nhiên việc xác định cảm xúc dựa vào sinh trắc học đòi hỏi nhiều cảm biến và phải gắn cảm biến lên người các đối tượng cần đo, sẽ không phù hợp để triển khai rộng rãi trong công nghiệp. Hơn nữa, nhóm nhận thấy tính ứng dụng rộng rãi của việc phát hiện được cảm xúc trong thời gian thực thông qua các video như thu được từ webcam hay camera. Chính vì vậy nhóm quyết định chọn đề tài nghiên cứu cho môn trí tuệ nhân tạo và ứng dụng là "Nhận diện khuôn mặt và phân loại cảm xúc người thời gian thực qua video".

Việc theo đuổi giải pháp hiệu quả và thời gian thực thông qua webcam là bước quan trọng trong việc đáp ứng nhu cầu ngày càng tăng về khả năng tương tác và phân tích cảm xúc một cách chính xác. Tuy đã có sự tiến bộ trong công nghệ nhận diện cảm xúc, nhưng vẫn còn những thách thức cần vượt qua, như việc nhận diện khuôn mặt ở trạng thái trung lập mạnh mẽ, thách thức về chất lượng hình ảnh thu được có thể bị ảnh hưởng bởi các yếu tố môi trường, hay việc nhận diện cảm xúc với độ chính xác cao đòi hỏi hiệu suất tính toán cao.

1.2 Mục đích và nhiệm vụ nghiên cứu

Mục tiêu của dự án là có thể ứng dụng kiến thức lý thuyết của môn trí tuệ nhân tạo và ứng dụng, xây dựng một mô hình cơ bản giúp nhận diện khuôn mặt và phân loại biểu cảm con người từ video trong thời gian thực, hướng tới việc cân bằng độ phức tạp của mô hình, tốc độ xử lý và độ chính xác. Đồng thời nhóm cũng triển khai mô hình trong thực tế và cho một nhóm người trải nghiệm thử.

Để thực hiện mục đích trên, chúng em sẽ thực hiện các nhiệm vụ sau:

- Xác định mục tiêu, đề tài của bài tập lớn và đưa ra kế hoạch triển khai chi tiết. Viết đặc tả (SPEC) của bài toán
- Khảo sát, thống kê, tổng hợp, phân tích và so sánh các phương pháp đang được nghiên cứu và sử dụng. xác định phương pháp nhóm triển khai

- Tiền xử lý dữ liệu, lập trình để huấn luyện các mô hình và triển khai thử mô hình
- Thu thập kết quả và đánh giá kết quả huấn luyện, chạy thử và đưa ra các phương án cải thiện.

1.3 Phạm vi nghiên cứu, phương pháp thực hiện

Nhóm sẽ nghiên cứu tập trung vào việc làm sạch, tiền xử lý dữ liệu và xây dựng mô hình phân loại cảm xúc, về mô hình để nhận diện khuôn mặt, nhóm sẽ tận dụng các mô hình đã được huấn luyện với lượng lớn dữ liệu.

Nhóm thực hiện huấn luyện và triển khai thử mô hình trên máy tính cá nhân 16GB RAM, core i7-8750H với tần số hoạt động 2.2GHz.

1.4 Kết cấu của báo cáo

Trong bài báo cáo này, nhóm chúng em sẽ trình bày về quá trình học hỏi, thực hành cũng như kết quả triển khai bài tập lớn của nhóm về mô hình học sâu giúp nhận diện cảm xúc con người trong thời gian thực qua video. Ngoài phần mở đầu, danh sách các hình vẽ, bảng biểu, phân công nhiệm vụ, kết luận, tài liệu tham khảo và chương 1- Đặt vấn đề, toàn bộ bội dung nghiên cứu được trình bày trong 3 chương sau, cụ thể:

- Đề xuất phương pháp
- Cụ thể phương pháp
- Kết quả, đánh giá và đề xuất cải thiện

CHƯƠNG 2
ĐỀ XUẤT PHƯƠNG PHÁP

Trong chương này, nhóm sẽ trình bày các bước nhóm khảo sát các bài báo, phương án đang được áp dụng, sau đó so sánh, phân tích ưu nhược điểm các phương pháp và lựa chọn phương pháp giải quyết vấn đề đặt ra (bao gồm cả phương án lựa chọn tệp dữ liệu và các mô hình sử dụng).

2.1 Đặc tả (SPEC) và yêu cầu bài toán

Sau khi xác định đề tài, nhóm xác định các yêu cầu của bài toán:

- Bài toán yêu cầu nhận diện biểu cảm thời gian thực qua video nên sẽ cần 2 mô hình tương ứng là mô hình dùng để nhận diện khuôn mặt người trong video và mô hình dùng để phân loại biểu cảm.
- Yêu cầu về tốc độ: >20fps để đáp ứng được yêu cầu là thời gian thực.
- Yêu cầu về độ chính xác: Độ chính xác của các mô hình sử dụng phải đạt ít nhất là 60%.
- Yêu cầu về số tham số: Nhóm ưu tiên xây dựng mô hình nhỏ và nhẹ nên đặt ra yêu cầu về tham số của mạng là nhỏ hơn 5,000,000 tham số.

2.2 Khảo sát giải pháp

Nhóm tiến hành tìm kiếm và khảo sát các bài báo có liên quan bằng Google Scholar, tìm kiếm các bộ data trên Kaggle với các từ khóa “Emotion Recognition”, “Emotion Classification”, “Face Detection”, “Facial Expression Recognition” cũng như từ các cuộc thi về AI. Sau đó nhóm tiến hành phân tích cụ thể các bộ dữ liệu, các phương pháp để phát hiện và phân loại cảm xúc để lựa chọn phương pháp triển khai cho đề tài.

2.1.1 Kết quả khảo sát các bài báo

Với các từ khóa như đã trình bày ở trên, nhóm tiến hành đọc và lập bảng khảo sát các bài báo theo các tiêu chí như tên, năm xuất bản, mô hình sử dụng, bộ dữ liệu và kết quả như bảng 1 dưới đây.

Bảng 1: Tóm tắt khảo sát các bài báo

No.	Title	Year	Keywords	Models	Dataset	Evaluation
1	Neutral face classification using personalized appearance models for fast and robust	2015	Neutral vs. emotion classification, Constrained Local Model, Key Emotion Points, Procrustes analysis, Local Binary Pattern Histogram, Statistical model,	CLM	Cohn-Kanade AU-Coded Facial Expression Image Database (CK+, ISL, SRID1, SRID2)	67 to 98%

	emotion detection		Structural similarity, Action units.			
2	Deep Learning Approaches in EEG-based Emotion Recognition	2018	Deep learning in EEG emotion recognition, focusing on CNNs	CNNs, DBNs, RNNs	DEAP database	above 84%
3	Emotional State Classification: An Additional Step in Emotion Classification through Face Detection	2019	Keywords-facial expression; emotion recognition; action units; computer vision	K nearest neighbor classifier (k-NN), Multilayer Perceptron (MLP), J48 Decision Tree	data from participants	0.981, 0.859
4	Emotion Recognition Based On CNN	2019	Emotion recognition, DEAP, EEG, PCA, CNN, classification	Convolutional neural network (CNN)	EEG signal characteristics in the DEAP data set	84.3±4.0% 81.2±3.0% (Valence and Arousal)
5	Facial Emotion Detection Using Deep Learning	2020	AI, Facial emotion recognition (FER), Convolutional neural networks (CNN), Rectified linear units (ReLu), Deep learning (DL).	convolutional neural networks (CNN)	Facial emotion recognition challenge (FERC-2013); Japaness female facial emotion (JAFFE).	70.14 and 98.65% FERC-2013 and JAFFE datasets
6	Emotion Detection and Sentiment Analysis of Images	2020	SVM, RESNET, Places205-VGG16, VGGImageNet	SVM on high level features of VGG-ImageNet, RESNET, Places205-VGG16 and VGGImageNet	5 emotional categories - Love, Happiness, Violence, Fear, and Sadness from Flick	70.14% and 98.65%

2.1.2 Phân tích và lựa chọn tệp dữ liệu

Nhóm thu được danh sách các bộ dữ liệu được sử dụng rộng rãi trong cả nghiên cứu (qua các bài báo) và trong việc luyện tập bao gồm Cohn-Kanade Database (CK+), Facial Expression

Recognition 2013 (FER-2013), Labeled Faces in the Wild (LFW), AffectNet, Ascertain. Sau đó, nhóm tiền hành phân tích và lập bảng so sánh sơ bộ ưu nhược các tệp dữ liệu tìm được như bảng 2 dưới đây.

