TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN - CƠ - TIN HỌC

BÁO CÁO CUỐI KÌ

KHAI PHÁ DỮ LIỆU

Đề tài:

Mô hình mạng nơ-ron tích chập trong nhận diện cảm xúc qua hình ảnh

Giảng viên	LÊ HOÀNG SƠN	
Sinh viên thực hiện	LÊ TRỌNG MINH	20002072
	BÙI ĐĂNG DƯƠNG	20001899
	NGUYỄN QUANG DŨNG	20001896

Hà Nội, 05-2024

LỜI NÓI ĐẦU

Ngày nay, phát hiện và hiểu biết cảm xúc là một khía cạnh quan trọng trong trí tuệ nhân tạo và tương tác con người - máy tính. Có thể thấy rằng cảm xúc tác động mạnh mẽ đến hành vi và quyết định của con người. Do đó, khả năng phát hiện cảm xúc của máy sẽ giúp cải thiện khả năng tương tác giữa máy và con người.

Hình ảnh khuôn mặt là một trong những dấu hiệu cảm xúc phổ biến và dễ quan sát nhất. Bộ dữ liệu FER 2013 cung cấp hơn 35.000 hình ảnh khuôn mặt được đánh dấu cảm xúc, cho phép phát triển và đánh giá các phương pháp phát hiện cảm xúc dựa trên hình ảnh.

Trong báo cáo này, nhóm em sẽ mô tả và chuẩn bị dữ liệu, đồng thời đề xuất các phương pháp máy học khác nhau để phát hiện bảy cảm xúc cơ bản - giận dữ, sợ hãi, buồn, bất ngờ, hạnh phúc,ghê tởm và bình thường - từ các hình ảnh khuôn mặt. Nhóm em hy vọng nghiên cứu này sẽ là gợi ý cho những nghiên cứu tiếp theo liên quan đến phát hiện cảm xúc trên hình ảnh.

Trong nội dung bài báo này, chúng em sẽ trình bày cách các mô hình CNN từ cơ bản cho đến phức tạp hơn phân tích cảm xúc từ hình ảnh . Phần còn lại của bài báo được tổ chức như sau:

Chương I: Giới thiệu chung

Chương II: Cơ sở lý thuyết

Chương III: Thiết kế và xây dựng mô hình

Chương IV: Thảo luận và đánh giá kết quả.

Chương V: Kết luận

Chương VI: Tài liệu tham khảo

Đây là một đề tài thú vị và khá mới mẻ, mặt khác các tài liệu cho nghiên cứu không nhiều, do đó kết quả đạt được chắc chắn chưa thể thỏa mãn được yêu cầu thực tế đặt ra. Chúng em kính mong các thầy/ cô góp ý thêm để luận văn của chúng em đạt gần với thực tế hơn. Chúng em xin chân thành cảm ơn.

Contents

1 Đặt vấn đề **4**

1.1 Đặt vấn đề 4

1.2 Đối tượng nghiên cứu 4

2 Cơ sở lý thuyết **7**

2.1 Tăng cường dữ liệu trong Deep Learning 7

2.2 Giới thiệu về các phương thức tăng cường dữ liệu 7

2.3 Giới thiệu về tính toán Class weights 9

2.4 Giới thiệu về Convolutional Neural Networks (CNN) 10

2.5 Các thành phần cơ bản của CNN 10

2.5.1 Lớp Convolutional Layer 10

2.5.2 Lớp Pooling Layer 11

2.5.3 Rectified Linear Unit - ReLU layer 13

2.5.4 Lớp Fully-Connected Layer 13

2.5.5 Output Layer 14

2.5.6 Drop out 14

2.6 Giới thiệu về Hyperparameter tuning (Tối ưu siêu tham số) 15

3 Thiết kế và xây dựng mô hình **16**

3.1 Kiến trúc mô hình nhận diện cảm xúc khuôn mặt sử dụng mạng nơ-ron tích chập(CNN) 16

3.2 Tối ưu siêu tham số của mô hình mạng nơ-ron tích chập 18

4 Thảo luận và đánh giá kết quả **21**

4.1 Kết quả của mô hình CNN 21

4.2 So sánh giữa hai mô hình CNN trước và sau khi tối ưu siêu tham số và hiệu chỉnh class weights 22

5 Kết Luận **24**

6 Tài liệu tham khảo **26**

1 Đặt vấn đề

1.1 Đặt vấn đề

Trong thời đại công nghệ số ngày nay, AI đang ngày càng được áp dụng nhiều hơn vào các mặt của đời sống, công việc. Việc nhận diện cảm xúc từ hình ảnh gương mặt đã trở thành một lĩnh vực nghiên cứu quan trọng và có nhiều ứng dụng thực tiễn. Nhận diện cảm xúc mang lại nhiều lợi ích trong các lĩnh vực như y tế, giáo dục, marketing và chăm sóc khách hàng. Công nghệ này giúp máy tính hiểu và phản hồi phù hợp với trạng thái cảm xúc của con người, tạo nên một môi trường tương tác thân thiện và hiệu quả hơn.

Một trong những công nghệ phổ biến để thực hiện nhận diện cảm xúc từ hình ảnh gương mặt là Convolutional Neural Network - CNN. CNN là một trong những mô hình học sâu (deep learning) mạnh mẽ nhất, đặc biệt hiệu quả trong việc xử lý và phân loại hình ảnh.

Với bộ dữ liệu FER2013, chúng em mong muốn mô hình có thể được huấn luyện và đưa ra những chỉ số tốt nhất.

1.2 Đối tượng nghiên cứu

Ở đề tài này nhóm em chọn bộ dữ liệu FER-2013 là đối tượng nghiên cứu.

Bộ dữ liệu FER-2013 được tạo ra bởi các nhà nghiên cứu tại trường đại học New York University vào năm 2013. Bộ dữ liệu này bao gồm tổng cộng 35,887 hình ảnh khuôn mặt với kích thước 48x48 pixel. Các hình ảnh này được thu thập từ nhiều nguồn khác nhau, bao gồm các trang web chia sẻ hình ảnh, các ứng dụng webcam và các bộ dữ liệu trước đó.

Các nhãn cảm xúc của từng hình ảnh được gán bởi các người đánh giá con người, bao gồm sáu loại cảm xúc chính: vui vẻ, buồn bã, tức giận, sợ hãi, ghê tởm, bất ngờ và bình thường. Mỗi hình ảnh được gán một nhãn cảm xúc duy nhất tương ứng với cảm xúc mà người đánh giá cho là phù hợp nhất.

Sau khi thu thập và gán nhãn cho các hình ảnh, các nhà nghiên cứu đã sử dụng các kỹ thuật xử lý ảnh để tiền xử lý và trích xuất đặc trưng của các hình ảnh này. Tiếp đó, các mô hình máy học và trí tuệ nhân tạo được huấn luyện trên bộ dữ liệu này để phân tích cảm xúc qua hình ảnh.

Bộ dữ liệu FER-2013 đã trở thành một trong những bộ dữ liệu phổ biến nhất được sử dụng để huấn luyện các mô hình phân tích cảm xúc qua hình ảnh. Nó cũng đã được sử dụng trong nhiều nghiên cứu khác nhau, chẳng hạn như trong lĩnh vực nhận dạng khuôn mặt, phân tích tâm trạng của khách hàng trong lĩnh vực bán lẻ, và trong các ứng dụng trí tuệ nhân tạo khác.

Đây là một trong những bộ dữ liệu phổ biến được sử dụng trong các nghiên cứu về nhận dạng cảm xúc từ khuôn mặt và học sâu. Nó đã được sử dụng để huấn luyện nhiều mô hình học máy và học sâu, bao gồm các mô hình sử dụng mạng nơ-ron tích chập (CNN).

Figure 1: 7 Cảm xúc cơ bản trong bộ dữ liệu FER-2013

FER-2013 chứa các ảnh cảm xúc khuôn mặt định dạng ở mức xám có kích thước 48x48. Các ảnh khuôn mặt được phân loại thành 7 cảm xúc cơ bản của con người với cách đánh số tương ứng như sau : Giận : 0, Ghê tởm : 1, Sợ : 2, Vui : 3 , Buồn : 4, Ngạc nhiên : 5, Tự nhiên :6

Bộ dữ liệu FER-2013 được chia thành 2 nhóm chính :

- Training set : là dữ liệu được dùng để huấn luyện mạng
- Test set : là nhóm dữ liệu dùng để đánh giá độ chính xác của mô hình, sau khi đã huấn luyện xong.

