HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

VÕ THỊ HỒNG NHUNG

PHÂN TÍCH BIỂU CẢM MẶT NGƯỜI DÙNG MẠNG NƠ RON TÍCH CHẬP

Chuyên ngành: Hệ Thống Thông Tin

Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

(Theo định hướng ứng dụng)

TP. HỒ CHÍ MINH - NĂM 2022

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **PGS.TS. Lê Hoàng Thái**

Phản biện 1: ……

Phản biện 2: ……

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng .... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

I. MỞ ĐẦU

1. Lý do chọn đề tài

Phân loại biểu cảm là lĩnh vực đã được nghiên cứu trong nhiều năm qua với nhiều ứng dụng trong nhiều lĩnh vực khác nhau gắn liền với các hệ thống tương tác người máy. Trong máy học, phân loại biểu cảm là một bài toán khó, tuy nhiên, đối với con người, vấn đề này có thể giải quyết ngay lập tức. Các thách thức chính là: hình ảnh biểu cảm của cùng một người ở cùng một biểu cảm vẫn có thể khác nhau ở những điều kiện ánh sáng, môi trường và góc quay. Những biến đổi này càng lớn khi các đối tượng nghiên cứu càng đa dạng.

- Nhận biết cảm xúc từ nét mặt có một số lợi thế như:
  - Tiếp cận theo hướng tự nhiên nhất để xác định trạng thái cảm xúc của khuôn mặt
  - Nhiều bộ dữ liệu có sẵn cho biểu hiện cảm xúc trên khuôn mặt.
  - Nhiều công cụ hỗ trợ xác định cảm xúc khuôn mặt có sẵn.
- Nhận biết cảm xúc từ nét mặt cũng có một số nhược điểm như:
  - Không thể cung cấp thông tin ngữ cảnh, do đó đôi khi kết quả bị sai lệch.
  - Kết quả phát hiện cảm xúc phụ thuộc vào chất lượng hình ảnh hoặc video
  - Chuyển động liên quan đến cảm xúc khuôn mặt có thể được đối tượng cố tình làm giả như các diễn viên …

Vì thế, nhận biết biểu cảm vẫn là một thách thức với thị giác máy tính. Trong luận văn này, đưa ra một hướng tiếp cận đơn giản cho

nhận biết biểu cảm khuôn mặt: kết hợp giữa Convolutional Neural Network (CNN) và các bước tiền xử lý đặc trưng. CNN sẽ đạt độ chính xác rất cao nếu học với bộ dữ liệu lớn. Tận dụng ưu điểm này, dự kiến đề xuất phương pháp áp dụng vài kỹ thuật tiền xử lý để chỉ rút trích các thành phần đặc trưng cho biểu cảm trên khuôn mặt và kết hợp với CNNs để thực hiện phân loại cảm xúc hiệu quả. Dự kiến sẽ thực nghiệm đánh giá trên 2 tập dữ liệu công khai lớn (CK+, JAFFE). Các thực nghiệm sẽ được thực hiện để đánh giá các ảnh hưởng của tiền xử lý và các một số ảnh hưởng của các yếu tố khác. Hy vọng xây dựng được hệ thống phân biệt cảm xúc có độ chính xác cao và đáp ứng các yêu cầu về thời gian thực.

2. Tổng quan về vấn đề nghiên cứu

2.1 Phân chia cảm xúc khuôn mặt

- Bảng dưới đây cho biết biểu cảm trên khuôn mặt thể hiện bảy cảm xúc chính của con người [1]:

Bảng 1. 1: Mô tả các cảm xúc cơ bản của con người

Cảm xúc	Biểu cảm khuôn mặt
Vui vẻ	Khóe môi hé mở,Má nâng cao
Buồn bã	Đôi mí mắt trên sụp xuống, mắt mất tập trung, mép kéo nhẹ xuống
Tức giận	Mắt Nhìn chằm chằm, Mũi nỡ ra, Môi ép chặt

Sợ hãi	Lông mày nhướng lên, Miệng mở ra
Ghê tởm	Đôi môi được nâng cao lên, Mũi nhăn
Ngạc nhiên	Lông mày cong cao hơn Tròng trắng của mắt rõ hơn, miệng há
Bình thường	Không biểu hiện gì

2.2 Tình hình nghiên cứu

3. Mục đích nghiên cứu

Nghiên cứu đề tài này nhằm mục đích tìm hiểu bài toán nhận biết cảm xúc từ nét mặt, từ đó xây dựng các hệ thống ứng dụng trong thực tiễn như: đánh giá cảm xúc nhân viên trong thời gian làm việc tại công ty, từ đó xác định hiệu quả công việc…

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: tập trung tìm hiểu một số phương pháp CNN phổ biến hiện nay, xác định một trong bảy trạng thái cảm xúc cơ bản của con người dựa vào hình ảnh đơn nhập vào.

Phạm vi nghiên cứu: thực hiện trên tập dữ liệu chuẩn CK+ [10] và JAFFE, trên hai giới tính nam lẫn giới tính nữ, độ tuổi từ 18 - 45 tuổi, với nhiều chủng tộc người khác nhau. Đồng thời, cũng thử nghiệm trên một số ảnh chụp webcam để minh hoạ tính khả thi của hệ thống về mặt ứng dụng.

