ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

FACIAL EXPRESSION RECOGNITION

Giảng viên: Nguyễn Vinh Tiệp

Lớp: CS331.M21

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Nguyễn Thị Như Ý	19522555
2	Nguyễn Tấn Tú	19522454
3	Đinh Hoàng Linh Đan	19521309
4	Trần Nguyễn Quỳnh Anh	19521217

TP. HỒ CHÍ MINH – 6/2022

Mục Lục

1. GIỚI THIỆU

Nhận diện biểu cảm biểu cảm là một trong những bản năng của con người

Con người thế hiện cảm xúc thông qua các biểu cảm khuôn mặt, luôn tồn tại một số đặc trưng chung trên khuôn mặt bất kể độ tuổi, vị trí địa lí hay điều kiện sống... Dựa vào đặc trưng này, ta rút trích ra các đặc điểm quan trọng của cảm xúc, mô hình hóa, và dạy cho máy tính hiểu được cảm xúc đó.

Ta chia biểu cảm khuôn mặt vào bảy loại sắc thái chính: tức giận (angry), sợ hãi (fear), ngạc nhiên (surprised), buồn (sad), chán ghét (disgusted), hạnh phúc (happy), trung lập (neutral).

Hình 1: Biểu cảm gương mặt người

2. NỘI DUNG

2.1. Input, Output

- Input: Video webcam hoặc bức ảnh chưa mặt người
  - Quay/ chụp chính diện

+ Ánh sáng tốt
+ Không vật cản lớn trên gương mặt

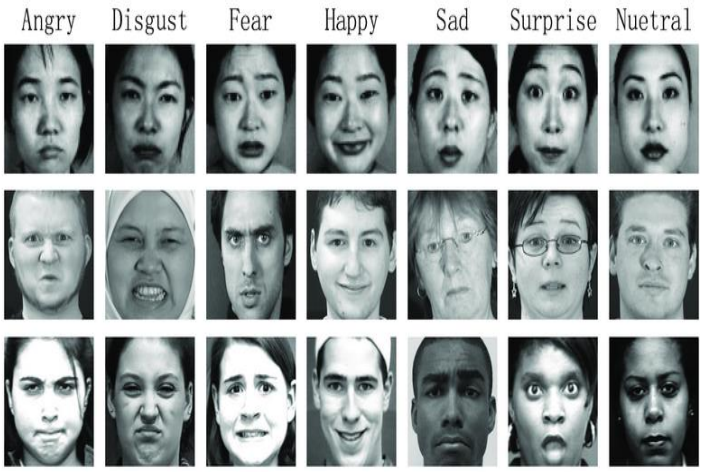
Hình 2: Input và Output

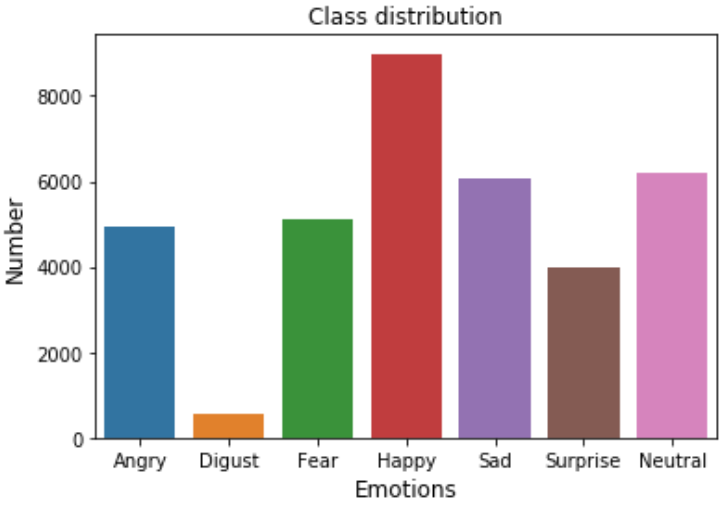
2.2. Ứng dụng

- Trong lĩnh vực tiếp thị: phân tích phản hồi của khách hàng
- Trong dịch vụ khách hàng: phân tích mức độ hài lòng của khách hàng.
- Trong trò chơi điện tử và trò chơi thực tế ảo: kiểm tra trải nghiệm người dùng.
- Trong y tế: Giúp bác sĩ và chuyên gia y tế đánh giá sức khỏe bệnh nhân.
- Xác định những người có **hành vi đáng ngờ** trong đám đông từ đó có thể ngăn chặn tội phạm, khủng bố.

2.3. Dataset

- Dataset: Fer2013 gồm 35887 ảnh xám kích cỡ 48x48 chia thành 2 tập training và validation theo tỉ lệ lần lượt là 0.8 và 0.2
- Có 7 lớp biểu cảm:

