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

BÁO CÁO CUỐI KÌ

PHƯƠNG PHÁP NGHIÊN CỨU KHOA HỌC

Đề tài:

ỨNG DỤNG CNN NHẬN DIỆN CẢM XÚC TRONG THỜI GIAN THỰC

Giảng viên	TS. NGUYỄN THỊ MINH HUYỀN	
Giảng viên hướng dẫn	TS. PHẠM HUY THÔNG	
Sinh viên thực hiện	VŨ MẠNH ĐỨC	MSV:20002046

Hà Nội, 05-2023

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 1 mô hình CNN phân tích cảm xúc từ hình ảnh . Phần còn lại của bài báo được tổ chức như sau:

Chương I: Tổng quan về phân tích cảm xúc.

Chương II: Cơ sở lý thuyết

Chương III: Phương pháp nghiên cứu.

Chương IV: Kết quả nghiên cứu

Chương V: Thảo luận và đánh giá kết quả.

Chương VI: Kết luận

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.

List of Figures

Contents

1 Tổng quan về phân tích cảm xúc

1.1 Đặt vấn đề

Trong cuộc sống hàng ngày, phân tích cảm xúc của con người là một kỹ năng quan trọng trong giao tiếp và tương tác xã hội. Tuy nhiên, việc phân tích cảm xúc của con người đôi khi không phải là điều dễ dàng, đặc biệt là khi chúng ta phải xử lý một lượng lớn thông tin và dữ liệu. Trong lĩnh vực trí tuệ nhân tạo, phân tích cảm xúc qua hình ảnh trở thành một lĩnh vực nghiên cứu được quan tâm rất nhiều trong thời gian gần đây.

Bộ dữ liệu FER-2013 là một trong những bộ dữ liệu phổ biến nhất được sử dụng để huấn luyện các mô hình máy học và trí tuệ nhân tạo để phân tích cảm xúc qua hình ảnh. Tuy nhiên, việc phân tích cảm xúc qua hình ảnh vẫn còn nhiều thách thức. Một trong những thách thức đó là khả năng nhận dạng chính xác các cảm xúc khác nhau trên khuôn mặt của con người.

Vì vậy, vấn đề là làm thế nào để sử dụng mạng neural tích chập (CNN) để phân tích cảm xúc qua hình ảnh trên bộ dữ liệu FER-2013 một cách chính xác và hiệu quả. Mục tiêu của đề tài là xây dựng một mô hình CNN để phân tích cảm xúc qua hình ảnh trên bộ dữ liệu FER-2013 và đánh giá hiệu suất của phương pháp này. Việc phân tích cảm xúc qua hình ảnh có thể ứng dụng trong nhiều lĩnh vực khác nhau, chẳng hạn như trong lĩnh vực giải trí, truyền thông, giáo dục, y tế, và nhiều lĩnh vực khác.

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.

1.3 Tổng quan về cảm xúc trên khuôn mặt con người

1.3.1 Đặc trưng khuôn mặt của con người

Khuôn mặt là một trọng tâm chính trong mối quan hệ giao tiếp trong xã hội, đóng vai trò quan trọng trong việc truyền tải bản sắc riêng và cảm xúc của con người. Chúng ta có thể nhận ra khuôn mặt của rất nhiều người trong suốt cuộc đời, và việc nhận diện khuôn mặt quen thuộc chỉ trong nháy mắt thậm chí sau nhiều năm không gặp mặt. Điều này khá là rõ nét, bất chấp những thay đổi lớn về thị giác, biểu hiện, lão hóa, hoặc những thay đổi về kiểu tóc, về kính,... Ngoài ra, ảnh khuôn mặt trong thực tế còn chứa đựng rất nhiều vấn đề như: độ sáng, độ nhòe, mờ , độ nhiễu, độ phân giải, góc ảnh, tầm này...

Khuôn mặt con người có tổng cộng 43 cơ được chi phối bởi các dây thần kinh mặt và được chia thành 5 nhóm cơ chính bao gồm: : Cơ trên sọ, các cơ quanh tai, các cơ quanh ổ mắt và mí, các cơ mũi và các cơ quanh miệng.

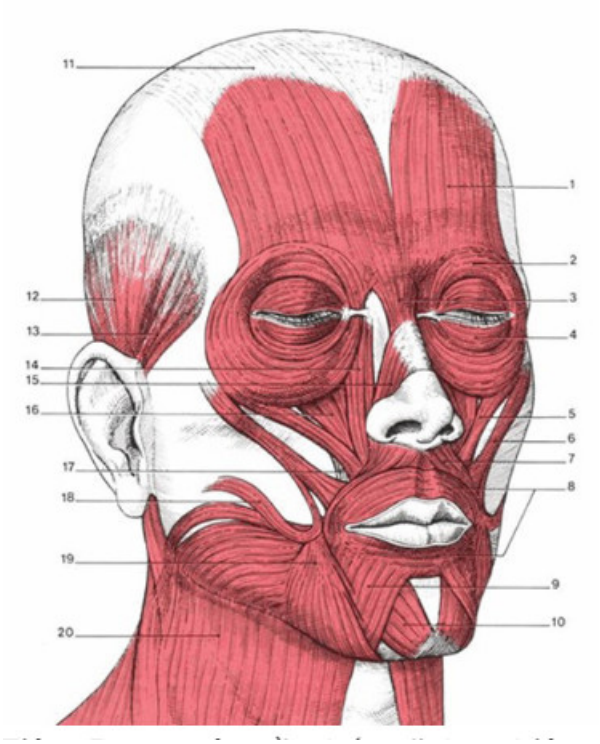

Figure 1: Các cơ trên khuôn mặt

1.3.2 Các loại cảm xúc cơ bản

Con người truyền đạt thông tin với nhau qua nhiều cách : ngôn ngữ, ngôn thể, cử chỉ, lời nói,.. Biểu hiện cảm xúc trên khuôn mặt cũng là một cách để truyền đạt thông tin một cách hiệu quả, nó có thể biểu hiện một nhận định của con người đối với những sự vật, hiện tượng xảy ra trước mặt họ. Biểu cảm trên khuôn mặt là những cử động tự nguyện và không chủ ý, xảy ra khi một hoặc nhiều cơ trên khuôn mặt được hoạt động.

Chúng là một nguồn giao tiếp phi ngôn ngữ phong phú, hiển thị một lượng lớn thông tin về cảm xúc và nhận thức của con người.

Hiện nay trên thế giới cũng ủng hộ rằng có tất cả 7 loại cảm xúc cơ bản bao gồm : Hạnh phúc (happiness), Buồn bã(sadness), Sợ hãi (fear), Ghê tởm (disgust), Giận dữ (anger), Khinh thường (contempt) và Ngạc nhiên (surprise). Mỗi loại đều có nét độc đáo, đặc tính và cách thể hiện riêng của nó.

Vào năm 1872, Charles Darwin là người đầu tiên cho rằng cảm xúc có tính phổ quát, ý tưởng của ông về cảm xúc là trọng tâm trong lý thuyết tiến hóa của ông, cho thấy rằng cảm xúc và các biểu hiện của chúng là bẩm sinh và thích nghi về mặt tiến hóa, và những điểm tương đồng trong chúng ta có thể được nhìn thấy về mặt phát sinh loài. Tuy nhiên, nghiên cứu ban đầu đã thử nghiệm những ý tưởng của Darwin không kết luận được. Những tuyên bố của Darwin đã được Tomkins nghiên cứu tiếp (1962, 1963), ông ấy cho rằng cảm xúc là cơ sở của động lực con người và khuôn mặt là nơi quan trong nhất để thể hiện các cảm xúc. Cuối thế kỷ XX khi Tiến sĩ Paul Ekman và nhóm của ông thực hiện nghiên cứu về tính phổ biến của các biểu hiện cảm xúc trên khuôn mặt, chúng ta mới bắt đầu thấy bằng chứng đáng kể rằng lý thuyết của Charles Darwin là đúng.