Đề xuất cách tiếp cận học sâu kết hợp với các kỹ thuật tiền xử lý như: chuẩn hóa hình ảnh và tăng cường mẫu học bằng các phép rotation, translation và scaling trên ảnh thật (synthetic training-samples generation), với hy vọng nâng cao độ chính xác trên các bộ dữ liệu thử nghiệm đã chọn. Tiến tới, xây dựng một hệ thống phân loại cảm xúc thoả các tiêu chí bên dưới:

- Hiệu suất cao và đáp ứng yêu cầu thời gian thực.
- Giảm tác động của môi trường và giải quyết vấn đề dữ liệu học quá ít (cải tiến khâu tiền xử lý).
- Phân tích đánh giá các Kết quả thử nghiệm để chỉ ra hiệu quả của đề xuất.

5. Phương pháp nghiên cứu

- Phương pháp chuyên gia:
- Phương pháp thực nghiệm:
- Phương pháp tổng kết kinh nghiệm:

6. Dự kiến nội dung của luận văn

Chương 1: Giới thiệu chung

Chương 2: Hệ thống nhận dạng biểu cảm khuôn mặt

Chương 3: Thử nghiệm và thảo luận

Chương 4: Ứng dụng

Chương 5: Kết luận và hướng phát triển

II. NỘI DUNG

CHƯƠNG 1: GIỚI THIỆU CHUNG

1.1 Mạng nơ ron nhân tạo

1.1.1 Giới thiệu mạng nơ ron nhân tạo

Mạng nơ ron nhân tạo (Artificial Neural Network ANN) là một chuỗi các giải thuật lập trình, mô phỏng dựa trên cách hoạt động của mạng lưới thần kinh trong não bộ các sinh vật sống. Mạng nơ ron nhân tạo được sử dụng để tìm ra mối quan hệ của một tập dữ liệu thông qua một thiết kế kiến trúc chứa nhiều tầng ẩn (hidden layer), mỗi tầng lại chứa nhiều nơ ron. Các nơ ron được kết nối với nhau và độ mạnh yếu của các liên kết được biểu hiện qua trọng số liên kết. [13]

1.1.2 Kiến trúc mạng nơ ron nhân tạo

Một mạng Neural nhân tạo có cấu trúc như sau: Tầng lớp đầu vào (Input Layer), Tầng lớp ẩn (Hidden Layer),Tầng đầu ra (Output layer)

1.2 Mạng nơ ron tích chập (Convolutional Neural Networks)

1.2.1. Khái niệm về mạng nơ ron tích chập

1.2.2. Mô hình mạng nơ ron tích chập

Một kiến trúc CNN bao gồm các lớp: convolution layer, pooling layer và fully connected layer. Ở giữa các lớp convolution và pooling thường có các hàm kích hoạt phi tuyến. Ảnh khi đưa vào mạng sẽ được

lan truyền qua tầng convolution layer, giá trị tính được từ các tầng convolution sẽ đi qua một hàm kích hoạt, sau đó giá trị này sẽ được lan truyền qua pooling layer. Cuối cùng ảnh sẽ được lan truyền đến tầng fully connected layer và đi qua hàm kích hoạt Softmax, thường thì cuối cùng sẽ thu được một vector chứa xác suất phần trăm thuộc về các lớp đối với các bài toán phân loại.

1.4 Kết luận chương 1

Mạng nơ ron nhân tạo là một chuỗi các thuật toán được sử dụng để tìm ra mối quan hệ của một tập dữ liệu thông qua cơ chế vận hành của bộ não sinh học. Mạng nơ ron nhân tạo thường được huấn luyện qua một tập dữ liệu chuẩn cho trước, từ đó có thể đúc rút được kiến thức từ tập dữ liệu huấn luyện, và áp dụng với các tập dữ liệu khác với độ chính xác cao.

Các phương pháp sử dụng để huấn luyện mạng nơ ron nhân tạo ngày càng tối ưu hơn về mặt tính toán và phục vụ cho nhiều mục đích khác nhau. Hiện nay, kiến trúc mạng nơ ron ngày càng được hoàn thiện cho nhiều nhiệm vụ, trong đó mạng nơ ron tích chập được chú ý rất nhiều vì tính hiệu quả trong thị giác máy tính. Mạng nơ ron tích chập với các cải tiến góp phần giảm thời gian tính toán và tăng độ chính xác hứa hẹn sẽ là một trong những phương pháp được áp dụng rất nhiều vào thực tế trong tương lai.

CHƯƠNG 2: HỆ THỐNG NHẬN DẠNG BIỂU CẢM KHUÔN MẶT

Ở chương này, học viên sẽ trình bày chi tiết hệ thống nhận dạng biểu cảm khuôn mặt dựa trên kiến trúc mạng tích chập Convolution Neural Network cũng như quy trình thí nghiệm trên hai bộ dữ liệu chuẩn cho bài toán này là CK+ và JAFFE. Quy trình thí nghiệm bao gồm hai bước là: bước huấn luyện và bước thử nghiệm.

Đầu vào: Bức ảnh của một người trong bộ dữ liệu CK+ hoặc JAFFE.

Đầu ra: Nhãn cảm xúc của người dùng theo các lớp như 0. angry, 1. disgust, 2. fear, 3. happy, 4 . neutral, 5. sad và 6. Surprise.