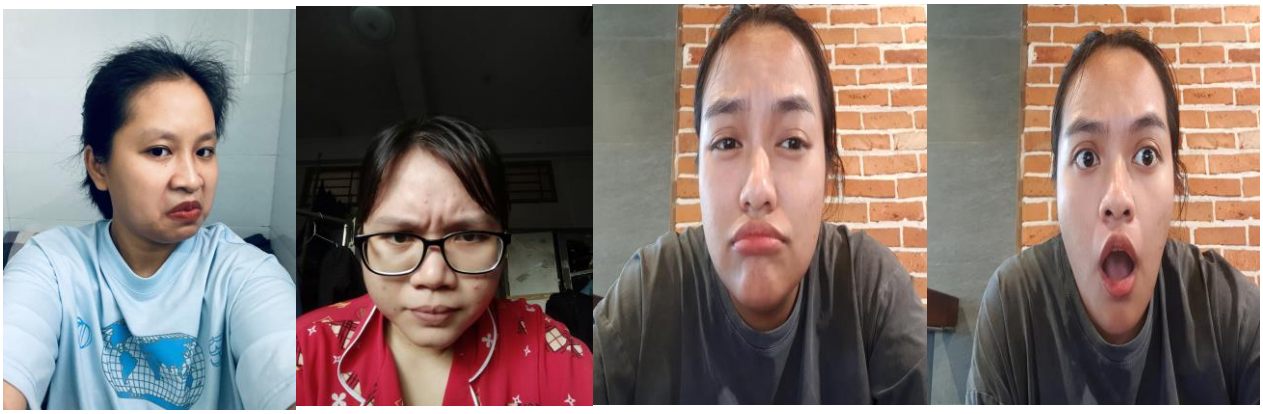




	emotion	number
0	Angry	4953
1	Digust	547
2	Fear	5121
3	Happy	8989
4	Sad	6077
5	Surprise	4002
6	Neutral	6198

Hình 3: Dataset Fer2013

Ngoài ra, còn sử dụng thêm tập private test gồm 883 ảnh được chia theo đúng tỉ lệ của tập training



Hình 4: Một số ảnh trong private test

2.4. Các nghiên cứu trước đây

- **Inception**

- Inception module là một mạng CNN giúp training wider (thay vì thêm nhiều layer hơn vì rất dễ xảy ra overfitting + tăng parameter người ta nghĩ ra tăng deeper ở mỗi tầng layer) so với mạng CNN bình thường.
- Tương đối phức tạp về mặt tính toán.
- Accuracy: 71.6%

- **Local learning deep + BOW**

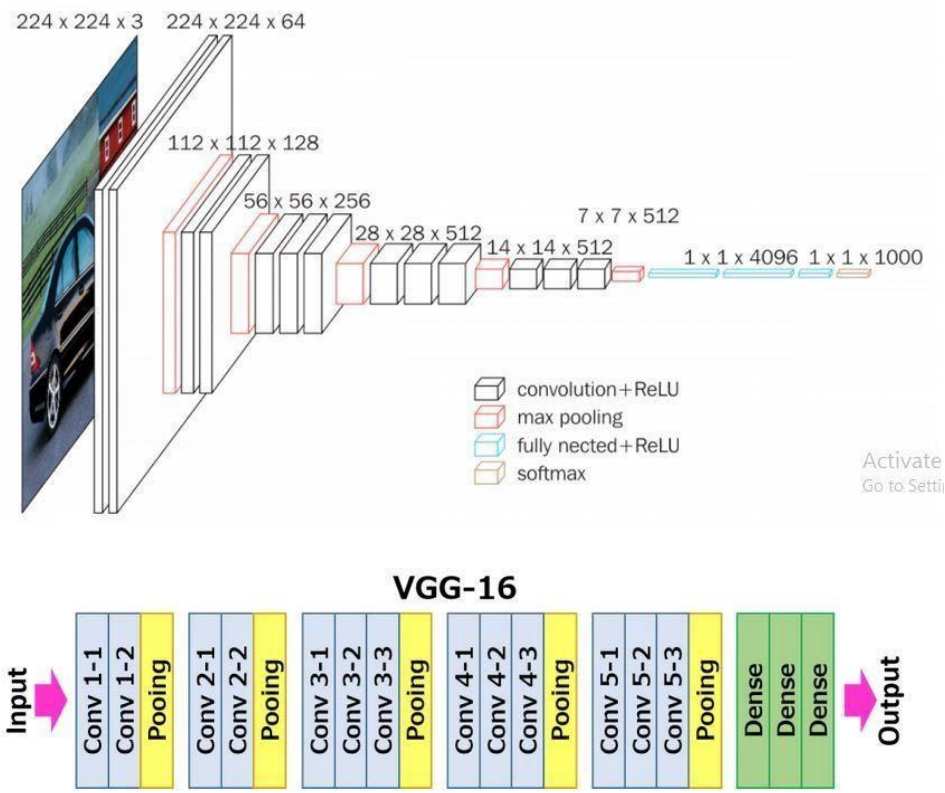
- Là sự kết hợp của feature được học bởi CNN và feature thủ công được tính toán bởi BOVW model.
- Chưa phân biệt được biểu cảm giả vờ và biểu cảm thật.
- Accuracy: 75.42%

- **LHC-Net**

- LHC là một khối xử lý được tích hợp vào một kiến trúc phức hợp đã có từ trước. Nó kế thừa seft-attention module từ kiến trúc Transformer nổi tiếng.
- Dataset không cân bằng.
- Không tối ưu hóa cấu trúc liên kết chung của LHC-Net với các hyper-parameter của các attention block.
- Accuracy: 74.42%

2.5. Model

2.5.1. VGG16



Hình 5: Kiến trúc mạng VGG16

Mạng VGG16 là một mạng có 13 lớp convolution, 5 lớp max pool và 3 lớp dense. Số 16 trong VGG16 là tượng trưng cho số lớp có trọng số (weights). Các lớp convolution trong mạng đều có kernel 3x3, sau mỗi layer conv là maxpooling downsize xuống 0.5, và 3 layer fully connection ở cuối cùng.

