

TRƯỜNG ĐẠI HỌC PHENIKAA KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO BÀI TẬP LỚN HỌC PHẦN THỊ GIÁC MÁY TÍNH

Nhóm 2

Đề tài: Nhận diện cảm xúc khuôn mặt

Giảng viên hướng dẫn: TS.Đặng Thị Thuý An

Nhóm sinh viên thực hiện

Họ và tên	MSSV
Nguyễn Việt Hoàng	21010664

Hà Nội, 11/2024

Lời cảm ơn

Được sự hướng dẫn của TS.Đặng Thị Thúy An – cán bộ trực tiếp giảng dạy học phần Thị giác máy tính, với sự nhiệt tình và tâm huyết đã trang bị những kiến thức quý giá, giải đáp thắc mắc giúp chúng em có được những kỹ năng trong việc ứng dụng kỹ thuật vào việc phát triển nghiên cứu đề tài: **Nhận diện khuôn mặt**. Chúng em mong rằng đề tài sau khi thực hiện sẽ được hoàn thiện và phát triển giúp cho việc quản lý cũng như vận hành trong lĩnh vực này sẽ trở nên số hóa, tiện lợi và dễ dàng hơn.

Do hạn chế về mặt thời gian và hiểu biết, đề tài của chúng em có thể còn nhiều thiếu sót, rất mong sẽ nhận được sự góp ý, bổ sung từ thầy và các bạn để nhóm chúng em có thể hoàn thiện vốn kiến thức của mình, tạo hành trang vững chắc cho việc phát triển trong tương lai.

Chúng em xin chân thành cảm ơn!

MỤC LỤC

LỜI CAM KẾT	5
1. GIỚI THIỆU	1
1.1 Đặt vấn đề.....	1
1.2 Giải pháp	2
2. THIẾT KẾ VÀ TRIỂN KHAI.....	3
2.1 Phân tích thuật toán CNN.....	3
2.1.1 CNN (Convolutional Neural Network).....	3
2.1.2 Convolutional	3
2.1.3 Các lớp cơ bản của mạng CNN	4
2.1.4 Cấu trúc của mạng CNN.....	8
2.2 Triển khai thuật toán.....	9
2.2.1 Yêu cầu chức năng.....	9
2.2.2 Yêu cầu phi chức năng	9
2.2.3 Các thư viện và hàm được sử dụng	10
2.2.4 Chuẩn bị dữ liệu	1
2.2.5 Xây dựng mô hình	2
2.2.5 Kết quả triển khai.....	5
2.3 Đánh giá kết quả mô hình.....	7
2.3.1 Đánh giá dựa trên ma trận nhầm lẫn (Confusion Matrix)	7
2.3.2 Đánh giá dựa trên kết quả thực nghiệm:	8
2.3.3 Hướng cải thiện.....	9
3. MỘT SỐ THÀNH PHẦN KHÁC.....	9
3.1 Ứng dụng của sản phẩm.....	10
3.2 Một số thách thức khi đưa vào các trường hợp thật.....	10
3.3 Kiến thức đã được học và chiến lược sắp tới.....	10
3.3.1 Kiến thức đã được học và ứng dụng.....	10
3.3.2 Chiến lược sắp tới.....	1
4. TÀI LIỆU THAM KHẢO.....	2

MỤC LỤC HÌNH ẢNH

Hình 1: Lớp convolutional	4
Hình 2: Lớp Pooling.....	5
Hình 3: Lớp Dropout.....	6
Hình 4: Lớp Fully Connected.....	7
Hình 5: Cấu trúc của mạng CNN	8
Hình 6: Kết quả sau khi chạy đủ 100 Epoch.....	5
Hình 7: Kết quả khi thử nghiệm với ảnh.....	6
Hình 8: Kết quả với video	6
Hình 9: Ma trận nhầm lẫn và báo cáo phân loại	7

LỜI CAM KẾT

Họ và tên nhóm sinh viên:

- Nguyễn Việt Hoàng

Điện thoại liên lạc(Nguyễn Việt Hoàng): 0911460573

Email: 21010664@st.phenikaa-uni.edu.vn

Lớp: Thị giác máy tính N01

Hệ đào tạo: Đại học chính quy

Em/Chúng em cam kết Bài tập lớn (BTL) là công trình nghiên cứu của bản thân/nhóm em. Các kết quả nêu trong BTL là trung thực, là thành quả của riêng em, không sao chép theo bất kỳ công trình nào khác. Tất cả những tham khảo trong BTL – bao gồm hình ảnh, bảng biểu, số liệu, và các câu từ trích dẫn – đều được ghi rõ ràng và đầy đủ nguồn gốc trong danh mục tài liệu tham khảo. Em/chúng em xin hoàn toàn chịu trách nhiệm với dù chỉ một sao chép vi phạm quy chế của nhà trường.

Hà nội, ngày 18 tháng 06 năm 2024

Tác giả/nhóm tác giả BTL

Họ và tên sinh viên

1. GIỚI THIỆU

1.1 Đặt vấn đề

Dự án nhận dạng cảm xúc khuôn mặt bằng mạng nơ-ron tích chập (CNN) hướng tới phát triển một hệ thống thông minh có khả năng tự động phân tích và nhận diện các trạng thái cảm xúc con người thông qua hình ảnh khuôn mặt. Nhận dạng cảm xúc không chỉ là một thách thức về kỹ thuật mà còn mang lại nhiều tiềm năng trong các lĩnh vực như:

- Giáo dục.
- Dịch vụ khách hàng.
- Bảo mật và an ninh.
- Chăm sóc sức khỏe tâm lý.

Với sự tiến bộ của học sâu và mạng nơ-ron tích chập, khả năng nhận dạng và phân loại cảm xúc từ hình ảnh đã đạt được những bước tiến vượt bậc. CNN, với khả năng học và nhận dạng các đặc trưng phức tạp từ dữ liệu hình ảnh, đã chứng minh hiệu quả vượt trội so với các phương pháp truyền thống. Dự án này không chỉ tập trung vào việc xây dựng một mô hình CNN với khả năng nhận dạng với độ chính xác và tốc độ cao. Bằng cách áp dụng những kỹ thuật và thực nghiệm trên một bộ dữ liệu nổi tiếng FER-2013 từ Kaggle, dự án kỳ vọng sẽ là bước nền tảng cho ý tưởng phát triển một hệ thống nhận dạng cảm xúc trong tương lai, mở ra triển vọng mới cho các ứng dụng trong đời sống hàng ngày.