Pha huấn luyện bao gồm các kỹ thuật: dùng phương pháp tạo ảnh tổng hợp (Synthetic images generation) để tăng số lượng mẫu học; sử dụng một chuỗi các kỹ thuật tiền xử lý ảnh (pre-processing) như: tính toán lại góc nghiêng của ảnh trong các điều kiện chụp không lý tưởng, chuẩn hóa giá trị cường độ ảnh, loại bỏ nền xung quanh gương mặt, điều chỉnh lại kích thước ảnh gương mặt (down sampling). Sau đó sử dụng một kiến trúc mạng học sâu tích chập (Convolution neural network) để huấn luyện

2.1 Tiền xử lý ảnh mặt người và tăng cường mẫu học

2.1.1 Tổng hợp tạo mẫu

Các thuật toán Học sâu hiện nay đôi khi cần phải được huấn luyện trên một bộ dữ liệu đủ nhiều, khi thiếu dữ liệu chúng ta sẽ phải gặp các vấn đề như: thuật toán hoạt động tốt trên tập huấn luyện nhưng mô hình lại dự đoán kém trên tập thử nghiệm (over-fitting); huấn luyện trở nên khó khăn, khó tìm ra được các trọng số tốt nhất cho mô hình, khó hội tụ. Do đó ta có thể giải quyết bằng các cách: thiết kế một mạng không quá phức tạp để thử nghiệm lại; sử dụng kỹ thuật transfer learning để tận dụng các trọng số đã được huấn luyện trên một bộ dữ liệu lớn và tiếp tục huấn luyện trên dữ liệu riêng; thu thập thêm nhiều dữ liệu.

2.1.2 Chỉnh sửa xoay (Rotation correction)

Các ảnh trong môi trường thật có thể khác nhau ở góc độ, ánh sáng, kích thước dù cùng 1 biểu cảm, những khác biệt đó có thể ảnh hưởng đến độ chính xác của hệ thống. Do đó, cần loại bỏ các ảnh hưởng của góc độ bằng cách căn chỉnh lại vùng mặt theo phương ngang bằng các phép quay và phép chuyển đổi. Để xoay lại ảnh gương mặt, cần phải có hai thông tin hình ảnh khuôn mặt và trung tâm của hai mắt.

2.1.3 Cắt ảnh gương mặt (Face cropping)

Với mô hình đề xuất, cần xác định khoảng cách giữa 2 mắt, sau đó thực hiện cắt ảnh gương mặt với các kích thước xác định như sau: chiều cao (height) vùng cắt có hệ số 4.5 gồm: 1.3 cho vùng trên mắt và 3.2 cho vùng dưới mắt, chiều rộng vùng cắt (width) có hệ số 2.4, các hệ

số trên nhân với khoảng cách từ giữa 2 mắt đến mắt phải để có kích thước thật.

2.1.4 Giảm kích thước ảnh gương mặt (Downsampling)

Với mô hình đề xuất, sử dụng nội suy tuyến tính để giảm kích thước ảnh về 32x32.

2.1.5 Chuẩn hóa cường độ

Các yếu tố về độ sáng (brightness) và độ tương phản (contrast) sẽ gây nhầm lẫn ở các ảnh của cùng lớp biểu cảm và cùng đối tượng. Do dó, cần giảm ảnh hưởng hai yếu tố này để giảm độ phức tạp cho bộ phân lớp bằng chuẩn hóa cường độ (Intensity normalization).

2.2 Mạng nơ ron tích chập cho phân lớp cảm xúc

2.2.1 Kiến trúc mạng nơ-ron tích chập (Convolutional Neural Network)

Mô hình đề xuất có kiến trúc mạng CNN như sau:

Layer (type)	Output Shape	Param #
Conv2D_1 (Conv2D)	(None, 28, 28, 32)	832
Activation_1 (Activation)	(None, 28, 28, 32)	0
MaxPooling2D_1 (MaxPooling2D	(None, 14, 14, 32)	0
Conv2D_2 (Conv2D)	(None, 8, 8, 64)	100416
Activation_2 (Activation)	(None, 8, 8, 64)	0
MaxPooling2D_2 (MaxPooling2D	(None, 4, 4, 64)	0
Flatten_1 (Flatten)	(None, 1024)	0
Dense_1 (Dense)	(None, 256)	262400
Activation_3 (Activation)	(None, 256)	0
Dense_2 (Dense)	(None, 7)	1799
Activation_4 (Activation)	(None, 7)	0

Total params: 365,447
Trainable params: 365,447
Non-trainable params: 0

Hình 2. 1: Thông số chi tiết mô hình CNN trong thí nghiệm của học viên

Trong quá trình huấn luyện, bộ tối ưu hóa được sử dụng là Stochastic Gradient Descent (SGD) và hàm kích hoạt ở mỗi lớp tích chập là Relu.

2.2.2 Huấn luyện

Trong phương pháp đề xuất, dữ liệu đã qua tiền xử lý và nhãn tương ứng sẽ được đưa vào kiến trúc mạng học sâu để tiến hành huấn luyện. Trước khi huấn luyện, học viên tiến hành trộn dữ liệu (Data shuffle), mục đích của việc này giúp dữ liệu không theo một thứ tự có trước nào, nên khi đưa một tập ảnh vào kiến trúc mạng thì mạng sẽ học được nhiều các trường hợp cảm xúc hơn. Nếu không trộn dữ liệu, các mẫu dữ liệu có cảm xúc giống nhau đưa vào mạng khiến mạng không

học được các mẫu dữ liệu có cảm xúc khác, điều này ảnh hưởng cực kỳ nặng đến độ chính xác của mô hình.

2.2.3 Kiểm thử

Sau quá trình huấn luyện sẽ thu được bộ trọng số tốt nhất. Bộ trọng số này được dùng để kiểm thử độ chính xác của mô hình trên tập dữ liệu thử nghiệm. Các ảnh thử nghiệm cũng được đưa qua hệ thống tiền xử lý giống như huấn luyện (trừ sử dụng Elastic Disortions để tăng cường dữ liệu), sau đó đầu ra của quá trình tiền xử lý là đầu vào của mạng học sâu. Các đặc trưng ảnh ngày càng tăng về độ sâu để học đặc trưng ngữ nghĩa cấp cao, đến các lớp kết nối đầy đủ (fully connected layer) sẽ được trải thẳng ra thành một dạng lớp ẩn gồm nhiều node, cuối cùng đi qua một lớp Softmax gồm 6 node chính là phân bố xác suất của 6 lớp cảm xúc đầu ra. Lúc này, xác suất dự đoán của cảm xúc nào cao nhất thì đầu ra cuối cùng sẽ là cảm xúc đó.