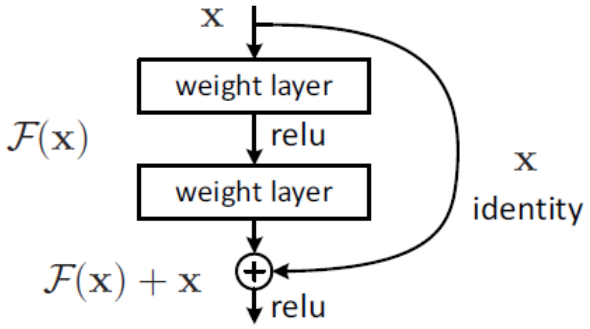
Với kích thước ảnh là 224x224 sau khi qua 2 lớp convolution đầu tiên sẽ tới lớp max pool làm giảm kích thước ảnh xuống 112x112. Tương tự như vậy sau mỗi lớp max pool kích thước ảnh sẽ giảm xuống một nửa.

2.5.2. ResNet50

Vanishing Gradient là việc các gradient trong quá trình Backpropagation khi được cập nhật trở lại các layers trước đó qua

càng nhiều lớp thì nó càng ít đổi nhiều. Dẫn đến kết quả là các cập nhật thực hiện bởi Gradients Descent không làm thay đổi nhiều weights của các layers đó và chúng không thể hội tụ và mạng sẽ không thu được kết quả tốt.

Giải pháp mà tác giả của bài báo về ResNet đưa ra đó chính là các kết nối tắt (shortcut connections) để xuyên qua một hay nhiều lớp. Một khối như vậy còn được gọi là các Residual Block.



Hình 6: Residual Block

Đầu vào các khối Residual Block là **x** và đầu ra của nó là **F(x)**.

Với **F(x)** có được từ **x** sau: **x → weight1 → ReLU → weight2**

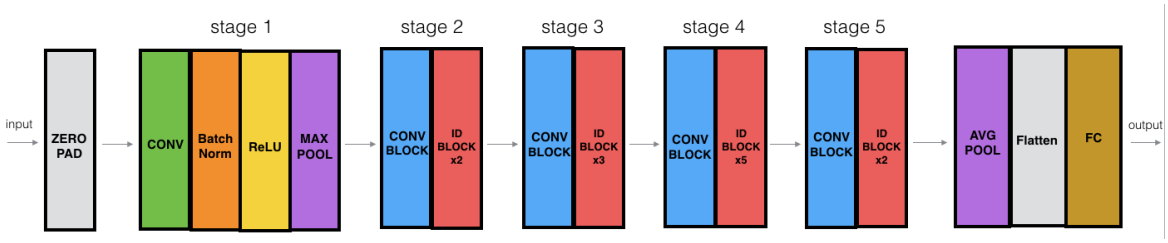
Với **output H(x)** có được bằng cách: **H(x) = F(x) + x**

Do đó ResNet có thể làm việc với hàng trăm hoặc hàng nghìn lớp tích chập mà vẫn có thể đạt được hiệu suất tốt.

⇒ Chính vì những lí do trên, ResNet đã trở thành mô hình mạng nổi tiếng trong các bài toán phân lớp (classification). Do đó việc lựa chọn ResNet cho bài toán này là tất yếu.

Kiến trúc Model:

- ResNet Model gồm 48 lớp Convolution, 1 lớp MaxPool và 1 lớp Average Pool. ResNet50 và ResNet152 là biến thể của ResNet Model, với con số đằng sau là số lớp convolution của model đó có. Như vậy thì ResNet50 sẽ có 50 lớp convolution, còn ResNet152 thì có 152 lớp convolution.



Hình 7: ResNet50 Model

Trước khi vào lớp convolution đầu tiên, dữ liệu đầu vào sẽ qua lớp Zero Padding 2D để đảm bảo sau khi qua lớp Conv, 7x7, 64, stride 2 thì dữ liệu sẽ giảm kích thước đi một nửa.

Ví dụ: Ảnh vào có kích thước (224, 224, 3) sau khi qua lớp Zero Padding 2D thì ảnh sẽ có kích thước (230,230,3). Sau khi qua lớp Conv, 7x7 , 64, stride 2 thì kích thước ảnh còn lại là (112, 112, 64)

Ở câu trúc residual block trên ta thấy các kết nối tắt này xuyên qua 2 weight layers. Từ ResNet50 trở đi thì nó xuyên qua tới 3 weight layers.

Điều đặc biệt ta cần nhắc đến ở mô hình ResNet50 này đó chính là các Convolution Block (trong ảnh là **CONV BLOCK**) và các Identity Block (trong ảnh là **ID BLOCK**). Cả Convolution Block và Identity Block đều là các residual block. Nhưng chúng có 1 điểm khác biệt nho nhỏ đó chính là:

Convolution Block: theo như cấu trúc các residual block thì các **output H(x) = F(x) + x.** Tuy nhiên ở các Convolution Block, giá trị **input x** sẽ đi qua 1 lớp convolution thứ 3. Điều này nhằm đảm bảo **F(x)** và **x** có cùng kích thước với nhau.