Mục tiêu nhóm chúng em đề ra trong dự án Nhận diện cảm xúc khuôn mặt bao gồm:

- Mô hình CNN: Nhận diện chính xác 7 loại cảm xúc cơ bản của con người.
- Độ chính xác, tốc độ xử lý: Mô hình phải có độ chính xác và tốc độ cao để tích hợp vào các phần cứng như camera, máy khám bệnh,...

1.2 Giải pháp

Giải pháp mà nhóm em đưa ra là sử dụng thuật toán CNN để training một model có khả năng xử lý và phân tích ảnh, video đầu vào để nhận diện được cảm xúc khuôn mặt và sử dụng hàm HaarCascades của cv2 để detect khuôn mặt:

CNN: Convolutional Neural Network (CNN) là một loại mạng nơ-ron nhân tạo đặc biệt hiệu quả trong việc xử lý dữ liệu hình ảnh. CNN có khả năng tự động học các đặc trưng từ dữ liệu đầu vào thông qua các lớp convolution và pooling, giúp tăng độ chính xác trong việc phân loại cảm xúc khuôn mặt.

HaarCascades: HaarCascades là một phương pháp dùng để phát hiện các đối tượng trong ảnh, được áp dụng rộng rãi trong việc phát hiện khuôn mặt.

Tích hợp CNN và HaarCascades: Quá trình nhận diện cảm xúc khuôn mặt được thực hiện qua các bước sau:

- Phát hiện khuôn mặt: Sử dụng HaarCascades để phát hiện và cắt các vùng chứa khuôn mặt từ ảnh hoặc video.
- Tiền xử lý: Chuẩn hóa kích thước ảnh khuôn mặt, chuyển đổi ảnh sang dạng grayscale (nếu cần) và thực hiện các bước tiền xử lý cần thiết khác.
- Nhận diện cảm xúc: Đưa ảnh khuôn mặt đã được tiền xử lý vào mô hình CNN để phân loại và nhận diện cảm xúc.

Với giải pháp này, nhóm hy vọng có thể xây dựng một hệ thống nhận diện cảm xúc khuôn mặt tự động, hiệu quả và chính xác, có thể ứng dụng trong nhiều lĩnh vực khác nhau như giáo dục, dịch vụ khách hàng, chăm sóc sức khỏe, bảo mật và an ninh.

2. THIẾT KẾ VÀ TRIỂN KHAI

2.1 Phân tích thuật toán CNN

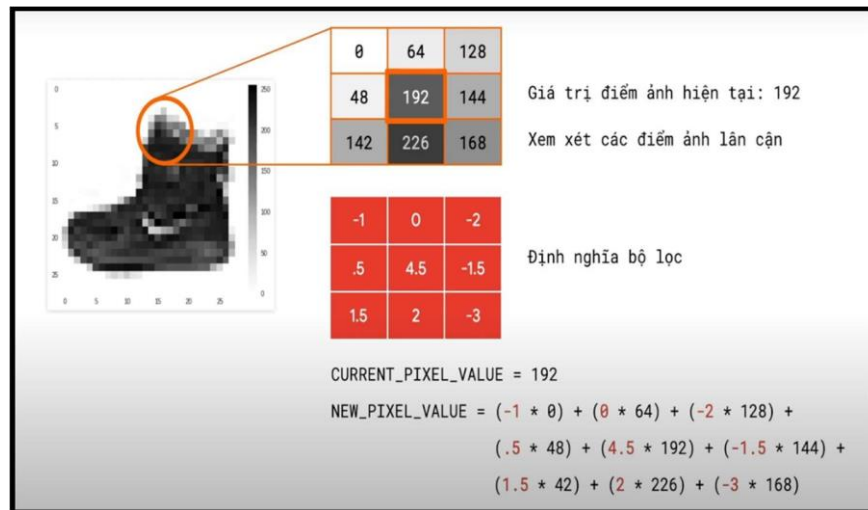
2.1.1 CNN (Convolutional Neural Network)

- Convolutional Neural Network (CNNs – Mạng nơ-ron tích chập) là một trong những mô hình Deep Learning tiên tiến. Nó giúp cho chúng ta xây dựng được những hệ thống thông minh với độ chính xác cao như hiện nay.
- CNN được sử dụng nhiều trong các bài toán nhận dạng các object trong ảnh. Để tìm hiểu tại sao thuật toán này được sử dụng rộng rãi cho việc nhận dạng (detection), chúng ta hãy cùng tìm hiểu về thuật toán này.

2.1.2 Convolutional

- Convolutional là một loại cửa sổ dạng trượt nằm trên một ma trận. Các convolutional layer sẽ chứa các parameter có khả năng tự học, qua đó sẽ điều chỉnh và tìm ra cách lấy những thông tin chính xác nhất trong khi không cần chọn feature.
- Convolution hay tích chập đóng vai trò là nhân các phần tử thuộc ma trận. Sliding Window, hay được gọi là kernel, filter hoặc feature detect, là loại ma trận có kích thước nhỏ.