Bảng 2: Khảo sát, phân tích các tệp dữ liệu thông dụng

Tệp dữ liệu	Mô tả, thông số	Ưu điểm	Nhược điểm
CK+ (Cohn-Kanade Database)	- Hình ảnh khuôn mặt của các người tham gia biểu diễn - Đã được gán nhãn 6 biểu cảm: vui, buồn, tức giận, sợ hãi, thất vọng, ngạc nhiên - Video: 30PFS, 640x490 hoặc 640x480 - Số mẫu: 327 video	- Chất lượng biểu cảm và số lượng hình ảnh chất lượng tốt. - Dễ dàng sử dụng cho phân loại biểu cảm	- Có thể không đủ đa dạng để tạo ra một model tổng quát hóa tốt
FER2013 (Facial Expression Recognition 2013)	- Ảnh mặt người đa dạng độ tuổi - Đã được gán nhãn 7 biểu cảm: vui, buồn, tức giận, sợ hãi, ghê tởm, ngạc nhiên, bình thường - Ảnh: ảnh xám 48x48 - Số mẫu: 35,887	- Số lượng hình ảnh lớn (hơn 35.000 hình ảnh) giúp mô hình có khả năng tổng quát hóa tốt hơn. - Đa dạng biểu cảm - Kích thước ảnh nhỏ, ảnh xám nên nhẹ - Được sử dụng rộng rãi	- Chất lượng dữ liệu không đồng đều: Một số hình ảnh chứa nhiễu. - Phân phối mất cân đối: một số lớp biểu cảm có số lượng hình ảnh ít hơn so với các lớp khác, trong khi có lớp lại có rất nhiều hình ảnh
LFW (Labeled Faces in the Wild)	- Hình ảnh khuôn mặt của nhiều người nổi tiếng trong môi trường thực tế. - Chưa được gán nhãn biểu cảm - Ảnh: Ảnh màu, 250x250 - Tổng số mẫu là 13233	- Tập dữ liệu lớn với đa dạng khuôn mặt	- Tập dữ liệu này phù hợp cho phát hiện khuôn mặt hơn là phân loại biểu cảm
AffectNet	- Hình ảnh khuôn mặt của con người - Được gán nhãn với 8 biểu cảm: vui, buồn, tức giận, sợ hãi, ghê tởm, ngạc nhiên, bình thường, coi thường - Ảnh: Ảnh màu, kích thước 128x128 - Tổng số mẫu của dataset là hơn 1 triệu	- Đa dạng biểu cảm - Tập dữ liệu có kích thước lớn (hơn 1 triệu hình ảnh) - Chất lượng hình ảnh tốt	- Kích thước một ảnh và bộ dữ liệu quá lớn dẫn đến cần nhiều tài nguyên để huấn luyện và xử lý - Khó khăn trong việc xây dựng mô hình nhỏ, nhẹ mà đem lại hiệu suất tốt

Vì mục tiêu của nhóm là xây dựng một mô hình nhận diện biểu cảm theo thời gian thực cũng như do hạn chế về tài nguyên huấn luyện và triển khai, nhóm ưu tiên một mô hình gọn nhẹ và đáp ứng nhanh, tuy nhiên vẫn cần lượng dữ liệu tương đối. Sau khi so sánh và cân nhắc giữa các tệp dữ liệu

trên, nhóm lựa chọ tệp dữ liệu FER-2013 cho bài tập lớn lần này. Đây là tập dữ liệu phổ biến, nhiều người sử dụng và đã từng được lấy làm tập dữ liệu cho cuộc thi AI. Mô tả rõ hơn về tệp dữ liệu này sẽ được trình bày cụ thể ở phần "*3.1 Phân tíchtập dữ liệu và tiền xử lý dữ liệu*".

2.1.3 Phân tích và lựa chọn mô hình phát hiện khuôn mặt từ video

Sau khi tìm kiếm và khảo sát, nhóm tìm đươc một vài phương pháp phát hiện khuôn mặt từ video. Nhóm tiếp tục tiến hành so sánh như bảng 3 dưới đây để lựa chọn phương pháp phù hợp

Bảng 3: Khảo sát, phân tích các phương pháp phát hiện khuôn mặt trong video

Phương pháp	Đặc trưng	Ưu điểm	Nhược điểm
Haar Cascade Classifier	- Phát hiện khuôn mặt dựa trên các tính năng Haar-like và một bộ phân lớp được đào tạo trước.	- Nhanh và hiệu quả cho việc phát hiện khuôn mặt trong thời gian thực. - Có thể phát hiện nhiều khuôn mặt trong một khung hình - Có thể chạy trên các thiết bị có tài nguyên hạn chế.	- Không thể phát hiện khuôn mặt với góc độ lớn hoặc biến đổi nghiêng
Viola-Jones Face Detection	- Phát hiện khuôn mặt sử dụng các tính năng Haar-like và AdaBoost để xác định vùng khuôn mặt trong hình ảnh.	- Tốc độ nhanh và yêu cầu tính toán thấp. - Thực hiện tốt trong điều kiện ánh sáng kém	- Độ chính xác hạn chế và khả năng phát hiện khuôn mặt đối với các biến thể không tốt
Deep Learning Models	- Các mô hình sâu hơn dựa trên deep learning như Single Shot MultiBox Detector (SSD), You Only Look Once (YOLO), và Faster R-CNN có khả năng phát hiện khuôn mặt trong video và ảnh tốt hơn	- Độ chính xác cao và khả năng phát hiện khuôn mặt trong nhiều điều kiện khác nhau. - Có thể phát hiện nhiều khuôn mặt trong cùng một khung hình.	- Yêu cầu tài nguyên tính toán mạnh hơn và có thể không đáp ứng nhanh trên các thiết bị có tài nguyên hạn chế. - Thời gian đào tạo mô hình dài hơn và đòi hỏi dữ liệu lớn.
OpenCV DNN Module	- Mô-đun DNN (Deep Neural Networks) của OpenCV cho phép sử dụng các mô hình deep learning	- Cung cấp tích hợp dễ dàng với các mô hình deep learning và có thể tận dụng tài	- Cần cấu hình mô hình và tối ưu hóa để đảm bảo độ chính xác và hiệu năng

	đã được đào tạo trước để phát hiện khuôn mặt và đối tượng.	nguyên tính toán của GPU	

Để phù hợp với bài toán đặt ra cũng như cân nhắc đến độ phức tạp các phương pháp, giới hạn tài nguyên của thiết bị huấn luyện nhóm quyết định sử dụng phương pháp HAAR (đơn giản, dễ triển khai và có tốc độ nhanh [>20 fps]). Sau khi tìm hiểu cùng với cân nhắc lượng tài nguyên (thời gian, tài nguyên máy tính, lượng dữ liệu), nhóm nhận thấy để nhận diện khuôn mặt qua video thu được độ chính xác cao thì cần lượng dữ liệu là video tương đối, do đó nhóm quyết định sử dụng mô hình HAAR của OpenCV đã được huấn luyện trên lượng lớn dữ liệu thay vì huấn luyện lại từ đầu mô hình này. Thay vào đó, nhóm sẽ tập trung xây dựng và huấn luyện mô hình phân loại biểu cảm khuôn mặt.

