**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO TIỂU LUẬN**

**Đề tài: “SỬ DỤNG MẠNG CNN TRONG BÀI TOÁN NHẬN DIỆN CẢM XÚC KHUÔN MẶT”**

**Sinh viên thực hiện : Hoàng Trung Kiên**

**Trần Hoài Nam**

**Nguyễn Minh Tuyền**

**Nguyễn Trường Xuân**

**Lớp : CNTT K18G**

**GVHD : TS. Trần Quang Quý**

**Thái Nguyên, tháng 3 năm 2023**

**Thái Nguyên, tháng 9 năm 2023**

**MỤC LỤC**

[LỜI NÓI ĐẦU 4](#_Toc146981199)

[CHƯƠNG 1: CƠ SỞ LÝ THUYẾT 5](#_Toc146981200)

[1.1 Mạng nơ-ron nhân tạo 5](#_Toc146981201)

[1.1.1 Giới thiệu về mạng nơ-ron 5](#_Toc146981202)

[1.1.2 Một số kiểu mạng nơ-ron 7](#_Toc146981203)

[1.1.3 Mạng nơ-ron lan truyền ngược MLP 10](#_Toc146981204)

[1.2 Mạng tích chập 14](#_Toc146981205)

[1.2.1 Phép tích chập 15](#_Toc146981206)

[1.2.2 Kiến trúc mạng tích chập 17](#_Toc146981207)

[1.2.3 Quá trình huấn luyện mạng 19](#_Toc146981208)

[CHƯƠNG 2: BÀI TOÁN NHẬN DẠNG BIỂU CẢM 20](#_Toc146981209)

[2.1 Các thách thức trong bài toán nhận dạng biểu cảm 20](#_Toc146981210)

[2.2 Phân loại 21](#_Toc146981211)

[2.3 Một số phương pháp nhận dạng biểu cảm khuôn mặt 21](#_Toc146981212)

[2.4 Các bước thực hiện bài toán nhận diện biểu cảm 22](#_Toc146981213)

[2.5 Các nghiên cứu gần đây 24](#_Toc146981214)

[CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU 28](#_Toc146981215)

[3.1 Khởi tạo mẫu tổng hợp (Synthetic sample generation) 29](#_Toc146981216)

[3.2 Phép quay (Rotation correction) 31](#_Toc146981217)

[3.3 Cắt tỉa hình ảnh (Image Cropping) 32](#_Toc146981218)

[3.4 Lấy mẫu xuống (Down-sampling) 33](#_Toc146981219)

[3.5 Chuẩn hóa cường độ (Intensity normalization) 34](#_Toc146981220)

[3.6 Kiến trúc mạng CNN 34](#_Toc146981221)

[CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM 40](#_Toc146981222)

[4.1 Môi trường thực nghiệm 40](#_Toc146981223)

[4.2 Cơ sở dữ liệu 40](#_Toc146981224)

[4.3 Kết quả thực nghiệm: 41](#_Toc146981225)

[KẾT LUẬN 43](#_Toc146981226)

[TÀI LIỆU THAM KHẢO 44](#_Toc146981227)

[NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN 46](#_Toc146981228)

# LỜI NÓI ĐẦU

Trong thời đại số hóa và công nghệ thông tin ngày nay, học máy đã trở thành một lĩnh vực nghiên cứu đầy tiềm năng và ứng dụng mạnh mẽ. Trong khuôn khổ bài tiểu luận này, chúng ta sẽ bước vào thế giới của Học máy và tập trung vào một khía cạnh cụ thể của nó: Sử dụng mạng CNN trong bài toán nhận diện cảm xúc khuôn mặt.

Cảm xúc là một phần quan trọng của cuộc sống con người, và khả năng nhận biết cảm xúc từ khuôn mặt là một khả năng quan trọng trong giao tiếp và tương tác xã hội. Điều này có ý nghĩa đặc biệt trong nhiều lĩnh vực như trí tuệ nhân tạo, công nghệ nhận diện khuôn mặt, giáo dục, y học và nhiều lĩnh vực khác.

CNN, mạng neural tích chập, đã đánh dấu một bước tiến lớn trong việc giải quyết các vấn đề liên quan đến thị giác máy tính. CNN không chỉ có khả năng học được các đặc trưng quan trọng từ dữ liệu hình ảnh mà còn giúp giảm thiểu lỗi và tăng hiệu suất trong nhiều nhiệm vụ thị giác máy tính, bao gồm cả việc nhận diện cảm xúc từ khuôn mặt.

Trong bài tiểu luận này, chúng ta sẽ khám phá sâu hơn về cách CNN hoạt động và làm thế nào chúng có thể được ứng dụng để nhận diện cảm xúc từ khuôn mặt. Chúng ta sẽ xem xét các phương pháp, kỹ thuật và các tài liệu nghiên cứu liên quan đến chủ đề này để hiểu rõ hơn về khả năng và hạn chế của CNN trong bài toán nhận diện cảm xúc.

Nhằm giúp tăng hiểu biết về sự quan trọng của việc ứng dụng CNN trong việc nhận diện cảm xúc từ khuôn mặt, chúng ta sẽ cùng nhau thảo luận về những tiềm năng và ứng dụng thực tiễn của nghiên cứu này trong các lĩnh vực khác nhau.

Chúng ta hy vọng rằng thông qua bài tiểu luận này, bạn sẽ có cái nhìn tổng quan và hiểu biết sâu hơn về cách mạng CNN có thể giúp chúng ta nhận diện và hiểu cảm xúc con người thông qua khuôn mặt, và đồng thời khám phá tiềm năng ứng dụng mạnh mẽ của nó trong nhiều lĩnh vực của cuộc sống và công nghiệp.

# CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

## Mạng nơ-ron nhân tạo

### 1.1.1 Giới thiệu về mạng nơ-ron

Mạng Nơ-ron nhân tạo (Artificial Neural Network- ANN) là mô hình xử lý thông tin được mô phỏng dựa trên hoạt động của hệ thống thần kinh của sinh vật, bao gồm số lượng lớn các nơ-ron được gắn kết để xử lý thông tin. ANN giống như bộ não con người, được học bởi kinh nghiệm (thông qua huấn luyện), có khả năng lưu giữ những kinh nghiệm hiểu biết (tri thức) và sử dụng những tri thức đó trong việc dự đoán các dữ liệu chưa biết.

Các ứng dụng của mạng nơ-ron được sử dụng trong rất nhiều lĩnh vực như điện, điện tử, kinh tế, quân sự, …để giải quyết các bài toán có độ phức tạp và đòi hỏi có độ chính xác cao như điều khiển tự động, khai phá dữ liệu, nhận dạng, …

Kiến trúc tổng quát của một ANN được mô tả trong hình 1.1 như sau:

A diagram of a network

Description automatically generated

*Hình 1.1: Kiên trúc mạng nơ-ron nhân tạo*

Trong đó, mỗi nơ-ron nhận các dữ liệu vào (Inputs) xử lý chúng và cho ra một kết quả (output) duy nhất. Kết quả xử lý của một nơ-roncó thể làm đầu vào cho các nơ-ron khác.

Đầu vào (Input): Là tập các tín hiệu đầu vào của nơ-ron, các tín hiệu này thường được đưa vào dưới dạng vector N chiều.

Lớp ẩn gồm các nơ-ron nhận dữ liệu input từ các nơ-ron ở lớp (Layer) trước đó và chuyển đổi các input này cho các lớp xử lý tiếp theo. Một nơ-ron như là mô hình ra quyết định bằng cách xem xét các dữ kiện đầu vào.

Đầu ra (Output): là tín hiệu đầu ra của một nơ-ron, với mỗi nơ-ron sẽ có tối đa 1 đầu ra.

Xét một nơ-ron đơn giản có nhiệm vụ đưa ra quyết định có nên đi chơi vào cuối tuần hay không như hình 1.2 dựa vào ba thông tin input: khả năng tài chính, hoàn thành nhiệm vụ được giao và số bạn bè đi cùng. Nơ -ron output tính toán kếtquả bằng cách tổ hợp tuyến tính các thông tin input, kèm theo trọng số của mỗi thông tin: 0.2 × 3 + 1 × 0.2 + 2 × (−0.05)

Quyết định được đưa ra theo một luật đặt trước tùy ý, ví dụ như output lớn hơn 0.6 thì đi chơi, và ngược lại.

A diagram of a mathematical equation

Description automatically generated

*Hình 1.2: Cách hoạt động của một nơ-ron*

Để cài đặt đơn giản hơn, giá trị ngưỡng này được đưa vào ngay sau phép tổ hợp tuyến tính và so sánh kết quả với 0. Giá trị ngưỡng (trong trường hợp này là -0.6) được gọi là hệ số bias. Giá trị ngưỡng này thường được đưa vào như thành phần của hàm truyền. Hàm truyền (Transfer function) dùng để giới hạn đầu ra của mỗi nơ-ron.

(0.2 × 3+1×0.2+2×(−0.05))−0.6

Mỗi thành phần (𝑤) ứng với mỗi nơ-ron gọi là trọng số của nơ-ron đó. Tập hợp tất cả các trọng số của toàn bộ nơ-ron ta có bộ trọng số. Bộ trọng số này là thành phần rất quan trọng của một ANN, nó thể hiện mức độ quan trọng của dữ liệu đầu vào đối với quá trình xử lý thông tin, ví dụ: trọng số ứng với thông tin có “tiền” là một số lớn, nghĩa là yếu tố tài chính có mức ảnh hưởng lớn hơn so với các yếu tố còn lại. Quá trình của ANN thực ra là quá trình điều chỉnh các trọng số của các input data để có được kết quả mong muốn.

Ban đầu bộ trọng số và bias được khởi tạo ngẫu nhiên, dần dần qua quá trình huấn luyện bằng dữ liệu, các trọng số và bias sẽ được cập nhật lại và đưa ra kết quả dự đoán chính xác hơn.