2.2.4 Mạng Deep Convolutional Neural Network (DCNN)

Tuy nhiên khi tăng độ sâu của mô hình CNN sẽ dẫn đến trường hợp overfitting trong quá trình huấn luyện mô hình. Do đó, học viên cũng áp dụng các phương pháp giảm overfitting trong quá trình huấn luyện như kỹ thuật DropOut hay kỹ thuật BatchNormalization để nhằm mục đích tăng hiệu quả dự đoán các nhãn trên hai bộ dữ liệu thực nghiệm.

```
Model: "DCNN"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 32, 32, 64)        1664
_________________________________________________________________
batchnorm_37 (BatchNormaliza (None, 32, 32, 64)        256
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 32, 32, 64)        102464
_________________________________________________________________
batchnorm_1 (BatchNormalizat (None, 32, 32, 64)        256
_________________________________________________________________
maxpool2d_1 (MaxPooling2D)   (None, 16, 16, 64)        0
_________________________________________________________________
dropout_1 (Dropout)          (None, 16, 16, 64)        0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 16, 16, 128)       73856
_________________________________________________________________
batchnorm_2 (BatchNormalizat (None, 16, 16, 128)       512
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 16, 16, 128)       147584
_________________________________________________________________
maxpool2d_2 (MaxPooling2D)   (None, 8, 8, 128)         0
_________________________________________________________________
dropout_2 (Dropout)          (None, 8, 8, 128)         0
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 8, 8, 256)         295168
_________________________________________________________________
batchnorm_7 (BatchNormalizat (None, 8, 8, 256)         1024
_________________________________________________________________
conv2d_6 (Conv2D)            (None, 8, 8, 256)         590080
_________________________________________________________________
maxpool2d_3 (MaxPooling2D)   (None, 4, 4, 256)         0
_________________________________________________________________
dropout_3 (Dropout)          (None, 4, 4, 256)         0
_________________________________________________________________
flatten (Flatten)            (None, 4096)              0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               524416
_________________________________________________________________
batchnorm_8 (BatchNormalizat (None, 128)               512
_________________________________________________________________
dropout_4 (Dropout)          (None, 128)               0
_________________________________________________________________
out_layer (Dense)            (None, 7)                 903
=================================================================
Total params: 1,738,695
Trainable params: 1,737,415
Non-trainable params: 1,280
_________________________________________________________________
```

Hình 2. 2: Chi tiết đầu vào và các thông số của mô hình DCNN được sử dụng

- **Nhận xét sự so sánh giữa hai mô hình DCNN và CNN**

Mô hình DCNN có tổng tham số là 1,738,695 tham số, trong khi số lượng này ở mô hình CNN chỉ là 365,447 tham số. Bởi vì ở mô hình DCNN, chúng ta sử dụng nhiều lớp tích chập Convolutional layer với các bộ lọc và kích thước khác nhau. Dẫn đến việc mô hình có nhiều tham số cần phải học trong quá trình huấn luyện hơn. Đối với mỗi epoch thì mô hình DCNN cần khoảng 12s để hoàn thành một epoch, trong khi đó với mô hình CNN thì chỉ cần khoảng 3s để hoàn thiện trên một epoch.

2.3 Kết luận của chương 2

Trong chương này, học viên đã trình bày tổng quan hệ thống nhận diện khuôn mặt và các bước tiền xử lý hình ảnh được áp dụng trong quá trình thí nghiệm nhằm mục đích nâng cao hiệu quả dự đoán trên cả hai bộ dữ liệu. Bên cạnh đó, học viên cũng đã trình bày chi tiết hai mô hình mạng học sâu CNN và mô hình học sâu DCNN bao gồm chi tiết các thông số, kiến trúc và số lượng tham số của mỗi mô hình. Đây là hai mô hình chính được sử dụng để chạy thí nghiệm trên hai bộ dữ liệu thực nghiệm và so sánh.

CHƯƠNG 3: THỬ NGHIỆM VÀ THẢO LUẬN

3.1 Cơ sở dữ liệu

3.1.1 Dữ liệu Cohn-Kanade mở rộng (CK+)[18]

Dữ liệu CK+ bao gồm các ảnh gương mặt (phần lớn là các ảnh đơn sắc) với các loại cảm xúc: buồn bã (sadness), bất ngờ (surprise), hạnh phúc (happiness), sợ hãi (fear), giận dữ (fear), khinh thường (contempt), ghê tởm (disgust). Mỗi ảnh trong tập dữ liệu có kích thước 48x48 và có tổng cộng 981 ảnh trong bộ dữ liệu.

3.1.2 The Japanese Female Facial Expression (JAFFE) Dataset

Dữ liệu JAFFE bao gồm ảnh của 10 phụ nữ Nhật Bản với 6 cảm xúc cơ bản (Giận dữ, khinh thường, sợ hãi, hạnh phúc, buồn bã, bất ngờ) và 1 nhãn bình thường. Dữ liệu bao gồm 213 ảnh đơn sắc với kích thước 256x256.

