HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

SOULINSOMPHOU Oupala

NGHIÊN CỨU PHÂN LOẠI ĐỘ TUỔI CỦA NGƯỜI BẰNG HÌNH ẢNH SỬ DỤNG MẠNG NƠ RON TÍCH CHẬP

Chuyên ngành : KHOA HỌC MÁY TÍNH

Mã số : 8.48.01.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

HÀ NỘI – NĂM 2020

Luận văn được hoàn thành tại:
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS. NGUYỄN ĐÌNH HÓA

Phản biện 1:
……

Phản biện 2:
……..

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

Vào lúc: ..... giờ ..... ngày ..... tháng .... năm 2020

Có thể tìm hiểu luận văn tại:
- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

PHẦN MỞ ĐẦU

Với sự phát triển phần cứng mạnh mẽ cho phép tính toán song song hàng tỉ phép tính, tạo tiền đề cho Mạng nơ-ron tích chập trở nên phổ biến và đóng vai trò quan trọng trong sự phát triển của trí tuệ nhân tạo nói chung và xử lý ảnh nói riêng. Một trong các ứng dụng quan trọng của mạng nơ-ron tích chập đó là cho phép các máy tính có khả năng "nhìn" và "phân tích" hình ảnh. Phân tích đặc điểm khuôn mặt người luôn là một chủ đề được quan tâm chủ yếu do tính ứng dụng của nó. Hiện nay kỹ thuật Deep Learning là kỹ thuật hiệu quả giúp phân tích những đặc điểm dựa trên khuôn mặt của con người. Trong đó tuổi tác và giới tính, hai trong số các đặc điểm quan trọng, đóng một vai trò rất cơ bản trong các tương tác xã hội. Mặc dù các vai trò cơ bản mà các thuộc tính này đóng góp trong cuộc sống hàng ngày, song khả năng tự động ước tính độ tuổi chính xác và đáng tin cậy từ hình ảnh khuôn mặt người vẫn chưa đáp ứng được nhu cầu của các ứng dụng thương mại.

Từ đó việc ước tính độ tuổi từ một hình ảnh khuôn mặt người là một nhiệm vụ quan trọng trong các ứng dụng thông minh, như kiểm soát truy cập, tương tác giữa người với máy tính, thực thi pháp luật, trí thông minh tiếp thị và giám sát trực quan.

Trong luận văn này em đề xuất xây dựng mô hình kiến trúc mạng nơ-ron tích chập để phân lớp dữ liệu hình ảnh mặt người để dự đoán ra độ tuổi của người đó.

Dựa vào thực trạng như trên kết hợp với các kỹ thuật khai phá dữ liệu đã được học hỏi và nghiên cứu để đưa ra đề tài "***Nghiên cứu phân loại độ tuổi của người bằng ảnh mặt người sử dụng mạng nơ ron tích chập***".

Nội dung của Luận văn được xây dựng thành 3 chương như sau:

Chương 1. Giới thiệu Tổng quan về Bài toán phân loại độ tuổi người bằng hình ảnh, bao gồm tổng quan về về bài toán phân loại ảnh mặt người, các nghiện cứu liên quan và một số ứng dụng thực tế của bài toán phân loại độ tuổi bằng ảnh mặt người.

Chương 2. Phân loại độ tuổi của người bằng hình ảnh sử dụng mạng nơ ron tích chập. Trên cơ sở xác định được hướng giải quyết của luận án ở Chương 1, Chương 2 sẽ giới thiệu về mạng nơ ron tích chập và kiến trúc của mạng này trong phương pháp học sâu. Chương này cũng trình bày về các kỹ thuật tiền xử lý dữ liệu đầu vào và việc xây dựng mô hình huấn luyện cho bài toán.

Chương 3. Cài đặt và thử nghiệm. Chương này giới thiệu về bộ dữ liệu được sử dụng trong bài toán, môi trường thực hiện và áp dụng mô hình tốt nhất được xây dựng ở chương 2 vào bộ dữ liệu và đánh giá kết quả phân loại độ tuổi.