Data	Happy	Disgust	Anger	Fear	Sad	Neutral	Suprise
Train	7215	436	3995	4097	4830	4965	3171
Test	1774	111	958	1024	1247	1233	831

Table 1: Thống kê dữ liệu trong bộ dữ liệu FER-2013

FER-2013 là một tập dữ liệu cảm xúc khuôn mặt lớn sử dụng cho việc huấn luyện mạng nơ-ron, tuy nhiên việc phân bố dữ liệu không đồng đều giữa các trạng thái cảm xúc với nhau. Việc này có ảnh hưởng đến kết quả huấn luyện, sẽ được trình bày ở phần kết quả thực nghiệm của mô hình.

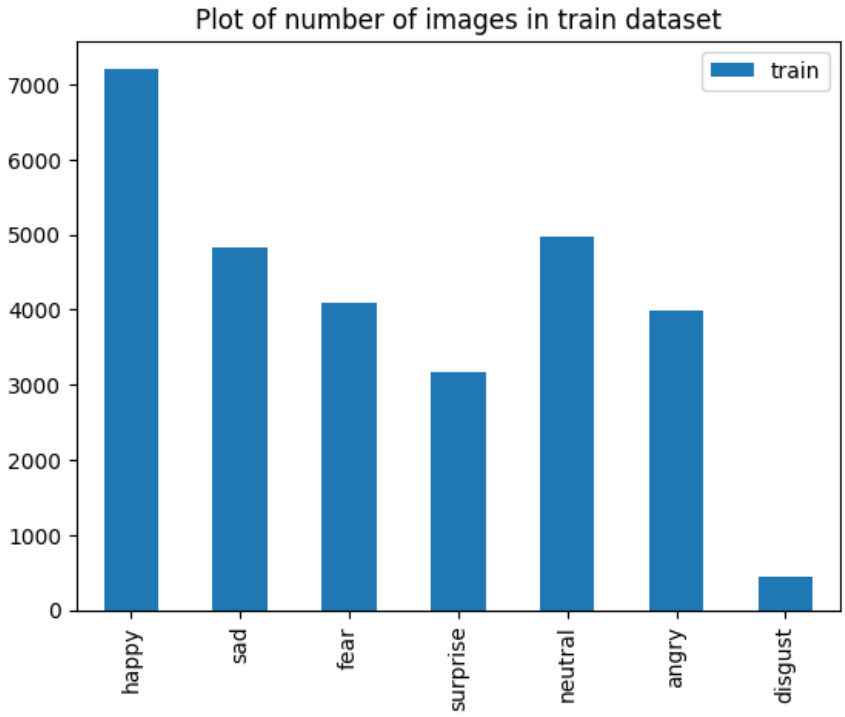


Figure 2: Biểu đồ phân bố dữ liệu trong tập train trong bộ dữ liệu FER-2013

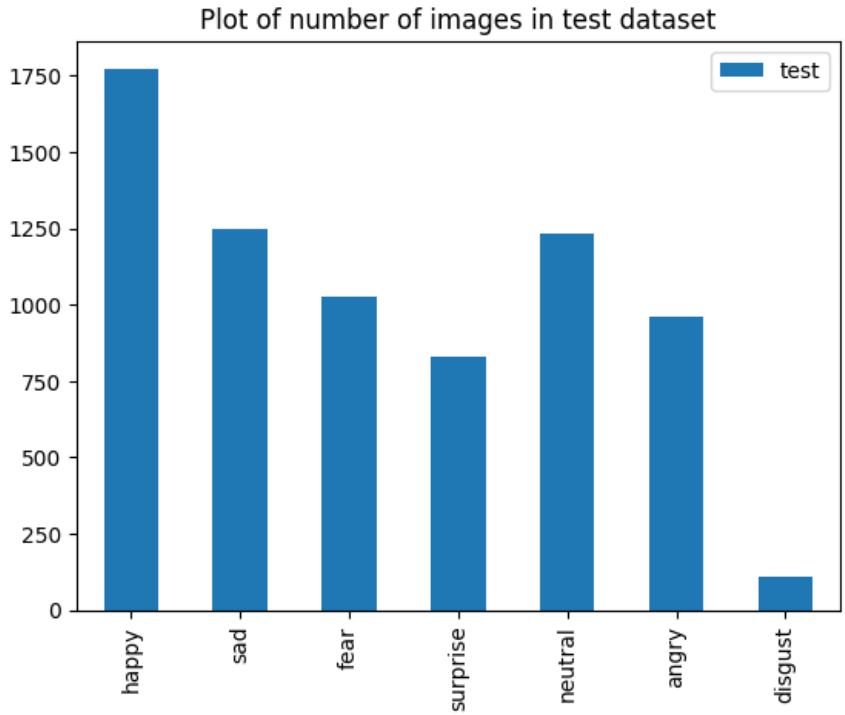


Figure 3: Biểu đồ phân bố dữ liệu trong tập test trong bộ dữ liệu FER-2013

2 Cơ sở lý thuyết

2.1 Tăng cường dữ liệu trong Deep Learning

Hiện nay trong deep learning thì vấn đề dữ liệu có vai trò rất quan trọng. Chính vì vậy có những lĩnh vực có ít dữ liệu để cho việc train model thì rất khó để tạo ra được kết quả tốt trong việc dự đoán. Do đó người ta cần đến một kỹ thuật gọi là tăng cường dữ liệu (data augmentation) để phục vụ cho việc nếu bạn có ít dữ liệu, thì bạn vẫn có thể tạo ra được nhiều dữ liệu hơn dựa trên những dữ liệu bạn đã có. Ví dụ như hình dưới, đó là các hình được tạo ra thêm từ một ảnh gốc ban đầu.

Ví dụ:

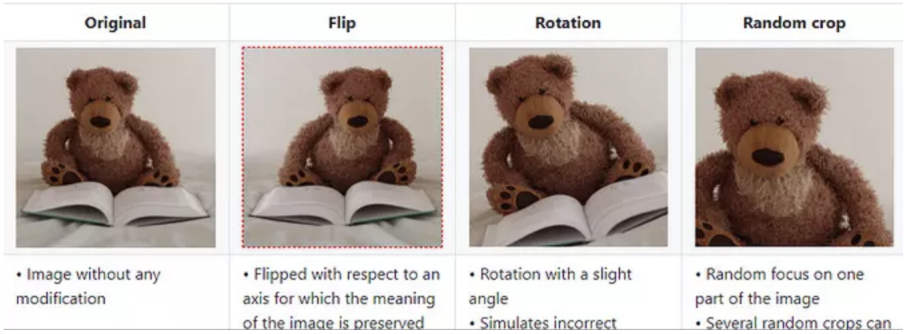


Figure 4: Phương thức tăng cường dữ liệu cơ bản

Bài báo cáo này sẽ trình bày phương pháp đơn giản nhất trong các phương pháp tăng cường dữ liệu. Phương pháp này là horizontal flip. Ta nhận thấy sự mất cân bằng dữ liệu trong bộ dữ liệu, từ đó sử dụng phương pháp tăng cường dữ liệu keras để có thể cân bằng lại bộ dữ liệu.

2.2 Giới thiệu về các phương thức tăng cường dữ liệu

Tăng cường dữ liệu (Data Augmentation) là một kỹ thuật quan trọng trong học máy và học sâu để tăng cường số lượng và đa dạng dữ liệu huấn luyện mà không cần thu thập thêm dữ liệu mới. Điều này giúp cải thiện hiệu suất của mô hình bằng cách giảm hiện tượng overfitting. Dưới đây là một số kỹ thuật phổ biến để tăng cường dữ liệu cho các hình ảnh:

Dịch chuyển (Translation): Di chuyển hình ảnh theo các hướng khác nhau (trái, phải, lên, xuống).

Xoay (Rotation): Xoay hình ảnh một góc nhất định.

Lật (Flip): Lật hình ảnh theo chiều ngang hoặc chiều dọc.