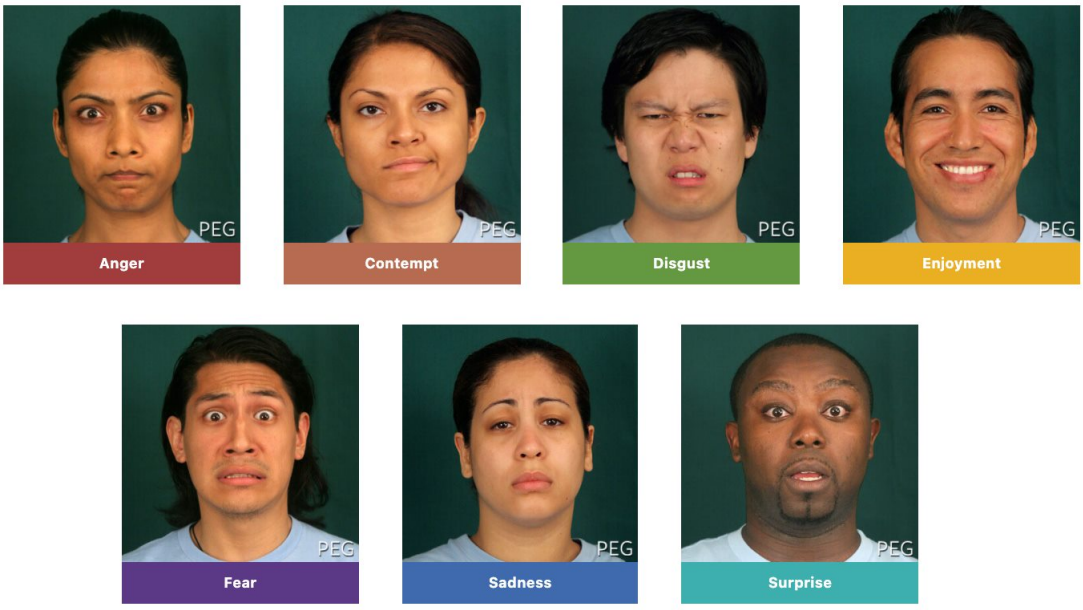

Figure 2: 7 Loại cảm xúc

1.4 Phân tích cảm xúc qua hình ảnh là gì?

Phân tích cảm xúc qua hình ảnh (Image Emotion Analysis) là quá trình sử dụng các công cụ và kỹ thuật xử lý hình ảnh để phân tích và nhận diện cảm xúc được truyền tải qua hình ảnh. Các cảm xúc phổ biến có thể được phân tích bao gồm vui vẻ, buồn bã, sợ hãi, tức giận, bất ngờ và trầm trồ.

Phân tích cảm xúc qua hình ảnh có nhiều ứng dụng trong cuộc sống thực, chẳng hạn như trong lĩnh vực giải trí, quảng cáo, marketing, tâm lý học, y tế và an ninh. Ví dụ, phân tích cảm xúc qua hình ảnh có thể giúp các nhà quảng cáo đánh giá hiệu quả của chiến dịch quảng cáo, phân tích hành vi khách hàng và cải thiện trải nghiệm người dùng. Ngoài ra, phân tích cảm xúc qua hình ảnh còn được sử dụng để giúp các chuyên gia tâm lý học nhận biết và điều trị các rối loạn tâm lý, bệnh trầm cảm và rối loạn lo âu.

Trong các ứng dụng thực tế, phân tích cảm xúc qua hình ảnh thường được thực hiện bằng cách sử dụng các kỹ thuật trích xuất đặc trưng từ hình ảnh, sau đó áp dụng các thuật toán máy học để phân loại các hình ảnh thành các cảm xúc khác nhau. Một số phương pháp phổ biến bao gồm sử dụng Convolutional Neural Networks (CNN) và Deep Learning

1.5 Tầm quan trọng của phân tích cảm xúc qua hình ảnh

Phân tích cảm xúc qua hình ảnh là một lĩnh vực nghiên cứu có tầm quan trọng rất lớn trong nhiều lĩnh vực khác nhau. Dưới đây là một số ví dụ về tầm quan trọng của phân tích cảm xúc qua hình ảnh:

1. Quảng cáo và Marketing: Phân tích cảm xúc qua hình ảnh giúp các nhà quảng cáo đánh giá hiệu quả của chiến dịch quảng cáo, cải thiện trải nghiệm người dùng và thu hút khách hàng. Điều này cũng giúp cho các doanh nghiệp tìm ra cách tiếp cận khách hàng một cách tốt nhất.

2. Giải trí: Phân tích cảm xúc qua hình ảnh có thể được sử dụng để đánh giá phản hồi của khán giả đối với các bộ phim, chương trình truyền hình và game. Điều này giúp các nhà sản xuất có thể cải thiện sản phẩm của họ và đưa ra trải nghiệm tốt hơn cho người dùng.

3. Y tế: Phân tích cảm xúc qua hình ảnh có thể được sử dụng để giúp các chuyên gia tâm lý học nhận biết và điều trị các rối loạn tâm lý, bệnh trầm cảm và rối loạn lo âu. Điều này giúp cải thiện chất lượng cuộc sống của các bệnh nhân và giúp họ phục hồi nhanh chóng hơn.

4. An ninh: Phân tích cảm xúc qua hình ảnh có thể được sử dụng để giúp các nhà lãnh đạo đánh giá mức độ đe dọa của các sự kiện, đặc biệt là trong lĩnh vực an ninh. Điều này giúp các nhà lãnh đạo có thể đưa ra các quyết định nhanh chóng và đúng đắn

để bảo vệ an ninh quốc gia.

5. Tâm lý học : Phân tích cảm xúc qua hình ảnh có thể giúp các nhà nghiên cứu tâm lý học nghiên cứu và hiểu sâu hơn về cảm xúc con người, cải thiện việc chẩn đoán và điều trị các rối loạn tâm lý.

1.6 Các phương pháp phân tích cảm xúc qua hình ảnh hiện nay

Hiện nay, có nhiều phương pháp và kỹ thuật được sử dụng để phân tích cảm xúc qua hình ảnh. Sau đây là một số phương pháp phổ biến:

1. Phân tích dựa trên đặc trưng : Phương pháp này sử dụng các phương pháp trích xuất đặc trưng từ hình ảnh, sau đó áp dụng các thuật toán máy học để phân loại các hình ảnh thành các cảm xúc khác nhau. Các phương pháp trích xuất đặc trưng phổ biến bao gồm Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG) và Scale-Invariant Feature Transform (SIFT).

2. Phân tích dựa trên Deep Learning : Deep Learning là một phương pháp học sâu có khả năng học các đặc trưng cấp cao của hình ảnh thông qua việc thực hiện các phép tích chập trên hình ảnh đầu vào. Các mô hình Deep Learning phổ biến được sử dụng trong phân tích cảm xúc qua hình ảnh bao gồm Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) và Deep Belief Networks (DBN).

3. Phân tích dựa trên phương pháp kết hợp: Phương pháp này sử dụng sự kết hợp giữa các phương pháp trích xuất đặc trưng và Deep Learning để cải thiện độ chính xác và hiệu quả của phân tích cảm xúc qua hình ảnh.

4. Phân tích dựa trên mô hình tâm lý học : Phương pháp này sử dụng các mô hình tâm lý học để đánh giá cảm xúc của con người dựa trên các đặc trưng như ánh mắt, môi, và biểu cảm khuôn mặt. Các phương pháp này thường được sử dụng trong các ứng dụng như y tế và tâm lý học.

Tuy nhiên, mỗi phương pháp đều có những ưu điểm và hạn chế riêng. Do đó, việc lựa chọn phương pháp phù hợp sẽ phụ thuộc vào mục đích và yêu cầu của ứng dụng cụ thể.

1.7 Các công cụ phát triển

Xây dựng mô hình bằng ngôn ngữ Python và các thư viện hỗ trợ

1.7.1 Ngôn ngữ Python

Python mang bản chất là ngôn ngữ lập trình bậc cao, được tạo ra bởi Guido van Rossum (2/1991). Ngôn ngữ Python được thiết kế hướng tới đối tượng với cấu trúc hàng và cách xử lý dữ liệu đơn giản, dễ đọc , dễ tiếp cận. Nó sẽ giúp người dùng tạo ra những chương trình hay với số lượng dòng code ít nhất, đơn giản và dễ hiểu.

Các cuộc khảo sát cho thấy Python hiện là một trong những ngôn ngữ lập trình hàng đầu chỉ đứng sau C và Java. Nó cho phép các nhà phát triển xây dựng các hệ thống phụ trợ mạnh mẽ cho các dự án Python AI. Ngôn ngữ lập trình Python có nhiều lợi ích đối với việc phát triển Máy học và AI

1.7.2 Các thư viện hỗ trợ

Các thư viện được sử dụng trong quá trình xây dựng model nhận diện cảm xúc khuôn mặt bao gồm:

- **TensorFlow:** là thư viện mã nguồn mở cho Machine Learning nổi tiếng nhất thế giới, được phát triển bởi các nhà nghiên cứu từ Google. Việc hỗ trợ mạnh mẽ các phéo toán học để tính toán trong Machine Learning và Deep Learning đã giúp việc tiếp cận các bài toán trở nên đơn giản, nhanh chóng và tiện lợi hơn nhiều.
- **Keras:** được coi là một thư viện 'high-level' với phần 'low-level' (còn được gọi là backend) có thể là TensorFlow, CNTK, hoặc Theano. Keras có cú pháp đơn giản hơn TensorFlow rất nhiều.
- **Scikit-learn (Sklearn) :** là thư viện mạnh mẽ nhất dành cho các thuật toán học máy được viết trên ngôn ngữ Python. Thư viện cung cấp một tập các công cụ xử lý các bài toán Machine Learning và Statistical Modeling gồm : Classification, Regression, Clustering và Dimensionality Reduction.