3.2 Môi trường thử nghiệm.

Để huấn luyện mô hình và cài đặt thuật toán này "Phân tích biểu cảm mặt người dùng mạng nơ ron tích chập", học viên sử dụng các công cụ, thư viện và ngôn ngữ lập trình như sau:

- Lập trình bằng Python: Python có cú pháp rất đơn giản, rõ ràng. Nó dễ đọc và viết ngắn hơn rất nhiều khi so sánh với

những ngôn ngữ lập trình khác như C++, Java, C#. Python làm cho việc lập trình trở nên thú vị, cho phép chúng ta tập trung vào những giải pháp chứ không phải cú pháp.

- Thư viện máy học Tensorflow – keras: Keras được coi là một thư viện mức độ cao của TensorFlow, Microsoft (CNTK), hoặc Theano. Keras có cú pháp đơn giản hơn TensorFlow rất nhiều. Các ưu điểm của thư viện này là: (1) Keras ưu tiên trải nghiệm của người lập trình, (2) Keras hỗ trợ huấn luyện trên nhiều GPU phân tán, (3) Keras đã được sử dụng rộng rãi trong doanh nghiệp và cộng đồng nghiên cứu.
- Thư viên Xử lý ảnh – OpenCV: OpenCV là tên viết tắt của Open Source computer vision library. Đây là một thư viện mã nguồn mở phục vụ cho hướng nghiên cứu xử lý hình ảnh, phát triển các ứng dụng đồ họa trong thời gian thực. OpenCV cho phép cải thiện tốc độ của CPU khi thực hiện các hoạt động real time. Nó còn cung cấp một số lượng lớn các mã xử lý phục vụ cho quy trình của thị giác máy tính hay các mô hình máy học khác nhau.

Cấu hình thử nghiệm cho các nghiên cứu được thực hiện trên máy tính cá nhân Window 10 Enterprise LTSC, Intel Core i7 Tiger Lake -1180H 4.6 GHz với NVIDA GeForce RTX 3050 Ti 4GB có khả năng có 2.3 Gb bộ nhớ trong GPU.

3.3 Cài đặt thử nghiệm và độ đo đánh giá

Đối với phương pháp mô hình mạng tích chập CNN, học viên sử dụng thư viện Tensorflow Keras trên ngôn ngữ lập trình Python để cài đặt. Giá trị dropout em sử dụng là 0.2 và hàm kích hoạt là Relu. Số lượng node của lớp đầy đủ là 128 với hàm kích hoạt là Relu. Gía trị dropout được sử dụng tại lớp này là 0.25 để giảm. Hàm tối ưu được sử dụng là SGD với giá trị học là 0.01 và tốc đô momentum là 0.95. Hàm mất mát được sử dụng là hàm Cross-entropy.

Để đánh giá hiệu quả của các phương pháp, học viên tiến hành các mô hình thử nghiệm đề xuất và các mô hình học máy cơ bản sử dụng ba độ đo là độ chính xác – Accuracy, độ chính xác - Precision, độ phủ - Recall và chỉ số F1 – F1 score giữa tập dự đoán và tập dữ liệu được gán nhãn. Các độ đo được tính bằng các công thức sau đây:

- Accuracy – Độ chính xác: Cách đơn giản nhất để đánh giá một mô hình phân lớp đó là sử dụng độ chính xác (Accuracy). Ý tưởng đơn giản là tỷ lệ giữa các mẫu dự đoán đúng trên tổng số mẫu của dữ liệu kiểm thử.
- Ma trận Confusion: Cách tính sử dụng accuracy chỉ cho chúng ta biết được bao nhiêu phần trăm dữ liệu được dự đoán đúng mà không chỉ ra các dữ liệu được dự đoán đúng/sai như thế nào. Do đó chúng ta cần một phương pháp đánh giá tốt hơn gọi là ma trận Confusion (Confusion matrix). Một cách tổng quát,

ma trận này sẽ thể hiện có bao nhiêu điểm dữ liệu thực sự thuộc về một lớp, và được được dự đoán là rơi vào một lớp.

3.4 Số liệu

3.4.1 Thử nghiệm bộ dữ liệu CK+ gốc

3.4.2 Thử nghiệm bộ dữ liệu CK+ khi tăng cường dữ liệu học

3.4.3 Thử nghiệm bộ dữ liệu JAFFE gốc

3.4.3 Thử nghiệm bộ dữ liệu JAFFE gốc

3.5 Kết quả thử nghiệm

Nhìn vào Bảng 3.1, chúng ta sẽ thấy kết quả chính của mô hình CNN khi kết hợp các phương pháp tiền xử lý khác nhau trên bộ dữ liệu CK+ tương ứng cho mỗi nhãn cảm xúc. Chúng ta dễ dàng nhìn thấy được sự hiệu quả của mô hình này đối với các nhãn cảm xúc “Disgust”, “Happy”, “Surprise” hay nhãn “Neutral” với chỉ số F1-score lần lượt là 0.88, 0.89, 0.93 và 0.81. Số lượng dữ liệu cho từng nhãn cảm xúc ít nhất trong bộ dữ liệu CK+ tương ứng là nhãn “Fear”, nhãn “Sad” và nhãn “Angry”, đó là chính là lý do tại sao kết quả trên ba nhãn này đạt kết quả thấp nhất trong toàn bộ nhãn cảm xúc