Việc tính tích các đầu vào với trọng số liên kết của nó gọi là hàm tổng của một nơ-ron. Hàm tổng đối với nhiều Nơ-ron trong cùng một lớp được tính theo công thức 1-1 như sau:

(1-1)

### 1.1.2 Một số kiểu mạng nơ-ron

Các nơ-ron trong mạng có thể kết nối đầy đủ (fully connected) tức là mỗi nơ-ron của lớp trước đều được kết nối với tất cả các nơ-ron ở lớp sau, hoặc kết nối cục bộ (partially connected) như chỉ kết nối nối giữa các nơ ron trong các lớp khác nhau.

* **Tự kết hợp (autoassociative):** là mạng có các nơ-ron đầu vào cũng là các nơ-ron đầu ra. Mạng Hopfield là một kiểu mạng tự kết hợp.

A diagram of a diagram

Description automatically generated

*Hình 1.3: Mạng tự kết hợp*

* **Kết hợp khác kiểu (heteroassociative):** là mạng có tập nơ-ron đầu vào và đầu ra riêng biệt. Perceptron, các mạng Perceptron nhiều tầng (MLP: MultiLayer Perceptron), mạng Kohonen, … thuộc loại này.

A diagram of a path

Description automatically generated with medium confidence

*Hình 1.4: Mạng kết hợp khác kiểu*

Ngoài ra tùy thuộc vào mạng có các kết nối ngược (feedback connections) từ các nơ-ron đầu ra tới các nơ-ron đầu vào hay không, người ta chia ra làm 2 loại kiến trúc mạng.

* **Kiến trúc truyền thẳng (feedforward architechture):** là kiểu kiến trúc mạng không có các kết nối ngược trở lại từ các nơ-ron đầu ra về các nơ-ron đầu vào; mạng không lưu lại các giá trị output trước và các trạng thái kích hoạt của nơ-ron. Các mạng nơ-ron truyền thẳng cho phép tín hiệu di chuyển theo một đường duy nhất; từ đầu vào tới 7 đầu ra, đầu ra của một tầng bất kì sẽ không ảnh hưởng tới tầng đó. Các mạng kiểu Perceptron là mạng truyền thẳng.

A diagram of a network

Description automatically generated

*Hình 1.5: Mạng truyền thẳng*

* **Kiến trúc phản hồi (Feedback architecture):** là kiểu kiến trúc mạng có các kết nối từ nơ-ron đầu ra tới nơ-ron đầu vào. Mạng lưu lại các trạng thái trước đó, và trạng thái tiếp theo không chỉ phụ thuộc vào các tín hiệu đầu vào mà còn phụ thuộc vào các trạng thái trước đó của mạng. Mạng Hopfield thuộc loại này.

A diagram of a diagram

Description automatically generated

*Hình 1.6: Mạng phản hồi*

### 1.1.3 Mạng nơ-ron lan truyền ngược MLP

Mô hình mạng nơ-ron MLP (Multi Layer Perceptron) là mô hình được sử dụng phổ biến nhất.

**1.1.3.1 Kiến trúc mạng**

Một mạng MLP thông thường có ít nhất 1 tầng ẩn.

**A diagram of a network

Description automatically generated**

*Hình 1.7: mạng nơ-ron lan truyền ngược MLP*

Đầu vào là các vector (x1, x2, ... xn) trong không gian n chiều, đầu ra là các vector (y1, y2, …ym) trong không gian m chiều. Xét bài toán nhận dạng biểu cảm khuôn mặt: đầu vào là các bức ảnh có kích thước 32x32, nhiệm vụ của mạng là phân loại các mẫu này vào 7 lớp biểu cảm tương ứng: giận, khinh, ghét, sợ, vui, buồn, ngạc nhiên.

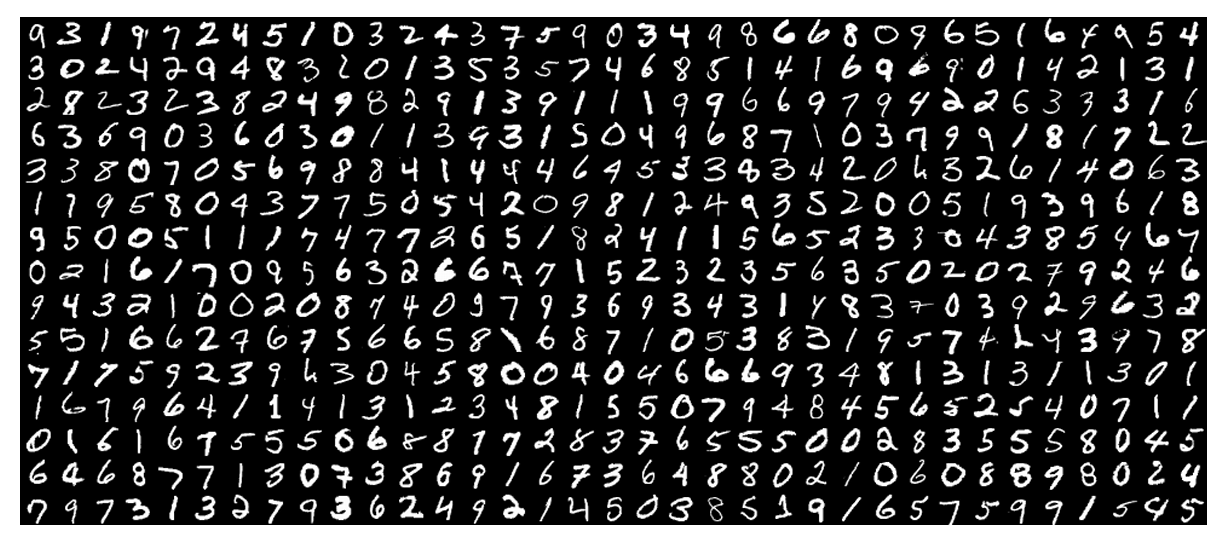
**1.1.3.2 Phương thức học**

Có ba phương thức học phổ biến là học có giám sát (supervised learning), học không giám sát (unsupervised learning) và học tăng cường (reinforcement learning).

**Học có giám sát (Supervised learning):** là thuật toán dự đoán đầu ra từ dữ liệu đầu dựa trên các cặp dữ liệu vào-ra đã biết trước.

Cho tập đầu vào 𝑋 = {𝑥1,𝑥2,…,𝑥𝑛} và một tập nhãn tương ứng 𝑌 = 𝑦1, 𝑦2, …,𝑦𝑛}, trong đó (𝑥𝑖,𝑦𝑖) ∈ 𝑋x𝑌 là các cặp dữ liệu biết trước, còn gọi là tập huấn luyện. Từ tập này, cần tạo ra một hàm số ánh xạ mỗi phần tử tập 𝑋 sang một phần tử xấp xỉ tương ứng của tập Y: 𝑦𝑖 ≈ 𝑓(𝑥𝑖),∀𝑖 = 1,2,…,𝑛

Ví dụ: trong nhận dạng chữ viết tay, ta có ảnh của hàng nghìn ví dụ của mỗi chữ số được viết bởi nhiều người khác nhau. Các bức ảnh này được đưa vào trong một thuật toán và chỉ cho nó biết mỗi bức ảnh tương ứng với chữ số nào. Sau khi 9 thuật toán tạo ra một mô hình, tức một hàm số mà đầu vào là một bức ảnh và đầu ra là một chữ số, khi nhận được một bức ảnh mới mà mô hình chưa nhìn thấy bao giờ, nó sẽ dự đoán bức ảnh đó chứa chữ số nào.



*Hình 2.8: Bộ cơ sở dữ liệu MNIST của chữ số viết tay*

Cách học này khá giống cách học chữ của con người khi còn nhỏ. Ta đưa một chữ “a” và “người thầy” bảo em nhỏ đây là chữ “a”. Việc này được thực hiện trên tất cả các mẫu chữ cái. Sau đó khi kiểm tra ta sẽ đưa ra một chữ cái bất kì (có thể viết hơi khác đi) và hỏi em đó là chữ gì?

Học có giám sát được chia thành 2 loai chính: Phân lớp (Classsification) và hồi quy (Regression). Ví dụ ứng dụng dự đoán giới tính và tuổi dựa trên khuôn mặt2 gần đây của Microsoft, phần giới tính có thể coi là thuật toán phân lớp, phần dự đoán tuổi có thể xem như thuật toán hội quy.

**Học không giám sát (Unsupervised learning):** Là thuật toán không biết được đầu ra hay nhãn mà chỉ có dữ liệu đầu vào. Quá trình học không giám sát không cần có bất kỳ một sự giám sát nào, số lớp các phân loại chưa được biết trước và tùy theo tiêu chuẩn đánh giá mà các mẫu có thể thuộc về các phân lớp khác nhau.

Trong bài toán học không giám sát, tập dữ liệu huấn luyện được cho dưới dạng: D={(𝑥1,𝑥2,…,𝑥𝑛) } với (𝑥1,𝑥2,…,𝑥𝑛) là vector đặc trưng của mẫu huấn luyện. Nhiệm vụ của thuật toán là chia tập dữ liệu D thành các nhóm con, mỗi nhóm con chứa các vector đầu vào có đặc trưng giống nhau.

**Học tăng cường (Reinforcement learning):** là sự tổ hợp của 2 phương pháp học trên, giúp cho hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao nhất. Ví dụ các thuật toán thuộc về Lý thuyết trò chơi (Game Theory) để xác định nước đi tiếp theo đạt được điểm số cao nhất.

1.1.3.3 Quỳ trình lan truyền tiến

Quá trình tính toán từ vector đầu vào x được vector dự đoán y gọi là quá trình làn truyền tiến.

1.1.3.4 Thuật toán Gradient Descent

Gradient descent là một trong những thuật toán phổ biến nhất được dùng để tính giá trị cực tiểu hàm chi phí.

Có 2 thuật toán gradient descent, phân biệt bởi số lượng mẫu dữ liệu được dùng để tính hàm chi phí” batch gradient descent và stochastic gradient descent.

**Batch gradient descent (GD)**

Trong thuật toán này Batch được hiểu là tất cả, tức là không cập nhật hàm chi phí 𝜃=w, ta sử dụng tất cả các điểm dữ liệu xi. Thuật toán này bị hạn chế với cơ sở dữ liệu có nhiều điểm, việc tính toán lại đạo hàm với tất cả các các điểm sau mỗi vòng lặp làm thuật toán trở nên cồng kềnh và không hiệu quả.