Identity Block: tương tự như các residual block bình thường, **output H(x) = F(x) + x.**

Để trực quan hơn, ta cùng quan sát ảnh của nhóm tác giả về kiến trúc các mạng ResNet18, ResNet34, ResNet50, ResNet101, ResNet152

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Hình 8: Kiến trúc các mạng ResNet

Như ta có thể thấy ở trong hình, kiến trúc mạng ResNet50 gồm như sau:

Một lớp convolution với 64 kernel 7x7, stride 2 → **1 lớp convolution**

Một lớp max pool stride 2

Residual block qua 3 lớp convolution lần lượt là 1x1, 64 kernel; 3x3, 64 kernel; 1x1, 256 kernel. Lặp lại 3 lần → **9 lớp convolution**

Tiếp đó là residual block qua 3 lớp convolution lần lượt là 1x1, 128 kernel; 3x3, 128 kernel; 1x1, 512 kernel. Lặp lại 4 lần → **12 lớp convolution**

Sau đó là residual block qua 3 lớp convolution lần lượt là 1x1, 256 kernel; 3x3, 256 kernel; 1x1, 1024 kernel. Lặp lại 6 lần → **18 lớp convolution**

Tiếp tục lại là residual block qua 3 lớp convolution lần lượt là 1x1, 512 kernel; 3x3, 512 kernel; 1x1, 2048 kernel. Lặp lại 3 lần → **9 lớp convolution**

Cuối cùng ta qua lớp average pool và kết thúc với lớp fully connected gồm 1000 nodes cùng softmax function → **1 lớp convolution**

Như vậy ResNet50 gồm 50 lớp convolution, ta không đếm các lớp max pooling và average pooling cũng như các activation functions.

2.5.3. Training Model

Nhóm transfer learning 2 model ResNet50 và VGG16 bằng cách : Dùng pre-trained model ResNet50 và VGG16.

- Thêm các lớp sau:

◆GlobalAveragePooling2D()

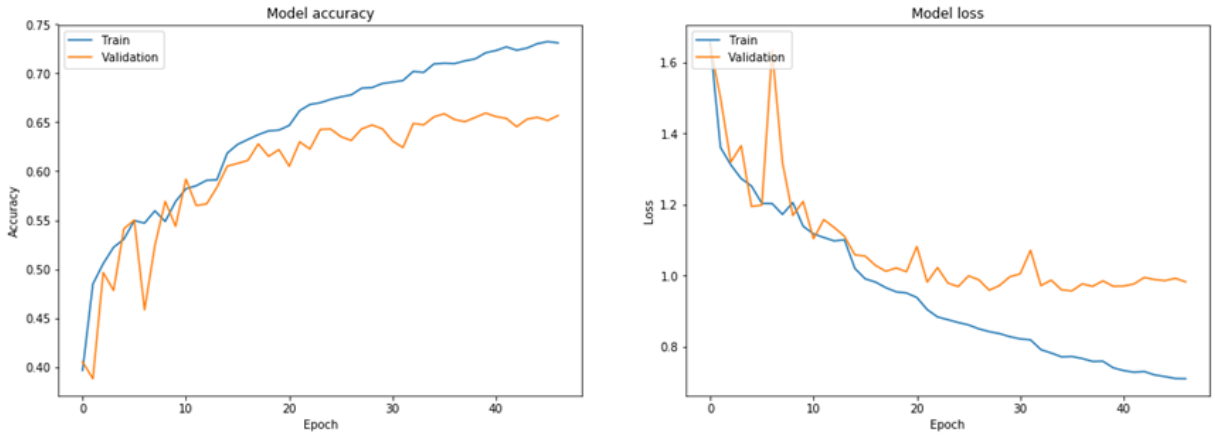
◆Dropout(0.2)

◆Dense(7, activation='softmax')

- Sử dụng learning rate thấp : 0.0005.
- Parameters: ResNet50 (23,602,055)

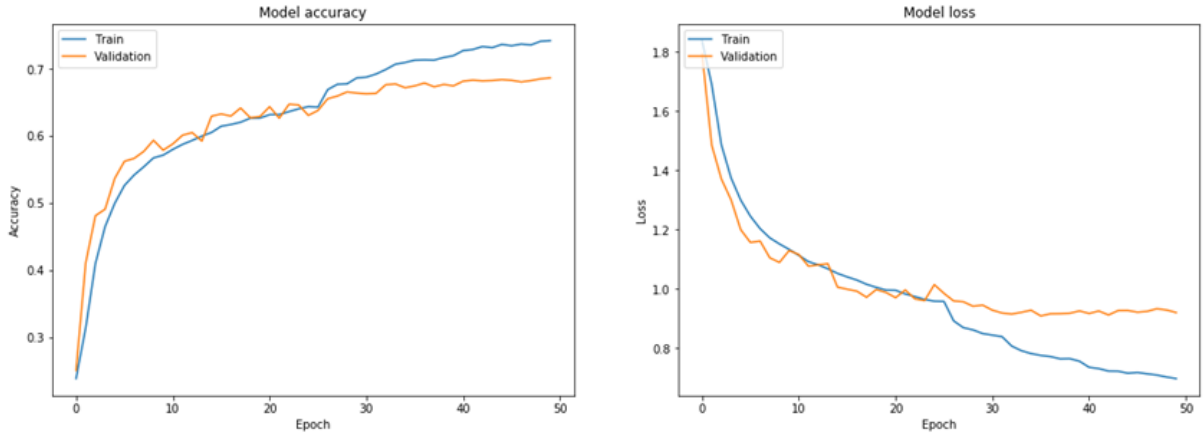
VGG16 (14,718,279)