3.6 Điều chỉnh tiền xử lý

3.7 So sánh kết quả mô hình CNN và DCNN

3.7.1 Tăng số lượng lớp tích chập – Convolution layer

3.7.2 Áp dụng kỹ thuật dropout và batch normalization

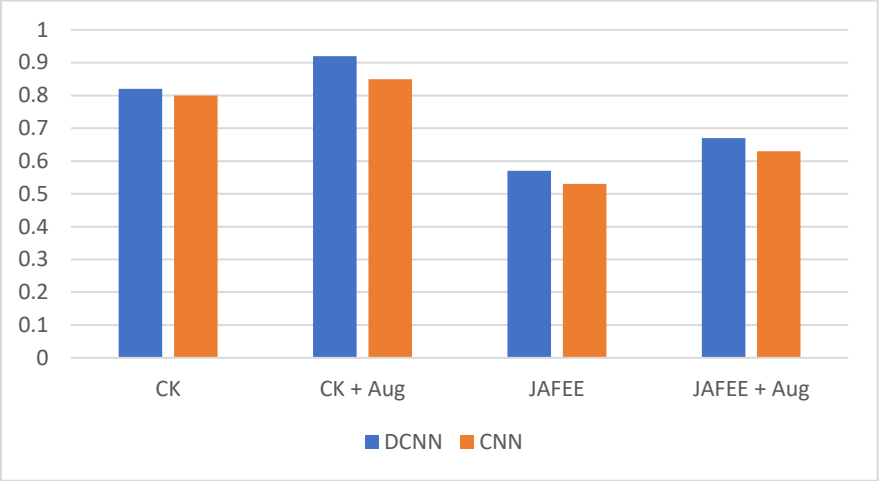
3.7.3 Mô hình

- **Kết quả thí nghiệm**

Ở phần này, học viên sẽ so sánh kết quả thí nghiệm mô hình cải tiến DCNN và mô hình CNN được trình bày ở chương 2 trên cả hai bộ dữ liệu CK+ và JAFFE gốc và khi tăng cương dữ liệu. Kết quả được báo cáo trên các độ đo như độ chính xác, độ phủ và chỉ số F1-score. Bảng 3.9 và Hình 3.8 trình bày kết quả so sánh giữa hai mô hình DCNN và mô hình CNN trên cả hai bộ dữ liệu trước và sau khi tăng cường dữ liệu.

Bảng 3. 1: Kết quả các độ đo DCNN trên hai bộ dữ liệu gốc và sau khi tăng cường dữ liệu

Bộ dữ liệu	Loại dữ liệu	Precision	Recall	F1-score
CK+	Dữ liệu gốc	0.83	0.82	0.82
	Dữ liệu tăng cường	0.92	0.92	0.92
JAFFE	Dữ liệu gốc	0.59	0.55	0.57
	Dữ liệu tăng cường	0.71	0.63	0.67

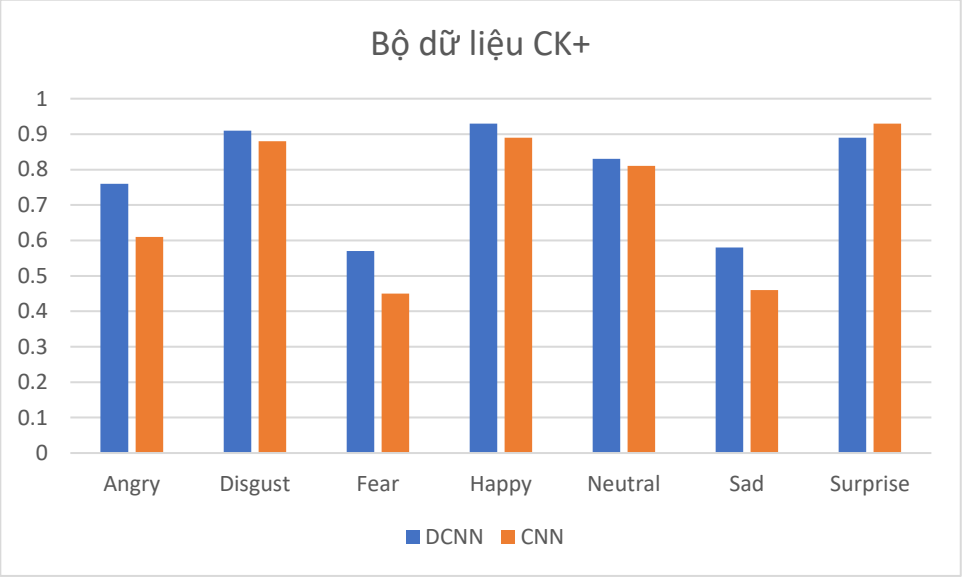


Hình 3. 1: Kết quả độ đo F1 giữa mô hình DCNN và CNN trên hai bộ dữ liệu gốc và tăng cường dữ liệu