2 Cơ sở lý thuyết

2.1 Mô hình học máy trong nhận dạng cảm xúc

2.1.1 Sử dụng học sâu (Deep Learning) trong nhận dạng cảm xúc

Học sâu là một nhánh máy học, dựa trên hoạt động của bộ não con người trong việc xử lý dữ liệu và tạo ra các mẫu để sử dụng cho việc đưa ra các quyết định. Học sâu là một nhánh của học máy trong AI, có các mạng lưới có khả năng **tự học** mà không bị giám sát từ dữ liệu không có cấu trúc hoặc không được gắn nhãn. Trong học sâu, máy tính sử dụng các lớp khác nhau để học hỏi từ dữ liệu. Học sâu có các mạng lưới có khả năng học không giám sát từ dữ liệu không có cấu trúc hoặc không được gắn nhãn.

Mặc dù học sâu có thể học tốt trên nhiều dữ liệu, nhưng có nhiều vấn đề trong đó không có đủ dữ liệu có sẵn để học sâu trở nên hữu ích. Trong học sâu, chúng ta có thể học nhóm hoặc sắp xếp dữ liệu không được gắn nhãn theo sự tương đồng giữa các mẫu trong dữ liệu này.

Các thuật toán học sâu đang dần được ứng dụng vào các bài toán thực tế trong nhiều lĩnh vực khác nhau. Các thuật toán học sâu phổ biến nhất là : Convolutional Neural Network (CNN), Recurrent neural network (RNN), Deep Boltzmann Machinea (DBM), Deep Belief Networks (DBN).

2.1.2 Các phương pháp nhận diện cảm xúc

1. Phương pháp nhận diện cảm xúc truyền thống

Hệ thống nhận diện cảm xúc qua khuôn mặt bằng phương pháp truyền thống sẽ xử lý bài toán với các giai đoạn : tiền xử lý hình ảnh khuôn mặt, trích chọn ra đặc trưng của hình ảnh và tiến hành phân loại.

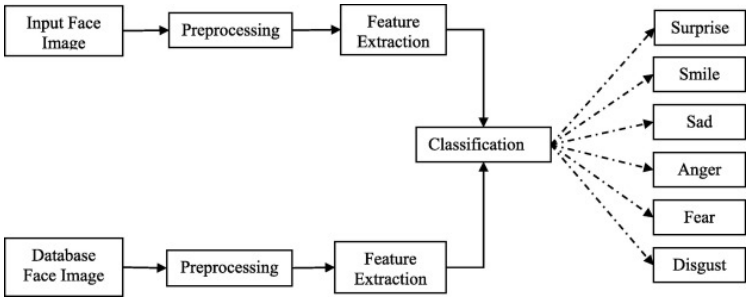

Figure 3: Kiến trúc hệ thống nhận dạng cảm xúc khuôn mặt phương pháp truyền thống

- **Tiền xử lý dữ liệu (Preprocessing) :** Là quá trình cải thiện hiệu suất của hệ thống nhận dạng cảm xúc qua khuôn mặt và được thực hiện các loại quy trình khác nhau

: căn chỉnh độ nét, điều chỉnh tỷ lệ hình ảnh, điều chỉnh kích thước hình ảnh, điều chỉnh độ tương phản, màu sắc của hình ảnh và sử dụng các quy trình để cải thiện chất lượng của dữ liệu.

- **Trích chọn đặc trưng (Feature Extraction) :** là một giai đoạn rất quan trọng, quá trình này giúp trích chọn ra những đặc trưng riêng nhất của hình ảnh, sau đó những đặc trưng của dữ liệu này có thể được sử dụng làm đầu vào cho bài toán phân loại.
- **Phân loại (Classification) :** đây là giai đoạn cuối cùng của hệ thống nhận diện cảm xúc qua khuôn mặt(FER) , để phân loại các loại cảm xúc trên khuôn mặt bao gồm : Hạnh phúc, buồn bã, bất ngờ, tức giận, sợ hãi, ghê tởm và bình thường. Sử dụng các phương pháp để phân loại như : Decision Tree (ID3), SVM (Support Vector Machine), HMM (Hidden Markov Model)... Trong đó, phương pháp SVM cho độ chính xác và có kết quả phân loại tốt nhất.

2. Phương pháp nhận diện cảm xúc hiện đại

Hệ thống nhận diện cảm xúc qua khuôn mặt thực hiện các giai đoạn sau : tiền xử lý hình ảnh khuôn mặt, phân lớp dữ liệu sử dụng học sâu. Các năm gần đây , học sâu đã mang lại độ chính xác, ổn định hơn phương pháp truyền thống vì nó không phải thông qua bước trích xuất các đặc trưng một cách tường minh mà thay vào đấy thực hiện đi kèm với phương pháp phân loại.

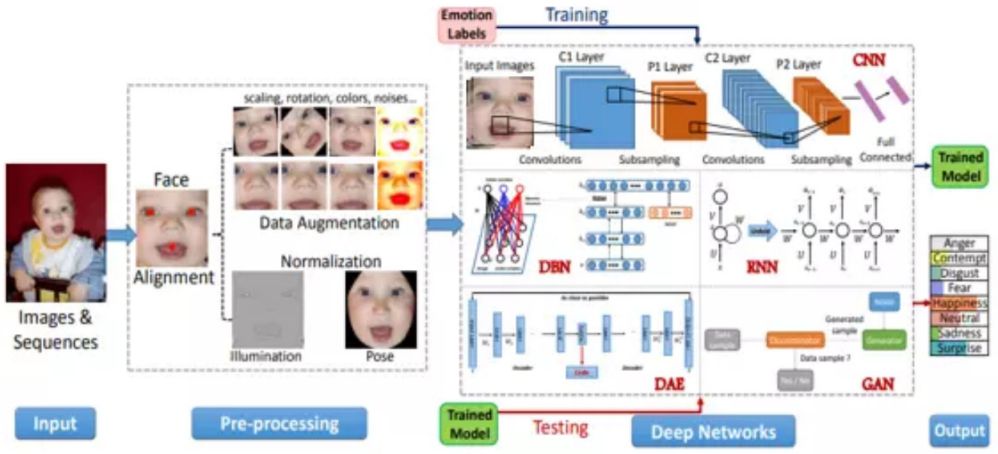

Figure 4: Kiến trúc hệ thống nhận dạng cảm xúc khuôn mặt của phương pháp học sâu

- **Tiền xử lý dữ liệu (Preprocessing) :** Xử lý một số vấn đề của ảnh đầu vào hệ thống, xử lý trước quá trình training. Các bước thực hiện : Căn chỉnh khuôn mặt để phát hiện khuôn mặt, tăng dữ liệu hình ảnh đảm bảo có đủ dữ liệu training, cuối cùng là chuẩn hóa dữ liệu khuôn mặt, Sử dụng các phương pháp như CNN, DBN, DAE, RNN, GAN,...

- **Phân loại (Classificatio) :** Trong phương pháp truyền thống , bước trích chọn đặc trưng và bước phân loại là 2 bước độc lập với nhau nhưng trong học sâu có thể thực hiện FER theo cách từ đầu đến cuối. Một lớp mất được thêm vào cuối mạng để điều chỉnh lỗi lan truyền ngược, sau đó xác suất dự đoán từng mầu có thể được mạng xuất ra trực tiếp.

2.2 Nơ-ron nhân tạo

2.2.1 Lịch sử của nơ-ron nhân tạo

Lịch sử của nơ-ron nhân tạo (tiếng Anh là artificial neural network - ANN hay neural network) bắt đầu từ những năm 1940, khi các nhà khoa học như Warren McCulloch và Walter Pitts đề xuất mô hình nơ-ron đầu tiên để mô phỏng hoạt động của hệ thần kinh trung ương trong não người. Mô hình này mô tả một nơ-ron nhân tạo bao gồm đầu vào, trọng số và hàm kích hoạt.

Tuy nhiên, phát triển mạng nơ-ron đã gặp trở ngại vào những năm 1990, khi các phương pháp khác như máy học dựa trên quy tắc trở nên phổ biến hơn.

Vào những năm 2000, mạng nơ-ron bắt đầu trở lại sự chú ý với sự phát triển của các thuật toán học sâu (deep learning) và khả năng tính toán mạnh mẽ hơn của máy tính. Ngày nay, nơ-ron nhân tạo đang được sử dụng rộng rãi trong các lĩnh vực như nhận dạng hình ảnh, ngôn ngữ tự nhiên, và xe tự lái. Trong quá trình phát triển, mạng nơ-ron đã đạt được những thành tựu đáng kể, bao gồm các kiến trúc mạng nơ-ron sâu, các thuật toán học sâu, và các ứng dụng thực tế đang được triển khai trên toàn thế giới.

Vào năm 1989, Yann LeCun đã áp dụng một phương pháp gọi là "convolutional neural network" (CNN) cho mạng nơ-ron. CNN là một loại mạng nơ-ron nhân tạo được thiết kế đặc biệt để xử lý dữ liệu đa chiều như hình ảnh và âm thanh. Ông đã thiết kế một kiến trúc mạng gọi là LeNet, với kiến trúc được tổ chức thành các lớp liên kết với nhau. Mỗi lớp có thể chứa một hoặc nhiều bộ lọc (filters) để trích xuất các đặc trưng từ dữ liệu đầu vào.