𝜃 =𝜃−𝛼⋅𝛻𝜃𝐿(𝜃) (2-2)

Trong đó 𝛻𝜃𝐿(𝜃) là gradient của hàm chi phí 𝐿 theo 𝜃, 𝛼 > 0 là hệ số học, hệ số học 𝛼 quyết định độ lớn “bước đi” trong mỗi lần cập nhật. Nếu 𝛼 quá lớn, 𝜃 có thể đi vượt qua khỏi cực trị, làm cho L tăng. Ngược lại, nếu 𝛼 quá nhỏ, quá trình huấn luyện sẽ rất chậm do 𝜃 chỉ thay đổi một lượng nhỏ sau mỗi lần cập nhật. Do đó, hệ số học cần được lựa chọn hợp lý, sao cho không quá nhỏ cũng không quá lớn.



*Hình 1.9: Quá trình tối ưu theo GD*

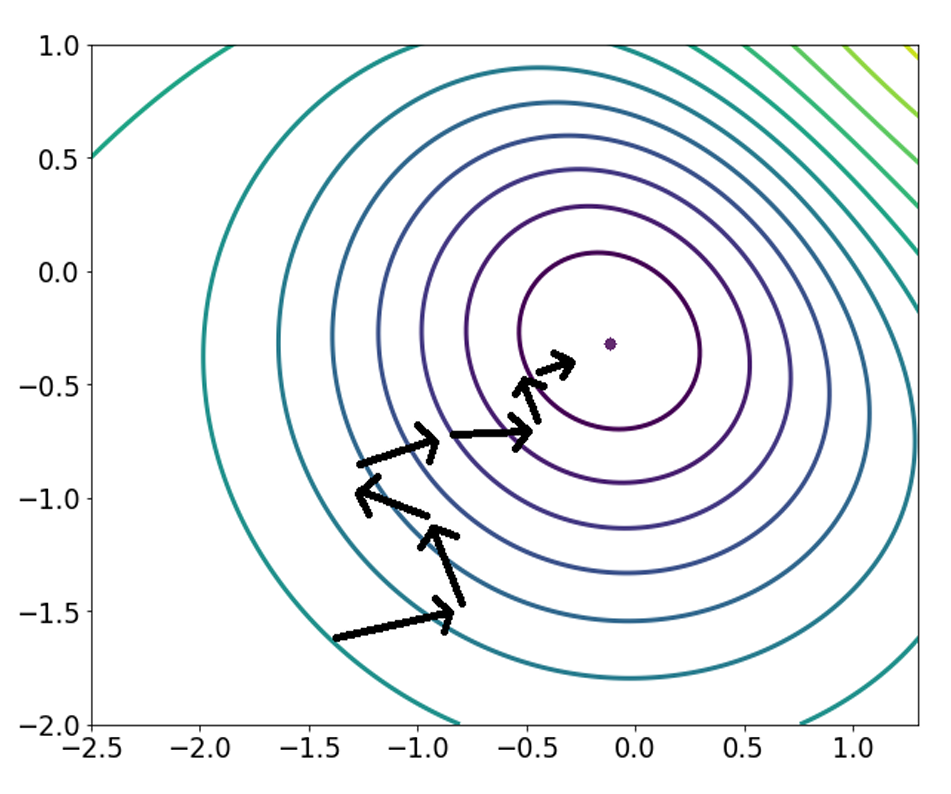
**Stochastic gradient descent (SGD)**

Trong thuật toán này, tại 1 thời điểm, ta chỉ tính đạo hàm của hàm mất mát dựa trên chỉ một điểm dữ liệu xi rồi cập nhật 𝜃 dựa trên đạo hàm này. Việc này được thực hiện với từng điểm trên toàn bộ dữ liệu sau đó lặp lại quá trình trên. Thuật toán này trên thực tế đem lại hiệu quả cao hơn, đặc biệt đối với các hệ thống học trực tuyến.

Mỗi lần duyệt qua tất cả các điểm trên toàn bộ dữ liệu được gọi là một epoch. Với GD thì mỗi epoch ứng với 1 lần cập nhật 𝜃, với SGD thì mỗi epoch ứng với n lần cập nhật 𝜃 với n là số điểm dữ liệu. Việc cập nhật từng điểm có thể làm giảm tốc độ thực hiện 1 epoch, tuy nhiên SGD chỉ yêu cầu một lượng epoch rất nhỏ (thường là 10 cho lần đầu tiên, sau khi có dữ liệu mới thì chạy thêm một 12 epoch đã đem lại kết quả tốt). Vì vậy SGD phù hợp với các bài toán có lượng cơ sở dữ liệu lớn.

(2-3)

Trong đó chỉ tập 𝑚 mẫu dữ liệu từ đến . Tập 𝑚 mẫu được gọi là một mini-batch và số lượng mẫu 𝑚 được gọi là kích thước của mini batch.



*Hình 1.10: Quá trình tối ưu hóa theo SGD*

**1.1.3.5 Quá trình lan truyền ngược**

Lan truyền ngược là một thuật toán để tính từng thành phần trong một cách hiệu quả. Thuật toán này áp dụng quy tắc tính đạo hàm của hàm hợp.

## 1.2 Mạng tích chập

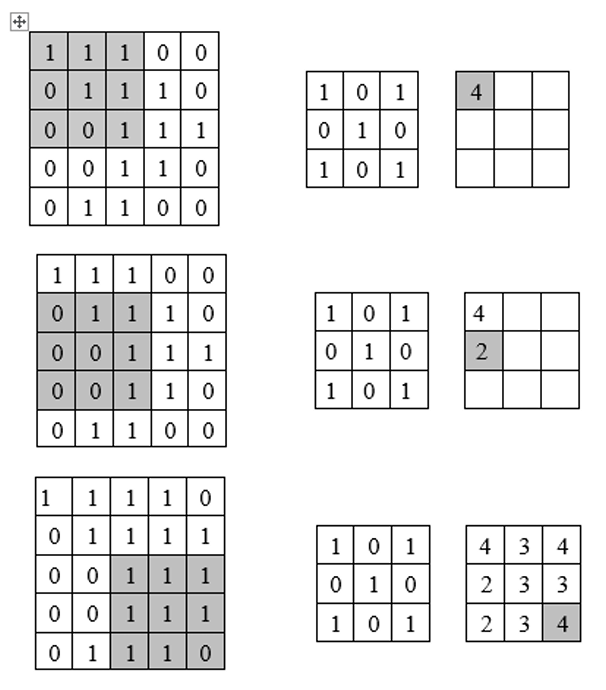
Mạng tích chập (Convolutional Neural Networks-CNN hoặc ConvNet) là một trong những thuật toán phổ biến của mô hình học sâu, kiến trúc mạng này 13 thường được dùng khi xử lý các bài toán dữ liệu có tổ chức dạng lưới như tín hiệu số, hình ảnh… đặc biệt là trong việc phân lớp dữ liệu hình ảnh, video, âm thanh.

### 1.2.1 Phép tích chập

Tích chập được sử dụng đầu tiên trong xử lý tín hiệu số. Nhờ vào nguyên lý biến đổi thông tin, các nhà khoa học đã áp dụng kĩ thuật này vào xử lý ảnh và video số. Để dễ hình dung, có thể xem phép tích chập như một cửa sổ trượt áp lên một ma trận. Phép tính tích chập của ảnh I với bộ lộc (filter) K được thể hiện trong công thức 1-4

(1-4)

Ví dụ: Phép tích chập 2D: I(5x5) và K(3x3) với độ dài bước di chuyển stride =1, ta có kết quả khi K\*I

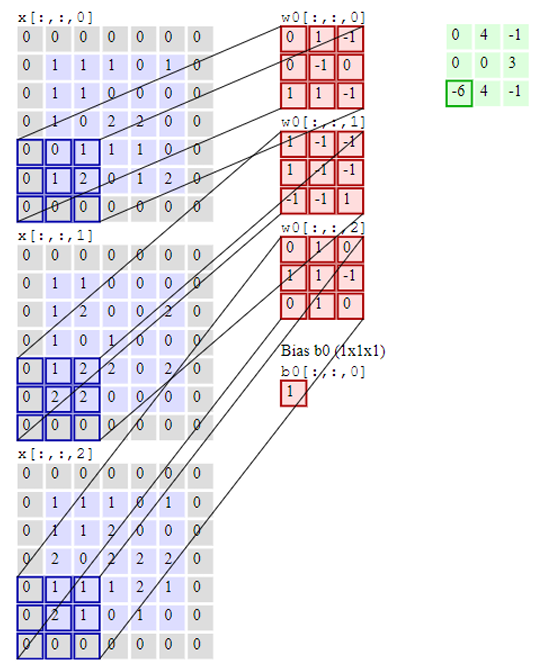


*Hình 1.11: Phép tích chập 2 chiều*

Trong trường hợp độ dài bước di chuyển dẫn đến việc thiếu hụt dữ liệu cho các nơ-ron ở vị trí ngoài cùng, có thể thêm vào các biên có giá trị 0. Kích thước của phép tích chập được tính bằng công thức 1-5

(1-5)

Ví dụ phép tích chập với ảnh 3D với I (5x5) và K(3x3) Padding:1 và Stride:2



*Hình 1.12: Phép tích chập 3 chiều*

Lấy ý tưởng từ việc các điểm ảnh nằm gần nhau sẽ có độ tương đồng (cường độ, màu sắc) nhiều hơn các điểm ảnh ở xa nhau, phép tích tích chập trên 𝐼 cho ta một kết quả là mối quan hệ giữa các phần từ nằm gần nhau trong 𝐼.

Vì phép tích chập phụ thuộc vào bộ lọc K nên khi thay đổi giá trị K thì kết quả đem lại sẽ khác nhau.

A close-up of a person's face

Description automatically generated

*Hình 1.13: Hình ảnh gốc bị biến đổi sau phép tích chập khác nhau*

### 1.2.2 Kiến trúc mạng tích chập

Mạng CNN được lấy ý tưởng từ việc sử dụng phép tính tích chập để rút trích đặc trưng và ngày càng phổ biến đem lại hiệu quả cao trong việc rút trích đặc trưng từ ảnh gốc và phân lớp chúng để nhận diện đối tượng, gương mặt và biểu cảm nhờ3 3 đặc điểm:

* Loại bỏ tính năng khai thác thủ công.
* Tự động rút trích đặc trưng nhờ các bộ lọc.
* Đem lại kết quả đáng công nhận cho bài toán nhận diện.
* Có thể xây dựng thêm đặc điểm nhận dạng mới từ mô hình đã có.