2.1.3 Các lớp cơ bản của mạng CNN



Hình 1: Lớp convolutional

Lớp Convolutional (Convolutional Layer):

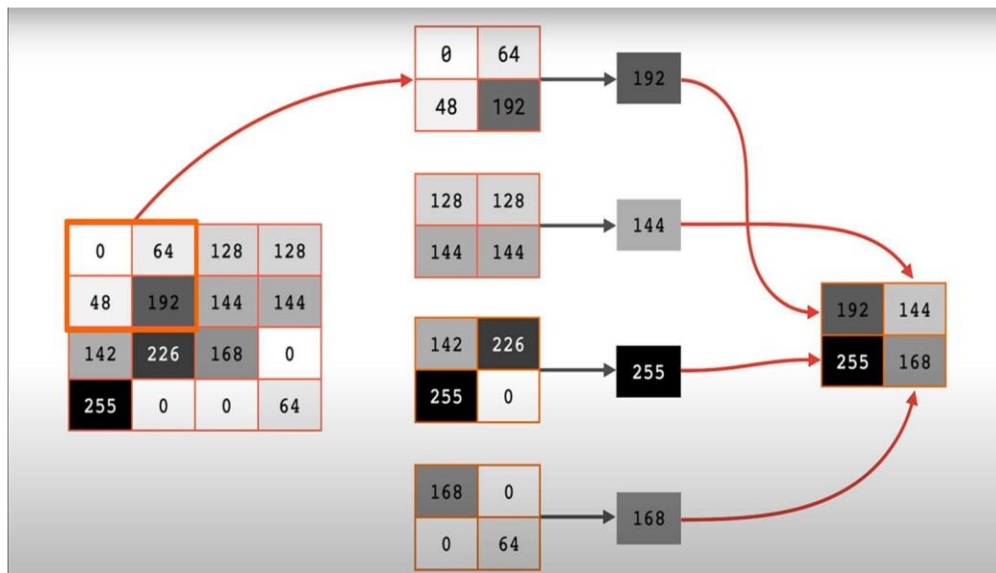
Đây là lớp chính của CNN, chịu trách nhiệm thực hiện các tính toán chính. Các yếu tố quan trọng của lớp này bao gồm stride, padding, filter map và feature map.

- Cơ chế hoạt động: CNN tạo ra các bộ lọc (filter) và áp dụng chúng lên từng vùng của hình ảnh. Các filter map này là các ma trận 3 chiều chứa các tham số dưới dạng số liệu.
- Stride: Là bước dịch chuyển của filter map trên hình ảnh theo từng pixel từ trái sang phải.
- Padding: Là các giá trị 0 được thêm vào xung quanh lớp đầu vào để giữ nguyên kích thước của đầu ra sau khi áp dụng filter.
- Feature Map: Sau mỗi lần quét của filter, một quá trình tính toán diễn ra và kết quả được thể hiện trong feature map. Feature map thể hiện các đặc trưng được trích xuất từ hình ảnh sau khi filter map quét qua.

Lớp ReLU (ReLU Layer):

- Còn được gọi là hàm kích hoạt (activation function), hàm này mô phỏng hoạt động của các neuron sinh học bằng cách kích hoạt tín hiệu truyền qua axon.
- Các hàm kích hoạt phổ biến bao gồm ReLU, Tanh, Sigmoid, Maxout và Leaky ReLU. Lớp ReLU thường được sử dụng rộng rãi trong huấn luyện nơ-ron do các ưu điểm nổi bật của nó.

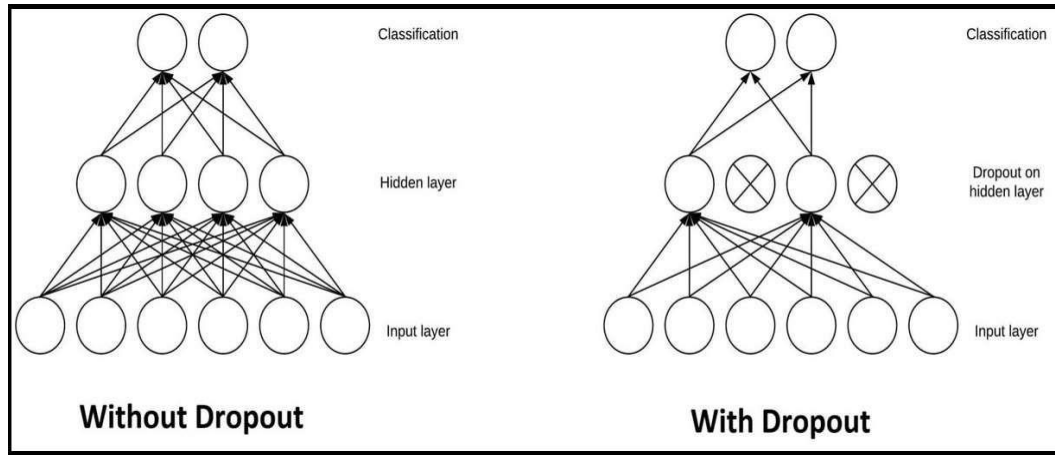
Lớp Pooling (Pooling Layer):



Hình 2: Lớp Pooling

Khi đầu vào quá lớn, các lớp pooling được chèn giữa các lớp convolutional để giảm số lượng tham số. Pooling layer có hai loại phổ biến là max pooling và average pooling.

Lớp Dropout (Dropout Layer):

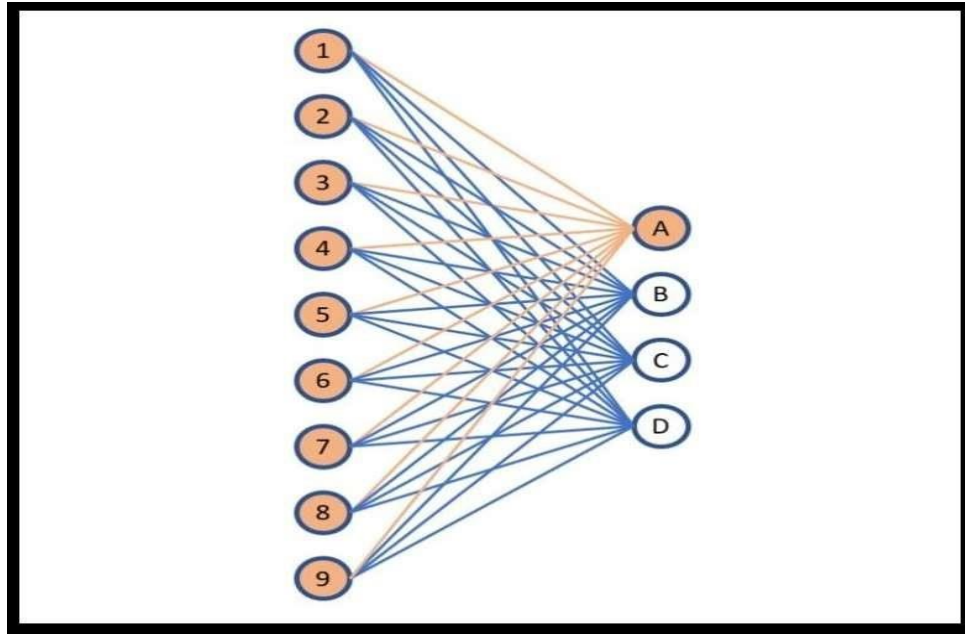


Hình 3: Lớp Dropout

Đây là lớp được sử dụng để giảm thiểu hiện tượng overfitting bằng cách ngẫu nhiên bỏ qua một số nơ-ron trong quá trình huấn luyện.