2.1.3 Lựa chọn mô hình phân loại biểu cảm

Qua các bài báo đã khảo sát cũng như tìm kiếm trên Github, nhóm tìm được một số mô hình để phân loại biểu cảm như Convolutional Neural Network (CNN), Local Binary Pattern (LBP), Support Vector Machine (SVM) hay học chuyển tiếp (Transfer Learning). Nhóm tiến hành phân tích, so sánh các mô hình như bảng 4 dưới đây.

Bảng 4: Khảo sát, phân tích các phương pháp phân loại biểu cảm qua khuôn mặt

Phương pháp	Mô tả	Ưu điểm	Nhược điểm
Convolutional Neural Network (CNN)	- Là mô hình học sâu đặc biệt phù hợp với xử lý ảnh, gồm nhiều lớp convolutional (tích chập) và pooling để học các đặc trưng cấp thấp đếp cấp cao trong hình ảnh	- Tự học đặc trưng. - Hiệu suất cao với dữ liệu lớn. - Phù hợp cho hình ảnh lớn.	- Đòi hỏi tài nguyên tính toán tương đối. - Yêu cầu dữ liệu đào tạo đủ lớn.
Local Binary Pattern (LBP)	- Mô hình học máy có giám sát. - Hoạt động bằng các so sánh giá trị của các điểm ảnh trong một vùng cục bộ với giá trị của điểm ảnh trung tâm và chuyển thành chuỗi nhị phân.	- Đơn giản và hiệu quả cho việc nhận diện biểu cảm cơ bản. - Tích hợp tốt với các bước tiền xử lý.	- Độ chính xác thấp với biểu cảm phức tạp. - Nhạy cảm với nhiễu.
Support Vector Machine (SVM)	- Mô hình học máy có giám sát sử dụng chủ yếu cho bài toán phân loại và hồi quy - Mục tiêu là tìm ra đường ranh giới hyperplane tốt nhất để phân tách các lớp dữ liệu.	- Hiệu suất tốt với dữ liệu lớn và nhiều chiều. - Khá ổn định và làm việc tốt khi dữ liệu không cân bằng.	- Đòi hỏi tiền đề về việc chọn lựa đặc trưng hiệu quả. - Cần điều chỉnh tham số cẩn thận.

			- Khó khăn với dữ liệu lớn.
Transfer Learning	- Sử dụng các mô hình đã được huấn luyện trên một tập dữ liệu lớn để chuyển giao nhằm thực hiện một nhiệm vụ tương tự.	- Hiệu suất cao với dữ liệu giới hạn. - Tiết kiệm thời gian huấn luyện.	- Phụ thuộc nhiều vào tương đồng của tác vụ. - Nguy cơ có thông tin từ nhiệm vụ trước không phù hợp với nhiệm vụ cụ thể, gây ra "truyền nhiễm" thông tin không mong muốn.

Sau khi khảo sát, nhóm thấy để thu được hiệu quả tương đối, phù hợp với người mới học, có hiệu suất cao và khả năng đáp ứng thời gian thực, mô hình gọn và không có quá nhiều tham số huấn luyện, dễ dàng triển khai trên các máy tính nhỏ và khả năng tính toán tương đối, nhóm quyết định sử dụng mô hình CNN để phân loại biểu cảm khuôn mặt.

CHƯƠNG 3
CỤ THỂ PHƯƠNG PHÁP

Sau khi lựa chọn được phương pháp thực hiện, nhóm tiến hành tìm hiểu cơ sở lý thuyết của các mô hình nhóm sử dụng, phân tích tệp dữ liệu đã lựa chọn. Sau đó nhóm tiền xử lý tệp dữ liệu và tiến hành xây dựng, huấn luyện mô hình.

3.1 Phân tích tập dữ liệu và tiền xử lý dữ liệu

Như đã trình bày ở chương 2, nhóm sẽ sử dụng tập dữ liệu FER-2013 cho bài tập lớn lần này. Đây là tập dữ liệu được thu bởi Google trong hơn 10 năm trên các nền tảng của họ, đã được cắt các khuôn mặt người và gán nhãn ứng với 7 loại biểu cảm: giân dữ, ghê tớm, sợ hãi, vui, buồn, ngạc nhiên, bình thường. Đây là tập dữ liệu phổ biến, được sử dụng rộng rãi cho các bài toán phân loại biểu cảm khuôn mặt [1] và từng được lấy làm tệp dữ liệu cho cuộc thi như ICML, "*Challenges in Representation Learning: Facial Expression Recognition Challenge*" [2]. Đây là một tệp dữ liệu mang tính thách thức tương đối cao khi độ chính xác khi thử nghiệm với việc dùng chính con người phân loại chỉ ở múc 60%. Các ảnh trong tệp dữ liệu sử dụng là ảnh xám, kích cỡ 48x48 pixel và lưu dưới dạng jpg. Ảnh 1 là vài hình ảnh lấy từ tệp dữ liệu FER-2013.

Ảnh 1: Ví dụ về các hình ảnh trong FERR-2013

3.1.1 Số lượng dữ liệu mỗi biểu cảm

Tệp dữ liệu có tổng số lượng ảnh là 35,887 ảnh, được phân phối tới các lớp như bảng 5.

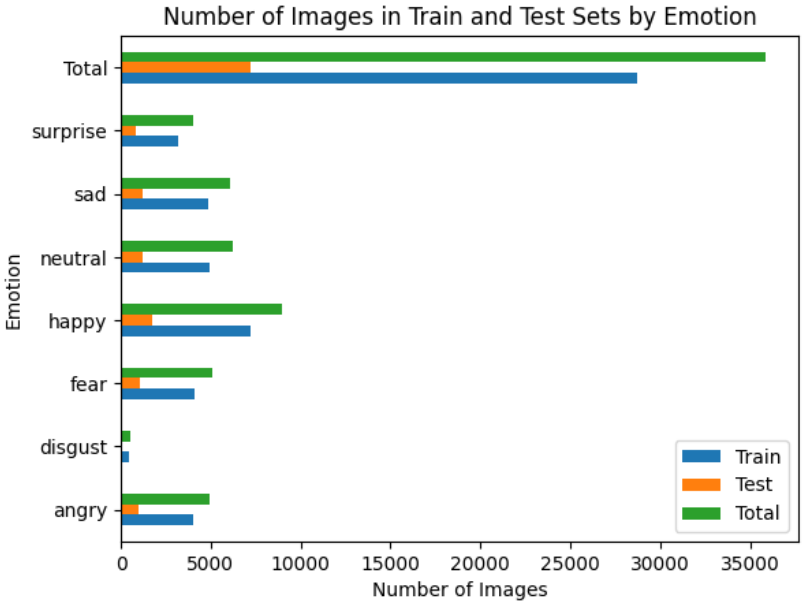
Bảng 5:Thống kê số lượng ảnh từng lớp của FER-2013

	angry	disgust	fear	happy	neutral	sad	surprise	Total
Train	3,995	436	4,097	7,215	4,965	4,830	3,171	28,709
Test	958	111	1,024	1,774	1,233	1,247	831	7,178
Total	4,953	547	5,121	8,989	6,198	6,077	4,002	35,887

3.1.2 Các vấn đề của FER-2013

Tuy được sử dụng rộng rãi song FER-2013 vẫn tồn tại những vấn đề lớn.

Đầu tiên là về việc dữ liệu mất cân bằng khi lượng ảnh cho lớp vui (happy) nhiều vượt trội so với các lớp dữ liệu khác trong khi dữ liệu cho lớp ghê tớm (disgust) được thể hiện rõ ở ảnh 2. Việc này có thể giải thích do Google thu thập ảnh từ nhiều nguồn của họ, song mọi người thường có xu hướng đăng ảnh mình vui vẻ hơn là những bức ảnh mang tính tiêu cực như ghê tớm.