A close-up of a person's face

Description automatically generated

*Hình 1.14: Quá trình tự học đặc trưng và phân lớp*

Cấu trúc mạng CNN cũng giống như cấu trúc mạng nơ-ron thông thường tuy nhiên thay vì các layer được kết nối trực tiếp với nhau thông qua trọng sô w thì nó được liên kết với nhau thông qua cơ chế tích chập. Như vậy đầu ra của tích chập lớp này sẽ là đầu vào của lớp liền kề sau nó, mỗi layer sẽ rút trích ra những đặc trưng khác nhau tùy thuộc vào bộ lọc mà nó chọn. Các bộ lọc này có thể đi từ đơn giản đến phức tạp như biên của hình ảnh, đến viền gương mặt, vị trí mắt, miệng, chân mày. Kiến trúc cơ bản của mạng tích chậm được mô tả trong hình 1.15.

A diagram of a diagram of a complex structure

Description automatically generated with medium confidence

*Hình 1.15: Ví dụ mạng tích chập nhiều tầng*

**Tầng tích chập (Convolution):** nhằm mục đích trích xuất đặc trưng cơ bản, như các cạnh định hướng, điểm kết thúc, các góc và hình dạng nói chung thông qua phép tích chập. Các bộ lọc thực hiện tích chập dữ liệu tầng trước thành dữ liệu ở tầng kế tiếp này được gọi là feature map, các bộ lọc có tác dụng khác nhau sẽ tạo ra các feature map khác nhau. Những bộ lọc này có vai trò như trọng số.

Trong vấn đề nhận dạng khuôn mặt, các đặc trưng được phát hiện chủ yếu là hình dạng, góc cạnh của mắt, lông mày và môi. Một khi các đặc trưng được phát hiện, vị trí chính xác của nó không quan trọng nhưng khoảng cách với các đặc trưng khác quan trọng. Ví dụ, vị trí chính xác của lông mày không quan trọng, nhưng khoảng cách từ mắt đến chúng đáng chú ý vì khoảng cách lớn có thể chỉ ra sự biểu hiện bất ngờ. Vị trí chính xác này không chỉ không chỉ không liên quan mà còn có thể gây ra vấn đề, bởi vì nó có thể thay đổi một cách tự nhiên đối với các đối tượng khác trong cùng một biểu hiện.

**Tầng ReLu (Rectified linear unit):** giúp quá trình huấn luyện mạng nhanh và hiệu quả hơn vì nó bỏ qua các giá trị bằng 0 và giữ lại các giá trị có mức tác động.

**Tầng giảm (Pooling):** làm giảm độ phân giải không gian của của bộ lọc để tăng các bất biến cục bộ. Phép pooling tính giá trị đặc trưng của một vùng dữ liệu và lấy giá trị đặc trưng đó để đại diện. Có hai loại pooling phổ biến là max pooling - lấy giá trị lớn nhất trên nhiều nơ-ron và average pooling - lấy giá trị trung bình của nhiều nơ-ron.

**Tầng kết nối toàn diện (Fully connected):** gần giống với mạng nơ ron truyền thống, các nơ-ron kết nối hoàn toàn với lớp trước.

### 1.2.3 Quá trình huấn luyện mạng

Tiến trình học của mạng tích chập bao gồm việc tìm trọng số tốt nhất. Học tăng cường có thể biểu diễn bằng cách sử dụng phương pháp gradient descent.

Thuật toán lan truyền tiến và lan truyền ngược cho mạng tích chập sẽ khác so với mạng kết nối đầy đủ (khác ở các tầng tích chập và pooling) nhưng tư tưởng cơ bản của hai thuật toán này vẫn không đổi.

Trong suốt quá trình huấn luyện, CNN sẽ tự động học các thông số cho các filter.

# CHƯƠNG 2: BÀI TOÁN NHẬN DẠNG BIỂU CẢM

## 2.1 Các thách thức trong bài toán nhận dạng biểu cảm

Việc nhận dạng biểu cảm là việc mà con người thực hiện dễ dàng hàng ngày. Tuy nhiên, nó lại không hề đơn giản với máy tính, mặc dù các phương pháp gần dây với độ chính xác cao hơn 95% trong một số điều kiện (mặt người chính diện, môi trường chuẩn, và độ phân giải ảnh cao). Các phương pháp nghiên cứu chưa thống nhất phương pháp đánh giá (không sử dụng cùng bộ dữ liệu huấn luyện và kiểm thử) và do đó gây nhầm lẫn độ chính xác đạt kết quả cao. Mặt khác, độ chính xác thấp cũng có thể do dữ liệu với môi trường không kiểm soát và việc đánh giá chéo dữ liệu. Vượt qua những thử thách này vẫn là một thử thách cho máy tính vì nó thực sự khó để tách riêng biệt những đặc trưng biểu cảm của mặt. Những đặc trưng trên gương mặt của một đối tượng với hai biểu cảm khác nhau có thể rất gần với nhau, trong khi đặc trưng mặt người của 2 đối tượng với cùng biểu cảm lại có thể khác xa nhau. Thêm vào đó, vài biểu cảm như “buồn”, “sợ hãi” cũng như “khinh” và “ghét” trong một số trường hợp lại giống nhau.



*Hình 2.1: 3 đối tượng với biểu cảm vui vẻ. Các tấm hình khác nhau rất nhiều không những ở cách họ biểu cảm mà còn do tác động cua ánh sáng, góc chụp và nền. Các tấm ảnh được lấy từ bộ dự liệu FER-2013, JAFF và BU-3DFE*

Hình 2.1 cho thấy ba người với biểu cảm vui vẻ. Các tấm hình khác nhau rất nhiều không những ở cách họ biểu cảm mà còn do tác động của ánh sáng, góc chụp và nền. Hình ảnh này cũng minh họa cho một thử thách khác trong đánh giá biểu cảm người đó là việc không kiểm soát được các kịch bản trong quá trình huấn luyện và kiểm thử (hình ảnh huấn luyện có thể rất khác nhau về điều kiện môi trường, dân tộc). Một cách thức để đánh giá nhận dạng biểu cảm gương mặt theo đúng kịch bản là huấn luyện với cùng 1 bộ dữ liệu và kiểm tra với 1 bộ dữ liệu khác (có thể từ một dân tộc khác).

Rộng hơn bài toán phân lớp biểu cảm là bài toán xác định biểu cảm khuôn mặt vì con người ngoài 7 cảm xúc cơ bản còn có rất nhiều cảm xúc đa dạng khác.

## 2.2 Phân loại

Hệ thống nhận diện biểu cảm mặt người có thể được chia thành 2 loại:

* Ảnh tĩnh
* Ảnh động

Phương pháp dựa trên ảnh tĩnh không cần sử dụng thông tin thời gian, vector đặc trưng chỉ lưu thông tin thời điểm hiện tại ảnh đầu vào.

Mặt khác, phương pháp ảnh động sử dụng thông tin về thời điểm của ảnh để ghi lại biểu cảm của một hoặc nhiều khung (frame) ảnh. Hệ thống nhận diện biểu cảm tự động nhận đầu vào mong muốn (thường là ảnh động hoặc ảnh tĩnh) và trả về đầu ra là một trong 6 biểu cảm cơ bản (giận, buồn, ngạc nhiên, vui, ghét, hoặc sợ); một vài hệ thống cũng nhận dạng những biểu cảm trung tính.

## 2.3 Một số phương pháp nhận dạng biểu cảm khuôn mặt

Có nhiều phương pháp đã được nghiên cứu để giải bài toán về biểu cảm khuôn mặt, sử dụng các đơn vị vận động trên khuôn mặt (Action units –AU), sử dụng PCA, AAM kết hợp tương quan điểm, sử dụng các phương pháp học tăng cường khác … Mỗi phương pháp đều có ưu nhược điểm riêng, đối với phương pháp sử dụng mạng tích chập để huấn luyện như trong đồ án nghiên cứu cũng tương đối khó khăn và tốn kém vì cần nhiều người làm mẫu, những người này có có khả năng diễn đạt cảm xúc không tốt.

Phương pháp dựa trên đặc trưng ảnh: phân tích thành phần chính PCA, sau đó huấn luyện bằng các thuận toán học.

Phương pháp sử dụng Action Unnits: nhận dạng biểu cảm dựa trên chuyển động của khuôn mặt. Có tất cả 64 AU, mỗi AU là sự kết hợp của một số các cơ trên khuôn mặt như trong hình 2.2. Cảm xúc được nhận dạng bằng việc xác nhận tại một thời điểm có bao nhiêu AU xuất hiện cùng nhau tương ứng với 1 cảm xúc.

A collage of different facial expressions

Description automatically generated

*Hình 2.2 Ví dụ về một số AU trên tập dữ liệu FER-2013*

Phương pháp dùng mô hình AAM kết hợp tương quan điểm: Phương pháp này dùng mô hình AAM để phát hiện khuôn mặt sau đó dựa vào tỷ lệ giữa 2 mắt, lông mày, miệng, mũi … để nhận dạng biểu cảm.

Một kĩ thuật khác được áp dụng thành công trong vấn đề nhận diện biểu cảm gương mặt là mạng đa lớp sâu (Deep multi-layer neural network). Kĩ thuật này bao gồm 3 bước cho nhận dạng biểu cảm gương mặt (học, chọn và phân lớp) vào 1 bước duy nhất.

## 2.4 Các bước thực hiện bài toán nhận diện biểu cảm

Mô hình tổng quan nhận dạng biểu cảm gương mặt có thể hiểu như sau:

A white arrow with black text

Description automatically generated

*Hình 2.3 Mô hình tổng quan nhận dạng biểu cảm*

Phân tích biểu cảm gương mặt tự động gồm ba bước:

*A grey rectangular sign with black text

Description automatically generated*

*Hình 3-4 Các bước phân tích biểu cảm gương mặt*

Thu thập có thể được chia thành 2 bước lớn như sau: định dạng gương mặt và định vị phần đầu.