Kết quả sau khi train của ResNet50:



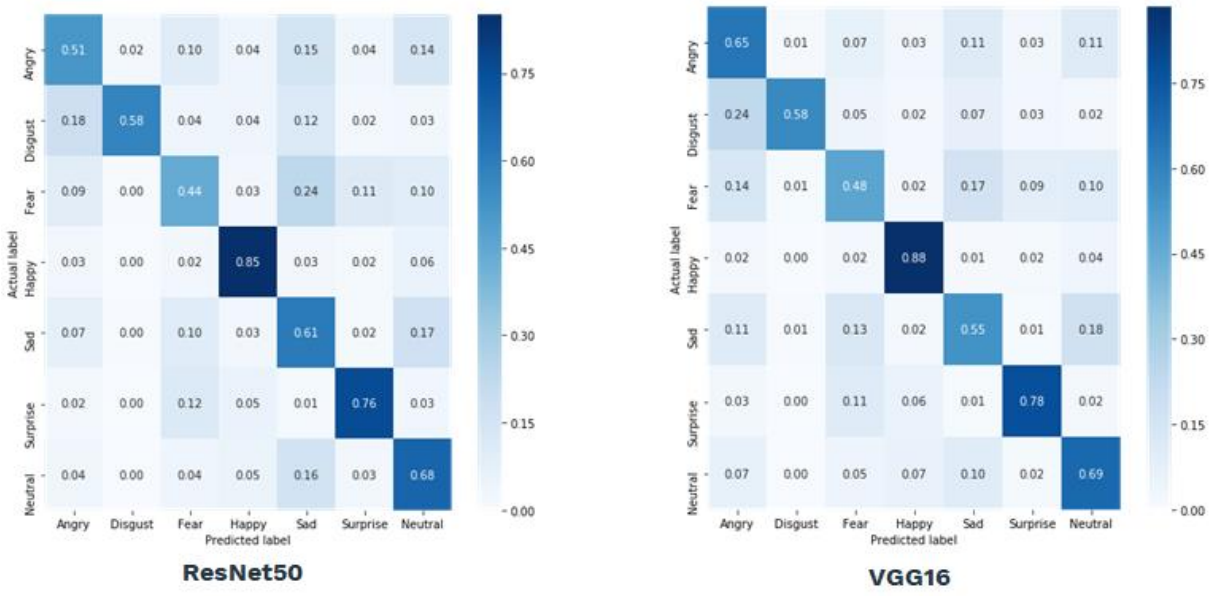
Hình 9: Kết quả của ResNet50

Kết quả sau khi train của VGG16:



Hình 10: Kết quả của VGG16

So sánh kết quả của 2 model bằng Normalized Confusion Matrix:



Hình 11: Normalized Confusion Matrix của 2 model ResNet50 và VGG16

2.6. Thách thức bài toán và hướng phát triển

2.6.1. Thách thức bài toán

Việc nhận diện biểu cảm khuôn mặt có thể khó khăn với cả chúng ta, có nghiên cứu cho rằng cùng một loại biểu cảm tuy nhiên những người khác nhau sẽ biểu hiện khác nhau. Vì thế đó cũng chính là thách thức đối với AI:

a. Bộ dữ liệu (Dataset)

Bất cứ thuật toán máy học và học sâu nào cũng đòi hỏi rất nhiều dữ liệu sử dụng cho training. Dữ liệu phải bao gồm video bao gồm những tốc độ khung hình khác nhau, từ nhiều góc độ khác nhau, nhiều bối cảnh khác nhau, và lấy dữ liệu từ những người thuộc các giới tính, quốc tính và chủng tộc khác nhau, v.v.

Tuy nhiên, hầu hết các tập dữ liệu công khai trong đó có bộ dữ liệu FER2013 được sử dụng trong đồ án được lấy từ người châu Âu, như chúng ta đã biết thì biểu cảm của người châu Âu được thể hiện rõ ràng hơn người châu Á, vì thế cũng giảm thiểu độ chính

xác của thuật toán, giới tính và có một vài biểu cảm còn hạn chế như số lượng ảnh disgust còn thiếu sót, hoặc biểu cảm surprise và fear khá giống nhau vì thế cũng giảm độ chính xác của thuật toán.

b. Các yếu tố ảnh hưởng

Ánh sáng: Sự thay đổi nhỏ trong điều kiện anh sáng cũng có thể gây ra thách thức đáng kể trong thuật toán nhận dạng biểu cảm gương mặt và cũng sẽ ảnh hưởng đến kết quả cảu thuậ toán. Nếu độ sáng khác nhau, cùng một cá nhân được chụp bằng cùng một máy ảnh và với biểu cảm và tư thế của khuôn mặt gần như giống hết nhau, kết quả thu được có thể khác nhau. Sự chiếu sáng làm thay đổi diện mạo khuôn mặt một cách đáng kể. Người ta nhận thấy rằng sự khác biệt giữa hai khuôn mặt giống nhau với các ánh sáng khác nhau nhiều hơn so với hai khuôn mặt khác nhau được chụp dưới cùng một ánh sáng.

Độ phân giải: Hình ảnh có độ phân giải thấp không cung cấp nhiều thông tin vì hầu hết đã bị mất đi. Đây cũng là một thách thức lớn trong quá trình nhận diện biểu cảm khuôn mặt.