Cắt (Cropping): Cắt một phần nhỏ của hình ảnh.

Thay đổi độ sáng (Brightness Adjustment): Thay đổi độ sáng của hình ảnh.

Thêm nhiễu (Adding Noise): Thêm nhiễu ngẫu nhiên vào hình ảnh.

Biến dạng (Distortion): Biến dạng hình ảnh theo các cách khác nhau (zoom, shear, stretch).

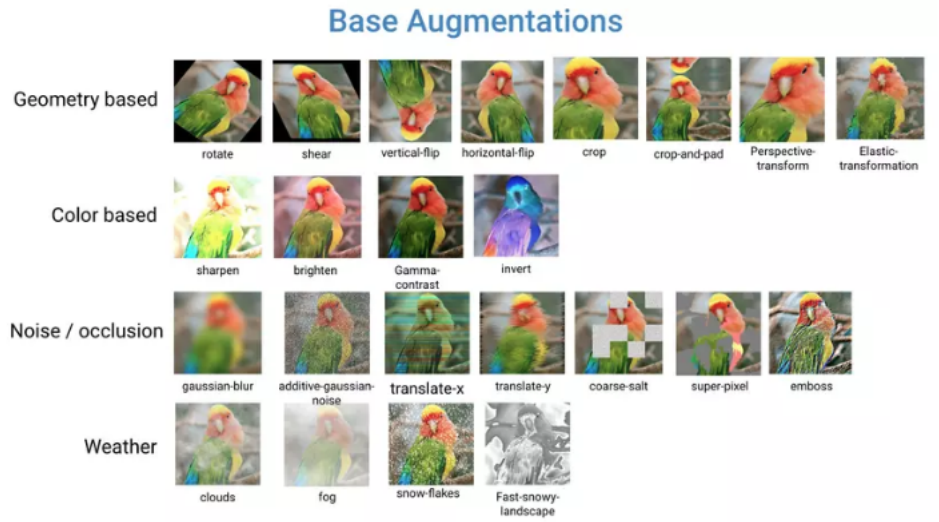


Figure 5: Phương thức tăng cường dữ liệu cơ bản

Chúng ta sẽ sử dụng kỹ thuật tăng cường dữ liệu để cải thiện độ cân bằng của mô hình. Do lớp Happy có lượng dữ liệu lớn và có lẽ có nhiều hình ảnh đa dạng, chúng ta sẽ chỉ áp dụng tăng cường dữ liệu cho các lớp còn lại.

Qúa trình Data Augmentation bao gồm các bước sau:

Bước 1 - Lấy các lớp cảm xúc khác không bao gồm happy: Tạo ra 1 tập con dữ liệu bằng cách lấy tất cả các 'emotion' ngoại trừ các dòng dữ liệu với emotion được label là happy.

Bước 2 - Xây dựng data generator để tăng cường dữ liệu: Gọi hàm ImageData-Generator, và lật hình theo chiều ngang

Bước 3 - Tăng chiều cho tập con dữ liệu: Để sử dụng hàm ImageDataGenerator, bộ dự liệu cần phải là 4 chiều, vì thế ta cần tăng thêm cho bộ dữ liệu 1 chiều (color).

Bước 4 - Áp dụng hàm tăng cường dữ liệu và giảm lại về 3 chiều: Sau khi áp dụng hàm tăng cường dữ liệu và giảm lại chiều, ta gộp bộ dữ liệu tăng cường vào lại bộ dữ liệu gốc.

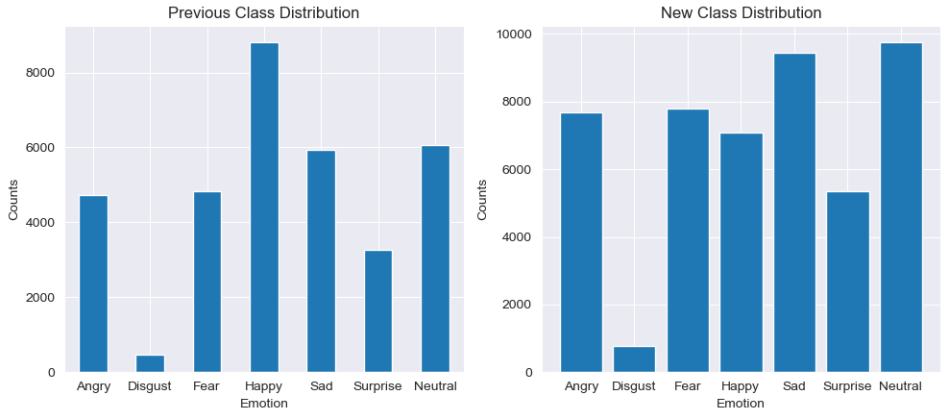


Figure 6: So sánh dữ liệu trước và sau khi tăng cường

2.3 Giới thiệu về tính toán Class weights

Class weights là một kỹ thuật được sử dụng trong huấn luyện mô hình học máy, đặc biệt là các mô hình phân loại, để xử lý vấn đề bất cân bằng dữ liệu. Khi một hoặc nhiều lớp trong tập dữ liệu chiếm số lượng ít hơn đáng kể so với các lớp khác, mô hình có thể bị thiên vị và ưu tiên dự đoán các lớp chiếm đa số. Việc sử dụng class weights giúp giảm thiểu hiện tượng này bằng cách gán trọng số cao hơn cho các lớp chiếm thiểu số trong quá trình huấn luyện.

Cách hoạt động của class weights: Trong quá trình tính toán hàm mất mát, class weights sẽ điều chỉnh đóng góp của mỗi mẫu vào tổng lỗi, dựa trên lớp của mẫu đó. Các mẫu từ lớp chiếm thiểu số sẽ có trọng số cao hơn, khiến chúng có ảnh hưởng lớn hơn đến quá trình cập nhật trọng số của mô hình.

Qúa trình tính toán Class weight bao gồm các bước sau:

Bước 1 - Tính toán trọng số lớp: Sử dụng $\mathrm{comput\_class\_weight}$ từ scikit-learn để tính toán trọng số cho mỗi lớp. Phương thức này nhận vào tham số $\mathrm{class\_weight} = \mathrm{'balanced'}$, giúp cân bằng các trọng số dựa trên tần suất xuất hiện của mỗi lớp trong tập dữ liệu.

Bước 2 - Chuyển đổi thành từ điển: Chuyển đổi kết quả từ $\mathrm{comput\_class\_weight}$ thành một từ điển, trong đó mỗi khóa là chỉ số của lớp và giá trị tương ứng là trọng số của lớp đó. Keras yêu cầu $\mathrm{class\_weight}$ phải ở dạng từ điển này.

Tính toán class weight giúp cho mô hình không bị thiên vị bởi các lớp chiếm đa số. Ngoài ra giúp mô hình học tốt hơn và có khả năng phân loại chính xác hơn đối với các lớp chiếm thiểu số. Việc sử dụng class weights là một kỹ thuật quan trọng và hiệu quả để xử lý các tập dữ liệu không cân bằng, đặc biệt là trong bài toán phân loại và nhận diện cảm xúc bằng mạng nơ-ron tích chập.

Nhận thấy sự mất cân bằng của dữ liệu, kể cả sau khi đã tăng cường dữ liệu, ta sử

dụng hàm tính toán class weight để có thể gia tăng trọng số cho lớp có số dữ liệu ít nhất 'disgust'. Hình sau đây sẽ là trọng số của các lớp trong bộ dữ liệu FER2013 được tính toán bằng hàm compute_class_weight.