- Cơ chế hoạt động: Trong mỗi bước huấn luyện, lớp Dropout ngẫu nhiên "tắt" (bỏ qua) một phần các nơ-ron với xác suất nhất định (thường được gọi là tỷ lệ dropout, ví dụ 0.5). Các nơ-ron bị tắt sẽ không tham gia vào quá trình tính toán forward pass và backward pass trong bước huấn luyện đó.
- Lợi ích: Giảm hiện tượng overfitting và tăng khả năng tổng quát của mô hình. Mạng nơ-ron học được các đặc trưng mạnh mẽ hơn, không phụ thuộc quá nhiều vào một số nơ-ron cụ thể.
- Sử dụng: Lớp Dropout thường được sử dụng sau các lớp fully connected hoặc convolutional. Tỷ lệ dropout thường được lựa chọn trong khoảng từ 0.2 đến 0.5, tùy thuộc vào cấu trúc và kích thước của mô hình.

Lớp Fully Connected (Fully Connected Layer):

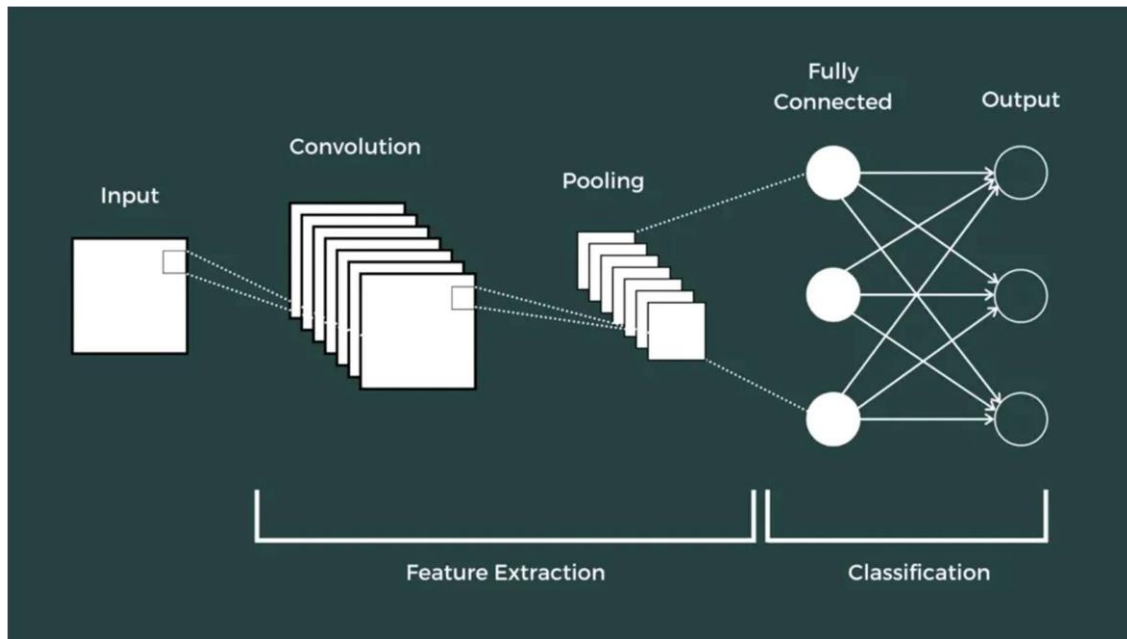


Hình 4: Lớp Fully Connected

Khi các lớp convolutional và pooling đã xử lý hình ảnh, lớp fully connected sẽ chịu trách nhiệm xuất ra kết quả cuối cùng. Lớp này liên kết tất cả các nơ-ron từ các lớp trước để tạo ra nhiều đầu ra hơn. Khi nhận được dữ liệu từ các lớp trước, lớp fully connected sẽ giúp chuyển đổi và phân loại các thông tin ảnh để đưa ra kết quả phân tích cuối cùng.

Nếu fully connected layer có dữ liệu về hình ảnh, nó sẽ chuyển đổi chúng thành các mục chưa được phân loại rõ ràng.

2.1.4 Cấu trúc của mạng CNN



Hình 5: Cấu trúc của mạng CNN

Mạng CNN bao gồm một tập hợp các lớp Convolution được chồng lên nhau, kết hợp với các hàm kích hoạt phi tuyến tính (như ReLU và Tanh) để kích hoạt trọng số trong các node.

Khi đi qua các hàm này, lớp Convolution sẽ cập nhật trọng số và tạo ra thông tin trừu tượng hơn cho các lớp kế tiếp. Đặc điểm chính của mô hình CNN:

- Tính bất biến: Độ chính xác của mô hình có thể bị ảnh hưởng nếu đối tượng được nhìn từ nhiều góc độ khác nhau. Pooling layer giúp tạo nên tính bất biến đối với phép dịch chuyển, co giãn và quay.
- Tính kết hợp: Các cấp độ biểu diễn thông tin từ mức độ thấp đến cao được thể hiện qua convolution từ các filter, giúp kết nối các layer với nhau. Mỗi nơ-ron ở lớp tiếp theo sẽ dựa trên kết quả filter áp đặt lên vùng ảnh cục bộ của nơ-ron trước đó.