2.2.2 Cấu tạo và quá trình xử lý thông tin của một nơ ron nhân tạo

Các nhà nghiên cứu đã tìm cách chuyển đổi những hiểu biết về cách thức hoạt động của các tế bão thần kinh sinh học thành các mô hình mạng nơ-ron nhân tạo (Artificial Neural Network) có thể hoạt động được trên máy tính. Hình bên dưới cho thấy mô hình của một nơ-ron đơn lẻ, được xem như đơn vị xử lý thông tin cơ bản của một mạng nơ-ron. Các nơ-ron này được sử dụng để xây dựng thành các mạng nơ-ron có kiến trúc phức tạp hơn sẽ được trình bày trong các phần sau. Chúng ta sẽ xem xét 3 thành phần cơ bản của một mạng nơ-ron:

- Một tập hợp các khớp thần kinh (synapse) hoặc còn được gọi là connecting link

dùng để kết nối các nơ-ron lại với nhau. Mỗi khớp thần kinh được đặc trưng bởi cường độ liên kết của nó. Cụ thể hơn, một khớp thần kinh dùng để chuyển tín hiệu từ nơ-ron có nhãn j sang nơ-ron có nhãn k với trọng số là w_{kj}. Không giống như trọng số của một khớp thần kinh trong hệ thần kinh sinh học, trọng số của một khớp thần kinh nhân tạo có thể mang giá trị âm hoặc dương.

- Một bộ cộng (adder) dùng để tổng hợp các tín hiệu đầu vào tại mỗi nơ-ron và gửi kết quả đi tiếp.
- Một hàm kích hoạt (activation function) dùng để đưa các tín hiệu đầu ra của nơ-ron vào một miền giá trị nhất định hoặc vào một tập hợp các giá trị cố định.

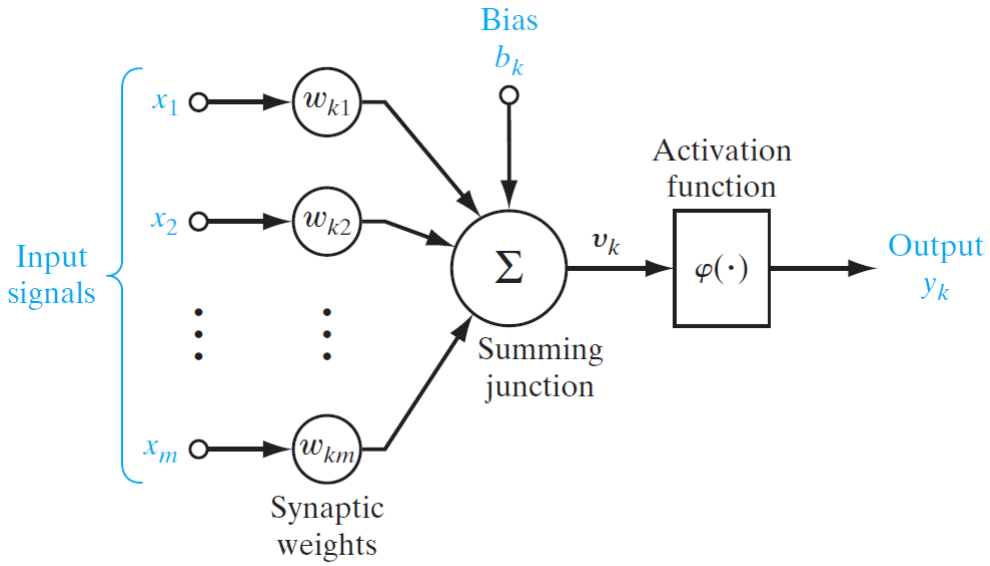

Figure 5: Mô hình của một nơ-ron nhân tạo được gán nhãn k

Chúng ta có thể mô tả hoạt động của nơ-ron có nhãn k trong hình trên bằng các phương trình toán học như sau:

$$u(k) = \sum_m j = 1(w_k x_j)$$

$$v_k = u_k + b_k$$

$$y_k = (v_k)$$

2.3 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. Các phép tích chập này sẽ tìm kiếm các đặc trưng nhất định trên hình ảnh, ví dụ như các đường viền, góc cạnh, vùng sáng và vùng tối. Sau đó, các đặc trưng được học sẽ được sử dụng để phân loại và nhận diện các hình ảnh.

CNN thường được thiết kế với 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. Hình 2.4 dưới đây là toàn bộ luồng CNN cơ bản để xử lý hình ảnh đầu vào và phân loại các đối tượng dựa trên giá trị của đối tượng đó.

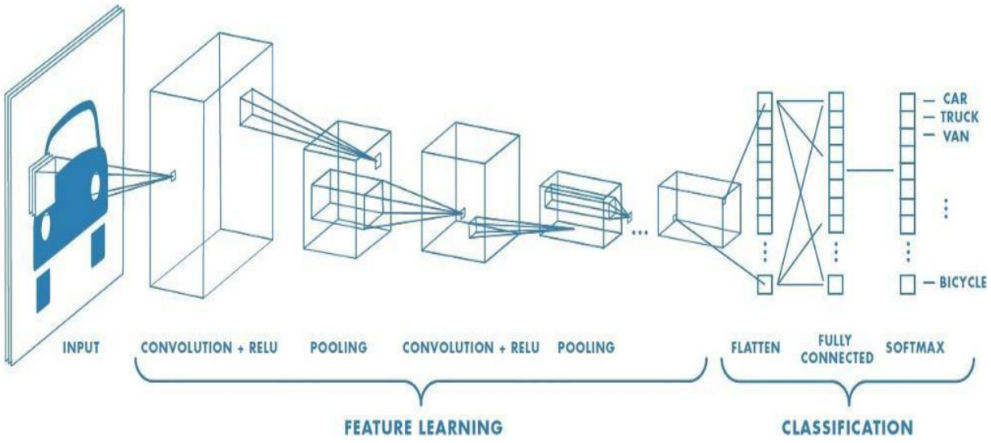

Figure 6: Luồng CNN

Với sự phát triển của công nghệ và sự ra đời của các kiến trúc CNN tiên tiến, CNN đã trở thành một công cụ mạnh mẽ trong việc xử lý hình ảnh và video trong nhiều lĩnh vực, chẳng hạn như y tế, an ninh, giải trí và kỹ thuật số hóa.

2.4 Các thành phần cơ bản của CNN

CNN có ba thành phần cơ bản:

2.4.1 *Lớp Convolutional Layer*

Lớp này sẽ thực hiện các phép tích chập trên hình ảnh đầu vào để tìm kiếm các đặc trưng như đường viền, góc cạnh, vùng sáng và vùng tối. 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 trong vùng được chỉ định đó. Bộ lọc sẽ lần lượt 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 :

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

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.

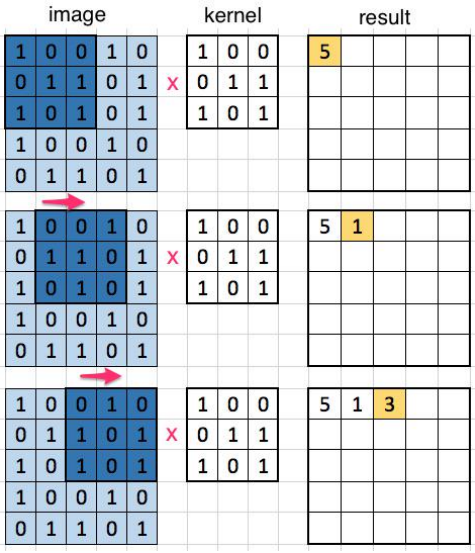

Figure 7: Tích chập chạy hàng đầu

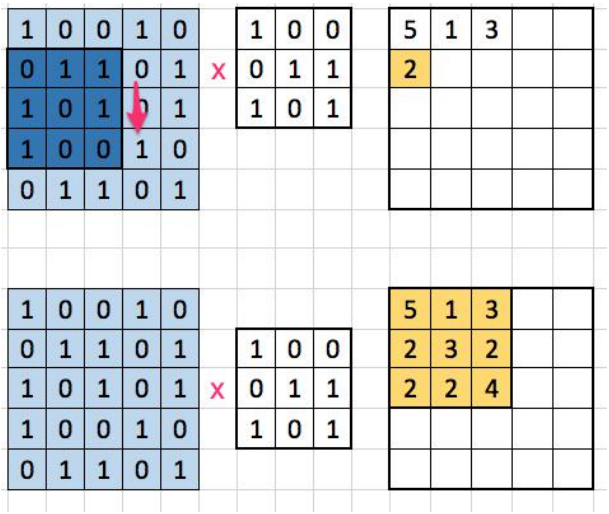

Figure 8: Tích chập kết quả cuối cùng

2.4.2 *Lớp Pooling Layer*

Lớp này sẽ 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 :

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

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

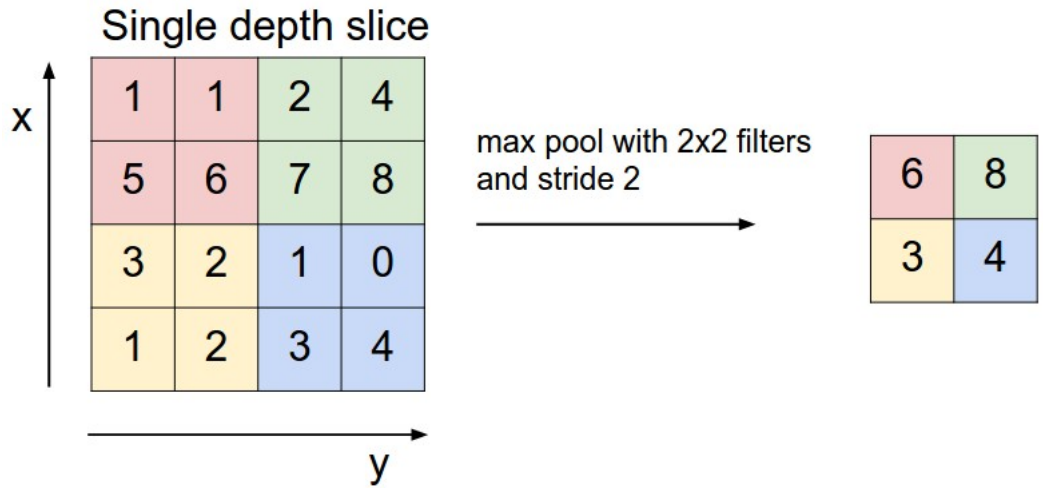

Figure 9: Hoạt động của max-pooling với cửa sổ trượt 2x2

2.4.3 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.