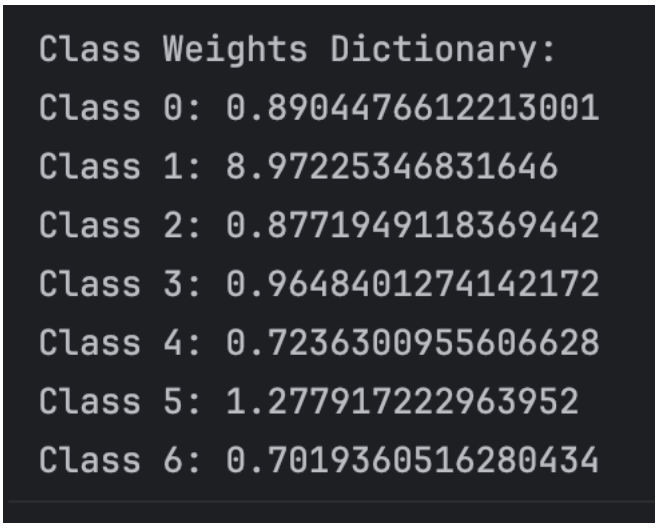


Figure 7: Class weights của các lớp trong bộ dữ liệu

Có thể thấy Class 1 hay là lớp cảm xúc "Disgust" có trọng số cao hơn hẳn khi lượng dữ liệu của lớp đó là thấp nhất.

2.4 Giới thiệu về Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNN) là một loại mạng neural được sử dụng trong việc xử lý hình ảnh và video. CNN có khả năng học và trích xuất các đặc trưng cấp cao của hình ảnh, giúp cho việc phân loại, nhận diện và phân tích hình ảnh trở nên hiệu quả hơn.

CNN được thiết kế để học các đặc trưng cấp cao của hình ảnh thông qua việc thực hiện các phép tích chập (convolution) trên hình ảnh đầu vào. CNN gồm các lớp chính như lớp Convolutional Layer, Pooling Layer và Fully-Connected Layer. Lớp Convolutional Layer sẽ thực hiện các phép tích chập trên hình ảnh đầu vào, tạo ra các đặc trưng cấp cao. Lớp Pooling Layer sẽ giảm kích thước của các đặc trưng bằng cách thực hiện các phép lấy mẫu trên các vùng của các đặc trưng. Lớp Fully-Connected Layer sẽ kết nối tất cả các đặc trưng với nhau để thực hiện phân loại và nhận diện các hình ảnh.

2.5 Các thành phần cơ bản của CNN

2.5.1 Lớp Convolutional Layer

Lớp này sẽ sử dụng một bộ gồm các bộ lọc có kích thước nhỏ so với ảnh áp vào một vùng nhất định trong ảnh và tiến hành tính tích chập giữa bộ filter và các giá trị điểm ảnh

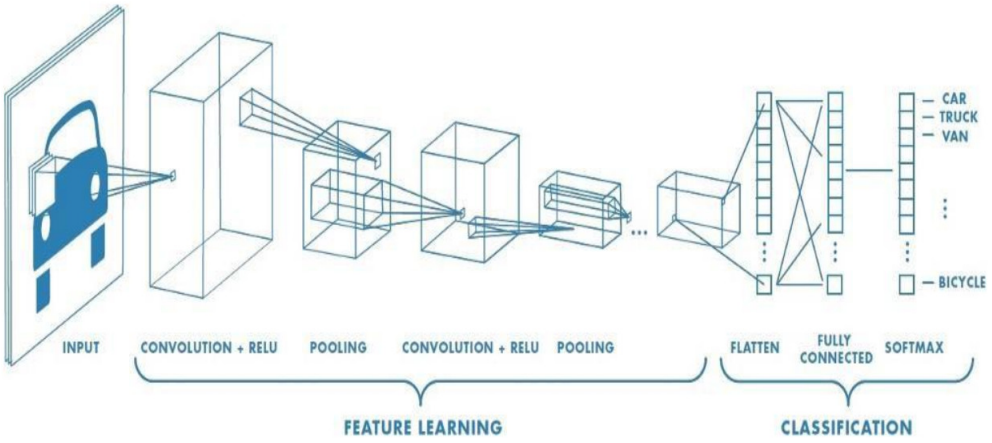


Figure 8: Luồng CNN

trong vùng được chỉ định đó. Bộ lọc sẽ lần lượt được di chuyển theo một giá trị bước trượt và quét trên toàn bộ ảnh. Các thông số của bộ lọc này sẽ được khởi tạo một cách ngẫu nhiên và sẽ được cập nhập thường xuyên trong quá trình huấn luyện cho mạng. Gỉa sử fk là bộ lọc có kích thước nxm được áp dụng trên đầu vào $x.nxm$ là số lượng liên kết đầu vào mà mỗi nơ-ron trong m có. Phép tích chập giữa fk và đầu vào x cho ta kết quả như sau :

$$(x_{u,v}) = \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} f_k(i,j) x_{u+i,v+j} \tag{1}$$

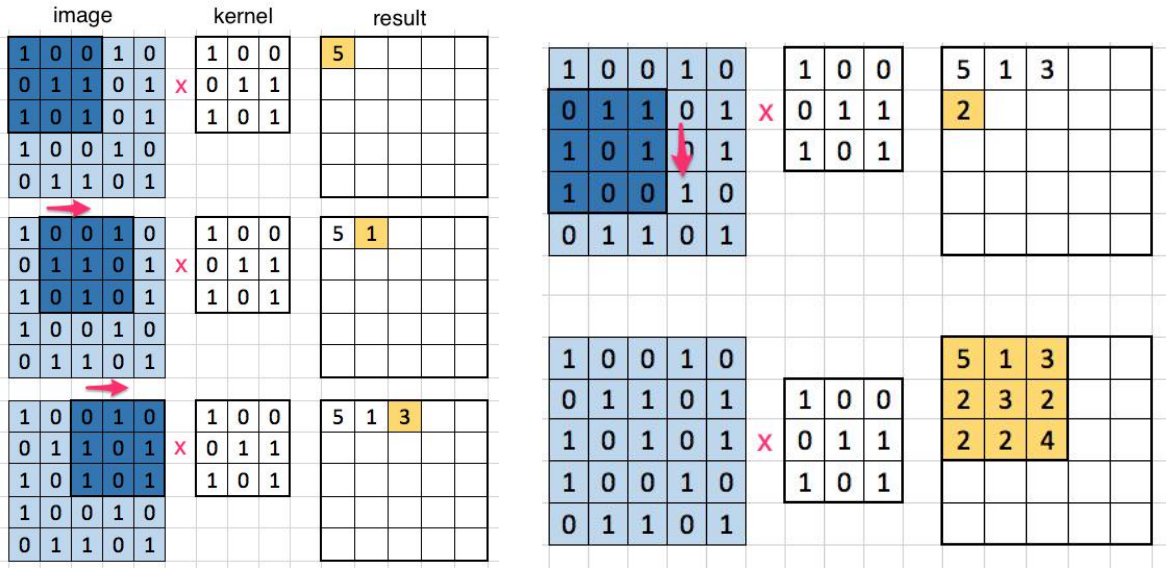
Để có được nhiều đặc trưng đại diện cho dữ liệu đầu vào, ta có thể áp dụng nhiều bộ lọc fk với $k \in N$. Bộ lọc fk được thực hiện bằng cách chia sẻ trọng số của các nơ-ron lân cận. Điều này có ý nghĩa tích cực cho việc cập nhật các trọng số thấp, trái ngược với mạng nơ-ron truyền thẳng, và các trọng số có sự ràng buộc với nhau.

2.5.2 Lớp Pooling Layer

Mục tiêu của lớp pooling là làm giảm kích thước của dữ liệu đầu vào nhưng vẫn giữ các thông tin quan trọng nhất trong đó. Sử dụng một hàm kích hoạt tương ứng với mục đích của người thiết kế đề ra. Các loại lớp lấy mẫu phổ biến bao gồm Max Pooling (lấy giá trị lớn nhất), Average Pooling (lấy giá trị trung bình) và Min Pooling (lấy giá trị nhỏ nhất).

Khác với lớp tích chập, lớp Pooling không tính tích chập mà tiến hành lấy mẫu (subsampling). Maxpooling là phương pháp giảm kích thước mẫu với hàm kích hoạt là Maximumn được áp dụng trên đầu vào x. Gỉa sử m là kích thước của cửa sổ trượt, kết quả thu được khi áp dụng hàm kích hoạt Maximun như sau :

$$M(x_i) = max\{x_{i+k,i+l} \mid |k| \leq \frac{m}{2}, |l| \leq \frac{m}{2}; k,l \in \mathbb{N}\} \tag{2}$$



(a) Tích chập chạy hàng đầu

(b) Tích chập kết quả cuối cùng

Figure 9

Lớp pooling này có tính bất biến đối với kích thước của cửa sổ trượt.