Ảnh 2: Biểu đồ thể hiện tương quan về số ảnh giữa các lớp trong FER-2013 gốc

Tiếp theo, trong tập dữ liệu vẫn còn chứa nhiều ảnh có nội dung không liên quan (ví dụ ảnh chữ, ảnh video bị lỗi), những ảnh quá tối hay những ảnh có khuôn mặt bị che bởi tay hoặc kính. Ngoài ra cũng có các ảnh khuôn mặt không phải của người mà là từ hoạt hình hay hình vẽ, từ đó có thể dẫn tới việc nhận dạng biểu cảm của con người không tốt. Thêm vào đó còn nhiều vấn đề về việc gán nhãn các ảnh như còn các ảnh bị gán nhãn nhầm hay có những ảnh bị cắt quá mức. Những vấn đề này được thể hiện ở ảnh 3.

Ảnh 3: Các vấn đề trong tập huấn luyện của FER-2013

3.1.3 Tiền xử lý và làm sạch dữ liệu

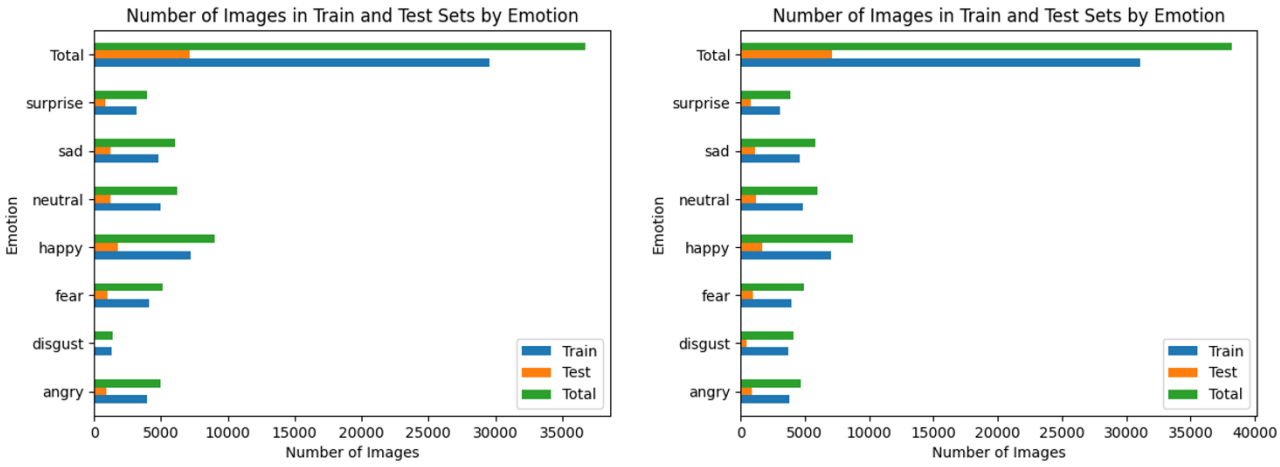
Chính vì tồn tại các vấn đề trên, nhóm có tiến hành các bước tiền xử lý dữ liệu. Đầu tiên, nhóm tiến hành augmentation cho lớp ghê tớm (disgust) để giảm sự mất cân bằng dữ liệu. Nhóm sử dụng các phương án augmentation được cho là phù hợp với bài toán phân loại hình ảnh như lật ảnh, padding ảnh, phương pháp Gauss thu được tập dữ liệu FER-2012-aug. Sau khi augmentation, nhóm còn làm thêm một phiên bản nữa của tập dữ liệu đó là làm sạch tập dữ liệu một cách thủ công, xóa bỏ các dữ liệu không mang thông tin như ảnh chữ, ảnh chụp không phải khuôn mặt và sau đó mới augmentation, thu được tập FER-2013-cleanaug. Tập dữ liệu sau khi làm các phương pháp trên được thể hiện ở bảng 6, bảng 7 và ảnh 8.

Bảng 6: Số lượng ảnh trong FER-2012-aug (sau khi augmentation)

	angry	disgust	fear	happy	neutral	sad	surprise	Total
Train	3,995	1,308	4,097	7,215	4,965	4,830	3,171	29,581
Test	958	111	1,024	1,774	1,233	1,247	831	7,178
Total	4,953	1,419	5,121	8,989	6,198	6,077	4,002	36,759

Bảng 7: Số lượng ảnh trong FER-2013-cleanaug (được làm sạch và augmentation)

	angry	disgust	fear	happy	neutral	sad	surprise	Total
Train	3,809	3,698	3,960	7,029	4,841	4,647	3,114	31,098
Test	898	448	962	1,692	1,182	1,170	791	7,143
Total	4,707	4,146	4,922	8,721	6,023	5,817	3,905	38,241



Ảnh 4: Biểu đồ số lượng ảnh của tập FER-2013-aug và FER-2013-cleanaug

Sau khi tiến hành các bước augmentation và làm sạch trên, chúng em tiếp tục chuẩn hóa giá trị các pixel từ 0 đến 255 về các giá trị trong khoảng [0,1]. Điều này là cần thiết do khi huấn luyện, nếu sử dụng các giá trị pixel ban đầu của dữ liệu thô sẽ rất phức tạp và chiếm dụng nhiều bộ nhớ, đồng thời sẽ tốn thời gian tính toán khi thực thi. [3].

3.2 Cơ sở lý thuyết về CNN và kiến trúc mô hình nhóm sử dụng

Một mô hình CNN cơ bản gồm các lớp cơ bản như lớp tích chập, lớp tổng hợp (pooling) và lớp kết nối đầy đủ (Fully Connection). Mạng sẽ gồm các lớp cơ bản chồng lên nhau và sử dụng các hàm kích hoạt phi tuyến để kích hoạt các trọng số trong các nốt. Mỗi lớp sau khi thông qua hàm kích hoạt sẽ tạo ra các thông tin trừu tượng hơn cho lớp tiếp theo, Tác dụng của các lớp này cũng như các hàm kích hoạt (activation function) và các thuật toán tối ưu (optimizers) sẽ được trình bày ở phần này.

3.2.1 Lớp tích chập

Lớp tích chập 2D là lớp chứa phép toán cơ bản nhất của mạng. Lớp này sẽ gồm nhân (kernel) là một ma trận nhỏ các trọng số (weights). Nhân này sẽ được cho "trượt" qua dữ liệu đầu vào và thực hiện phép nhân ma trận để tổng hợp thành một kết quả là giá trị cho một điểm ảnh đầu ra. Kích thước của ảnh đầu ra sẽ phụ thuộc vào các tham số như kích cỡ của nhân, số pixel padding và số nhân. Kích cỡ nhân thường được sử dụng và được chứng minh đem lại hiệu quả cao là 3x3, chính vì thế nhóm cũng sẽ sử dụng nhân với kích cỡ này.

3.2.2 Lớp tổng hợp (pooling layer)

Lớp tổng hợp thưởng được sử dụng để giảm chiều dữ liệu và đơn giản hóa thông tin đầu ra, qua đó giảm bớt số lượng nơ ron và yêu cầu tính toán. Lớp tổng hợp thường được sử dụng sau lớp tích chập. Trong lớp tổng hợp cũng có một nhân với kích thước nxn di chuyển qua ma trận đầu vào và tùy theo thuật toán mà thu được duy nhất 1 giá trị đầu ra. Không giống lớp tích chập, lớp tổng hợp không cần nhân ma trận. Có hai thuật toán phổ biến cho lớp tổng hợp đó là Max pooling (trong đó với mỗi cửa sổ sẽ chọn giá trị pixel lớn nhất để đưa ra đầu ra) và Average pooling (trong đó sẽ tính giá trị trung bình của các pixel ở mỗi của số để đưa ra đầu ra). Để tăng tính nổi bật của đặc trưng quan trọng trong hình ảnh , thường hay sử dụng Max pooling cho mô hình và nhóm cũng sẽ sử dụng max pooling.