Hệ thống nhận diện biểu cảm gương mặt cơ bản sử dụng 3 trạng thái huấn luyện: học đặc trưng, chọn đặc trưng và xây dựng phân lớp. Việc học đặc trưng đảm nhận khai thác hầu hết các đặc trưng liên quan đến biểu cảm gương mặt. Quá trình chọn đặc trưng chọn những đặc trưng tốt nhất để thể hiện biểu cảm gương mặt. Để tăng tính chính xác của bài toán nhận diện biểu cảm cần hạn chế biến thể nội bộ của biểu cảm trong khi tăng cường biến thể liên bộ. Việc làm nhỏ biến thể nội bộ của biểu cảm là cả một vấn đề vì hình ảnh của những cá thể khác nhau với cùng một biểu cảm khác xa nhau cho mỗi điểm ảnh. Việc mở rộng các biến thể liên bộ cũng khó vì hình ảnh của cùng một người với các biểu cảm khác nhau có thể rất gần với một người khác. Cuối tiến trình, các đặc trưng được chọn để sử dụng vào việc phân lớp suy luận biểu cảm khuôn mặt.

## 2.5 Các nghiên cứu gần đây

Những nghiên cứu về nhận dạng biểu cảm gương mặt được phát triển trong thập niên gần đây với việc cải tiến tiến độ trong hiệu suất nhận dạng. Một phần quan trọng của tiến trình nghiên cứu gần đây là sự nổi lên của các phương pháp học sâu đặc biệt mạng tích chập. Những hướng nghiên cứu này bắt đầu khả thi do: dữ liệu hiện nay hầu như lớn để áp dụng phương pháp học huấn luyện và sử dụng kĩ thuật GPU. Số lượng dữ liệu lớn rất quan trọng cho việc huấn luyện mạng với kiến trúc sâu, trong khi yếu tố thứ hai quan trọng đối với việc tính toán hiệu suất cao đòi hỏi chi phí thấp cho quá trình huấn luyện.

Một số hướng nghiên cứu khác lại tập trung vào môi trường không thể kiển soát được (mặt không chính diện, hình ảnh chồng chéo, biểu hiện tự phát và những vấn đề khác), đây vẫn là những vấn đề có nhiều thách thức. Hướng nghiên cứu này tập trung vào kiểm soát môi trường và đánh giá các nhóm dân tộc khác nhau.

Một phương pháp khác tạo ra danh sách phân lớp, được các tác giả đặt tên là các phân lớp yếu. Mỗi phân lớp yếu chị trách nhiệm cho một phân lớp biểu cảm gọi là Boosted deep belief network (BDBN) [10]. Quá trình học qua ba bước (học đặc trưng, chọn đặc trưng và xây dựng phân lớp) trong một lớp duy nhất. Các thực nghiệm được thực hiện qua hai bộ dữ liệu tĩnh công khai Cohn-Kanade [9] và JAFFE và đạt được độ chính xác lần lượt là 96.7% và 91.8%. Tất cả hình ảnh được xư lý dựa vào tọa độ mắt đã cho, canh chỉnh và cắt tỉa. Quá trình đào tạo và kiểm tra được thực hiện bằng phân lớp nhị phân. Thời gian để huấn luyện mạng là tám ngày.

Một hệ thống nhận diện biểu năm cảm gương mặt (giận, vui, buồn, ngạc nhiên và trung tính) sử dụng CNN và thực thi trên điện thoại thông minh [6]. Đề xuất mạng bao gồm 5 lớp và 65.000 tế bào thần kinh. Một số kĩ thuật để tăng số lượng mẫu huấn luyện và phương pháp drop-out đã được sử dụng trong quá trình huấn luyện mạng. Các thực nghiệm cũng được thực hiện bộ dữ liệu FER-2013 với độ chính xác đạt 99.2% (giả định có sự chồng chéo các đối tượng trong tập huấn luyện) và một số bộ dữ liệu khác được tạo ra bởi tác giả.

Kiến trúc mạng CNN mới được đề xuất bởi Burket [8] độc lập hoàn toàn với các đặc trưng hình ảnh gốc đầu vào (hand-craffed) có bốn phần. Phần thứ nhất chịu trách nhiệm cho việc tiền xử lý dữ liệu tự động, trong khi những phần còn lại tham gia vào quá trình trích xuất đặc trưng. Các đặc trưng trích xuất được phân lớp bởi một kết nối hoàn chỉnh cuối mạng. Kiến trúc được đề xuất bao gồm 15 lớp (7 lớp tích chập, 5 poolings, 2 lớp kết nối và 1 lớp thông thường). Họ đánh giá phương pháp với CK+ và MMI, tỉ lệ chính xác đạt được là 99.6% và 98.63%. Mặc dù độ chính xác cao, phương pháp của họ không đảm bảo rằng các đối tượng được sử dụng trong quá trình huấn luyện không được sử dụng trong quá trình kiểm tra. Như đã thảo luận trong phần 4, đây là một giới hạn quan trọng cần được thực thi để thực hiện một đánh giá hợp lý các phương pháp nhận diện khuôn mặt.

Liu và cộng sự cũng đề xuất một mạng học sâu, trong đó biểu cảm khuôn mặt được cho rằng nó được phân hủy từ các đơn vị hành động (AU) trên khuôn mặt [7]. Phương pháp này dùng để học sự biến đổi thông tin cục bộ, tối ưu kết hợp các biến đổi cục bộ và miêu tả cụ thể hơn việc nhận dạng biểu cảm. Các thực nghiệm đã được thực hiện bằng cách sử dụng các phướng pháp kiểm tra chéo không trùng lắm giữa các nhóm huấn luyện và kiểm tra để đánh giá sáu biểu cảm cơ bản. Phương pháp này đạt độ chính xác cao với bộ dữ liệu FER-2013 là 93.7%.

Một bộ sưu tập các mạng nơ-ron tăng cường cho việc nhận biết biểu hiện khuôn mặt đa sắc tộc [5] được đề xuất bởi Ali. Mô hình được đề xuất bao gồm ba bước: đầu tiên là tập hợp các mạng nơ-ron nhị phân được huấn luyện, thứ hai là dự đoán kết hợp tích hợp các mạng nơ-ron này thành một bộ duy nhất và cuối cùng là sử dụng các bộ để phát hiện sự có mặt của một biểu hiện. Cơ sở dữ liệu biểu hiện khuôn mặt đa sắc tộc được tạo ra bởi các tác giả với hình ảnh từ ba CSDL khác nhau, chứa hình ảnh người Nhật (JAFFE), Đài Loan (TFEID), người Caucasians (RAFD) và người Ma rốc. Các tác giả bào báo nhận ra năm biểu hiện (tức giận, vui, buồn, bất ngờ và sợ hãi) theo hai cách tiếp cận khác nhau. Trong lần đầu tiên, họ huấn luyện và đánh giá hệ thống trong CSDL sắc tộc đạt được độ chính xác cao 93.75%. Thực nghiệm thứ hai được thực hiện để đánh giá phương pháp được đề xuất trong môi trường ít được kiểm soát. Phương pháp được huấn luyện với hai CSDL (TFIID và RaFD) và đánh giá trong JAFF đạt độ chính xác 48.67%.

Shan và cộng sự [4] đã thực hiện một nghiên cứu sử dụng mô hình nhị phân cục bộ (Local binary patterns-LBP) để trích xuất đặc trưng. Họ đã kết hợp và so sánh các kỹ thuật máy học khác nhau như kết hợp mẫu, support vector machine (SVM), Linear discriminant analysis và lập trình tuyến tính để nhận dạng các biểu cảm trên gương mặt. Các tác giả cũng tiến hành nghiên cứu để phân tích tác động của độ phân giải hình ảnh trong đánh giá chính xác và kết luận rằng các phương pháp dựa trên các tính năng hình học không xử lý hình ảnh có độ phân giải kém, trong khi những người dựa trên hình dáng, như biểu đồ Gabor và LBP, không nhạy cảm với độ phân giải hình ảnh. Kết quả đạt được tốt nhất trong nghiên cứu là 95.1% khi sử dụng SVM và LBP với bộ dữ liệu CK+. Sử dụng biến đổi dữ liệu chéo (huấn luyện với CK+ và kiểm tra với JAFFE) để đánh giá hệ thống đề xuất với kịch bản ít kiểm soát, đạt được chính xác 41.3%. Hình ảnh đầu tiên được cắt tỉa dựa trên vị trí mắt. Thiết lập thực nghiệm sử dụng kế hoạch kiểm tra chéo 10 lần mà không trùng nhau. Việc huấn luyện và thời gian nhận dạng đã không đượcc đề cập.

Hệ thống nhận diện biểu cảm gương mặt dựa trên video được đề xuất bởi Byeon và Kwak [12]. Họ phát triển mô hình 3D-CNN với 1 chuỗi hình ảnh (từ trung tính đến biểu hiện cuối cùng) sử dụng 5 khung hình liên tiếp như đầu vào 3D. Do đó, đầu vào CNN là H\*W\*5 (với H và W là độ cao và bề rộng của hình ảnh, và 5 là số khung). Các tác giả cho rằng phương pháp 3D-CNN có thể xử lý một số mức độ thay đổi và biến dạng bất biến. Với cách tiếp cận này, chúng đạt được độ chính xác 95%, nhưng phương pháp dựa vào một trình tự có chứa toàn bộ chuyển động từ trung tính đến biểu hiện. Thực nghiệm tiến hành với 10 người chỉ dựa trên số liệu thông thường. Việc huấn luyện và thời gian nhận dạng không được đề cập đến.