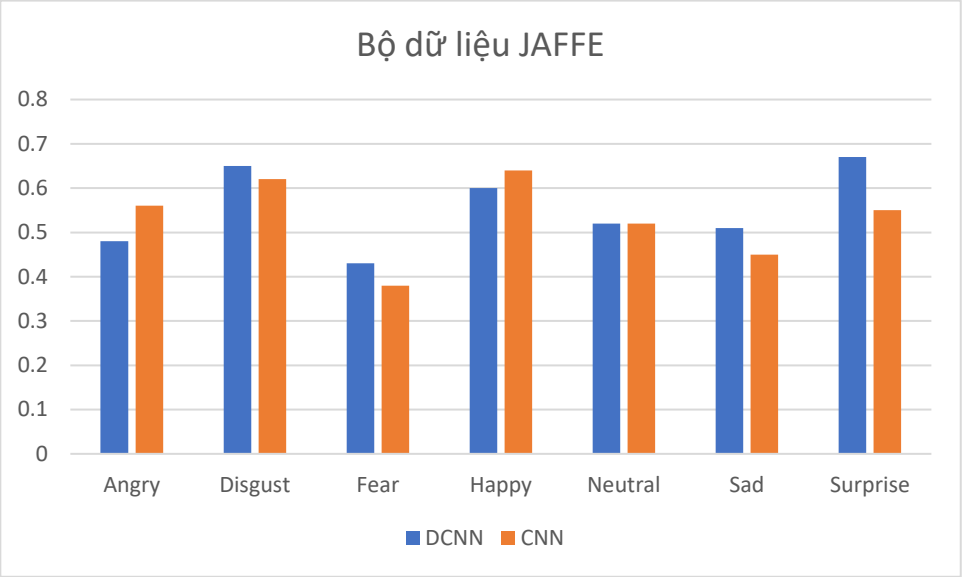
Nhìn vào bảng số liệu 3.9 và hình 3.12, chúng ta thấy mô hình cải tiếp DCNN với thêm các lớp tích chập để rút nhiều thông tin hơn từ bức ảnh và áp dụng 2 kỹ thuật giảm overfiting trong quá trình huấn luyện mô hình đều cho kết quả tốt hơn mô hình CNN. Cụ thể đối với bộ dữ liệu gốc CK+, mô hình DCNN đạt độ chính xác 0.83, độ phủ là 0.82 còn giá trị F1 là 0.82, so với mô hình CNN thì mô hình cải tiến này cao hơn +0.02. Còn đối với bộ dữ liệu CK+ khi tăng cường dữ liệu thì mô hình DCNN cao hơn mô hình CNN +0.07 về độ đo F1. Trong khi đó với bộ dữ liệu JAFFE, mô hình DCNN cũng chứng minh sự hiệu quả khi đối với dữ liệu gốc, mô hình này cao hơn mô hình CNN +0.03, còn trên dữ liệu tăng cường thì mô hình +0.04. Nhìn một cách tổng quan, chúng ta thấy mô hình DCNN đều cho kết quả tốt hơn mô hình CNN. Hình 3.13 và Hình 3.14 trình bày kết quả chi tiết các độ đo của mô hình

DCNN và mô hình CNN trên cả hai bộ dữ liệu CK+ và bộ dữ liệu JAFFE.

Nhìn vào hình 3.13, chúng ta thấy rằng mô hình DCNN giúp tăng hiệu qua hơn các mô hình CNN hầu hết trên tất cả các nhãn trừ nhãn Surprise. Đặc biệt là đối với hai nhãn có tỷ lệ nhãn ít và độ chính xác thấp như "Fear" và nhãn "Sad". Mô hình DCNN giúp tăng hiệu quả +0.12 cho cả hai nhãn này so với mô hình CNN.



Hình 3. 2: Kết quả các độ đo của mô hình DCNN và mô hình CNN trên bộ dữ liệu gốc CK+



Hình 3. 3: Kết quả các độ đo của mô hình DCNN và mô hình CNN trên bộ dữ liệu gốc JAFFE

3.8 Kết luận của chương 3

Học viên đã trình bày chi tiết hai bộ dữ liệu được sử dụng để thực nghiệm trong đề tài này là bộ dữ liệu CK và bộ dữ liệu JAFFE. Sau đó, học viên đã nhận xét và so sánh sự hiệu quả của hai mô hình CNN và DCNN trên cả hai bộ dữ liệu theo các độ đo độ chính xác – accuracy, độ phủ - recall, độ chính xác - precision và độ đo F1. Ngoài ra học viên còn phân tích xem tỷ lệ nhầm lẫn giữa các nhãn trong hai bộ dữ liệu để kiểm tra xem là nhãn nào hay bị dự đoán sai nhất. Cuối cùng học viên so sánh kết quả của hai mô hình DCNN và mô hình CNN trên hai bộ dữ liệu

CHƯƠNG 4: ỨNG DỤNG

4.1 Ứng dụng phát hiện cảm xúc khuôn mặt

Để xây dựng ứng dụng phát hiện cảm xúc khuôn mặt trên webcam của màn hình laptop, em sử dụng ngôn ngữ Python kết hợp với thư viện OpenCV để viết chương trình minh họa hỗ trợ cho việc xử lý dữ liệu đầu vào từ webcam. Quá trình xử lý qua 5 bước như sau:

Bước 1: Ảnh đầu vào được chuyển thành đa cấp xám.

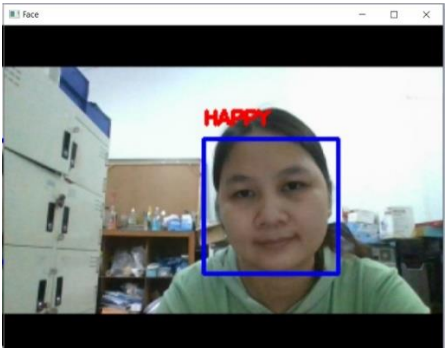
Bước 2: Dùng haar cascade của thư viện OpenCV để tìm kiếm vùng mặt người trên ảnh đầu vào, sau khi cắt được vùng khuôn mặt thì sẽ được chuyển đến bước 3.

Bước 3: Vùng ảnh mặt người được chuyển đổi về kích thước 32x32 bởi vì đây là kích thước đầu vào của việc huấn luyện các mô hình CNN trên các bộ dữ liệu.

Bước 4: Ảnh sau khi đã được chuẩn hóa thành kích thước 32x32 đa cấp xám chuyển đổi về miền [0, 1] sau đó đưa vào mô hình CNN đã được huấn luyện sẵn để tiến hành dự đoán nhãn cảm xúc.