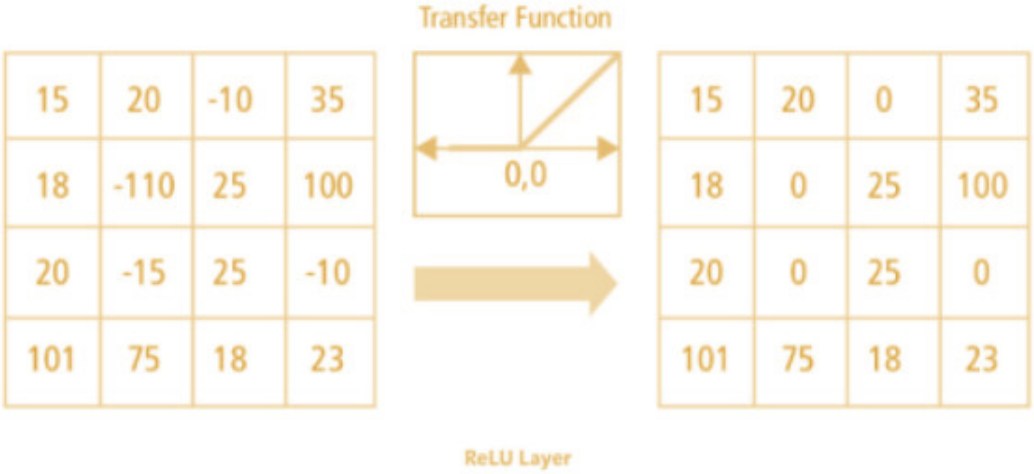

Figure 10: Hoạt động của lớp ReLU

2.4.4 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 $w_l xk$:

$$F(x) = \sigma(W * x)$$

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

Ví dụ : Sau khi các lớp Convolutional Layer và Pooling Layer nhận được các ảnh đã truyền qua chúng, sẽ thu được kết quả là Model đã đọc được khá nhiều thông tin về ảnh. Do đó, để liên kết các đắc điểm này lại và cho ra Output, thì cần dùng đến Fully Connected Layer.

Bên cạnh đó, khi có được các dữ liệu về hình ảnh, Fully Connected Layer sẽ chuyển đổi chúng thành những mục có phân chia chất lượng.

2.4.5 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.

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

2.4.6 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.

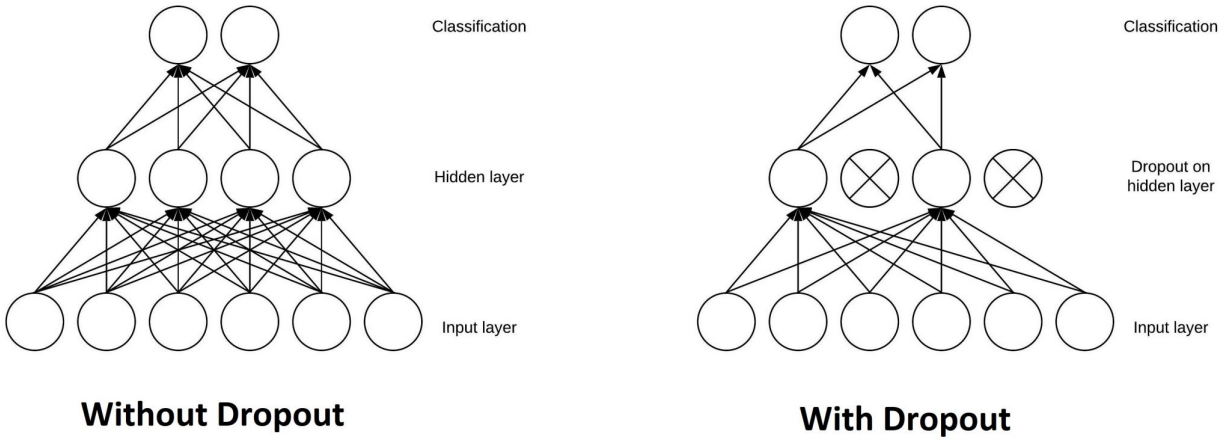

Figure 11: Dropout trong CNN

Tùy thuộc vào mục đích sử dụng, kiến trúc CNN có thể được thiết kế với các lớp khác nhau để đáp ứng yêu cầu của ứng dụng cụ thể. Tuy nhiên, các thành phần cơ bản của CNN vẫn được giữ nguyên và là những thành phần quan trọng trong việc xử lý hình ảnh và video.

3 Thiết kế, xây dựng hệ thống

3.1 Bộ dữ liệu thử nghiệm

Trong đề tài này, nhóm em sử dụng bộ dữ liệu FER-2013 cho việc huấn luyện mô hình CNN.Bộ dữ liệu FER-2013 là một bộ dữ liệu chứa các hình ảnh khuôn mặt của con người được gán nhãn với các cảm xúc khác nhau. FER là viết tắt của cụm từ "Facial Expression Recognition", có nghĩa là nhận diện biểu cảm khuôn mặt. Được tạo ra bởi nhóm nghiên cứu của Pierre-Luc Carrier và Aaron Courville thuộc Đại học Montreal, Canada vào năm 2013.

Bộ dữ liệu này được thu thập từ các hình ảnh chụp khuôn mặt của con người từ các tài khoản Facebook khác nhau, và được gán nhãn với 7 cảm xúc khác nhau: Tức giận, Ghê tởm, Sợ hãi, Hạnh phúc, Buồn bã, Bất ngờ và Trung tính. Tổng cộng, bộ dữ liệu này chứa chứa 35,887 hình ảnh khuôn mặt đơn lẻ, độ phân giải 48x48 pixel đã được gán nhãn.

Một số hình ảnh trong bộ dữ liệu có chứa nhiễu, mờ hoặc vô hình, dẫn đến việc nhận diện cảm xúc trên khuôn mặt trở nên khó khăn. Tuy nhiên, điều này cũng làm cho bộ dữ liệu này trở nên đa dạng và thực tế hơn.

Figure 12: 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 khuôn mặt được trích xuất một cách tự động sao cho khuôn mặt tập trung ở phần trung tâm và chiếm cùng không gian trong mỗi ảnh. 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 ch mạng
- Test set : là nhóm dữ liệu dùng để đánh giá độ chính xác của hệ thống, 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 hệ thống.