Một hướng nghiên cứu khác dựa trên video được đề xuất bằng việc sử dụng khung thời gian-không gian dựa trên biểu đồ gradient và dòng chảy quang học [11]. Phương pháp tích hợp ba bước: tiền xử lý, trích xuất đặc trưng và phân lớp. Trong bước tiền xử lý, việc phát hiện vùng trên khuôn mặt đã được thực hiện và canh chỉnh khuôn mặt đã được thực hiện (để giảm biến đổi đầu ra). Trong giai đoạn khai thác đặc trưng, một khuôn khổ tích hợp các thông tin động lấy ra từ biến thể trong khuôn mặt do các biểu thức đã được sử dụng. Trong giai đoạn cuối- phân lớp, một phân lớp SVM với một RBF kernel được sử dụng. Các thực nghiệm được thực hiện bằng CSDL FER-2013 và MMI. Độ chính xác đạt được của tác giả trong CSDL FER-2013 cho 7 biểu cảm là 83.7% và CSDL MMI là 74.3%. Thời gian đào tạo không được đề cập, trong khi thời gian nhận diện biểu cảm khoản 350 ms trên mỗi hình ảnh trong CSDL FER-2013 và 520 ms trong CSDL MMI.

Phương pháp được đề xuất bởi André Teixeira Lopes [3] trong bài báo nghiên cứu có tính chính xác cao trong cơ sở dữ liệu FER-2013 và JAFFE gần 100% mà không có sự trùng lắp dữ liệu giữa tập huấn luyện và kiểm tra; thời gian huấn luyện, đánh giá khá nhỏ.

|  |  |  |
| --- | --- | --- |
| **Phương pháp** | **Độ chính xác** | **Nhận xét** |
| BDBN [10] | 96.7% |  |
| CNN trên Smartphone [6] | 99.2% | Chồng chéo các đối tượng |
| CNN: 15 lớp [8] | 99.6% | Chồng chéo các đối tượng |
| Bộ sưu tập các mạng nơ-ron [5] | 93.75% |  |
| LBP-SVM [4] | 95.1% |  |
| 3D-CNN [12] | 95% |  |
| Dòng chảy quang học video [11] | 83.7% |  |
| Phương pháp nghiên cứu [3] | ≈100% |  |

*Bảng 1: Bảng so sánh độ chính xác giữa các phương pháp nghiên cứu*

# CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU

Trong chương này, nội dung chính của tiểu luận – Phương pháp tiền xử lý làm tăng kích thước cơ sở dữ liệu trong huấn luyện và kiến trúc mạng huấn luyện CNN để làm tăng tính chính xác của bài toán nhận dạng biểu cảm khuôn mặt.

Hệ thống nhận diện biểu cảm trong nghiên cứu biểu diễn ba bước học chỉ trong một phân lớp (CNN). Hệ thống đề xuất hoạt động trong hai bước chính: huấn luyện và kiểm tra. Trong suốt quá trình huấn luyện, hệ thống nhận dữ liệu huấn luyện bao gồm hình ảnh trắng đen của gương mặt với biểu cảm tương ứng với vị trí trung tâm mắt và trọng số học cho mạng. Trong suốt quá trình kiểm tra, hệ thống nhận hình ảnh trắng đen của gương mặt cùng với vị trí trung tâm mắt tương ứng của nó và đưa ra các biểu hiện dự đoán bằng cách sử dụng trọng số mạng cuối cùng đã học được trong quá trình huấn luyện.

Tổng quan về hệ thống được minh họa trong hình 3.1. Trong giai đoạn huấn luyện, hình ảnh mới được tạo ra để tăng kích thước cơ sở dữ liệu. Sau đó, một điều chỉnh quay được thực hiện để canh thẳng mắt theo trục ngang. Sau đó, hình ảnh được cắt loại bỏ thông tin nền và chỉ giữ lại các biểu hiện đặc trưng. Tiến trình lấy mẫu xuống được thực hiện để lấy đặc trưng ở một hình ảnh khác cho cùng vị trí. Sau đó, chuẩn hóa cường độ hình ảnh. Chuẩn hóa hóa hình ảnh được sử dụng để huấn luyện mạng CNN. Đầu ra của giai đoạn huấn luyện là một danh sách trọng số của lượt đạt kết quả tốt nhất với dữ liệu xác nhận sau một vài vòng huấn luyện xem xét dữ liệu theo thứ tự khác nhau. Giai đoạn kiểm tra sử dụng phương pháp tương tự như giai đoạn huấn luyện: Chuẩn hóa không gian, cắt xén, lấy down-sampling và chuẩn hóa cường độ. Đầu ra của nó là một số duy nhất – id- của một trong 7 biểu thức cơ bản. Các biểu thức được biểu diễn dưới dạng số nguyên (1-Giận, 2-Khinh, 3-Ghét, 4-Sợ, 5-Vui, 6-Buồn, 7-Ngạc nhiên)

A diagram of a network

Description automatically generated

*Hình 3.1: Tổng quan hệ thống nhận diện biểu cảm*

## 3.1 Khởi tạo mẫu tổng hợp (Synthetic sample generation)

Một trong những vấn đề chính của phương pháp học tập sâu là cần rất nhiều dữ liệu trong giai đoạn đào tạo để thực hiện huấn luyện. Thật không may, số lượng dữ liệu có sẵn trong bộ dữ liệu công cộng không đủ để thi hành trong ứng dụng. Để giải quyết vấn đề này ta thực hiện tăng cường dữ liệu bằng cách tạo ra các hình ảnh tổng hợp (tức là các hình ảnh thực với các phép quay nhân tạo, các phương pháp dịch chuyển và co giãn).

Theo ý tưởng này, bài báo cáo sử dụng các điểm mắt (trái và phải) được tạo ra từ mô hình AAM để tạo ra nhiễu ngẫu nhiên tại các vị trí xung quanh trung tâm mắt. Hình ảnh tổng hợp được tạo ra bằng cách gây nhiễu với 7 giá trị điểm cho mỗi mắt (kể cả trung tâm mắt) ta được 49 hình ảnh tổng hợp bổ sung được tạo ra.

Như có thể thấy trong hình 3.2, điểm được tạo ra cho hai mắt sau sự kết hợp các điểm xung quanh vị trí trung tâm ban đầu. Vị trí trung tấm mắt được tạo ra bằng cách tính trọng tâm các điểm lân cận. Vì các giá trị mới được đưa ra cho quy trình chuẩn hóa không phải là trung tâm mắt thực, nên các hình ảnh kết quả sẽ bị làm nhiễu bởi một một chuyển đổi, một vòng quay/ một co giãn hoặc không bị gì hết sau tất cả. Mặt khác, độ lệch lớn cho mỗi mắt có thể gây ra quá nhiều sự dịch chuyển, quay hoặc co giãn trong các hình ảnh làm cho kịch bản phức tạp hơn cho việc tìm các đặc trưng phân lớp.

A blue dotted line with numbers

Description automatically generated

*Hình 3.2: Minh họa khởi tạo mẫu từ các điểm đặc trưng mắt*

*A collage of pictures of a person

Description automatically generated*

*Hình 3.3: Minh họa khởi tạo mẫu tổng hợp*

## 3.2 Phép quay (Rotation correction)

Các hình ảnh trong CSDL, thậm chí là môi trường thực thi, thay đổi xoay, độ sáng và kích thước, ngay cả đối với hình ảnh của cùng một đối tượng. Những biến thể này không liên quan nhiều đến biểu hiện khuôn mặt nhưng có thể ảnh hưởng đến tỷ lệ chính xác của hệ thống. Để giải quyết vấn đề này, vùng khuôn mặt được canh chỉnh (với chuẩn) với phương ngang và điểm trung tâm để sửa các vấn đề hình học có thể xảy ra như phép quay và phép chuyển đổi. Để thực hiện căn chỉnh này cần phải có hai thông tin hình ảnh khuôn mặt và trung tâm của hai mắt. Đã có rất nhiều phương pháp trong tài liệu có thể tìm thấy đôi mắt, và những điểm khác của gương mặt với độ chính xác cao và đây không phải trọng tâm của bài báo cáo.

Để thực hiện việc căn chỉnh mặt, một phép biến đổi quay được áp dụng để căn chỉnh mắt với trục ngang của hình ảnh và một dịch chuyển affine được xác định bởi các vị trí của mắt để tập trung vào một điểm cụ thể của mặt. Phép quay làm cho góc được hình thành bởi đoạn thẳng từ trung tâm mắt này đến điểm kia, và trục ngang là 0. Quay và chuyển đổi trong những hình ảnh không liên quan đến biểu cảm trên khuôn mặt do đó cần được loại bỏ để tránh tiêu cực đến độ chính xác của hệ thống. Thủ tục hiệu chỉnh xoay được thể hiện ở hình 3.3.

A diagram of a person's face

Description automatically generated

*Hình 3.4: Ví dụ hiệu chỉnh quay*

Các đầu vào cho quy trình này có thể là một hình ảnh nguyên bản hoặc đã được tổng hợp. Việc điều chỉnh quay cho tổng hợp hình ảnh, có thể không thực hiện được canh chỉnh tốt hoàn hảo với trục ngang vì trung tâm mắt là các vị trí thực bị nhiễu bởi điểm ngẫu nhiên. Do đó, nó sẽ tạo ra hình ảnh bị xáo trộn bởi phép quay và phép chuyển đổi, làm tăng biến thể trong các ví dụ huấn luyện.

## 3.3 Cắt tỉa hình ảnh (Image Cropping)

Như trình bày trong hình 3.1, hình ảnh ban đầu có rất nhiều thông tin cơ bản đã làm là không quan trọng đối với các quá trình phân lớp biểu cảm. Những thông tin này có thể làm giảm độ chính xác của việc phân lớp vì các lớp có thêm nhiều khó khăn để giải quyết, ví dụ: phân biệt giữa lớp nền và lớp làm nổi. Sau khi cắt tỉa, tất cả các phần hình ảnh đã không có thông tin biểu cảm cụ thể được loại bỏ. Do đó, khu vực quan tâm được xác định dựa trên một tỷ lệ khoảng cách giữa mắt. Kết quả là phương pháp có khả năng xử lý những người khác nhau và kích thước hình ảnh mà không cần sự can thiệp của con người. Vùng cắt được giới hạn bởi một nhân tố dọc là 4.5 (xem như 1.3 cho vùng phía trên mắt và 3.2 đối với khu vực bên dưới). Áp dụng cho khoảng cách trung bình giữa hai điểm mắt và trung tâm phải. Vùng cắt ngang được giới hạn bởi yếu tố của 2.4 áp dụng khoảng cách tương tự. Những giá trị yếu tố được xác định khai thác thực nghiệm. Một ví dụ về quy trình này được minh họa trong hình 3.4.