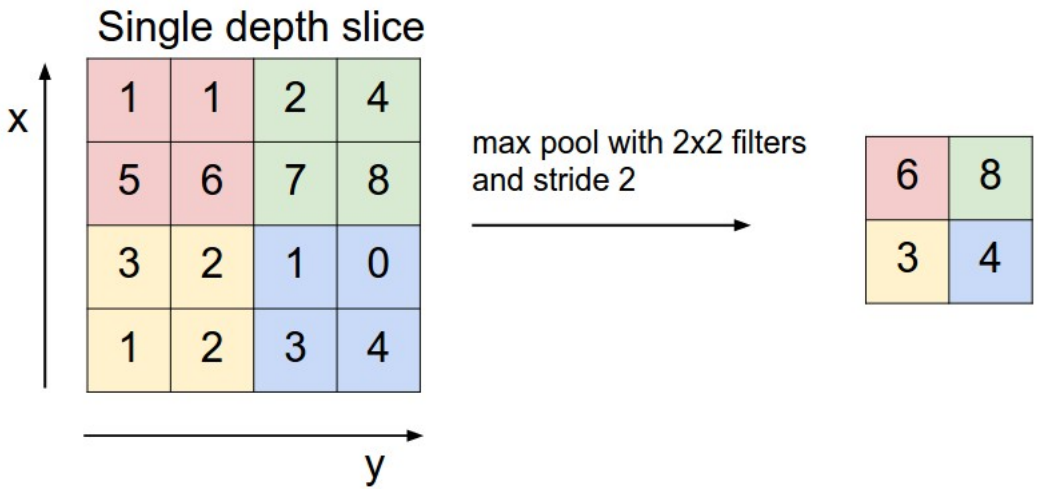


Figure 10: Hoạt động của max-pooling với cửa sổ trượt 2x2

2.5.3 Rectified Linear Unit - ReLU layer

Lớp này có nhiệm vụ chuyển toàn bộ giá trị âm trong kết quả lấy từ lớp tích chập thành giá trị 0 mà vẫn giữ được sự tin cậy toán học của mạng . Ý nghĩa của lớp này chính là tạo nên tính phi tuyến cho mô hình. Ngoài ra, nó còn có tác dụng giảm lượng tính toán cho các lớp tiếp theo, và ngăn chặn việc triệt tiêu sai số gradient vì gradient là một hàm tuyến tính hoặc là 0. Tương tự như trong mạng truyền thẳng, việc xây dựng dựa trên các phép biến đổi tuyến tính sẽ khiến việc xây dựng đa tầng đa lớp trở nên vô nghĩa. Có rất nhiều cách để khiến mô hình trở nên phi tuyến như sử dụng các hàm kích hoạt signmoid, tanh,... nhưng hàm $R(x) = max(0,x)$ dễ tính toán nhanh mà vẫn hiệu quả.

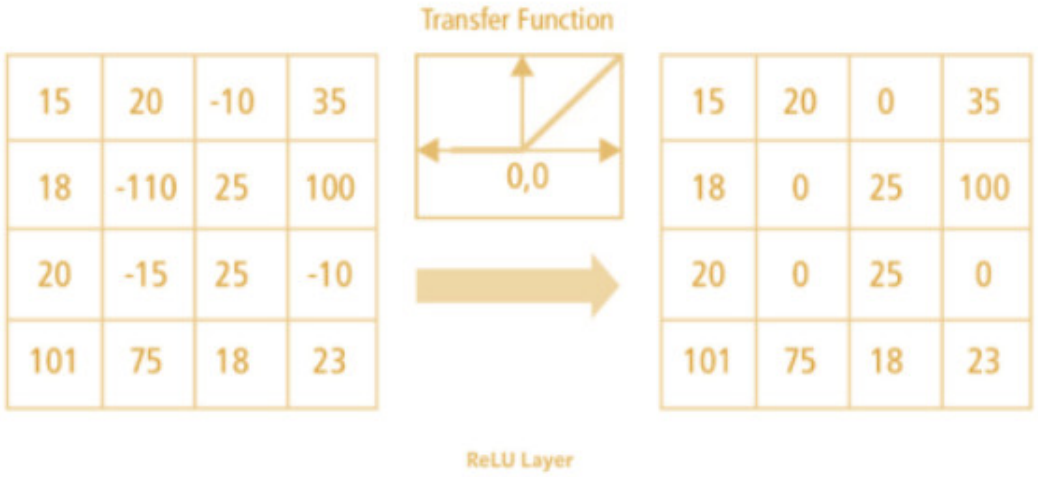


Figure 11: Hoạt động của lớp ReLU

2.5.4 Lớp Fully-Connected Layer

Lớp này được biết đến như là mạng nơ-ron nhiều tầng, các nơ-ron trong mạng kết nối tất cả các nơ-ron của lớp trước. Điều này cho phép các thông tin được truyền từ các vùng

đặc trưng được trích xuất từ các lớp tích chập đến các lớp đầy đủ, giúp mô hình học được các đặc trưng phức tạp và có khả năng phân loại chính xác hơn.

Gỉa sử đầu vào x có kích thước k và l là số lượng nơ-ron có trong lớp kết nối đầy đủ này. Kết quả trong ma trận $w_l xk$:

$$F(x) = \sigma(W * x) \tag{3}$$

Trong đó σ là hàm kích hoạt. Lớp này thường được sử dụng để đưa ra các kết quả.

2.5.5 Output Layer

Lớp output là một vector biểu diễn các lớp được định nghĩa hình ảnh đầu vào. Ở đề tài này, output là một vector bao gồm dữ liệu đại diện cho 7 cảm xúc ở trên khuôn mặt của hình ảnh cần nhận dạng cảm xúc.

$$C(x) = \{i | \exists i \forall j \neq i : x_j \leq x_i\} \tag{4}$$

2.5.6 Drop out

Mạng nơ-ron có nhiều thành công nhưng vẫn tồn tại nhược điểm, bao gồm sự phức tạp và tốn nhiều tài nguyên do sự tồn tại của các lớp ẩn phi tuyến. Quá trình huấn luyện cũng mất nhiều thời gian và sự phù hợp quá mức gây cản trở cho sự phát triển của mạng. Drop-out là một kỹ thuật được sử dụng để loại bỏ ngẫu nhiên một số thành phần và kết nối của chúng ra khỏi mạng trong quá trình huấn luyện nhằm ngăn chặn sự quá khớp của mạng.

Ngoài ra, CNN còn sử dụng một số kỹ thuật như Dropout và Batch Normalization để giảm overfitting và tăng hiệu quả của mô hình. Dropout sẽ loại bỏ một số đơn vị đầu vào hoặc đầu ra của một lớp để giảm thiểu quá trình overfitting. Batch Normalization sẽ chuẩn hóa đầu vào của một lớp để đảm bảo rằng các giá trị đầu vào có phân phối chuẩn và ổn định.

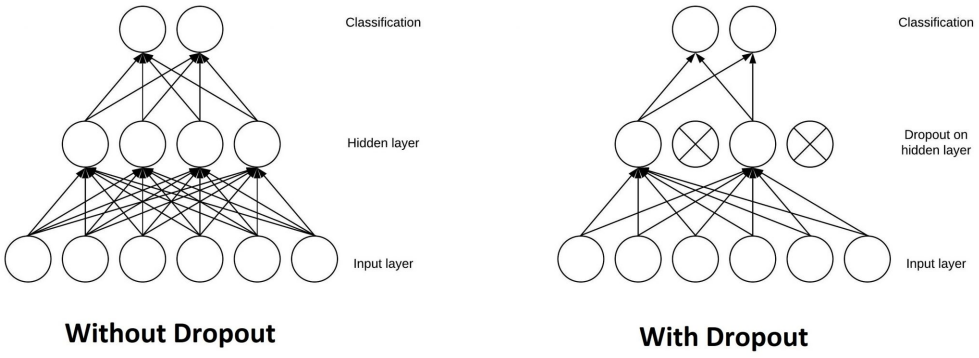


Figure 12: Dropout trong CNN

2.6 Giới thiệu về Hyperparameter tuning (Tối ưu siêu tham số)

Tối ưu siêu tham số (Hyperparameter Tuning) là quá trình điều chỉnh các siêu tham số của mô hình học máy nhằm cải thiện hiệu suất và độ chính xác của mô hình. Siêu tham số là những tham số không được học trực tiếp từ dữ liệu huấn luyện mà cần phải được thiết lập trước khi quá trình huấn luyện bắt đầu. Ví dụ về các siêu tham số bao gồm tỷ lệ học (learning rate) trong các mô hình mạng nơ-ron, số lượng cây trong rừng ngẫu nhiên (random forest), hoặc số lượng cụm trong thuật toán k-means.