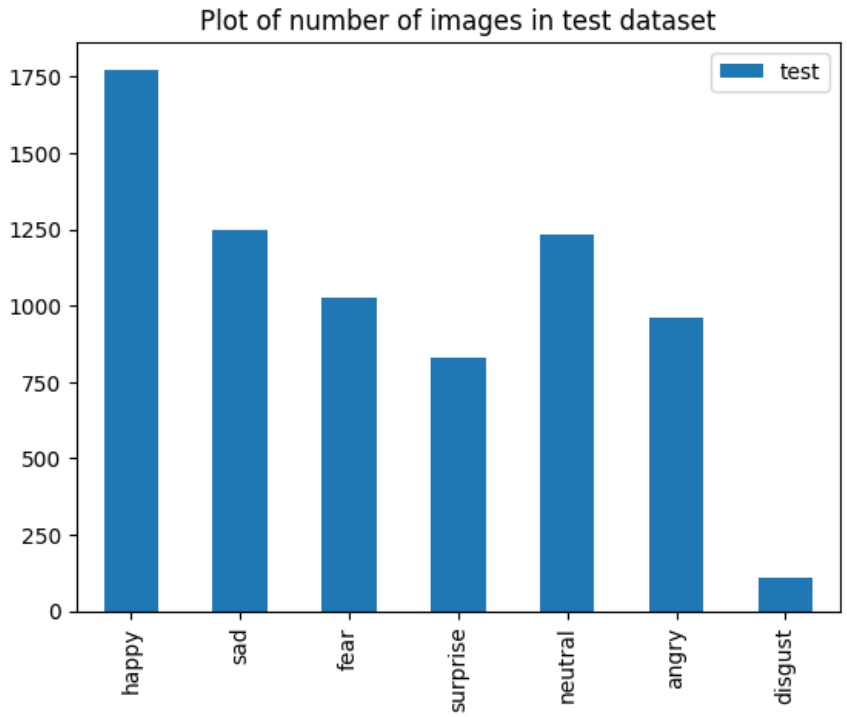

Figure 13: Biểu đồ phân bố dữ liệu trong tập train của FER-2013

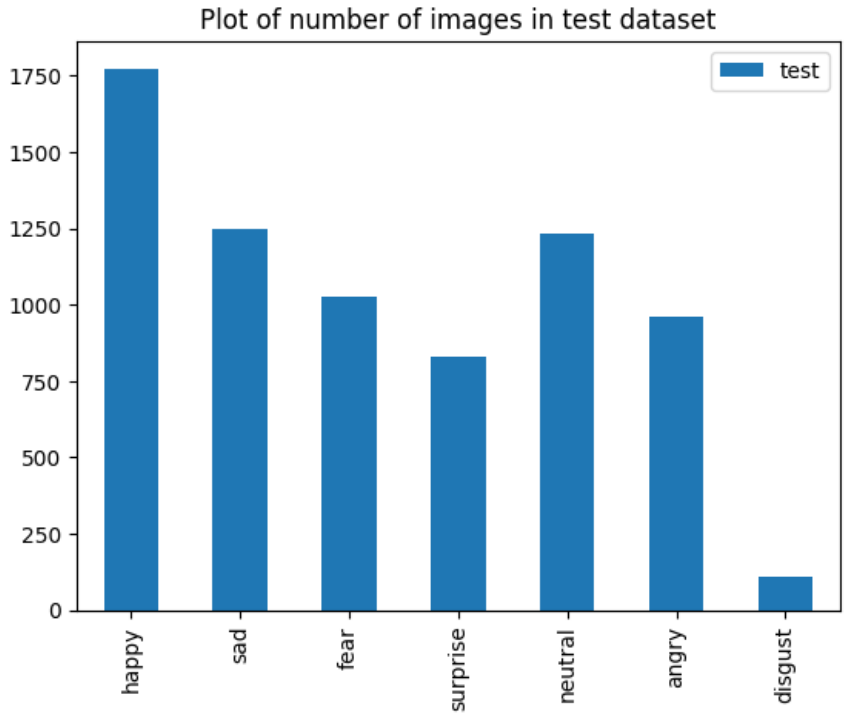

Figure 14: Biểu đồ phân bố dữ liệu trong tập test của FER-2013

3.2 Kiến trúc hệ thống nhận diện cảm xúc khuôn mặt sử dụng mạng nơ-ron tích chập(CNN)

Trong đề tài này, hệ thống được thiết kế dựa trên kiến trúc LeNet5 và cải tiến ở một số điểm để tăng hiệu suất cũng như thời gian huấn luyện. Một số thay đổi 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 over-fitting.
- 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 1 kích thước bộ lọc 5x5 như LeNet. Kích thước bộ lọc nhỏ sẽ giúp giảm số lượng tham số cho mô hình và mang lại 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.

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

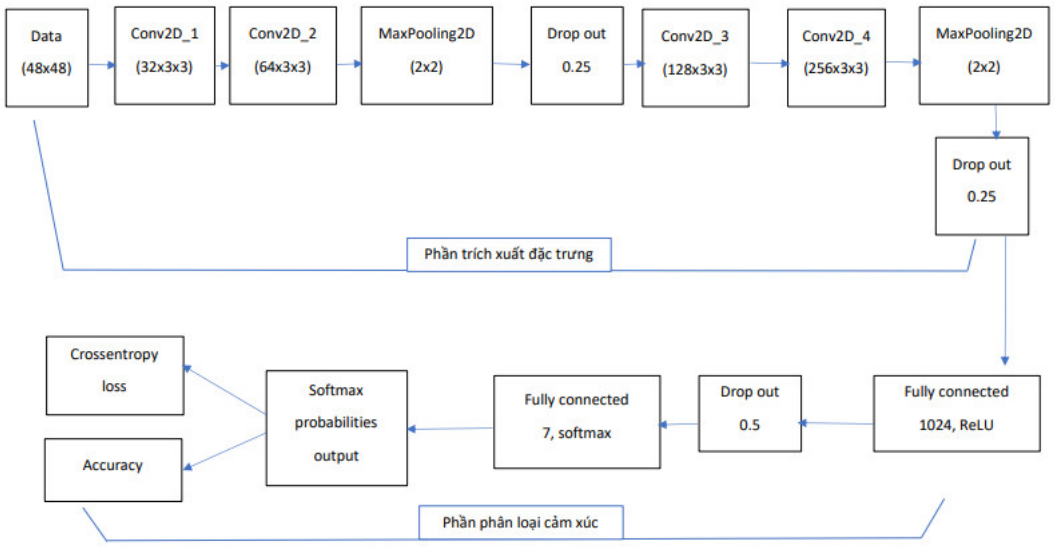

Figure 15: Cấu trúc của mạng nơ-ron tích phân chập

Phân tích đặc trưng : bao gồm 4 lớp tích phân chập, 2 lớp pooling và 2 lớp drop out, cụ thể như sau :

- Data : là dữ liệu đầu vào của hệ thống. 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_1 : là lớp tích chập đầu tiên của hệ thống, sử dụng 32 bộ lọc với cùng kích thước 3x3 pixels. Đi cùng với nó là hàm kích hoạt ReLU.
- Conv2D_2 : 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.

- Sau đó , 2 lớp tích chập liên tiếp với kích thước tương ứng là 128 bộ lọc 3x3 và 256 bộ lọc 3x3 . Cả 2 lớp tích chập này đều sử dụng hàm ReLU để làm hàm kích hoạt.
- Tiếp tục dùng lớp pooling thứ 2 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 1024 và 7. Kỹ thuật drop out được áp dụng trên với tỉ lệ 0.5,tức là có 50% nơ-ron của 2 lớp này bị tắt trong quá trình huấn luyện, nhằm hạn chế overfitting. Hàm mất mát softmax cross-entropy được sử dụng để phản hồi thông tin trong quá trình huấn luyện mạng.

a. Input layer

Những phương thức trong module tf.Keras cho việc tạo ra lớp convolutional và lớp pooling đối với dữ liệu ảnh 2 chiều chấp nhận đầu vào là một tensor có shape là [batch_size, image_width, image_height, channels]. Trong đó :

- batch_size : là số lượng mẫu được dùng khi thực hiện thuật toán gradient descent cho việc tối ưu trong quá trình huấn luyện.
- image_width : chiều rộng của bức ảnh.
- image_height : chiều dài bức ảnh.
- channels : là số kênh màu của ảnh. Nếu ảnh trắng đen thì channels = 1, nếu là ảnh RGB thì channels = 3.

  Trong dữ liệu huấn luyện FER-2013, một mẫu là một hình ảnh trắng đen (channels = 1) có kích thước 48x48 pixels, do đó chọn shape cho input layer là [batch_size, 48, 48, 1] và đồng thời input layer phải resize các mảng 1 chiều về mảng 2 chiều có kích thước 48x48 tương ứng với kích thước không gian của ảnh.

b. Lớp tích chập

Trong lớp tích chập đầu tiên, 32 bộ lọc 3x3 được sử dụng để áp dụng cho input layer, với activation function là hàm ReLU. Phương thức conv2d() trong module layer được sử dụng để tạo lớp tích chập đầu tiên

```
model.add(
    Conv2D(32, kernel_size=(3, 3), padding='same',activation='relu',
   input_shape=input_size))
```

Tham số input tiếp nhận đầu vào của lớp Conv2D và chỉ nhận tensor hợp lệ nếu có shape là [batch_size, image_width, image_height, channels] . Ở đây conv2D layer được kết nối với input_layer có shape [batch_size, 48, 48, 1]. Tham số filters chỉ ra số lượng bộ lọc được sử dụng, ở đây là 32. Tham số padding nhận 2 giá trị hoặc "same" hoặc "valid". Nếu giá trị truyền vào là "valid" , input được giữ nguyên để xử lý và dẫn đến các feature map ở đầu ra của lớp conv có kích thước nhỏ hơn đầu vào. Nếu giá trị truyền vào là "same", TensorFlow sẽ thêm các giá trị 0 vào input sao cho khi tính tích chập hoàn thành sẽ cho ra feature map có cùng kích thước với input. Tham số activation chỉ ra hàm activation function nào sẽ được áp dụng sau khi tính tích chập để cho đầu ra cuối cùng, ở đây activation function là hàm ReLU. Đầu ra của lớp conv2D sẽ có shape là [batch_size, 48, 48, 32] có cùng chiều dài và chiều rộng với input nhưng số channels ở đầu ra cho mỗi input là 32.