*A collage of a person's face

Description automatically generated*

*Hình 3.5: Minh họa cắt tỉa hình ảnh*

## 3.4 Lấy mẫu xuống (Down-sampling)

Lấy mẫu xuống là hoạt động được thực hiện để giảm kích thước hình ảnh cho mạng và để đảm bảo bình thường hóa quy mô, ví dụ: cùng một vị trí cho các thành phần khuôn mặt (mắt, miệng, lông mày, vv) trong tất cả các hình ảnh. Lấy mẫu xuống sử dụng phương pháp nội suy tuyến tính. Sau khi lấy mẫu, người ta có thể đảm bảo trung tâm mắt xấp xỉ vị trí trung tâm. Tiến trình này giúp CNN có thể học những vùng có liên quan đến biểu hiện đặc biệt. Những hình ảnh cuối cùng được lấy mẫu xuống, giống như một suy tuyến tính, có kích thước 32x32 pixel.

## 3.5 Chuẩn hóa cường độ (Intensity normalization)

Độ sáng và độ tương phản của hình ảnh có thể khác nhau thậm chí trong hình ảnh của cùng một người trong cùng một biểu cảm, do đó làm tăng sự biến thiên của vector đặc trưng. Các biển thể như vậy làm tăng tính phức tạp của vấn đề trong khi phân lớp chỉ giải quyết vấn đề biểu cảm gương mặt. Để giảm bớt những vấn đề này, một chuẩn mực cường độ đã được áp dụng.

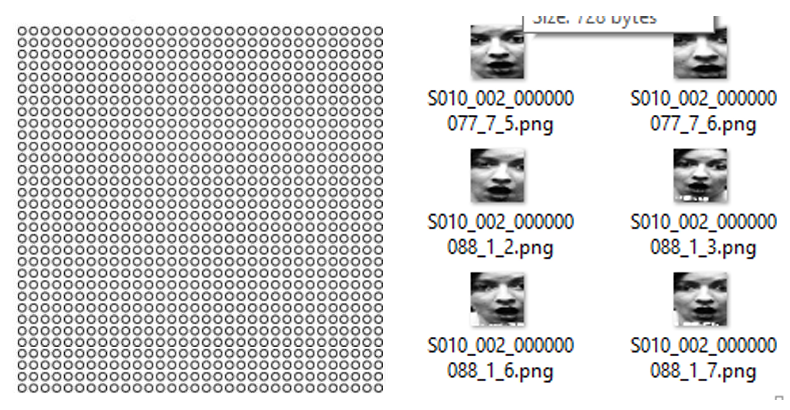
A close-up of a person's face

Description automatically generated

*Hình 3.6: Minh họa phép chuẩn hóa cường độ*

## 3.6 Kiến trúc mạng CNN

Cấu trúc mạng CNN được trình bày trong hình 3.13. Đầu vào của mạng là một ảnh xám 32x32 và đầu ra là biểu cảm. Lớp có giá trị lớn nhất được sử dụng như biểu cảm trong hình ảnh. Kiến trúc CNN bao gồm 2 lớp tích chập, 2 lớp lấy mẫu phụ và một lớp kết nối hoàn chỉnh.

**

*Hình 3.7: Đầu vào của mạng*

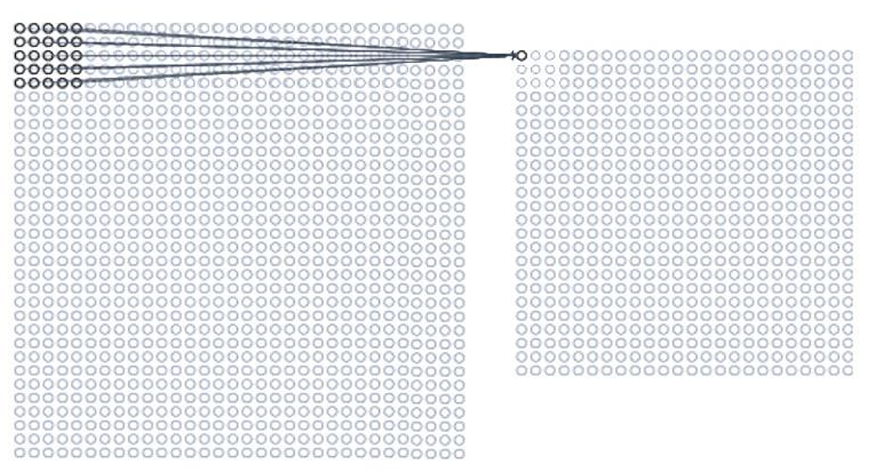
Lớp đầu tiên của CNN là một lớp tích chập, sử dụng tích chập kernel 5x5 và xuất ra 32 hình ảnh có 28x28 điểm ảnh.

*A diagram of a number of circles

Description automatically generated*

*Hình 3.8: Hình ảnh đầu vào kết hợp tích chập 5x5*

Như thường lệ, các điểm đầu vào của ảnh sẽ kết nối với tầng ẩn, nhưng thay vì kết nối mỗi điểm ảnh đầu vào với mỗi nơ-ron tầng ẩn, ta chỉ kết nối trong phạm vi nhỏ có kích thước 5x5 như hình 3.7. Sau đó một cửa sổ trượt feature map trên toàn bộ bức ảnh, mỗi lần trượt sẽ thu về một giá trị nơ-ron tương ứng.

**

*Hình 3.9: Minh họa trượt cửa sổ kernel lần 1*

*A close-up of a needle

Description automatically generated*

*Hình 3.10: Minh họa trượt cửa sổ kernel lần 2*

Như vậy, từ một ảnh đầu vào có kích thước 32x32 và 1 kernel có kích thước 5x5 ta sẽ có 28x28 nơ-ron trong lớp ẩn.

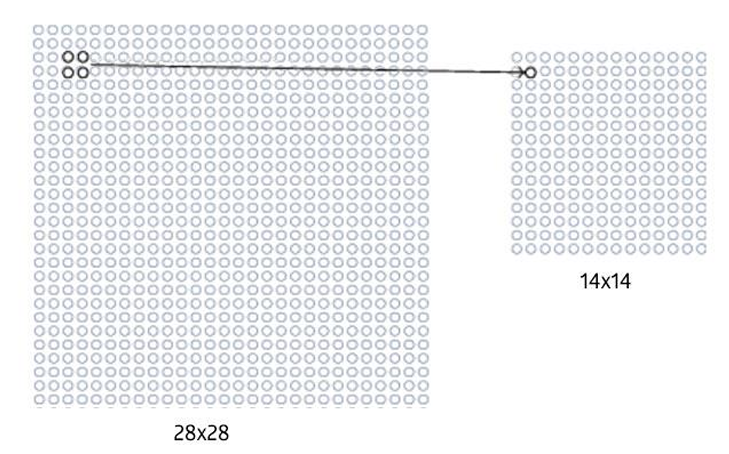
Để nhận dạng được biểu cảm ta cần nhiều hơn một bản đồ đặc trưng, vì vậy một tích chập cần nhiều bản đồ đặc trưng, trong trường hợp này là 32 bản đồ đặc trưng.

*A diagram of a number of squares

Description automatically generated*

*Hình 3.11: Qua tích chập ta tạo được 32x28x28 nơ-ron*

Lớp này được theo sau bởi một lớp mẫu nhỏ sử dụng max-pooling (kích thước kernel 2x2) để giảm hình ảnh xuống một nửa kích thước của nó. Như vậy, từ 28x28 nơ-ron đầu ra cho mỗi feature map từ lớp tích chập trước, sau khi pooling nhận được nơ-ron có kích thước 14x14.

**

*Hình 3.12: Kích thước ảnh đầu vào sau lớp tích chập và giảm chiều*

Sau đó, một lớp tích chập mới thực hiện 64 tích chập với kernel 7x7 để ánh xạ lớp trước và lớp mẫu phụ với kernel 2x2. Các đầu ra được cấp cho một lớp ẩn kết nối đấy đủ với 256 nơ-ron. Cuối cùng, mạng có 7 đầu ra (mỗi đầu ra đại diện cho 1 biểu cảm), các đầu ra này được kết nối hoàn toàn với lớp trước đó.

*A diagram of a diagram

Description automatically generated*

*Hình 3.13: Kiến trúc mạng đề xuất*

Lớp đầu tiên của mạng (lớp tích chập) nhằm mục đích trích xuất đặc trưng cơ bản, như các cạnh định hướng, điểm kết thúc, các góc và hình dạng nói chung được mô tả bởi Lecun. Trong vấn đề nhận dạng khuôn mặt, các đặc trưng được phát hiện chủ yếu là hình dạng, góc cạnh của mắt, lông mày và môi. Một khi các đặc trưng được phát hiện, vị trí chính xác của nó không quan trọng nhưng khoảng cách với các đặc trưng khác quan trọng. Ví dụ, vị trí chính xác của lông mày không quan trọng, nhưng khoảng cách từ mắt đến chúng đáng chú ý vì khoảng cách lớn có thể chỉ ra sự biểu hiện bất ngờ. Vị trí chính xác này không chỉ không liên quan mà còn có thể gây ra vấn đề, bởi vì nó có thể thay đổi một cách tự nhiên đối với các đối tượng khác trong cùng một biểu hiện.

Lớp thứ hai (lớp lấy mẫu phụ) làm giảm độ phân giải không gian của bản đồ đặc trưng.

Hai lớp kế tiếp, một tích chập một mẫu phụ, nhằm thực hiện các hoạt động tương tự như các thao tác đầu tiên, nhưng xử lý các tính năng thấp hơn, nhận ra các phần tử theo ngữ cảnh thay vì hình dạng đơn giản, cạnh và góc. Sự ghép nối các tập hợp các lớp chập lại và mẫu phụ đạt được mức độ bất biến cao đối với chuyển đổi hình học đầu vào.

Lớp ẩn cuối cùng (một lớp kết nối hoàn chỉnh) nhận được tập hợp các đặc trưng đã học được và cho ra mức độ tin cậy của các đặc trưng nhất định trong mỗi một trong những biểu hiện đã được xem xét.