Tư thế của khuôn mặt: Tư thế của một khuôn mặt thay đổi khi chuyển động của đầu và góc nhìn của người đó thay đổi. Các chuyển động của đầu hoặc góc nhìn khác nhau gây ra những thay đổi về diện mạo khuôn mặt và tạo ra các biển thể trong nội bộ lớp làm cho độ chính xác của thuật toán nhận diện khuôn mặt giảm từ đó độ chính xác của thuật toán nhận diện biểu cảm khuôn mặt cũng giảm đáng kể.

Những vật làm che khuất khuôn mặt: Đây có thể được coi là một trong những thách thức quan trọng nhất của hệ thống nhận dạng khuôn mặt. Những thứ làm che khuất khuôn mặt có thể kể đến

như râu, ria mép làm che khuất miệng của khuôn mặt, kính râm che khuất mắt, mũ lưỡi trai, khẩu trang, ...

2.6.2. Hướng phát triển

- Tăng số lượng mẫu của các biểu cảm để hệ thống đánh giá tốt hơn và chính xác hơn
- Thử nghiệm với nhiều mô hình mạng học sâu khác hơn.
- Cải thiện web API.
- Hướng đến phát hiện biểu cảm khuôn mặt thông qua video.
- Thu thập thêm tập dữ liệu người Châu Á để model nhận diện biểu cảm khuôn mặt của người Châu Á tốt hơn vì tập dữ liệu ***FER-2013*** là người Châu Âu.

2.7. Demo

2.7.1. Xây dựng web API

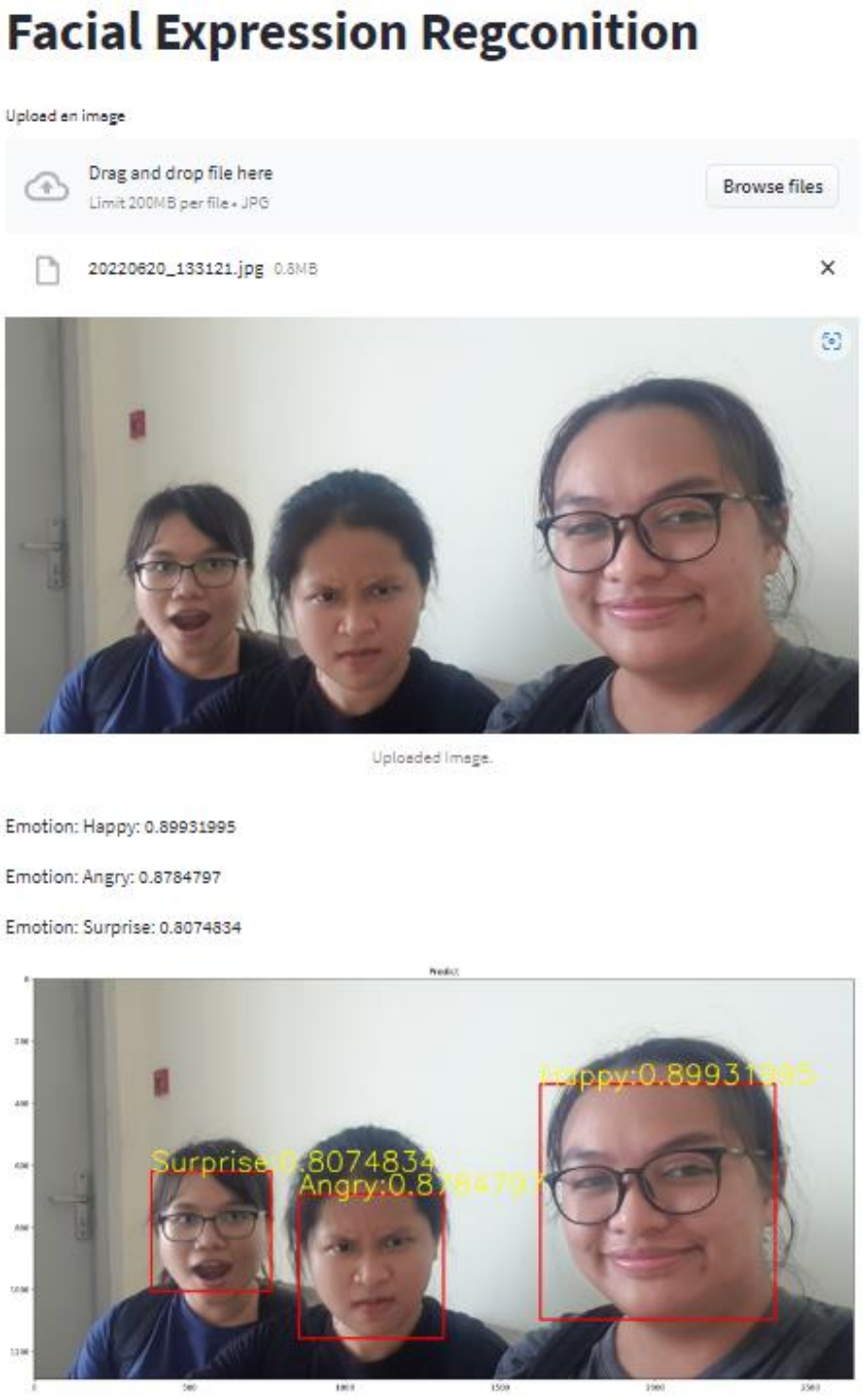
Sử dụng Spaces của Hugging Face và thư viện Streamlit để xây dựng web API. Trang web được sử dụng để người dùng có thể upload một ảnh lên, sau đó trang web sẽ dự đoán được biểu cảm gương mặt trong bức ảnh và in ra màn hình kèm theo độ chính xác.