Cũng với phương thức conv2d() , lớp conv2d() tiếp theo thuộc phần trích xuất đặc trưng được sử dụng với cấu hình khác :

```
model.add(
    Conv2D(64, kernel_size=(3, 3),activation='relu',padding='same'))
```

c. Lớp pooling

Lớp pooling thứ nhất sẽ kết nối với lớp conv2D_2 vừa được tạo ra. Phương thức max_pooling2d() được sử dụng để xây dựng lớp pooling sử dụng thuật toán max_pooling với bộ lọc 2x2, bước trượt bằng 1.

```
model.add(MaxPooling2D(2, 2))
```

Tham số inputs chỉ ra đầu vào của pool_1 với shape hợp lệ là [batch_size, image_width, image_height, channels] . Ở đây, input là đầu ra của conv2D_2 có shape là [batch_size, 48, 48, 32]. Tham số chỉ ra kích thước của bộ lọc lớp pooling, ở đây kích thước được sử dụng là 2x2. Và vì kích thước bộ lọc là 2x2 nên mỗi vùng nhỏ sẽ chỉ áp dụng bộ lọc 1 lần (không có pixels chung giữa các vùng trong mỗi lần học).

Đầu ra của lớp Pooling có shape là [batch_size, 24, 24, 32] : mỗi feature map đã được giảm đi 50% kích thước.

d. Lớp tích chập thứ 3,4 pooling thứ 2

Cũng với phương thức conv2d() và max_pooling2d(), 4 lớp tiếp theo thuộc phần trích xuất đặc trưng được sử dụng với cấu hình khác nhau:

```
model.add(
```

```
    Conv2D(128, kernel_size=(3, 3),activation='relu',
           padding='same',kernel_regularizer=regularizers.l2(0.01)))
model.add(
    Conv2D(256, kernel_size=(3, 3),activation='relu',
           kernel_regularizer=regularizers.l2(0.01)))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.25))
```

e. Lớp kết nối đầy đủ 1(FC layer)

Lớp FC (với 1024 nơ-ron và ReLU activation) được thêm vào mạng CNN để thực hiện nhiệm vụ phân loại dựa trên các đặc trưng đã được trích xuất bởi lớp convolutional và pooling ở phía trước. Lưu ý ở đây là các feature map ở đầu ra của lớp pooling thứ 2 có shape là [batch_size, 24, 24, 256], tương ứng 1 mẫu ở đầu ra của lớp conv2_c được trích xuất thành 256 ma trận 2 chiều kích thước 24x24. Trước khi đưa vào lớp FC, 64 ma trận 2 chiều (tương ứng với 1 đặc trưng) được phương thức flatten chuyển thành vector 1 chiều có 36864 phần tử, cũng chính là 36864 nơ-ron.

```
model.add(Flatten())
model.add(Dense(1024, activation='relu'))
model.add(Dropout(0.5))
```

Ở lớp fully connected thứ nhấy, 1024 nơ-ron được sử dụng tham gia vào quá trình phân loại của mạng, drop out cũng được áp dụng ở lớp này với tỉ lệ là 50%. Đầu ra của lớp FC thứ nhất này (sau khi áp dụng drop out) là một tensor có shape là [batch_size, 1024].

g. Lớp kết nối đầu đủ 2 (FC layer)

Lớp này là lớp cuối cùng trong mạng CNN là lớp sẽ trả về một vector hàng chứa các xác suất dự đoán của mô hình. Lớp FC cuối cùng gồm 7 nơ-ron tương ứng với 7 cảm xúc mà hệ thống xử lý nhận dạng và activation function là softmax.

```
model.add(Dense(classes, activation='softmax'))
```

Đầu ra cuối cùng của mạng CNN là tensor có shape là [batch_size, 7].

h. Cấu hình hôi quy regression

Đầu ra của lớp kết nối đầy đủ cuối cùng được đăng ký hồi quy với TensorFlow, Keras hỗ trợ phương thức đăng ký hồi quy cùng với việc định nghĩa tối ưu hàm mất mát (loss function) và tốc độ học của mạng. Ở đây, hàm mất mát là hàm categorical_crossentrop và tốc độ học được đặt là 0.0001.

```
model.compile(
    optimizer=Adam(learning_rate=0.0001, decay=1e-6),
    loss='categorical_crossentropy',metrics=['accuracy'])
```

3.3 Thiết lập mô hình CNN cho bài toán phân tích cảm xúc qua hình ảnh

3.4 Tiến hành huấn luyện và đánh giá mô hình

3.4.1 Dữ liệu huấn luyện

Sau khi kết thức quá trình xây dựng kiến trúc mạng nơ-ron tích chập. Tiến hành cài đặt huấn luyện mô hình. Dữ liệu FER-2013 được chia thành 2 thư mục training và test như đoạn code dưới đây.

```
import pandas as pd
import os
train_dir = "E:/FER_2013/train/"
test_dir = "E:/FER_2013/test/"

row, col = 48, 48
classes = 7

def count_exp(path, set_):
    dict_ = {}
    for expression in os.listdir(path):
        dir_ = path + expression
        dict_[expression] = len(os.listdir(dir_))
    df = pd.DataFrame(dict_, index=[set_])
    return df
train_count = count_exp(train_dir, "train")
test_count = count_exp(test_dir, "test")
print(train_count)
print(test_count)
```

Đoạn code trên được xây dựng để đếm số lượng ảnh có trong hai thư mục. Sau khi thực thi sẽ trả về các giá trị như sau:

```
       happy   sad  fear  surprise  neutral  angry  disgust
train   7215  4830  4097      3171     4965   3995      436
      happy   sad  fear  surprise  neutral  angry  disgust
test   1774  1247  1024       831     1233    958      111
```

Figure 16: Kết quả trả về của hàm count_expression

3.4.2 *Tăng cường dữ liệu hình ảnh*

Trong Keras có hỗ trợ class ImageDataGenerator cho phép tạo thêm dữ liệu. Cùng với đó là phương thức flow_from_directory() để đọc các ảnh từ thư mục chứa ảnh.

```
train_datagen = ImageDataGenerator(rescale=1./255,
                                   zoom_range=0.3,
                                   horizontal_flip=True)
train_set = train_datagen.flow_from_directory(train_dir,
                                              batch_size=64,
                                              target_size=(48, 48),
                                              shuffle=True,
                                              color_mode="grayscale",
    class_mode='categorical')
test_datagen = ImageDataGenerator(rescale=1./255)
test_set = test_datagen.flow_from_directory(test_dir,
                                            batch_size=64,
                                            target_size=(48, 48),
                                            shuffle=True,
                                            color_mode="grayscale",
    class_mode='categorical')
```

Đoạn code trên dùng để thực hiện lấy hình ảnh từ thư mục và tăng cường hình ảnh.

Thực hiện lấy ảnh từ thư mục với các cấu hình như sau : kích thước tiêu chuẩn (target_size = (48,48)), số lượng hình ảnh (batch_size = 64), màu của hình ảnh (color_mode = "grayscale"), vì số lượng lớp của mô hình là 7 nên giá trị của class_mode = "categorical".

ImageDataGenerator thực hiện tăng cường hình ảnh bằng việc thực hiện zoom ngẫu nhiên trong một phạm vi 0.3, Lật ảnh ngẫu nhiên theo chiều ngang và thay đổi tỉ lệ gí trị pixel từ phạm vi 0-255 đến phạm vi 0-1.

3.4.3 *Xây dựng mô hình huấn luyện*

Dựa vào kiến trúc CNN đã làm ở trên sẽ xây dựng được mô hình huấn luyện với các lớp tương đương như đoạn code dưới đây.

```
import tensorflow as tf
from keras.layers import Conv2D, Dense, BatchNormalization, Activation
    , Dropout, MaxPooling2D, Flatten
def get_model(input_size, classes=7):
    #Initialising the CNN
    model = tf.keras.models.Sequential()

```

```
    model.add(Conv2D(32, kernel_size=(3, 3), padding='same',
activation='relu', input_shape =input_size))
    model.add(Conv2D(64, kernel_size=(3, 3), activation='relu',
padding='same'))
    model.add(BatchNormalization())
    model.add(MaxPooling2D(2, 2))
    model.add(Dropout(0.25))

    model.add(Conv2D(128, kernel_size=(3, 3), activation='relu',
padding='same', kernel_regularizer=regularizers.l2(0.01)))
    model.add(Conv2D(256, kernel_size=(3, 3), activation='relu',
kernel_regularizer=regularizers.l2(0.01)))
    model.add(BatchNormalization())
    model.add(MaxPooling2D(pool_size=(2, 2)))
    model.add(Dropout(0.25))

    model.add(Flatten())
    model.add(Dense(1024, activation='relu'))
    model.add(Dropout(0.5))

    model.add(Dense(classes, activation='softmax'))

    #Compliling the model
    model.compile(optimizer=Adam(learning_rate=0.0001, decay=1e-6),
                  loss='categorical_crossentropy',
                  metrics=['accuracy'])
    return model
```

Với Input_size là một ảnh có size = (48,48) và có image channel = 1 (GRAYSCALE). Thêm các lớp như lớp tích chập, lớp pooling, lớp Fully Connected.