Mạng này sử dụng phương pháp SGD để tính toán trọng lượng synaptic giữa các nơ-ron. Giá trị ban đầu của các khớp thần kinh cho các tích chập và cho các lớp kết nối đầy đủ được tạo ra bằng cách sử dụng các bộ lọc tự động khởi tạo dựa trên số lượng nơ-ron đầu vào và đầu ra. Hàm mất mát được tính bằng cách sử dụng một hàm logistic của đầu ra soft-max.

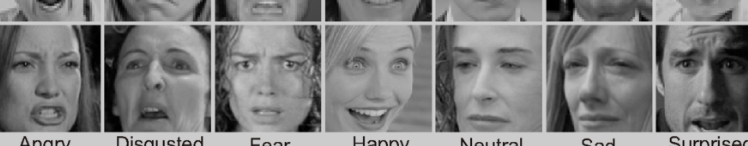
# CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM

## 4.1 Môi trường thực nghiệm

Để triển khai huấn luyện và thử nghiệm mô hình đề xuất, ngôn ngữ Python và thư viện Keras/TensorFlow được sử dụng cho việc xây dựng mô hình mạng CNN. Dữ liệu FER-2013 được tiền xử lý không đáng kể, ngoại trừ việc chuyển đổi đa cấp xám từ dạng số nguyên 0 đến 255 về miền số thực [0, 1] nhằm hỗ trợ tốt hơn cho dữ liệu đầu vào của mạng tích chập. Tất cả các thực nghiệm được thực hiện bằng cách sử dụng một bộ xử lý Ryzen 7 3.2 GHz. Môi trường của các thực nghiệm là Window 11.

## 4.2 Cơ sở dữ liệu

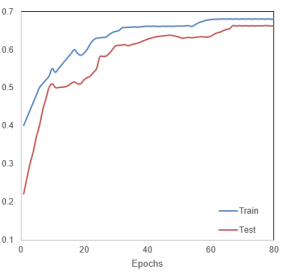
Dữ liệu FER-2013 được công bố bởi trang Kaggle trong khuôn khổ workshop của hội thảo ICML 2013. Dữ liệu gồm các ảnh đa cấp xám cỡ 48x48 chỉ gồm khuôn mặt hầu như được căn giữa ảnh và tỉ lệ khuôn mặt được điều chỉnh chiếm phần lớn diện tích của ảnh. Một ảnh sẽ được gán nhãn nằm một trong bảy loại cảm xúc giá trị từ 0 đến 6 (0: giận dữ, 1: căm phẫn, 2: sợ hãi, 3: hạnh phúc, 4: buồn rầu, 5: bất ngờ, 6: trung lập). Bộ dữ liệu này gồm 28.709 mẫu huấn luyện, mẫu kiểm tra công khai có 3.589 ảnh. Khi thực hiện đánh giá mô hình, Kaggle sẽ sử dụng một bộ kiểm tra khác cũng có 3.589 ảnh, vì vậy kết quả đánh giá của ban giám khảo có thể có sai lệch so với sử dụng bộ test công khai, một số trường hợp đặc biệt sai lệch có thể lên đến 5% [4]. Chúng tôi sử dụng bộ dữ liệu này cho mô hình thử nghiệm vì bộ dữ liệu có số mẫu khá lớn, phù hợp với việc huấn luyện với mạng học sâu, vốn đòi hỏi nhiều mẫu hơn các phương pháp học máy thông thường. Ngoài ra, bộ dữ liệu được cấu trúc dễ dàng xử lý bởi thư viện Keras/TensorFlow và có nhiều kết quả đối chứng khi thực hiện so sánh mô hình của chúng tôi với các kết quả của những nhóm nghiên cứu khác.

**

*Hình 4.1: Dữ liệu cảm xúc khuôn mặt*

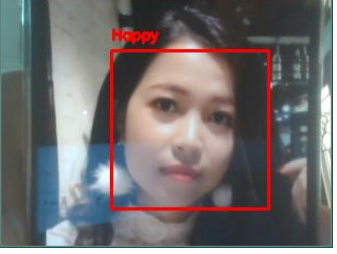
## 4.3 Kết quả thực nghiệm:

Mô hình được huấn luyện với epochs = 100, tuy nhiên kết quả về độ chính xác trên tập huấn luyện và tập kiểm tra gần như ổn định sau bước 70 khi kiểm nghiệm thực tế. Độ chính xác trên tập dữ liệu kiểm tra không bị giảm sau khi mạng đã ổn định, như vậy có thể thấy mô hình không bị hiện tượng quá khớp. Muốn tăng độ chính xác của mô hình, chúng tôi điều chỉnh phù hợp về số filter trên mỗi lớp và có thể tăng thêm một số lớp ẩn trong mạng CNN nhằm tăng khả năng nhận biết các cấu trúc phức tạp trên khuôn mặt.



*Hình 4.2:* Biến động về độ chính xác của mô hình trên tập huấn luyện và tập kiểm tra theo số lượt huấn luyện

Kết quả thử nghiệm thực tế cho thấy mô hình khá nhạy khi nhận biết cảm xúc hạnh phúc (happy), khá kém với cảm xúc căm phẫn (disgust). Việc hầu hết các mô hình được công bố với tập dữ liệu FER-2013 đều chỉ đạt độ chính xác thấp (dưới 70%), điều này có thể cho thấy bộ dữ liệu này có những yếu tố mất cân bằng hoặc nhiễu khi gán nhãn dữ liệu.

**

*Hình 4.3:* Một kết quả phát hiện cảm xúc khuôn mặt

# KẾT LUẬN

Với sự nỗ lực của cả nhóm và đặc biệt là sự giúp đỡ nhiệt tình của thầy giáo TS. Trần Quang Quý đã giúp nhóm em hoàn thành được báo cáo “Sử dùng mạng CNN trong bài toán nhận diện cảm xúc khuôn mặt”. Trong quá trình thực hiện nhóm em đã thu được những kết quả sau:

Tìm hiểu về bài toán nhận diện biểu cảm khuôn mặt, khảo sát các phương pháp nhận diện khuôn mặt đã có.

Tìm hiểu nghiên cứu về mạng nơ-ron CNN (Convolution Neuron Network). Tìm hiểu áp dụng kỹ thuật mạng CNN cho bài toán nhận diện biểu cảm khuôn mặt, từ đó phân tích để tối ưu một vài tham số của CNN sao cho có thể tăng tốc trên máy đơn CPU.

Cải tiến khâu tiền xử lý trên ảnh mặt người nhằm tăng hiệu quả việc phân lớp cảm xúc dùng CNN.

Thử nghiệm trên cơ sở dữ liệu chuẩn để chỉ ra tính khả thi của mô hình kết hợp (khâu tiền tử lý và mạng CNN).

Mặc dù đã rất nỗ lực, cố gắng nhưng do thời gian và kiến thức có hạn nên đề tài còn nhiều sai sót, hạn chế khiến cho mô hình xây dựng chưa đạt được kết quả cao. Em xin nhận được sự thông cảm và góp ý của thầy cô trong ban cố vấn và các thầy cô trong hội đồng, để báo cáo của nhóm em được hoàn thiện hơn nữa. Em xin chân thành cảm ơn.

Sinh Viên

Hoàng Trung Kiên

Trần Hoài Nam

Nguyễn Minh Tuyền

Nguyễn Trường Xuân

# TÀI LIỆU THAM KHẢO

[1] Lê Thị Thu Hằng: “Nghiên cứu về mạng neural tích chập và ứng dụng cho bài toán nhận dạng biển số xe”. Luận án thạc sĩ, Đại học Công nghệ, Hà Nội, 2016.

[2] Nguyễn Việt Nhật, Trương Lê Minh Nhựt: “Tăng tốc quá trình huấn luyện mạng nơ-ron sâu bằng Batch Normalization”, Khóa luận. Đại học Khoa học Tự nhiên, Hồ Chí Minh, 2017.

[3] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816

[4] G. Ali, M.A. Iqbal, T.-S. Choi, Boosted NNE collections for multicultural facial expression recognition, Pattern Recognit. <http://dx.doi.org/10.1016/j.patcog.2016.01.032>

[5] I. Song, H.-J. Kim, P.B. Jeon, Deep learning for real-time robust facial expression recognition on a smartphone, in: International Conference on Consumer Electronics (ICCE), Institute of Electrical & Electronics Engineers (IEEE), Las Vegas, NV, USA, 2014 54

[7] M. Liu, S. Li, S. Shan, X. Chen, Au-inspired deep networks for facial expression feature learning, Neurocomputing <http://dx.doi.org/10.1016/j.neucom.2015.02.011>

[8] P. Burkert, F. Trier, M.Z. Afzal, A. Dengel, M. Liwicki, Dexpression: Deep Convolutional Neural Network for Expression Recognition, CoRR abs/1509.05371 (URL 〈http://arxiv.org/abs/1509.05371〉).

[9] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended Cohn–Kanade dataset (CK+): a complete dataset for action unit and emotionspecified expression, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, pp. 94 101.

[10] P. Liu, S. Han, Z. Meng, Y. Tong, Facial expression recognition via a boosted deep belief network, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1805–1812

[11]X. Fan, T. Tjahjadi, A spatial–temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences, Pattern Recognit. 48 (11) (2015) 3407–3416.

[12] Y.-H. Byeon, K.-C. Kwak, Facial expression recognition using 3d convolutional neural network. International Journal of Advanced Computer Science and Applications(IJACSA), 5 (2014) Appl.

[13] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based Learn. Doc. Recognit. <http://dx.doi.org/10.1109/5.726791>.

[14] [Machine Learning cơ bản (machinelearningcoban.com)](https://machinelearningcoban.com/2016/12/27/categories/#supervised-learning-hoc-co-giam-sat)

[15] [CS231n Convolutional Neural Networks for Visual Recognition](https://cs231n.github.io/convolutional-networks/)

[[16] An intuitive guide to Convolutional Neural Networks (freecodecamp.org)](https://www.freecodecamp.org/news/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050)

[17] [Options for training deep learning neural network - MATLAB trainingOptions - MathWorks Australia](https://au.mathworks.com/help/deeplearning/ref/trainingoptions.html)

# NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

Thái Nguyên, ngày…..tháng…..năm 2023

Giảng viên hướng dẫn