2.2 Triển khai thuật toán

2.2.1 Yêu cầu chức năng

- Phát hiện khuôn mặt:
 - Hệ thống phải có khả năng phát hiện và theo dõi khuôn mặt trong khung hình video hoặc từ nguồn webcam thời gian thực.
- Dự đoán cảm xúc:
 - Hệ thống phải có khả năng phân loại cảm xúc từ các khuôn mặt đã phát hiện được.
 - Mô hình CNN phải có khả năng phân loại chính xác các cảm xúc bao gồm: "Angry", "Disgusted", "Fearful", "Happy", "Neutral", "Sad", và "Surprised".
- Hiển thị cảm xúc:
 - Hệ thống phải hiển thị nhãn cảm xúc trên khuôn mặt tương ứng trong khung hình video.
 - Nhãn cảm xúc phải rõ ràng và dễ đọc.
- Xử lý video từ nhiều nguồn:
 - Hệ thống phải có khả năng xử lý video từ các tệp video hoặc từ webcam thời gian thực.
- Lưu trữ và tải mô hình:
 - Hệ thống phải có khả năng lưu trữ cấu trúc và trọng số của mô hình đã huấn luyện.
 - Phải có khả năng tải mô hình từ các tệp lưu trữ để sử dụng.

2.2.2 Yêu cầu phi chức năng

- Hiệu suất:
 - Hệ thống phải phản hồi và xử lý dữ liệu video trong thời gian thực với độ trễ tối thiểu.

- Dự đoán cảm xúc phải được thực hiện nhanh chóng để đảm bảo tính thời gian thực.
- Độ chính xác:
 - Hệ thống phải đạt độ chính xác cao trong việc phát hiện khuôn mặt và phân loại cảm xúc.
 - Mô hình CNN phải được huấn luyện và tối ưu hóa để đạt kết quả chính xác nhất có thể.
- Khả năng mở rộng:
 - Hệ thống phải dễ dàng mở rộng để thêm các cảm xúc mới hoặc cải thiện mô hình dự đoán.
 - Cấu trúc hệ thống phải hỗ trợ việc cập nhật và nâng cấp các thành phần một cách dễ dàng.
- Tương thích:
 - Hệ thống phải tương thích với nhiều nền tảng và thiết bị khác nhau, bao gồm Windows, macOS, và Linux.
 - Phải hỗ trợ các phiên bản OpenCV và Keras phổ biến để đảm bảo tính linh hoạt.

2.2.3 Các thư viện và hàm được sử dụng

CV2 (OpenCV):

- Chức năng: OpenCV là một thư viện mã nguồn mở chuyên về xử lý ảnh và video. Nó cung cấp các công cụ mạnh mẽ để phát hiện và nhận diện khuôn mặt, cùng với nhiều chức năng khác liên quan đến xử lý ảnh.
- Ví dụ sử dụng: `cv2.CascadeClassifier` được sử dụng để phát hiện khuôn mặt trong ảnh hoặc video.

Numpy:

- Chức năng: NumPy là một thư viện quan trọng cho tính toán số học và xử lý mảng trong Python. Nó cung cấp các công cụ để thao tác và xử lý dữ liệu dưới dạng mảng.
- Ví dụ sử dụng: `cropped_img = np.expand_dims(cropped_img, axis=0)` thêm một chiều vào mảng dữ liệu để phù hợp với định dạng đầu vào của mô hình.


Matplotlib.pyplot:

- Chức năng: Matplotlib là một thư viện vẽ đồ thị mạnh mẽ cho Python. Pyplot là một module trong Matplotlib cung cấp các hàm để vẽ đồ thị một cách dễ dàng.
- Ví dụ sử dụng: `plt.plot(history.history['accuracy'])` vẽ đồ thị hiển thị độ chính xác qua các epoch huấn luyện.

2.2.4 Chuẩn bị dữ liệu

FER-2013

Learn facial expressions from an image



Fear Happy Neutral

[Data Card](#) [Code \(378\)](#) [Discussion \(6\)](#) [Suggestions \(2\)](#)

About Dataset

The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centred and occupies about the same amount of space in each image.

The task is to categorize each face based on the emotion shown in the facial expression into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The training set consists of 28,709 examples and the public test set consists of 3,589 examples.

Usability ⓘ
7.50

License
Database: Open Database, Cont...

Expected update frequency
Not specified

Tags

Arts and Entertainment

Art

Tập dữ liệu được lấy từ dataset FER-2013 của kaggle. Tập dữ liệu có 32298 ảnh xám kích thước 48 x 48 pixel và chia thành hai thư mục chính:

- Thư mục training có 28,709 ảnh, chiếm 88% tập dữ liệu
- Thư mục test có 3,589 ảnh, chiếm 12% tập dữ liệu
- Mỗi thư mục có 7 thư mục Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral chứa các ảnh theo đúng loại cảm xúc.

Tập dữ liệu đã phù hợp để đưa vào training nên không cần phải xử lý các bước ban đầu

2.2.5 Xây dựng mô hình

Cấu trúc mạng CNN:

- **Lớp tích chập đầu tiên với 32 bộ lọc kích thước 3x3:** lớp này sẽ trích xuất các đặc trưng cơ bản từ ảnh.
- **Lớp tích chập thứ hai với 64 bộ lọc 3x3:** Kết quả từ lớp tích chập đầu tiên sẽ được truyền đến lớp này để trích xuất thêm các đặc trưng phức tạp hơn.
- **Lớp MaxPooling:** Kết quả từ lớp tích chập thứ hai sẽ được giảm kích, lấy giá trị lớn nhất từ mỗi vùng 2x2 để giảm số lượng tham số và tránh overfitting.
- **Lớp Dropout:** một phần ngẫu nhiên (25%) các nơ-ron sẽ bị bỏ qua trong quá trình huấn luyện để tăng cường khả năng tổng quát hóa của mô hình.
- **Lớp tích chập thứ ba với 128 bộ lọc 3x3:** để trích xuất các đặc trưng sâu hơn.
- **Lớp MaxPooling:** Kết quả từ lớp tích chập thứ ba sẽ được giảm kích thước bằng lớp MaxPooling, giúp giảm số lượng tham số.
- **Lớp tích chập thứ tư với 128 bộ lọc 3x3:** dữ liệu sẽ tiếp tục được trích xuất các đặc trưng sâu hơn.