CHƯƠNG 1: TỔNG QUAN VỀ BÀI TOÁN PHÂN LOẠI ĐỘ TUỔI CỦA NGƯỜI BẰNG HÌNH ẢNH

1.1. Giới thiệu bài toán phân loại độ tuổi người qua hình ảnh

1.1.1. Tổng quan

Việc phân tích và trích xuất các thông tin có thể có từ các ảnh mặt người đã được các nhà khoa học nghiên cứu từ đầu những năm 90 của thế kỷ trước. Điều này là do có rất nhiều các thông tin có ích có thể khai thác từ một bức ảnh khuôn mặt, ví dụ như danh tính, giới tính, độ tuổi, cảm xúc, cử chỉ tương tác, dân tộc, tình trạng sức khỏe,… Trong số các thông tin có thể suy ra từ ảnh mặt người, độ tuổi là một thuộc tính quan trọng vì nó có khá nhiều ứng dụng trong thực tế, ví dụ như trong tương tác người máy, trong quảng cáo có định hướng, trong thống kê dân số.

Khuôn mặt là một đối tượng trong cơ thể con người và hình ảnh khuôn mặt mang rất nhiều thông tin quan trọng như: tuổi tác, giới tính, trạng thái cảm xúc, dân tộc,... Trong đó, việc xác định tuổi tác và giới tính là hết sức quan trọng, đặc biệt trong giao tiếp, chúng ta cần sử dụng những từ ngữ phù hợp với giới tính của người nghe ví dụ trong tiếng Việt chúng ta có: anh/chị, chú/cô... Hay với nhiều ngôn ngữ khác nhau trên thế giới, chẳng hạn như tiếng Việt thì lời chào hỏi dành cho người lớn tuổi khác với người trẻ tuổi. Do đó, việc xác định tuổi và giới tính dựa trên khuôn mặt là một bài toán hết sức quan trọng, có ý nghĩa thực tế to lớn.

1.2. Hướng tiếp cận và giải quyết bài toán

Phương pháp giải quyết bài toán này có thể được phân làm hai loại phương pháp học, là Phương pháp học máy truyền thống và Phương pháp học sâu.

1.2.1. Phương pháp học sâu

Trong luận văn này tôi áp dụng phương pháp học sâu để giải quyết bài toán phân loại độ tuổi người bằng hình ảnh. Học sâu (Deep Learning) hay viết tắt DL là một thuật toán dựa trên một số ý tưởng từ não bộ tới việc tiếp thu nhiều tầng biểu đạt, cả cụ thể lẫn trừu tượng, qua đó làm rõ nghĩa của các loại dữ liệu. DL được ứng dụng trong nhận diện hình ảnh, nhận diện giọng nói, xử lý ngôn ngữ tự nhiên Hiện nay rất nhiều các bài toán nhận dạng sử dụng DL để giải quyết do DL có thể giải quyết các bài toán với số lượng lớn, kích thước đầu vào

lớn với hiệu năng cũng như độ chính xác vượt trội so với các phương pháp phân lớp truyền thống.

Những năm gần đây, khi mà khả năng tính toán của các máy tính được nâng lên một tầm cao mới và lượng dữ liệu khổng lồ được thu thập bởi các hãng công nghệ lớn, Machine Learning đã tiến thêm một bước dài và một lĩnh vực mới được ra đời gọi là DL (Học Sâu). DL đã giúp máy tính thực thi những việc tưởng chừng như không thể vào 10 năm trước: phân loại cả ngàn vật thể khác nhau trong các bức ảnh, tự tạo chú thích cho ảnh, bắt chước giọng nói và chữ viết của con người, giao tiếp với con người, hay thậm chí cả sáng tác văn hay âm nhạc.

1.3. Kết luận chương

Trong chương I, luận văn đã trình bày tổng quan về bài toán phân loại độ tuổi qua ảnh mặt người, những ứng dụng của bài toán trong thực tế và hướng tiếp cận giải quyết bài toán dựa trên phương pháp học sâu sử dụng mạng nơ ron tích chập CNN.