4 Thảo luận và đánh giá kết quả

Với thiết kế như đã trình bày ở trên, hệ thống được huấn luyện sử dụng tập dữ liệu FER-2013 bao gồm một tập hình ảnh với kích thước 48x48 pixels được định dạng ở mức xám. Qúa trình huấn luyện được thực hiện qua 60 chu kì dữ liệu (60 epoch) với tổng thời gian huấn luyện gần 1 giờ.

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 0.0001. Độ chính xác trong quá trình huấn luyện cũng tăng theo từng chu kì dữ liệu, đạt đến khoảng 91.42% đối với tập huấn luyện, và độ chính xác đối với tập dữ liệu validation đạt 66.24% 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.

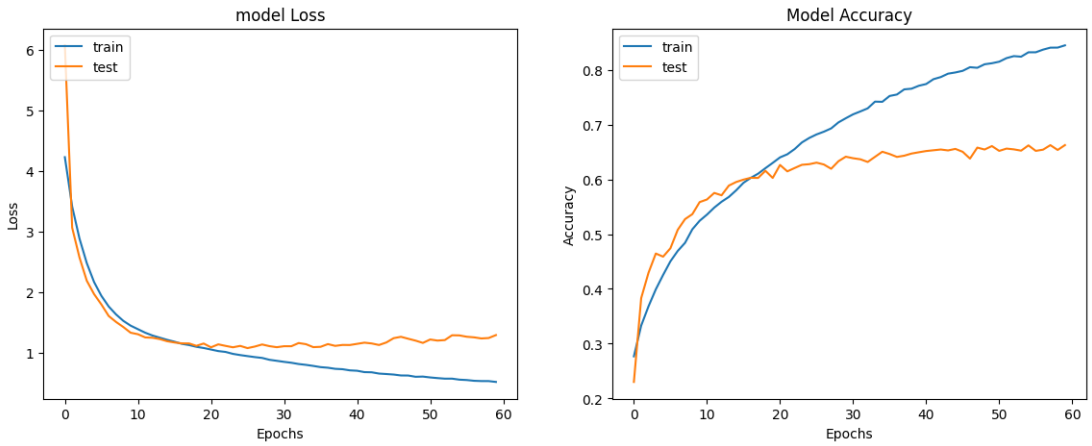

Figure 17: Biểu đồ Model Loss và Model Accuracy

```
449/449 [==============================] - 21s 47ms/step - loss: 0.3549
113/113 [==============================] - 5s 45ms/step - loss: 1.2945
final train accuracy = 91.42 , validation accuracy = 66.24
```

Figure 18: Tỷ lệ train accuracy đạt 91.42% và validation accuracy đạt 66.24%

Confusion Matrix

```
[[ 565   52  508 1040  699  674  457]
 [  61    2   70  100   76   89   38]
 [ 577   51  478 1080  677  758  476]
 [ 964   98  975 1899 1182 1301  796]
 [ 665   69  669 1331  828  862  541]
 [ 708   60  615 1263  854  799  531]
 [ 455   40  385  772  551  597  371]]
```

Classification Report

	precision	recall	f1-score	support
angry	0.14	0.14	0.14	3995
disgust	0.01	0.00	0.00	436
fear	0.13	0.12	0.12	4097
happy	0.25	0.26	0.26	7215
neutral	0.17	0.17	0.17	4965
sad	0.16	0.17	0.16	4830
surprise	0.12	0.12	0.12	3171
accuracy			0.17	28709
macro avg	0.14	0.14	0.14	28709
weighted avg	0.17	0.17	0.17	28709

Training Set

Confusion Matrix

```
[[137   9 116 239 168 173 116]
 [ 17   1  11  29  22  15  16]
 [148   8 114 266 164 215 109]
 [256  22 184 484 317 317 194]
 [188   9 112 337 202 247 138]
 [172   8 141 327 201 246 152]
 [116   8  87 223 159 138 100]]
```

Classification Report

	precision	recall	f1-score	support
angry	0.13	0.14	0.14	958
disgust	0.02	0.01	0.01	111
fear	0.15	0.11	0.13	1024
happy	0.25	0.27	0.26	1774
neutral	0.16	0.16	0.16	1233
sad	0.18	0.20	0.19	1247
surprise	0.12	0.12	0.12	831
accuracy			0.18	7178
macro avg	0.15	0.15	0.14	7178
weighted avg	0.18	0.18	0.18	7178

Testing Set

Figure 19: Confusion Matrix và Classification Report của tập dữ liệu FER-2013

- **Training Set :** Trong confusion matrix, các giá trị trên đường chéo chính (tương ứng với số lượng các trường hợp được phân loại đúng) chưa cao, chỉ nằm trong khoảng từ 14% đến 26%. Các giá trị nằm ngoài đường chéo chính (tương ứng với số lượng các trường hợp được phân loại sai) cũng khá cao, đặc biệt là ở các lớp "angry", "fear" và "happy".

  Trong classification report, các giá trị precision, recall và f1-score đều rất thấp, chỉ nằm trong khoảng từ 0.01 đến 0.25. Điều này cho thấy mô hình không thể phân loại các cảm xúc một cách chính xác và đáng tin cậy.

- **Testing Set :** Kết quả confusion matrix và classification report cho thấy hiệu suất của mô hình phân loại cảm xúc chưa tốt.

  Trong confusion matrix, các giá trị trên đường chéo chính (tương ứng với số lượng các trường hợp được phân loại đúng) không cao, chỉ nằm trong khoảng từ 11% đến 27%. Các giá trị nằm ngoài đường chéo chính (tương ứng với số lượng các trường hợp được phân loại sai) cũng khá cao, đặc biệt là ở các lớp "angry", "fear" và "happy".

  Trong classification report, các giá trị precision, recall và f1-score đều rất thấp, chỉ nằm trong khoảng từ 0.02 đến 0.25. Điều này cho thấy mô hình chưa thể phân loại các cảm xúc một cách chính xác.

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 hệ thống xử lý tương đối ổn định. Trong đó, bộ nhận dạng khuôn mặt và nhận diện cảm xúc trên khuôn mặt cho kết quả khá nhanh và khá chính xác. Thời gian phản hồi giữa các mô hình với hệ thống khá

nhanh, gần như thời gian thực. Đây là một gợi ý cho hướng 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.

Tuy đề tài bước đầu đã chỉ ra những cảm xúc cơ bản trên khuôn mặt người, nhưng vẫn còn nhiều bất cập :

- Đối với ảnh có các cảm xúc như : Vui, ngạc nhiên, bình thường và giận dữ thì hệ thống nhận diện với tỉ lệ khá tốt (>90%) , tuy nhiên đối với cảm xúc buồn và sợ thì tỉ lệ dự đoán thấp hơn. Riêng cảm xúc ghê tởm, hệ thống có tỉ lệ dự đoán thấp nhất.
- Phát hiện khuôn mặt : Với thuật toán phát hiện khuôn mặt hiện tại, 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 hệ thống chưa thể nhận dạng đúng hoàn toàn cảm xúc khuôn mặt người.

5 Kết Luận

Sau khi tìm hiểu và thực hiện đề tài "Ứng dụng CNN phân tích cảm xúc qua hình ảnh". Nhóm chúng em thực hiện xây dựng mô hình nhận diện cảm xúc dựa trên thuật toán CNN. 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 phân loại 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, em xin chân thành cảm ơn cô Nguyễn Thị Bích Thủy đã tận tình quan tâm, giúp đỡ và hướng dẫn chúng em hoàn thành báo cáo này.

6 Tài liệu tham khảo

[1]. Ngô Tứ Thành, Nguyễn Thế Dũng, Một số vấn đề về hướng của điểm đặc trưng trong ảnh vân tay. Hội thảo quốc gia về tin học ứng dụng, Quy Nhơn tháng 8/1988

[2]. Nguyễn Kim Sách, "Xử lý ảnh và video số", NXB Khoa học và kỹ thuật,1997.

[3]. Hộ chiếu điện tử và mô hình đề xuất tại Việt Nam (2007)– tạp chí KH ĐHQG Hà Nội.

[4]. Lương Mạnh Bá (1999), "Nhập môn xử lý ảnh số ", NXB Khoa học và Kỹ thuật.

[5]. Một số vấn đề khoa học đường vân tay trên cơ thể người (2017) , Tạp chí Y - Dược học Quân sự.

[6]. Handbook of Fingerprint Recognition (2009),

[7]. Parallel processing for Fingerprint feature extraction (2004) - G. Indrawan, B. Sitohang, S. Akbar

[8]. Gonzalez, Rafael C, 2008. "Digital Image Processing", Pearson Education, Inc., publishing as Prentice Hall.

[10] Segmentation of Fingerprint Images - Asker M. Bazen, Sabih H. Gere

[11] Enhancement of Latent Fingerprints using Morphological Filters - Nirmal. K, Madhubala. G

[12] Estimation of fingerprint orientation field by utilizing prior knowledge and self-organizing map - Sri Suwarno, Subanar, Agus Harjoko, Sri Hartati

[13] International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 15 97 Fingerprint Core Point Detection Algorithm Using Orientation Field Based Multiple Features - H B Kekre, V A Bharadi

https://trantheanh.github.io/2018/02/15/ML-21/