- **Lớp MaxPooling:** kết quả từ lớp tích chập thứ tư sẽ được giảm kích thước lần cuối bằng lớp MaxPooling.
- **Lớp Dropout thứ hai:** khi qua lớp Dropout thứ hai một phần ngẫu nhiên (25%) các nơ-ron sẽ bị bỏ qua để ngăn chặn overfitting.
- **Lớp Flatten:** kết quả từ các lớp Convolutional và MaxPooling sẽ được chuyển đổi từ ma trận hai chiều thành một vector một chiều.
- **Lớp fully Connected đầu tiên (Dense):** Vector đầu vào sẽ được truyền qua lớp Dense với 1024 nơ-ron và hàm kích hoạt ReLU để tạo ra các đặc trưng kết hợp.
- **Lớp Dropout thứ ba:** Một phần ngẫu nhiên (50%) các nơ-ron sẽ bị bỏ qua để ngăn chặn overfitting.
- **Lớp fully Connected thứ hai (Dense):** Lớp Dense cuối cùng với 7 nơ-ron và hàm kích hoạt softmax sẽ phân loại dữ liệu đầu vào thành một trong bảy loại cảm xúc: "Angry", "Disgusted", "Fearful", "Happy", "Neutral", "Sad", "Surprised".
- **Adam optimizer:** Trình tối ưu hóa Adam sẽ điều chỉnh các trọng số trong mô hình dựa trên hàm mất mát “categorical_crossentropy” và tốc độ học (learning rate) được điều chỉnh bởi “ExponentialDecay”.

Quá trình huấn luyện:

- Mô hình lấy dữ liệu từ bộ dữ liệu để huấn luyện.
- Số bước huấn luyện trong mỗi epoch được xác định là 28709 // 32, trong đó 28709 là số lượng mẫu huấn luyện và 32 là kích thước batch.
- Mô hình được huấn luyện trong 100 epochs, Batch size: 32- giúp cân bằng giữa tốc độ huấn luyện và hiệu quả của mô hình, Learning rate: Tốc độ học được giảm dần trong quá trình huấn luyện, giúp mô hình hội tụ nhanh hơn.

Lưu trữ thông tin huấn luyện:

- Kết quả của quá trình huấn luyện, bao gồm độ chính xác và hàm mất mát trên tập huấn luyện và tập validation sau mỗi epoch, được lưu trữ để đánh giá hiệu suất của mô hình.

- Cấu trúc của mô hình được chuyển thành định dạng JSON và lưu vào tệp "emotion_model.json" để có thể tái sử dụng cấu trúc mô hình mà không cần phải huấn luyện lại.
- Trọng số đã được huấn luyện của mô hình được lưu vào tệp "emotion_model.weights.h5" để có thể tái sử dụng mô hình đã được huấn luyện.

Kết hợp với Haar Cascade để thực nghiệm với video hoặc webcam

- Khởi tạo và Tải mô hình:
 - Tải cấu trúc và trọng số mô hình: Đọc cấu trúc từ tệp "emotion_model.json" và nạp trọng số từ "emotion_model.h5".
- Khởi động nguồn video:
 - Mở webcam hoặc video: Sử dụng cv2.VideoCapture(0) cho webcam hoặc chỉ định đường dẫn tới tệp video..
- Sử dụng Haar Cascade để phát hiện khuôn mặt:
 - Tải Haar Cascade: Sử dụng bộ phân loại Haar Cascade của OpenCV để phát hiện khuôn mặt.
 - Chuyển đổi sang ảnh xám: Video khung hình được chuyển đổi sang ảnh xám để quá trình phát hiện khuôn mặt trở nên hiệu quả hơn.
- Xử lý khuôn mặt:
 - Cắt vùng ảnh khuôn mặt: Xác định và cắt vùng chứa khuôn mặt từ ảnh gốc.
 - Chuẩn bị dữ liệu đầu vào: Thay đổi kích thước ảnh khuôn mặt thành 48x48 pixel và chuẩn bị định dạng phù hợp cho mô hình CNN.
- Dự đoán cảm xúc:
 - Dự đoán cảm xúc: Đưa ảnh khuôn mặt đã xử lý vào mô hình để dự đoán cảm xúc.
 - Xác định cảm xúc: Chọn cảm xúc có xác suất cao nhất từ kết quả dự đoán.

- Hiện thị kết quả:
 - Dán nhãn cảm xúc lên khung hình: Hiện thị nhãn cảm xúc lên vị trí khuôn mặt trong khung hình video.
 - Hiện thị khung hình: Hiện thị khung hình với các nhãn cảm xúc.
- Kết thúc
 - Giải phóng tài nguyên: Đóng các tài nguyên như video và cửa sổ hiện thị khi hoàn tất.

2.2.5 Kết quả triển khai

Kết quả sau khi chạy đủ 100 Epoch:

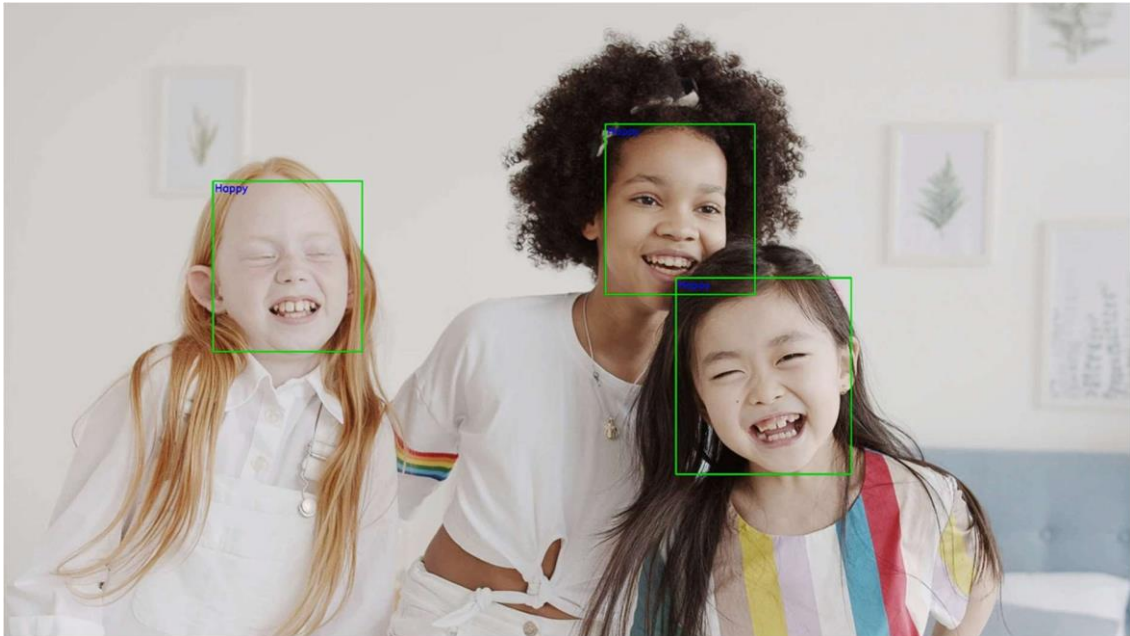
897/897	65s	73ms/step	- accuracy: 0.9105	- loss: 0.2584	- val_accuracy: 0.6244	- val_loss: 1.3389
Epoch 98/100						
897/897	0s	58us/step	- accuracy: 0.8438	- loss: 0.3383	- val_accuracy: 0.8000	- val_loss: 1.4085
Epoch 99/100						
897/897	64s	71ms/step	- accuracy: 0.9098	- loss: 0.2555	- val_accuracy: 0.6293	- val_loss: 1.3153
Epoch 100/100						
897/897	0s	51us/step	- accuracy: 0.9375	- loss: 0.1878	- val_accuracy: 0.6000	- val_loss: 1.2582

Hình 6: Kết quả sau khi chạy đủ 100 Epoch

Nhận xét khi chạy:

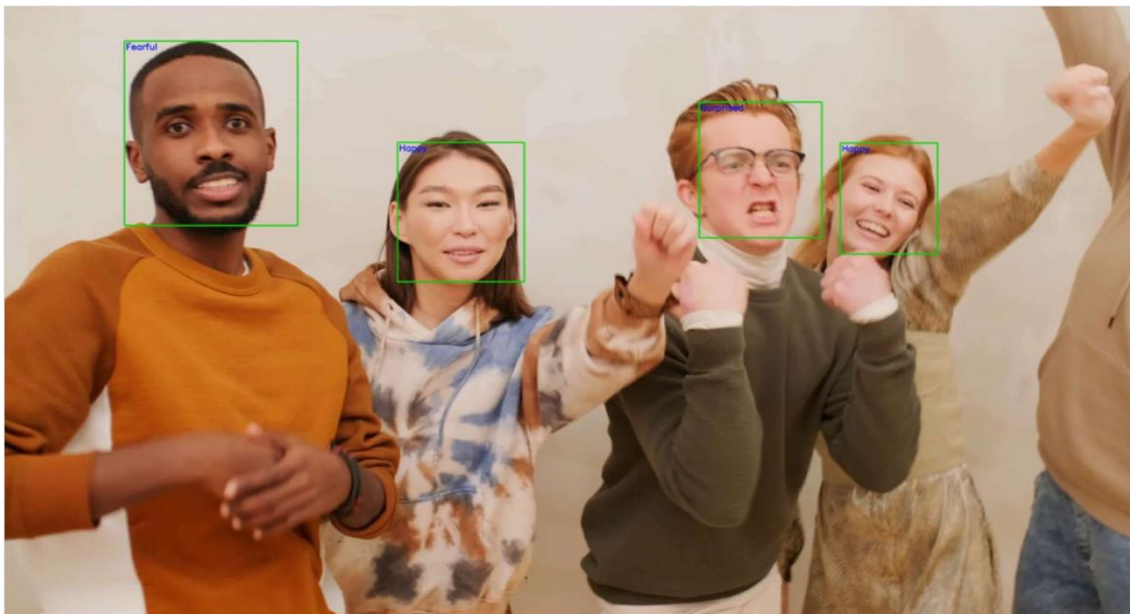
- Khi epoch đạt khoảng từ 90 trở lên thì bắt đầu xuất hiện overfitting có thể dừng tại đây vì chạy thêm đến 100 epoch thì kết quả cải thiện không đáng kể
- Độ chính xác khá ổn với accuracy: 0.9 trở lên và loss: 0.18

Kết quả khi thử nghiệm với ảnh:



Hình 7: Kết quả khi thử nghiệm với ảnh

Kết quả với video:



Hình 8: Kết quả với video

2.3 Đánh giá kết quả mô hình

2.3.1 Đánh giá dựa trên ma trận nhầm lẫn (Confusion Matrix)

Confusion Matrix:							
750	20	50	50	40	30	18	
15	80	5	2	3	4	2	
30	5	820	30	60	50	29	
50	10	30	1450	100	100	34	
40	5	50	70	1000	50	18	
30	2	40	70	50	1000	55	
20	4	20	40	20	20	707	
Classification Report:							
		precision		recall	f1-score	support	
	0	0.75		0.78	0.76	958	
	1	0.65		0.72	0.68	111	
	2	0.80		0.80	0.80	1024	
	3	0.81		0.82	0.81	1774	
	4	0.78		0.81	0.80	1233	
	5	0.79		0.80	0.80	1247	
	6	0.79		0.85	0.82	831	
	accuracy				0.80	7178	
	macro avg	0.77		0.80	0.78	7178	
	weighted avg	0.80		0.80	0.80	7178	

Hình 9: Ma trận nhầm lẫn và báo cáo phân loại

- Precision (Độ chính xác): Tỷ lệ các dự đoán dương tính đúng so với tổng số dự đoán dương tính.
- Recall (Độ nhạy): Tỷ lệ các trường hợp dương tính thực sự được mô hình nhận diện đúng.
- F1-Score: Trung bình điều hòa của độ chính xác và độ nhạy.
- $F1\text{-Score} = 2 * (Precision * Recall) / (Precision + Recall)$.
- Support (Số lượng mẫu): Số lượng các trường hợp thực sự của mỗi lớp trong tập dữ liệu.
- Accuracy (Độ chính xác tổng thể): Tỷ lệ tổng thể của các dự đoán đúng (bao gồm cả dự đoán đúng dương tính và âm tính).
- Macro Avg: Trung bình không trọng số của độ chính xác, độ nhạy và F1-score cho tất cả các lớp.
- Weighted Avg: Trung bình có trọng số của độ chính xác, độ nhạy và F1-score, có tính đến số lượng mẫu của mỗi lớp (support).