Streamlit là một framework Python có mã nguồn mở và miễn phí. Người dùng có thể dễ dàng sử dụng để xây dựng và chia sẻ những dashboard tương tác cũng nhưng các ứng dụng web máy học của mình.

Hugging Face Spaces giúp người dùng dễ dàng tạo và triển khai ứng dụng web máy học. Hugging Face Spaces cho phép host và demo ứng dụng máy học. Spaces lưu trữ mô hình máy học và tập dữ liệu bên trong kho lưu trữ git.

2.7.2. Demo trang web sử dụng model VGG16:

Link trang web: [Facial Emotion Regconition - a Hugging Face Space by linhdan412]



Facial Expression Regconition

Upload an image

Drag and drop file here
Limit 200MB per file • JPG

Browse files

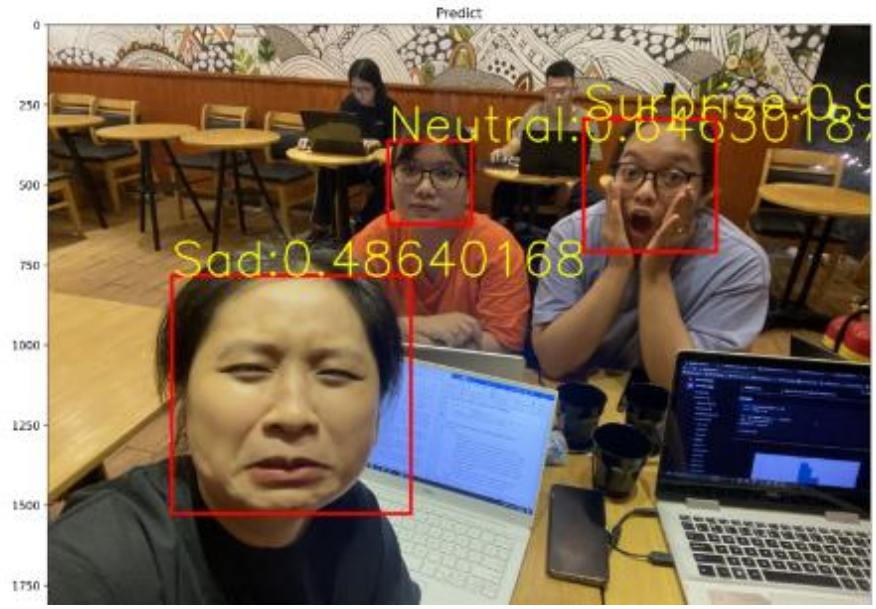
people.jpg 1.0MB

Uploaded Image.