Việc lựa chọn các giá trị phù hợp cho các siêu tham số này có thể ảnh hưởng lớn đến hiệu suất của mô hình. Tối ưu siêu tham số giúp chúng ta tìm ra các giá trị tốt nhất cho những tham số này, từ đó cải thiện khả năng dự đoán và tính tổng quát của mô hình.

Có vô số thuật toán điều chỉnh siêu tham số, mặc dù các thuật toán được sử dụng phổ biến nhất là tối ưu hóa Bayes, tìm kiếm lưới và tìm kiếm ngẫu nhiên.

Tối ưu hóa Bayes

Tối ưu hóa Bayes là kỹ thuật dựa trên định lý Bayes, trong đó mô tả xác suất diễn ra một sự kiện tương ứng với kiến thức hiện tại. Khi áp dụng kỹ thuật này để tối ưu hóa siêu tham số, thuật toán sẽ xây dựng một mô hình xác suất từ một bộ siêu tham số, qua đó tối ưu hóa một chỉ số cụ thể. Mô hình này sử dụng phân tích hồi quy để liên tục chọn bộ siêu tham số phù hợp nhất.

Tìm kiếm lưới

Với kỹ thuật tìm kiếm lưới, bạn chỉ định một danh sách siêu tham số và một chỉ số hiệu năng và thuật toán sẽ nghiên cứu tất cả các tổ hợp khả thi để xác định tổ hợp phù hợp nhất. Tìm kiếm lưới có hiệu quả cao nhưng tương đối dài dòng và yêu cầu cao về điện toán, đặc biệt là với số lượng lớn các siêu tham số.

Tìm kiếm ngẫu nhiên

Mặc dù dựa trên các nguyên tắc tương tự như tìm kiếm lưới nhưng tìm kiếm ngẫu nhiên chọn các nhóm siêu tham số một cách ngẫu nhiên trên mỗi lần lặp. Kỹ thuật này có hiệu quả cao khi kết quả của mô hình được quyết định phần lớn bởi số lượng tương đối ít các siêu tham số.

3 Thiết kế và xây dựng mô hình

3.1 Kiến trúc mô hình nhận diện cảm xúc khuôn mặt sử dụng mạng nơ-ron tích chập(CNN)

Kiến trúc mô hình nhận diện cảm xúc khuôn mặt sử dụng mạng nơ-ron tích chập(CNN) như sau :

- Sử dụng activation là ReLU thay cho Sigmoid. Trong đó ReLU là hàm có tốc độ tính toán nhanh nhờ đạo hàm chỉ có 2 giá trị $0, 1$ và không có lũy thừa cơ số e như hàm Sigmoid nhưng vẫn tạo ra được tính phi tuyến.
- Sử dụng dropout layer giúp giảm số lượng liên kết neural và kiểm soát được overfitting.
- Sử dụng các block dạng [Conv2D*n + Max Pooling]
- Xếp nhiều layers CNN + Max Pooling thay vì xen kẽ chỉ một layer CNN + Max Pooling. Các layers CNN sâu hơn có thể trích lọc đặc trưng tốt hơn so với chỉ 1 layer CNN
- Sử dụng các bộ lọc kích thước nhỏ 3x3 thay vì kích thước bộ lọc 5x5 như LenNet. Kích thước bộ lọc nhỏ sẽ giúp giảm số lượng tham số cho mô hình và hiệu quả tính toán hơn.

  Ví dụ : Nếu sử dụng 2 bộ lọc kích thước 3x3 trên một feature map (là output của một layer CNN) có độ sâu là 3 thì ta sẽ cần:

  n_filters x kernel_size x kernel_size x n_channels = 2 x 3 x 3 x 3 = 54 tham số. Nhưng nếu sử dụng 1 bộ lọc có kích thước 5x5 sẽ cần 5 x 5 x 3 = 75 tham số. Hai bộ lọc 3x3 vẫn mang lại hiệu quả hơn so với 1 bộ lọc 5x5.

Mô hình bao gồm 2 thành phần chính : Phần trích xuất đặc trưng và phần phân loại cảm xúc.

Phân tích đặc trưng : bao gồm 5 lớp tích phân chập, 3 lớp pooling và 2 lớp drop out, cụ thể như sau :

- Data : là dữ liệu đầu vào của mô hình. Dữ liệu huấn luyện cho mạng là tập dữ liệu gồm các hình ảnh được định dạng ở mức xám (channel 1) , có kích thước 48x48 pixels.
- Conv2D : là lớp tích chập đầu tiên của mô hình, sử dụng 32 bộ lọc với cùng kích thước 3x3 pixels. Đi cùng với nó là hàm kích hoạt ReLU.

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 48, 48, 32)	320
batch_normalization (BatchNormalization)	(None, 48, 48, 32)	128
conv2d_1 (Conv2D)	(None, 48, 48, 64)	18,496
batch_normalization_1 (BatchNormalization)	(None, 48, 48, 64)	256
max_pooling2d (MaxPooling2D)	(None, 24, 24, 64)	0
dropout (Dropout)	(None, 24, 24, 64)	0
conv2d_2 (Conv2D)	(None, 24, 24, 64)	36,928
batch_normalization_2 (BatchNormalization)	(None, 24, 24, 64)	256
max_pooling2d_1 (MaxPooling2D)	(None, 12, 12, 64)	0
dropout_1 (Dropout)	(None, 12, 12, 64)	0
conv2d_3 (Conv2D)	(None, 12, 12, 128)	73,856
batch_normalization_3 (BatchNormalization)	(None, 12, 12, 128)	512
conv2d_4 (Conv2D)	(None, 12, 12, 128)	147,584
batch_normalization_4 (BatchNormalization)	(None, 12, 12, 128)	512
max_pooling2d_2 (MaxPooling2D)	(None, 6, 6, 128)	0
dropout_2 (Dropout)	(None, 6, 6, 128)	0
flatten (Flatten)	(None, 4608)	0
dense (Dense)	(None, 128)	589,952
dense_1 (Dense)	(None, 7)	903

Figure 13: Cấu trúc của mạng nơ-ron tích phân chập

- Conv2D_1 : là lớp tích chập thứ 2 sử dụng 64 bộ lọc có kích thước 3x3 pixels. Đi cùng với nó là hàm kích hoạt ReLU.
- Tiếp theo sau là lớp pooling với kích thước là 2x2 pixels.
- Drop out được áp dụng trên 3 lớp đầu tiên với tỉ lệ 0.25 , tức là có 25% nơ-ron của 2 lớp này bị tắt trong quá trình huấn luyện.
- Conv2D_2 : là lớp tích chập thứ 3 sử dụng 64 bộ lọc có kích thước 3x3 pixels.
- Tiếp theo sau là lớp pooling thứ 2 với kích thước là 2x2 pixels.
- Drop out được áp dụng trên lớp này với tỉ lệ 0.25,tức là có 25% nơ-ron của 2 lớp này bị tắt trong quá trình huấn luyện. Nhằm hạn chế overfitting.
- Sau đó , 2 lớp tích chập liên tiếp với kích thước 128 bộ lọc 3x3 . Cả 2 lớp tích chập này đều sử dụng hàm ReLU để làm hàm kích hoạt và sau mỗi tích chập là một hàm chuẩn hoá BatchNormalization.
- Tiếp tục dùng lớp pooling thứ 3 ngay sau đó với kích thước là 2x2 pixels.
- Drop out được áp dụng trên lớp này với tỉ lệ 0.25,tức là có 25% nơ-ron của 2 lớp này bị tắt trong quá trình huấn luyện. Nhằm hạn chế overfitting. Đến đây đã kết thúc quá tringh Phân tích đặc trưng.