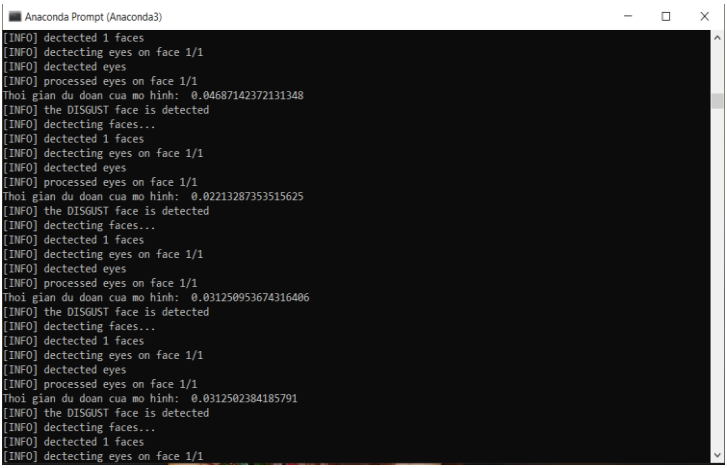
Bước 5: Đầu ra của CNN là xác suất của các cảm xúc, chọn cảm xúc có xác xuất cao nhất làm kết quả cuối cùng.

Ở trong phần này học viên lựa chọn mô hình CNN được huấn luyện trên bộ CK+, JAFFE bởi vì mô hình này đạt kết quả tương đối tốt và số lượng tham số ít hơn mô hình DCNN. Điều này sẽ đảm bảo việc xử lý khi sử dụng Webcam trực tiếp từ máy tính với cấu hình của laptop đang sử dụng.Dưới đây là hình ảnh minh họa chạy thực tế của học viên.



Hình 4. 1: Kết quả dự đoán mô hình CNN trên thử nghiệm thực tế đối vỡi nhãn "Happy"

Thời gian dự đoán mô hình CNN trên thử nghiệm thực tế khi nhận diện một khung hình để xác định cảm xúc, tính từ lúc webcam nhận hình đưa vào mô hình đến lúc ra kết quả mất trung bình 0.03s. Hình 4.2 là thời gian chạy thực nghiệm khi nhận dạng cảm xúc khuôn mặt qua webcam.



Hình 4. 2: Thời gian dự đoán mô hình CNN trên thử nghiệm thực tế

4.2 Kết luận chương 4

Trong chương 4 này, học viên đã trình bày một ứng dụng minh họa sử dụng trên webcam của máy tính để phát hiện và phân loại cảm xúc khuôn mặt. Với sự hỗ trợ của thư viện OpenCV trong việc xác định vùng chứa khuôn mặt, học viên lấy kết quả này đưa qua mô hình CNN đã được huấn luyện sẵn để tiến hành dự đoán nhãn cảm xúc rồi xuất lên màn hình.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết quả nghiên cứu của luận văn

Trong luận văn này, học viên đã tìm hiểu, khảo sát các kỹ thuật Học máy, Học sâu cũng như các kỹ thuật tiền xử lý. Dựa trên các tìm hiểu, học viên kiểm chứng một số phương pháp tiền xử lý ảnh và tăng cường mẫu học và mạng CNN sâu cải tiến cho bài toán nhận diện cảm xúc gương mặt đạt hiệu quả thời gian thực, hai bộ dữ liệu dùng để huấn luyện và đánh giá là JAFFE và CK+. Trong đó, các kết quả thực nghiệm ở Bảng 3.5 cho thấy việc áp dụng kỹ thuật tăng cường dữ liệu có ảnh hưởng tích cực rõ rệt thì độ chính xác thu được cao hơn khi huấn luyện với tập ảnh gốc ở cả hai bộ dữ liệu. Phương pháp tiền xử lý cũng chứng minh sự hiệu quả, trong đó các kết quả cho thấy việc áp dụng tất cả các kỹ thuật tiền xử lý được trình bày ở phần 2.1 sẽ hiệu quả hơn so với thực hiện riêng lẽ từng phương pháp.

Các thử nghiệm trình bày ở Bảng 3.9 cũng cho thấy kiến trúc mạng DCNN sâu đề xuất bởi học viên cho kết quả tốt trên cả hai bộ dữ liệu CK+ và JAFFE, tuy nhiên học viên sử dụng kiến trúc mạng đơn giản ở Chương 2 để xây dựng hệ thống demo phát hiện cảm xúc để phù hợp với yêu cầu thời gian thực với cấu hình máy laptop cá nhân trong môi trường thực tế.

Học viên cũng đã xây dựng một ứng dụng minh họa phát hiện cảm xúc khuôn mặt chạy trong môi trường thực tế với thời gian thực.

5.2 Những hạn chế trong luận văn

Bên cạnh những gì học viên đã đạt được trong quá trình làm luận văn cũng tồn tại những hạn chế mà cần phải được nghiên cứu và phát triển trong tương lai. Học viên hiện tại chỉ kiểm tra và đánh giá trên bộ dữ liệu chuẩn đã được công bố cho nghiên cứu là bộ dữ liệu CK và JAFFE. Kết quả thử nghiệm trên bộ dữ liệu JAFFE chưa đạt kết quả như mong đợi, còn thấp khá nhiều so với bộ dữ liệu CK+.

5.3 Hướng phát triển

Thử nghiệm trên nhiều bộ dữ liệu để đạt kết quả cao hơn như Fer 2013, VGG 16, Resnet … với hệ thống máy có cấu hình mạnh hơn.

Xây dựng được hệ thống hỗ trợ phân biệt cảm xúc bệnh nhân có độ chính xác cao và đáp ứng các yêu cầu về thời gian thực, phục vụ cho việc khám chữa bệnh tại Bệnh Viện Đa Khoa Tây Ninh để nâng cao trải nghiệm người dùng.