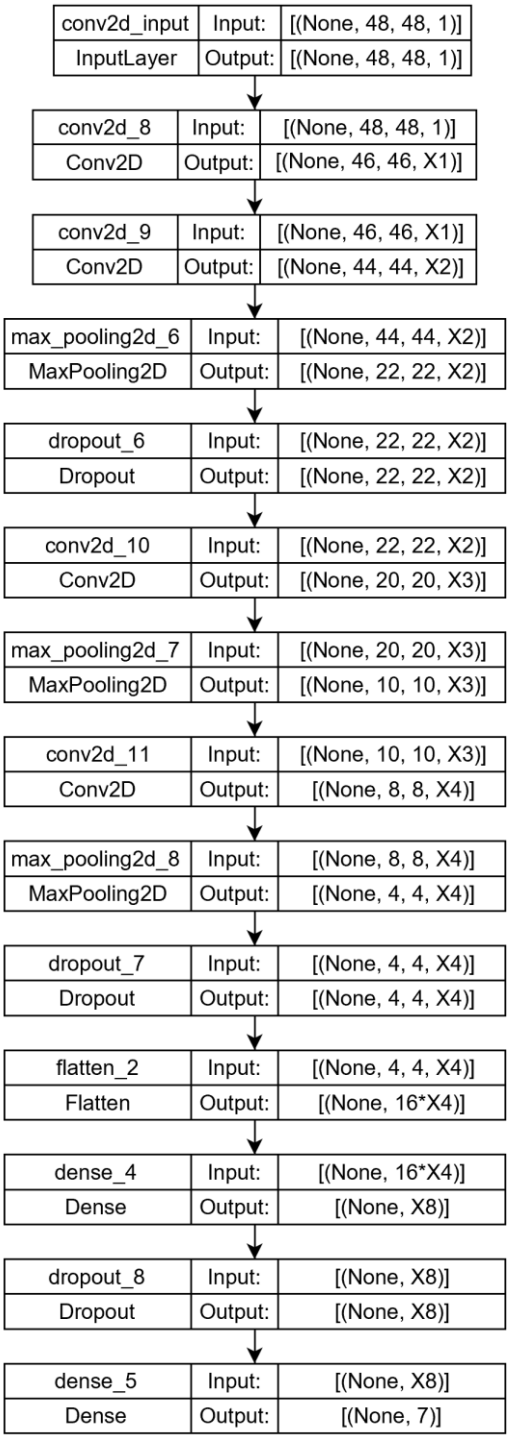
3.2.3 Lớp kết nối đầy đủ (Fully Connection)

Lớp kết nối đầy đủ được sử dụng để kết nối tất cả các nốt (nơ-ron) từ lớp trước với tất cả các nốt của lớp đang xét, tạo ra một mạng kết nốt đầy đủ. Điều này giúp mô hình có khả năng học được các mối quan hệ phức tạp, kết hợp các đặc trưng từ các phần khác nhau của đầu vào để đưa tạo ra các dự đoán. Trong bài toán phân loại, thường lớp cuối cùng của mạng sẽ là lớp Fully Connection với số lượng nơ-ron bằng số lớp đầu ra để phân loại, dự đoán kết quả.

3.2.4 Kiến trúc mô hình nhóm sử dụng, cách chọn các siêu tham số

Sau khi tìm hiểu các lớp cơ bản của CNN cũng như tham khảo các dự án có trên Kaggle và Github cũng như cân nhắc về tài nguyên, nhóm xây dựng mạng gồm các lớp như ảnh 5. Bên cạnh các lớp cơ bản như trên, để tránh hiện tượng quá khớp, nhóm có thêm các lớp dropout cũng như thêm một lớp flatten để "trải" dữ liệu ra một chiều sau các lớp tích chập và tổng hợp. Các siêu tham số (hyper parameter) đang được để dưới dạng ẩn (X1, X2, …) do nhóm sẽ sử dụng Keras tuner để tiến hành

chọn bộ siêu tham số phù hợp, đem lại kết quả cao nhất thay vì cố định hết siêu tham số từ đầu. Keras tuner cho phép người dùng đặt ra các rảng buộc cho các siêu tham số (ví dụ như kiểu dữ liệu, giá trị min, max của dữ liệu, các bước quét) và truyền các siêu tham số này như một ẩn vào hàm build và tiến hành tìm kiếm (search) bộ tham số mang lại hiệu quả cao nhất dựa theo các thuật toán và hàm mất mát, phương thức đánh giá mà người dùng muốn.



Ảnh 5: Kiến trúc mô hình CNN nhóm sử dụng

3.2.5 Hàm tối ưu (optimizer)

Hàm tối ưu là thành phần quan trọng trong quá trình huấn luyện, có vai trò là tối ưu hóa hàm mất mát bằng cách điều chỉnh các trọng số của mạng. Có nhiều hàm tối ưu, trong đó Adam (adaptive moment Estimation) thường là lựa chọn phổ biến do khả năng kết hợp thông tin từ cả gradient hiện tại, gradient quá khứ để điều chỉnh tỷ lệ học (learning rate) cho từng trọng số riêng lẻ và đạt hiệt suất tốt trong nhiều tình huống. Chính vì vậy nhóm sẽ sử dụng Adam optimizer.

3.2.6 Hàm kích hoạt (activation)

Hàm kích hoạt đóng vai trò quan trọng, có nhiệm vụ biến đổi đầu vào theo hàm kích hoạt để thu được đầu ra. Hàm kích hoạt thường là hàm phi tuyến, giúp mô hình có khả năng học các mối quan hệ phi tuyến và làm mạng linh hoạt hơn. Các hàm kích hoạt cơ bản có thể kể đến như hàm sigmoid, hàm tanH, hàm ReLU (Rectified Linear Unit) và hàm Softmax. Trong mô hình của nhóm, nhóm sử dụng hàm kích hoạt ReLU cho các lớp và hàm Softmax cho đầu ra.

Hàm ReLU có các ưu điểm khi đạo hàm của nó là đạo hàm của nó liên tục bằng 1, từ đó tránh hiện tượng vanishing gradient và các giá trị âm ở đầu vào được chuyển đổi thành 0 trong khi các giá trị dương sẽ giữ nguyên giá trị. ReLU thường cho hiệu suất đào tạo tốt hơn và hội tụ nhanh hơn so với các hàm kích hoạt khác trong khi có đạo hàm đơn giản và dễ tính, giảm độ phức tạp tính toán trong quá trình học.

Trong khi đó hàm softmax được chọn làm hàm kích hoạt ở đầu ra cho nó được ứng dụng cho bài toán phân loại đa lớp với phân phối xác suất đa thức, phù hợp với yêu cầu bài toán của nhóm.

3.2.7 Hàm mất mát

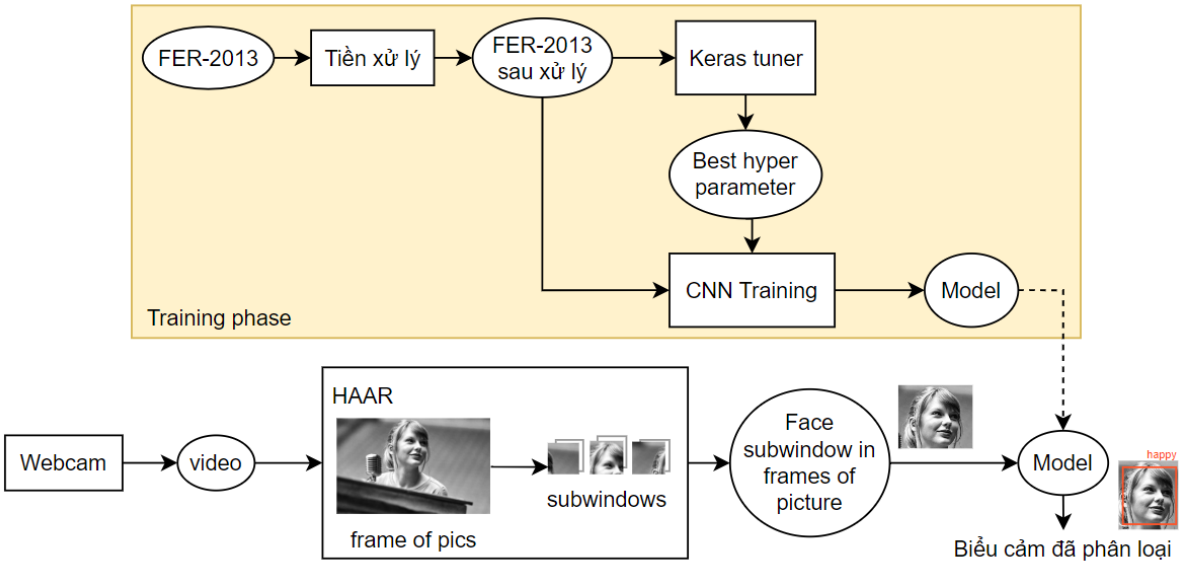
Hàm mất mát được sử dụng để tính toán sai số (sự khác biệt) giữa dự đoán của mô hình và giá trị thực tế trong tệp đánh giá, được sử dụng để điều chỉnh tham số của mô hình (vì mục đích của mô hình sẽ là tối thiểu hóa hàm mất mát).