CHƯƠNG 2: PHÂN LOẠI ĐỘ TUỔI CỦA NGƯỜI BẰNG HÌNH ẢNH SỬ DỤNG MẠNG NƠ RON TÍCH CHẬP

2.1. Giới thiệu về mạng nơ ron tích chập

Mạng nơ-ron tích chập (CNN hay ConvNet) là mạng nơ-ron (Wikipedia, bài báo, video) phổ biến nhất được dùng cho dữ liệu ảnh. Bên cạnh các lớp liên kết đầy đủ (FC layers), CNN còn đi cùng với các lớp ẩn đặc biệc giúp phát hiện và trích xuất những đặc trưng - chi tiết (patterns) xuất hiện trong ảnh gọi là Lớp Tích chập (Convolutional Layers). Chính những lớp tích chập này làm CNN trở nên khác biệt so với mạng nơ-ron truyền thống và hoạt động cực kỳ hiệu quả trong bài toán phân tích ảnh khi so sánh với mạng Nơ-ron truyền thống (Neural Network) - hoạt động không thực sự hiệu quả với dữ liệu đầu vào là hình ảnh: nếu coi mỗi điểm ảnh là một thuộc tính (feature), một ảnh RBG kích thước (64×64) có 12288 (=64×64×3) thuộc tính. Nếu kích thước ảnh tăng lên 1000×10000, chúng ta có 3triệu (3M) thuộc tính cho mỗi ảnh đầu vào. Nếu sử dụng mạng liên kết đầy đủ (fully connected NN) và giả sử lớp thứ 2 có 1000 thành phần (units/ neurons), ma trận trọng số sẽ có kích thước 1000×3M tương đương với 3B trọng số cần huấn luyện (learning). Điều này yêu cầu khối lượng tính toán cực lớn (expensive computational cost) và thường dẫn đến overfitting do không đủ dữ liệu huấn luyện

CNN là một trong những mô hình DL tiên tiến giúp cho chúng ta xây dựng được những hệ thống thông minh với độ chính xác cao như hiện nay. CNN được lấy cảm hứng từ vỏ não thị giác của con người, mỗi khi chúng ta nhìn thấy một vật nào đó, một loại các lớp tế bào thần kinh được kích hoạt, và mỗi lớp sẽ phát hiện ra một đặc trưng của đồ vật đó (hình dạng, kích thước, màu sắc,…). Lớp thần kinh mà nhận dạng được càng nhiều đặc điểm của đồ vật thì việc nhận dạng hoặc phân loại đồ vật đó đối với con người sẽ trở nên dễ dàng hơn.

Ý tưởng đằng sau của mạng nơ ron tích chập là nó thực hiện quá trình trích lọc hình ảnh trước khi đưa vào quá trình huấn luyện, sau quá trình trích lọc thì chúng ta sẽ nhận được các đặc trưng trong hình ảnh đó, và từ các đặc trưng đó chúng ta có thể phát hiện ra những gì mình muốn trong hình ảnh đó.

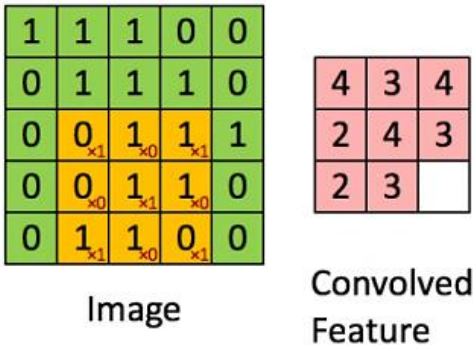
So với các thuật toán phân loại hình ảnh khác, mạng nơ ron tích chập sử dụng quá trình tiền xử lý tối thiểu, nghĩa là mạng học các bộ lọc thường được thiết kế bằng tay trong các hệ thống khác. Bởi vì CNN hoạt động với sự độc lập như vậy khỏi nỗ lực của con người, chúng mang lại nhiều lợi thế hơn các thuật toán khác.

Mục đích của CNN là giảm hình ảnh thành một hình thức dễ xử lý hơn và không mất đi các chi tiết hoặc tính năng quan trọng để hỗ trợ trong việc đưa ra các dự đoán. Điều này rất quan trọng khi chúng ta thiết kế mô hình không chỉ giỏi về các tính năng học tập mà còn xử lý được bộ dữ liệu lớn.