Tiếp theo là phần phân loại cảm xúc : bao gồm thành phần chính là 2 lớp fully connected với kích thước khác nhau, tương ứng là 128 và 7. Drop out được áp dụng sau lớp fully connected thứ nhất với tỉ lệ 0.25,tức là có 25% nơ-ron của lớp này. Hàm mất mát softmax cross-entropy được sử dụng để phản hồi thông tin trong quá trình huấn luyện mạng.

3.2 Tối ưu siêu tham số của mô hình mạng nơ-ron tích chập

Sau quá trình tìm hiểu, nhóm lựa chọn các tham số sau để đưa vào quá trình tối ưu siêu tham số cùng thuật toán Bayesian Optimization:

- learning rate - lr: tốc độ học của mô hình, với bước nhảy 0.0001 trong khoảng từ 0.0001 đến 0.0004.
- lớp fully connected đầu tiên với kích thước từ 128 cho đến 1024, với cấp số nhân là 2.

Sau quá trình tối ưu siêu tham số với mục tiêu đạt "val_loss" (độ mất mát ở bộ dữ liệu validation) tối thiểu, mô hình cho ra kết quả tốt nhất là mô hình với learning rate 0.0004 và lớp fully connected đầu tiên với 512 lớp.

Sau lần hiệu chỉnh và training đầu tiên, ta thấy rõ sự overfitting của mô hình với biểu đồ hiển thị loss và val_loss qua từng epoch của quá trình huấn luyện.

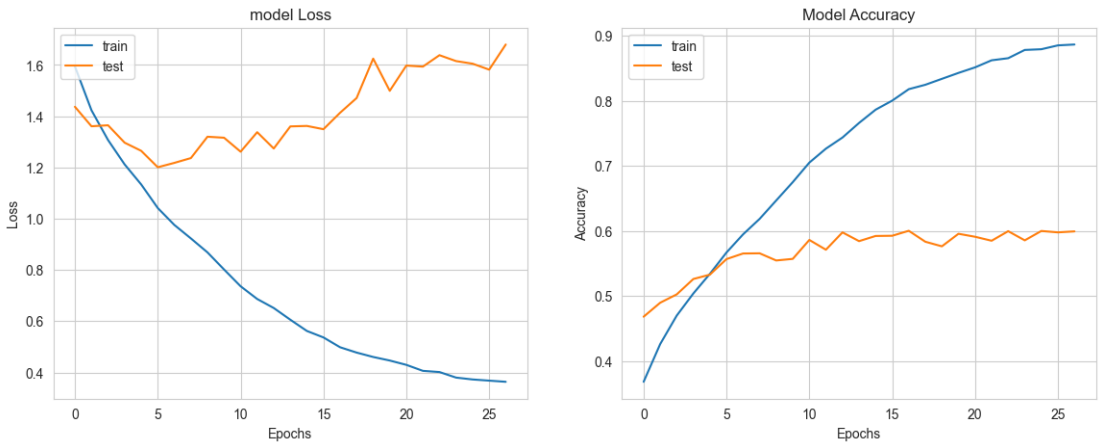


Figure 14: Model loss tăng đột biến

Dễ nhận thấy, đây là dấu hiệu của việc learning rate quá cao, ta sẽ tối ưu mô hình với learning rate giảm xuống còn 0.0001 và thử chạy tối ưu bayes với lớp drop out sau lớp fully connected.

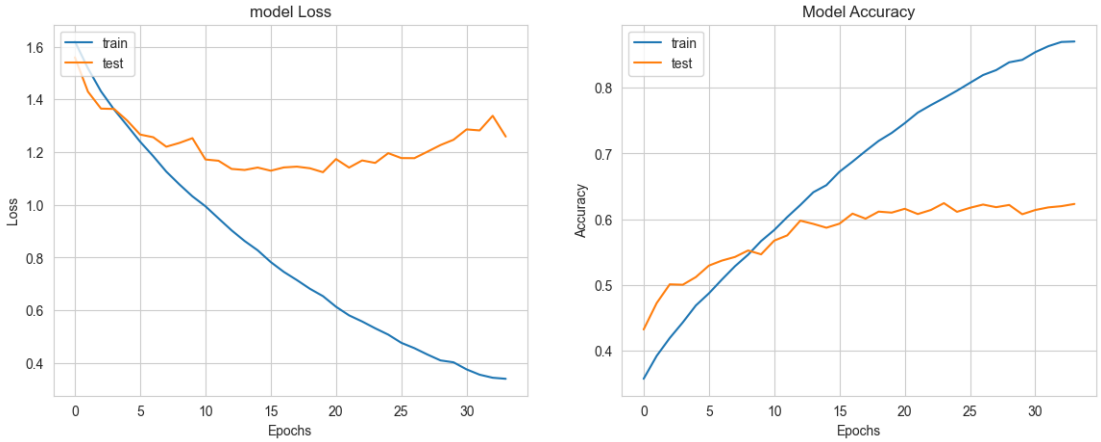


Figure 15: val loss đã có sự ổn định đáng kể

Ngoài ra, một hàm earlystop được gọi là checkpointer được sử dụng để tránh tình trạng overfitting. Hàm checkpointer sẽ dừng quá trình training lại khi mô hình có dấu hiện của overfitting.

Tuy nhiên, sau 2 lần quan sát mô hình train, ta có thể thấy rõ dấu hiệu của việc overfitting trong mô hình. Từ đó, ngoài việc tối ưu tham số sử dụng Bayesian Optimization, mô hình còn có thể cải thiện thêm nếu chúng em cân nhắc thêm về những siêu tham số khác như kích cỡ của các lớp tích chập, lớp Pooling, hay số lớp sử dụng.

Vì đây là một chủ đề còn rất nhiều không gian phát triển, chúng em mong có thể có thêm thời gian để tối ưu mô hình một cách chỉn chu hơn.

4 Thảo luận và đánh giá kết quả

4.1 Kết quả của mô hình CNN

Trong quá trình huấn luyện, các thông số của mạng được cập nhập liên tục sao cho sai số ở đầu ra đạt đến mức nhỏ nhất. Tốc độ học của mạng được đặt là 0.0004. Độ chính xác trong quá trình huấn luyện cũng tăng theo từng chu kì dữ liệu, đạt đến khoảng 88.86% đối với tập huấn luyện, và độ chính xác đối với tập dữ liệu validation đạt 60.95% cho toàn bộ quá trình huấn luyện.

Đánh giá mô hình thông qua bốn chỉ số Accuracy (độ chính xác), Precision, Recall, F1-Score và Confusion matrix.

Kết quả cho thấy giá trị accuracy tăng dần sau mỗi epoch. Gía trị accuracy trên tệp training khá cao xấp xỉ 88%. Tuy nhiên, trên test set lại chỉ đạt được xấp xỉ 61% . Điều này có thể do bộ dữ liệu phân bố không đồng đều.

(a) Untune Model (b) Tuned Model

Figure 16: Confusion Matrix của tập dữ liệu FER-2013

Kết quả cho thấy : Đối với ảnh có các cảm xúc như : Vui, ngạc nhiên, bình thường và giận dữ thì được mô hình phân loại chính xác cao và có các giá trị recall, precisison khá tốt .Tuy nhiên đối với cảm xúc buồn và ghê tởm thì tỉ lệ dự đoán thấp hơn và sợ hãi có tỉ lệ dự đoán thấp nhất.