Hàm mất mát được sử dụng phổ biến cho bài toán phân loại đa lớp là Categorical Cross-entropy, giúp đo lường sự tương thích giữa phân phối xác suất dự đoán và phân phối xác suất thực tế của các lớp. Hàm mất mát này thu được kết quả tương đối tốt với các bài toán phân loại đa lớp.

Trong quá trình tìm hiểu, nhóm còn tìm hiểu được hàm mất mát Focal Loss, một hàm mất mát được ứng dụng nhiều cho những bài toán có tệp dữ liệu mất cân bằng do Focal Loss còn tình toán dựa theo xác suất xuất hiện của từng lớp. Chính vì vậy nhóm cũng sẽ thử xây dựng mô hình sử dụng cả Focal Loss song song với Categorical Cross-entropy để so sánh hiệu quả.

3.3 Huấn luyện và triển khai

Sau khi xác nhận được kiến trúc mạng, nhóm tiền hành huấn luyện theo nhiều phương án và triển khai thực tế theo các bước như ảnh 6.



Ảnh 6: Sơ đồ các bước huấn luyện và triển khai mô hình

Các bước huấn luyện như tiền xử lý, sử dụng Keras tuner để thu được bộ siêu tham số cũng như kiến trúc mô hình CNN đã được đề cập ở phần trước. Nhóm tiến hành huấn luyện nhiều mô hình có cùng kiến trúc như được mô tả ở ảnh 5 nhưng với các tập dữ liệu, hàm mất mát khác nhau cũng như ứng dụng thêm một kỹ thuật để giảm tốc độ học (learning rate) động trong quá trình train dựa vào giá trị của hàm mất mát để xem ảnh hưởng và so sánh các mô hình. Mô hình CNN6, kiến trúc tương đối giống các mô hình khác (chỉ trừ việc có thêm các lớp BatchNormalization). Cụ thể về các mô hình được mô tả như bảng 8.

Bảng 8: Các mô hình nhóm huấn luyện

	Bộ dữ liệu	**Hàm mất mát**	**Tốc độ học động**
CNN1	FER-2013 gốc (bảng 5)	Categorical Cross-entropy	Không
CNN2	FER-2013 gốc (bảng 5)	Categorical Cross-entropy	Có
CNN3	FER-2013 gốc (bảng 5)	Focal Loss	Không
CNN4	FER-2013 đã augment (bảng 6)	Focal Loss	Có
CNN5	FER-2013 đã augment (bảng 6)	Categorical Cross-entropy	Có
CNN6	FER-2013 clean và augment (bảng 7)	Categorical Cross-entropy	Có

Khi triển khai, nhóm sẽ sử dụng camera hoặc webcam để thu hình ảnh người dùng, sau đó cho đi qua khối HAAR để nhận diện khuôn mặt. Sau khi qua khối HAAR sẽ thu được từng frame của ảnh từ video. Những khung ảnh này được đưa vào mô hình phân loại biểu cảm khuôn mặt đã được huấn luyện (nhóm sẽ chọn mô hình cho kết quả cao nhất) và hiển thị kết quả lên màn hình video sau khi thực hiện tính toán và phân loại.

Nhóm tiến hành huấn luyện và triển khai trên máy tính cá nhân 16GB RAM, core i7-8750H với tần số hoạt động 2.2GHz.

CHƯƠNG 4
KẾT QUẢ, ĐÁNH GIÁ VÀ ĐỀ XUẤT CẢI THIỆN

Trong phần này sẽ tập trung trình bày kết quả huấn luyện các mô hình, sau đó đánh giá về hiệu suất cũng như đề xuất thêm các phương án cải thiện.

4.1 Các tiêu chí đánh giá mô hình học máy, học sâu

Có rất nhiều tiêu chí để đánh giá mô hình học máy và học sâu cho bài toán phân loại nhưng ở đây nhóm sẽ đánh giá kết quả chủ yếu dựa vào accuracy, bên cạnh đó còn xét thêm precision, recall. Để hiểu về các tiêu chí này, ta cần biết về confusion matrix trước (chú ý là confusion matrix không phải một tiêu chí đánh giá).

4.1.1 Confusion matrix

Confusion matrix là kết quả thống kê dưới dạng bảng như ảnh 7, thể hiện được bao nhiêu mẫu dữ liệu test thực sự thuộc 1 lớp và được mô hình dự đoán là một lớp.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Ảnh 7: Mô hình confusion matrix

Các chỉ số TP, FP, TN, FN lần lượt có ý nghĩa như sau:

- TP (True Positive): Số trường hợp mô hình dự đoán là Positive và thực tế là Positive
- TN (True Negative): Số trường hợp mô hình dự đoán là Negative và thực tế là Negative
- FP (False Positive): Số trường hợp mô hình dự đoán là Positive và thực tế là Negative
- FN (False Negative): Số trường hợp mô hình dự đoán là Negative và thực tế là Positive

4.1.2 Công thức tính các tiêu chí đánh giá mô hình huấn luyện

Như đã trình bày, nhóm sẽ sử dụng các tiêu chí như accuracy, precision và recall. Công thức tính các tham số này như sau:

- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$

Ta có thể thấy từ các công thức, accuracy sẽ đánh giá khả năng của tổng thể mô hình với tất cả các lớp nhưng chỉ đánh giá lượng dữ liệu được phân loại đúng mà không chỉ ra cụ thể mỗi loại phân loại thế nào, hay lớp nào nay bị phân loại nhầm. Trong khi precision sẽ thường được sử dụng khi muốn tránh việc dự đoán sai một mẫu là positive trong khi nó là negative, còn recall được sử dụng khi muốn tránh việc dự đoán mẫu là negative trong khi mẫu đó positive.

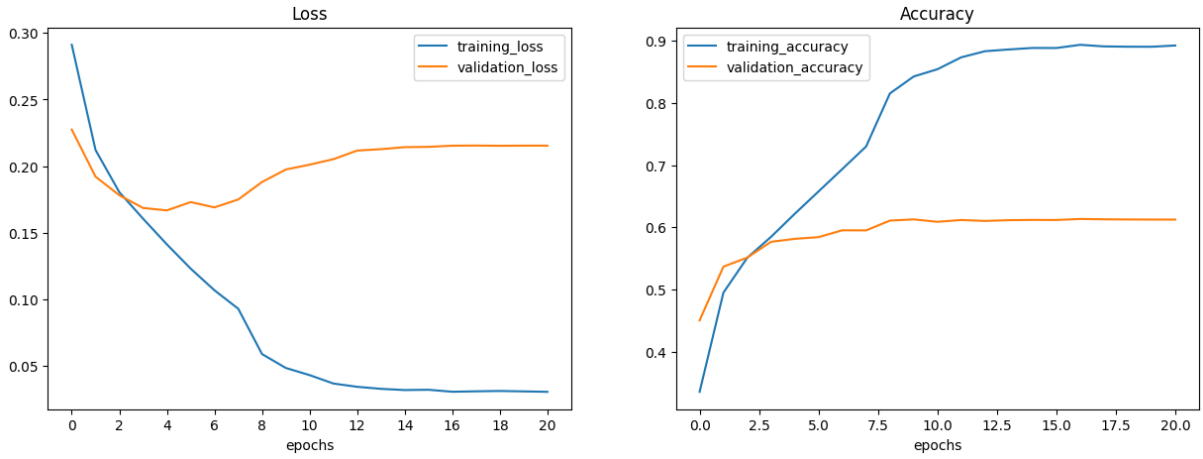
4.2 Kết quả huấn luyện các mô hình phân loại biểu cảm

Kết quả huấn luyện của các mô hình được tổng hợp ở bảng 9.