Emotion: Neutral: 0.64630187

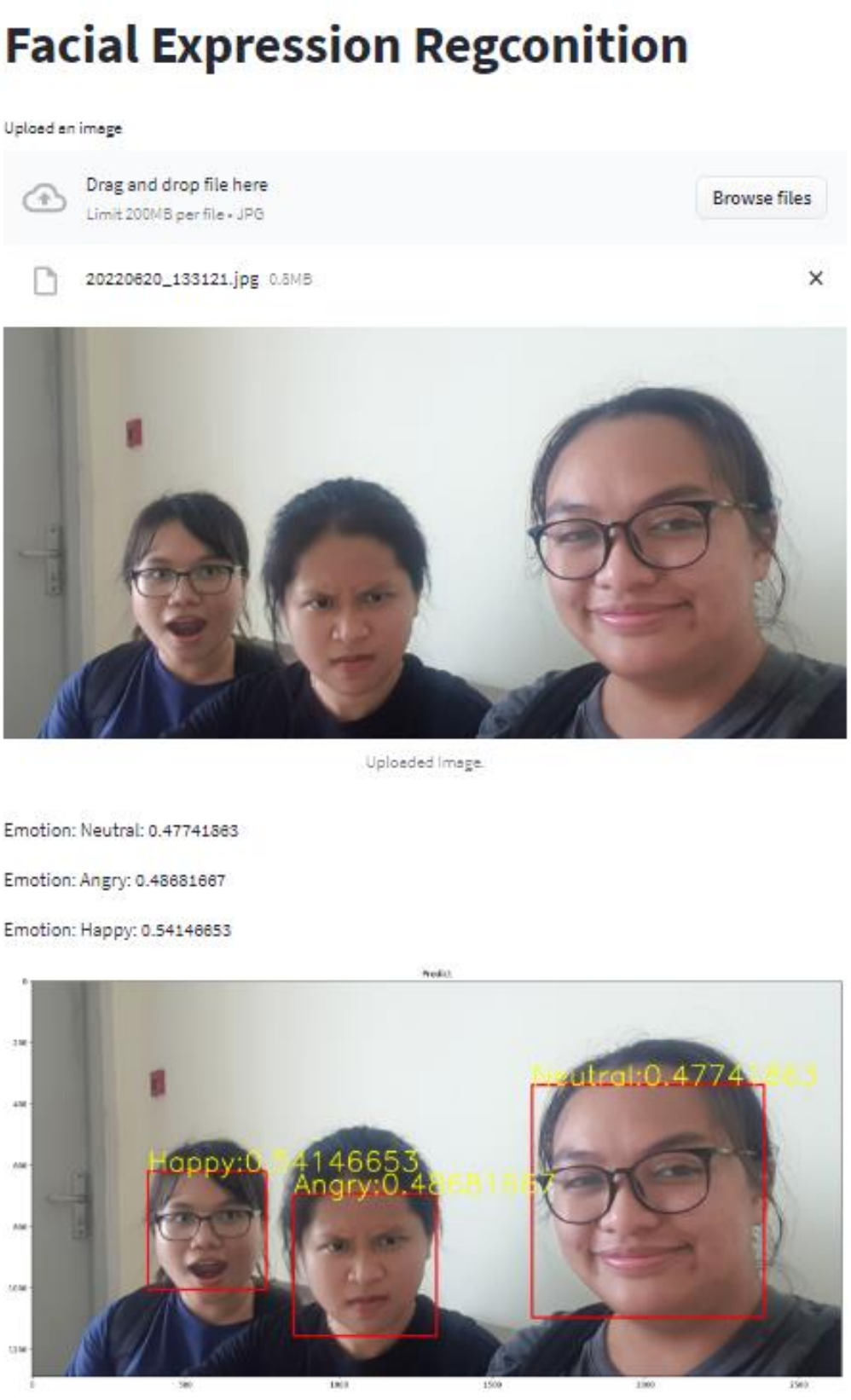
Emotion: Sad: 0.48640168

Emotion: Surprise: 0.98483145



2.7.3. Demo trang web sử dụng model ResNet50

Link trang web: [Facial Emotion Regconition ResNet50 - a Hugging Face Space by linhdan412]



Facial Expression Regconition

Upload an image

Drag and drop file here
Limit 200MB per file • JPG

Browse files

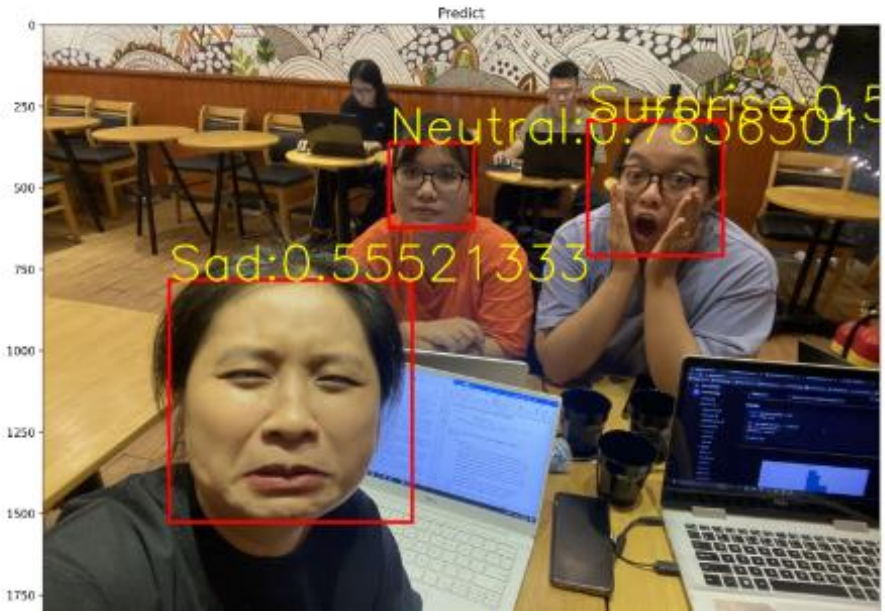
people.jpg 1.0MB

Uploaded image.

Emotion: Neutral: 0.7856301

Emotion: Sad: 0.55521333

Emotion: Surprise: 0.57603947



3. KẾT LUẬN

- Cả 3 model đều nhận diện được biểu cảm của người với với độ chính xác trên 60%.
- Kết quả thử nghiệm thực tế cho thấy mô hình khá nhạy khi nhận biết cảm xúc Happy, khá kém với cảm xúc Disgust.
- **Fear** với **Angry**, **Neutral** với **Sad** có biểu cảm khá giống nhau nên model thường nhầm lẫn dẫn đến cho ra độ chính xác không cao.
- Việc hầu hết các mô hình được công bố với tập dữ liệu ***FER-2013*** đều chỉ đạt độ chính xác thấp (dưới 70%), điều này có thể cho thấy bộ dữ liệu này có những yếu tố mất cân bằng hoặc nhiễu khi gán nhãn dữ liệu.

4.TÀI LIỆU THAM KHẢO

[1] Nghiên cứu nhận dạng biểu cảm khuôn mặt bằng phương pháp học sâu sử dụng kiến trúc ResNet - Tài liệu, ebook, giáo trình, hướng dẫn (timtailieu.vn)

[2] FER2013 | Kaggle

[3] facial-emotion-recognition/emotion_recognizer.ipynb at main · esra-polat/facial-emotion-recognition · GitHub

[4] 37-151.pdf (vap.ac.vn)

[5] 1804.08348.pdf (arxiv.org)

[6] Streamlit documentation

[7] Spaces (huggingface.co)

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Nguyễn Thị Như Ý	Train model VGG16, làm Powerpoint
2	Nguyễn Tấn Tú	Train model ResNet-50, Detect face and Predict
3	Đinh Hoàng Linh Đan	Deploy model lên web API
4	Trần Nguyễn Quỳnh Anh	Thu thập private test dataset, làm Powerpoint