Sau khi xây dựng mô hình và đưa vào thử nghiệm với các ảnh được chụp trực tiếp thông qua camera của laptop, có thể nhận xét chung là mô hình xử lý tương đối ổn định. Trong đó độ nhận diện cảm xúc trên khuôn mặt cho kết quả khá nhanh và chính xác.

	precision	recall	f1-score	support
0	0.53	0.47	0.50	435
1	0.00	0.00	0.00	41
2	0.44	0.34	0.38	445
3	0.81	0.81	0.81	865
4	0.50	0.54	0.52	625
5	0.78	0.64	0.70	296
6	0.47	0.63	0.54	586
accuracy			0.59	3293
macro avg	0.51	0.49	0.49	3293
weighted avg	0.59	0.59	0.59	3293

(a) Untune Model

	precision	recall	f1-score	support
0	0.48	0.56	0.52	435
1	0.67	0.39	0.49	41
2	0.49	0.37	0.42	445
3	0.83	0.79	0.81	865
4	0.51	0.47	0.49	625
5	0.62	0.78	0.69	296
6	0.52	0.58	0.55	586
accuracy			0.60	3293
macro avg	0.59	0.56	0.57	3293
weighted avg	0.60	0.60	0.60	3293

(b) Tuned Model

Figure 17: Classification Report của tập dữ liệu FER-2013

4.2 So sánh giữa hai mô hình CNN trước và sau khi tối ưu siêu tham số và hiệu chỉnh class weights

So sánh giữa hai mô hình CNN về

1. Độ chính xác (accuracy): Mô hình CNN trước khi tối ưu có chỉ số val_accuracy xấp xỉ 61.29% , test_accuracy xấp xỉ 59.36%. Trong khi đó, mô hình CNN sau khi đã tối ưu siêu tham số và hiệu chính class weights có độ chỉ số val_accuracy và test_accuracy thấp hơn, tương ứng là 62.45% và 59.99%.

2. Precision, recall và f1-score: CNN trước và sau khi tune không có quá nhiều khác biệt. Điểm đáng chú ý nhất chính là Độ chính xác (precision), độ phủ (recall) và F1-score của class "disgust" ở mô hình CNN chưa tối ưu đều rất nhỏ so sánh với con số xấp xỉ 50% của mô hình CNN đã tối ưu và điều chỉnh trọng số class weight

3. Tổng quan :

Tóm lại, mô hình CNN đã hiệu chỉnh có chỉ số tốt hơn so với mô hình CNN chưa hiệu chỉnh trong bài toán phân loại dữ liệu này, dựa trên các metric như độ chính xác, precision, recall và f1-score. Tuy nhiên, để đạt được hiệu suất tốt hơn, mô hình CNN cần được cải thiện và tinh chỉnh các siêu tham số khác ngoài những tham số đã nêu trên để có thể đạt hiệu năng tốt hơn.

Một số đánh giá rút ra sau khi training và thử nghiệm :

- Đối với ảnh có các cảm xúc như : Các cảm xúc vẫn còn chưa thể phân biệt rõ ràng, tuy nhiên các cảm xúc như vui, ngạc nhiên, neutral có thể được nhận diện tốt nhất, kế đến là tức giận, ghê tởm và buồn, thấp nhất là sợ hãi.

- Phát hiện khuôn mặt : Một vài hình ảnh vẫn còn ghi nhận sai vị trí khuôn mặt hoặc không phát hiện được khuôn mặt.

- Sự phức tạp trên cảm xúc khuôn mặt người là quá lớn (gần giống nhau giữa các cảm xúc) nên mô hình chưa thể nhận dạng đúng hoàn toàn.

5 Kết Luận

Sau khi tìm hiểu và thực hiện đề tài " Mô hình mạng nơ-ron tích chập trong nhận diện cảm xúc qua hình ảnh". Nhóm chúng em thực hiện xây dựng mô hình nhận diện cảm xúc dựa trên mô hình CNN, từ đó tối ưu một số siêu tham số và hiệu chỉnh trọng số của các class.

Đây là một lĩnh vực nghiên cứu đầy triển vọng, mang lại nhiều lợi ích cho con người trong nhiều lĩnh vực khác nhau. Việc áp dụng các kỹ thuật và công nghệ như CNN đang giúp cho phân tích cảm xúc qua hình ảnh trở nên chính xác hơn và tiên tiến hơn. Tuy nhiên, việc phân tích cảm xúc qua hình ảnh vẫn còn gặp một số thách thức nhất định, chẳng hạn như sự đa dạng về cảm xúc, độ chính xác của các kỹ thuật phân tích và độ tin cậy của dữ liệu đầu vào. Do đó, các nhà nghiên cứu cần phải tiếp tục nghiên cứu và phát triển các phương pháp và công nghệ mới để giải quyết các thách thức này.

Ưu điểm:

• Bộ dữ liệu FER-2013 đa dạng và đúng với thực tế trong đời sống, các dữ liệu đầu vào của người dung có thể được thu thập lại để cải thiện hiệu năng của mô hình.

• Mô hình hoạt động ổn định, kết quả hiển thị nhanh từ đó có thể làm các chức năng nâng cao hơn như nhận diện cảm xúc trong thời gian thực. Có thể tích hợp nhận diện cảm xúc vào những lĩnh vực khác như tiếp thị, chơi game, ngành dịch vụ, chăm sóc sức khỏe, giáo dục.

• Mô hình CNN cho kết quả nhận diện cảm xúc tương đối tốt, thời gian phản hồi nhanh, gần như thời gian thực.

Nhược điểm: Mặc dù số lượng hình ảnh trong tập dữ liệu FER-2013 rất lớn nhưng với đặc trưng của tập dữ liệu, rất khó để có được mô hình với độ chính xác cao trên tập dữ liệu này. Do đó để nâng cao độ chính xác, cần thay đổi tập dữ liệu (CP+, MMI,..) hoặc kết hợp thêm các bước tiền xử lý

Do thời gian nghiên cứu có hạn, khả năng cũng như kinh nghiệm của chúng em còn ít, nên báo cáo không tránh khỏi những thiếu sót. Báo cáo mới chỉ dừng ở mức nghiên cứu và tổng hợp. Xác xuất sai số trong khi phân tích cảm xúc là khá lớn. Để có thể đưa chương trình thực nghiệm vào áp dụng và phát triển trong thực tế một cách có hiệu quả, chắc chắn phải có thời gian để tiến hành khảo sát chi tiết, cụ thể hơn nữa mới đáp ứng đầy đủ các yêu cầu nghiệp vụ.

Nhưng kết quả nghiên cứu này sẽ là bước khởi đầu rất quan trọng, là nền tảng cơ bản để chúng em tiếp tục nghiên cứu cho những công trình khoa học tiếp theo. Rất mong những ý kiến đóng góp của các thầy, cô và các bạn.

Một lần nữa, nhóm em xin chân thành cảm ơn thầy Lê Hoàng Sơn đã tận tình quan tâm, giúp đỡ và hướng dẫn chúng em hoàn thành báo cáo này.

6 Tài liệu tham khảo

[1]. H. T. T. Nguyễn and N. T. P. Nguyễn, "Nhân dạng cảm xúc trong video sử dụng Mạng Nơ Ron Tích Chập," Khoa học Công Nghệ, vol. 181, no. 5, pp. 211-216, 2018.

[2]. Ole Helvig Jensen, “Implementing the Viola-Jones Face Detection Algorithm”, Technical University of Denmark, 2008.

[3]. Keiron Teilo O’Shea và Ryan Nash, “An Introduction to Convolutional Neural Networks”, ResearchGrate, 2015.

[4]. H.-C. Chu, W. W.-J. Tsai, M.-J. Liao and Y.-M. Chen, "Facial emotion recognition with transition detection for students with high-functioning autism in adaptive e-learning," Soft computing (Berlin, Germany), vol. 22, no. 9, pp.2973- 2999, 2017.

[5]. Lopes, A.T., de Aguiar, E., De Souza, A.F., Oliveira-Santos, T.: “Facial expression recognition with Convolutional Neural Networks: coping with few data and the training sample order”. Pattern Recogn. 61, 610–628, 2017.

[6]. Đinh Xuân Nhất, “Nghiên cứu các thuật toán nhân dạng cảm xúc trên khuôn mặt 2D”, Đại học Công nghệ - Đại học Quốc gia Hà Nội, 2010.

[7]. Le Thi-Lan và Dong Van-Thai, “Toward a Vietnamese facial expression recognition system for human-robot interaction”, Tại Hội nghị Quốc tế về Công nghệ tiên tiến áp dụng cho Truyền thông, 2011

[8]. Zafar, Sahar Ali, Fayyaz Guriro, Subhash Ali, Irfan Khan, Asif Zaidi, Adnan. (2019). Facial Expression Recognition with Histogram of Oriented Gradients using CNN. Indian Journal of Science and Technology. 12. 10.17485/ijst/2019/v12i24/145093.