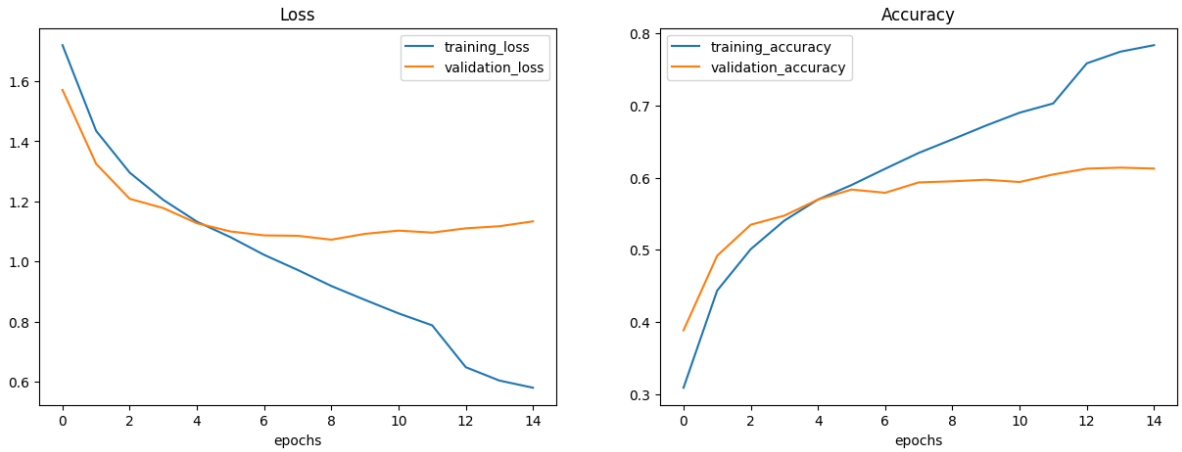
Bảng 9: Kết quả huấn luyện các mô hình

Mô hình	Loss	Accuracy	Lượng tham số
CNN1	1.53511	60.76%	6,002,719
CNN2	1.16322	61.63%	6,002,719
CNN3	0.2454	60.27%	5,774,273
CNN4	**0.21538**	61.24%	5,774,273
CNN5	1.15496	62.16%	4,478,087
CNN6	0.9498	**68.00%**	**2,347,015**

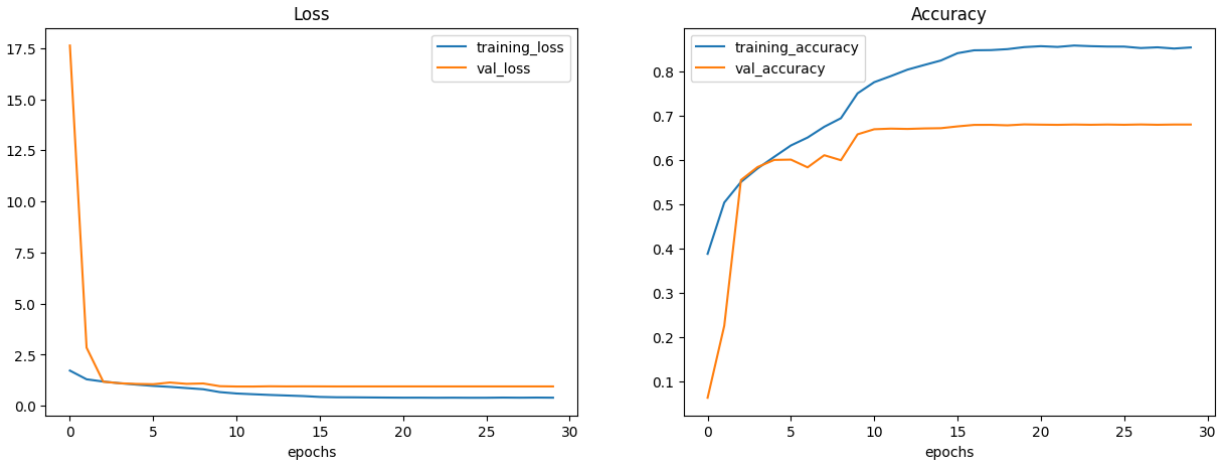
Biểu đồ ở ảnh 8, 9, 10 lần lượt thể hiện loss và accuracy qua từng epoch train của các mô hình CNN4, CNN5, CNN6.



Ảnh 8: Biểu đồ loss và accuracy theo từng epoch cho mô hình CNN4



Ảnh 9: Biểu đồ loss và accuracy theo từng epoch cho mô hình CNN5

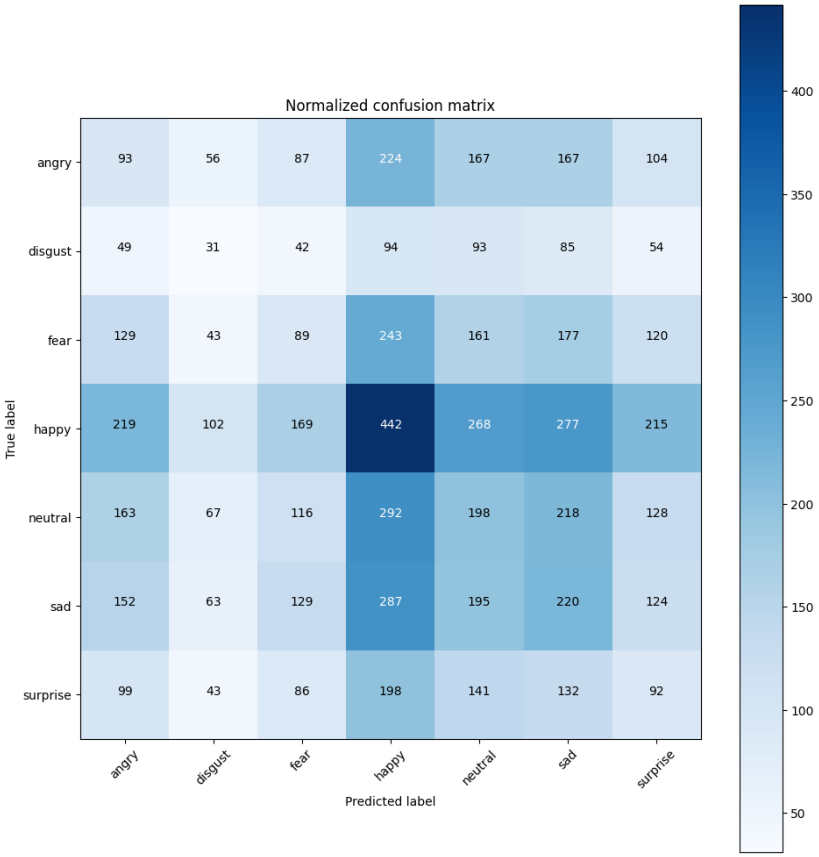


Ảnh 10: Biểu đồ loss và accuracy theo từng epoch cho mô hình CNN6

Ta nhận thấy trong các mô hình, mô hình CNN6 (mô hình lấy dữ liệu đã được làm sạch thủ công và augmentation, sử dụng Categorical Cross-entropy cũng như giảm learning rate tự động) cho được kết quả accuracy cao nhất. Các mô hình còn lại kết quả dao động ở mức 60.2% tới 62.2%. Cụ thể mô hình CNN6 với các hyper-parameter được trình bày ở bảng 10 và confusion matrix của mô hình CNN6 được thể hiện ở ảnh 11.