Đánh giá:

- Tổng thể mô hình hoạt động ở mức độ khá: Với độ chính xác khoảng 80%, mô hình có thể được xem là hoạt động tốt cho việc nhận diện cảm xúc từ khuôn mặt.
- Cảm xúc khó nhận diện: Emotion 1 (Disgusted) là cảm xúc mà mô hình gặp nhiều khó khăn nhất, với độ chính xác và F1-score thấp hơn so với các cảm xúc khác. Điều này có thể do số lượng mẫu của cảm xúc này ít hơn, hoặc đặc điểm nhận diện của cảm xúc này không rõ ràng.
- Cảm xúc dễ nhận diện: Emotion 6 (Surprised) là cảm xúc mà mô hình nhận diện tốt nhất, với độ chính xác và F1-score cao nhất.
- Nhầm lẫn giữa các cảm xúc: Có sự nhầm lẫn đáng kể giữa các cảm xúc như Angry, Fearful, Happy và Sad, điều này có thể do các cảm xúc này có các đặc điểm hình ảnh tương tự nhau trong dữ liệu huấn luyện.

2.3.2 Đánh giá dựa trên kết quả thực nghiệm:

Giống như kết quả của ma trận nhầm lẫn, tỉ lệ nhận diện đúng khoảng từ 80 - 90% trong điều kiện hình ảnh rõ nét rõ ràng.

Những lần bị nhầm lẫn sai thì có những lý do như:

- Chất lượng ảnh kém khiến nhận dạng bị sai.
- Góc mặt không nhìn từ chính diện khiến cho model không nhận diện được
- Những cảm xúc khó đoán như quá khích, kích động,... nói chung là khi một cảm xúc trở nên dâng cao lên thì đôi lúc sẽ có những đặc điểm bị nhầm lẫn sang các cảm xúc khác (có thể nhầm lẫn giữa Angry, Fearful, Happy, Surprise).
- Cảm xúc nhận diện thuộc trong nhóm nhầm lẫn cao (Angry, Fearful, Happy và Sad) cũng có thể là một lý do.

Đánh giá trên các yêu cầu chức năng và phi chức năng:

- Đáp ứng được đầy đủ các yêu cầu chức năng về hiệu năng nhanh, chính xác chạy được với hai loại dữ liệu hình ảnh và video,...
- Chưa đáp ứng được đầy đủ các yêu cầu phi chức năng do quy mô vẫn còn nhỏ cần mở rộng quy mô và tiếp tục cải tiến để đáp ứng được đầy đủ yêu cầu phi chức năng như dễ sử dụng, tính tương thích với các thiết bị,...

2.3.3 Hướng cải thiện

- Tăng số lượng mẫu cho cảm xúc ít xuất hiện: Đặc biệt là Emotion 1 (Disgusted), để mô hình có thể học tốt hơn.
- Lọc lại những ảnh đầu vào sao cho hạn chế sự giống nhau giữa các loại cảm xúc để bị nhầm lẫn
- Augmentation cho dữ liệu huấn luyện: Sử dụng các kỹ thuật tăng cường dữ liệu để làm phong phú thêm dữ liệu huấn luyện.
- Fine-tuning mô hình: Tối ưu hóa các tham số và cấu trúc của mô hình để cải thiện độ chính xác.
- Kiểm tra và điều chỉnh dữ liệu huấn luyện: Đảm bảo rằng dữ liệu huấn luyện có chất lượng cao và đại diện tốt cho các cảm xúc cần nhận diện.

3. MỘT SỐ THÀNH PHẦN KHÁC

3.1 Ứng dụng của sản phẩm

Giúp giải quyết các bài toán về phân tích cảm xúc con người trong các lĩnh vực y tế, tiếp thị, học tập,...

3.2 Một số thách thức khi đưa vào các trường hợp thật

Cảm xúc con người đôi lúc rất khó đoán nhất là trong trường hợp của các bệnh nhân tâm lý nên chỉ có thể dùng tham khảo cho các bác sĩ tâm lý nếu được ứng dụng trong việc chữa trị tâm lý cho bệnh nhân.

Một số cảm xúc đôi khi rất giống nhau tức giận, hồi hộp, bất ngờ,... và có những tình trạng tâm lý khá phức tạp như quá khích, vui đùa,... khiến cho độ chính xác thuật toán đôi khi chỉ mang tính chất tham khảo và dự đoán chứ không thể hoàn toàn đúng.

Hạ tầng camera phải đạt một chất lượng nhất định để có thể đưa ra được hình ảnh chuẩn xác nhất có thể từ đó mô hình mới đạt được độ chính xác cao.

3.3 Kiến thức đã được học và chiến lược sắp tới

3.3.1 Kiến thức đã được học và ứng dụng

- Kiến thức tổng quan và thực nghiệm thành công thuật toán cnn trong việc nhận diện cảm xúc khuôn mặt.
- Kết hợp các kiến thức với nhau: phát hiện khuôn mặt, tiền xử lý ảnh, giảm kích thước ảnh,...
- Cách phối hợp nhóm để thực hiện một dự án.

3.3.2 Chiến lược sắp tới

- Mở rộng mô hình: nhúng vào một số phần mềm để thực nghiệm độ ứng dụng (theo dõi bệnh tình bệnh nhân qua camera, ...)
- Gia tăng độ chính xác với tập dữ liệu lớn hơn.
- Tìm hiểu thêm về các phương pháp về nhận diện cảm xúc, thử nghiệm thêm một số mô hình khác (yolo, ...).

4. TÀI LIỆU THAM KHẢO

1. <https://arxiv.org/abs/1511.08458>
2. <https://paperswithcode.com/task/emotion-recognition>
3. <https://vietnix.vn/cnn-la-gi/>