Bảng 10: Hyper-parameter của mô hình CNN6

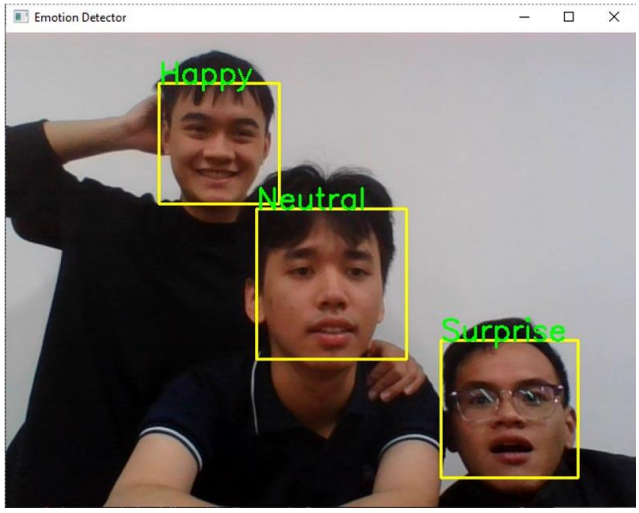
Tham số	**Filter Conv1**	**Filter Conv2**	**Drop1 rate**	**Filter Conv3**	**Filter Conv4**	**Drop2 rate**	**Dence neuron**	**Drop3 rate**
Giá trị	32	128	0.25	128	128	0.25	1024	0.5



Ảnh 11: Biểu đồ loss và accuracy theo từng epoch cho mô hình CNN6

4.3 Kết quả triển khai toàn bộ hệ thống

Nhóm tiến hành chạy chương trình toàn bộ hệ thống trên các máy tính cá nhân của các thành viên cũng như gửi chương trình đi cho nhiều người dùng thử. Ảnh 12 mô tả màn hình khi chạy thử hệ thống.



Ảnh 12: Chạy thử chương trình trong thực tế

Nhóm tiến hành đo tốc độ nhận diện và xử lý của chương trình thì thu được kết quả thu được từ 50fps tới 66fps.

4.4 Đánh giá và đề xuất cải thiện

Tất cả các mô hình có độ chính xác trên 60% và loss dao động quanh 1. Các mô hình bắt đầu hội tụ ở vòng lặp thứ 25. Độ chính xác tuy đã đáp ứng được yêu cầu (mô hình có độ chính xác cao nhất là CNN6 với 68%) do nhóm đặt ra song còn chưa được cao. Điều này một phần do mô hình mạng còn khá đơn giản cũng như dataset còn nhẹ và nhỏ (cớ 48x48 gray scale) cùng sự mất cân bằng, nhiều khuyết điểm (gán nhãn sai) và hình ảnh bất thường kể cả trong tập huấn luyện cũng như tập kiểm tra. Tham khảo thống kê từ [4], nhóm thống kê và so sánh với một số bài nghiên cứu khác sử dụng CNN trên cùng tập dữ liệu thấy mang lại kết quả tương đối tốt. Bảng so sánh được thể hiện tại bảng 11.

Bảng 11: So sánh kết quả với các mô hình khác trên cùng tập dữ liệu

Methodology	Data set	Year	Accuracy (%)
CNN	FER-2013	2021	72.16
CNN + SVM	FER-2013	2020	71.16
Proposed CNN	**FER-2013**	**2023**	**68**
CNN	FER-2013	2018	67.76
CNN	FER-2013	2021	67.18
ResNet 50	FER-2013	2017	65.1
CNN	FER-2013	2020	61.4
VGG-16	FER-2013	2017	59.2
CNN	FER-2013	2016	48
SVM	FER-2013	2017	31.8

Các mô hình có khả năng nhận dạng tốt các cảm xúc phổ biến như vui vẻ, tức giận, bất ngờ. Cảm xúc như ghê tởm, sợ hãi khó phân loại hơn vì dataset cho các cảm xúc này còn ít và không đa dạng. Mô hình sử dụng Focal Loss có loss tối ưu (loss nhỏ hơn) so với Categorical Cross-entropy, song kết quả huấn luyện lại cho độ chính xác kém hơn

Về mô hình để nhận diện khuôn mặt cũng chưa thực sự chính xác và không hoạt động nếu ngửa mặt về sau nhiều. Tuy nhiên, với tốc độ đo được của hệ thống là khoảng 50 tới 66 fps, hệ thống đã đạt tốt mục tiêu đề ra (>20fps) để thỏa mãn tính thời gian thực.

Để cải thiện hiệu suất, nhóm có đề ra một số phương án. Đầu tiên có thể tự thu thập dữ liệu và gán nhãn để đảm bảo độ chất lượng của tập dữ liệu, đồng thời, áp dụng các kỹ thuật khai thác để thu

được nhiều dữ liệu hơn từ các nguồn trực tuyến như hình ảnh, video chứa đựng nhiều tình huống biểu đạt cảm xúc khác nhau. Quá trình này giúp mô hình cập nhật và làm phong phú thêm kiến thức về những biểu hiện cảm xúc mới, từ đó nâng cao độ chính xác và khả năng khái quát hóa.

Một phương án nữa là sau khi có dữ liệu, thay vì dùng mô hình HAAR của OpenCV, có thể sử dụng YOLO hoặc các mô hình mạng khác như kết hợp nhiều loại mạng để thực hiện nhận diện khuôn mặt trong các hình ảnh và video được thu thập. Mô hình YOLO có thể sẽ xác định chính xác hơn khuôn mặt trong ảnh và do đó gắn nhãn các vùng chứa khuôn mặt và biểu cảm cảm xúc. Sử dụng CNN kết hợp với các loại mạng khác hoặc có thể thử phương án học chuyển tiếp để tiết kiện thời gian huấn luyện.

Tóm lại, trong tương lai mô hình cần được cải thiện nhiều hơn về chất lượng dữ liệu, mô hình huấn luyện để tăng khả năng áp dụng vào thực tế.

KẾT LUẬN

Nhóm đã tiến hành tìm hiểu và triển khai mô hình nhận diện khuôn mặt và biểu cảm con người trong thời gian thực qua video. Tuy kết quả chưa cao song nhóm cũng học được nhiều phương pháp xử lý một bài toán đặt ra để xây một mô hình, từ bước đặt vấn đề, tìm kiếm giải pháp tới việc tiền xử lý, làm sạch dữ liệu, tuning tham số, triển khai và đánh giá.

Để hoàn thành báo cáo này, chúng em xin cảm ơn sự hướng dẫn của TS. **Võ Lê Cường,** anh trợ giảng **Nguyễn Đoàn Khuê** cũng như các tài liệu tham khảo (đề cập cuối báo cáo).

Tuy đã tìm hiểu và cố gắng song chúng em không tránh khỏi những thiếu sót, em mong nhận được những lời nhận xét và góp ý của thầy để hoàn thiện hơn trong tương lai.

Nhóm 23

TÀI LIỆU THAM KHẢO

[1] Z. C. Yousif Khaireddin, "Facial Emotion Recognition: State of the Art Performance on FER2013," *Boston University,* 2021.

[2] https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/overview, "Kaggle," 2013. [Online]. Available: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/overview.

[3] S. Lonkar, "Facial Expressions Recognition with Convolutional Neural Networks," *SSN No: 2456-2165,* vol. 6, pp. 51-55, 2021.

[4] J. L. J. Santhosh P. Mathew, "Facial Expression Recognition for the Blind Using Deep Learning," *IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON),* 2021.