Trước khi tìm hiểu về kiến trúc, mô hình của mạng nơ ron tích chập CNN em sẽ trình bày những khái niệm thường được sử dụng khi làm việc với mạng nơ ron CNN.

a. Tích chập (Convolutional)

Tích chập được sử dụng đầu tiên trong xử lý tín hiệu số (Signal processing). 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, ta có thể xem tích chập nhoị một cửa sổ trượt (sliding window) áp đặt lên một ma trận. Bạn có thể theo dõi cơ chế của tích chập qua hình minh họa bên dưới.



Hình 0.1.1 Minh học phép toán tích chập

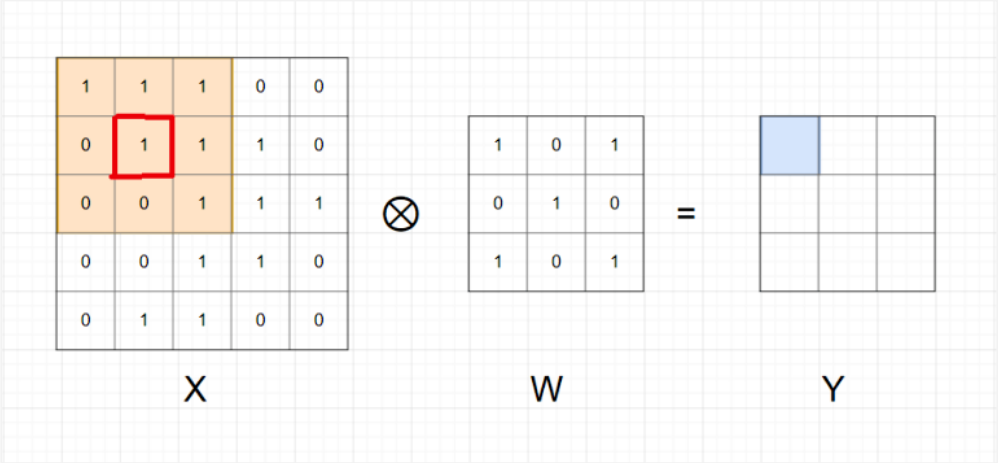
Ma trận bên trái là một bức ảnh đen trắng. Mỗi giá trị của ma trận tương đương với một điểm ảnh (pixel), 0 là màu đen, 1 là màu trắng (nếu là ảnh grayscale thì giá trị biến thiên từ 0 đến 255). Sliding window còn có tên gọi là kernel, filter hay feature detector. Ở đây, ta dùng một ma trận filter 3×3 nhân từng thành phần tương ứng (element-wise) với ma trận ảnh bên trái. Gía trị đầu ra do tích của các thành phần này cộng lại. Kết quả của tích chập là một ma trận (convoled feature) sinh ra từ việc trượt ma trận filter và thực hiện tích chập cùng lúc lên toàn bộ ma trận ảnh bên trái.

b. Bộ lọc (Kernel/Filter)

Độ phức tạp của đặc trưng được phát hiện bởi bộ lọc tỉ lệ thuận với độ sâu của lớp tích chập mà nó thuộc về. Nghĩa là bộ lọc ở lớp tích chập càng sâu thì phát hiện các đặc trưng càng phức tạp. Trong mạng CNN, những lớp tích chập đầu tiên sử dụng bộ lọc hình học (geometric filters) để phát hiện những đặc trưng đơn giản như cạnh ngang, dọc, chéo của bức ảnh. Những lớp tích chập sau đó được dùng để phát hiện đối tượng nhỏ, bán hoàn chỉnh như

mắt, mũi, tóc, v.v. Những lớp tích chập sâu nhất dùng để phát hiện đối tượng hoàn hỉnh như: chó, mèo, chim, ô tô, đèn giao thông, v.v.

Mục đích của việc tích chập (Convolutional) là để lấy ra được các hình dạng (pattern) trong hình ảnh bằng cách sử dụng các bộ lọc (Filter/Kernel). Kernel có thể được coi là tham số của mô hình CNN và được sử dụng để tính toán tích chập (convolve) trên ảnh. Chúng ta có thể thấy thao tác tích chập được mô tả trong hình dưới (Hình 2.1).



Hình 2.4 Bộ lọc W (kernel)

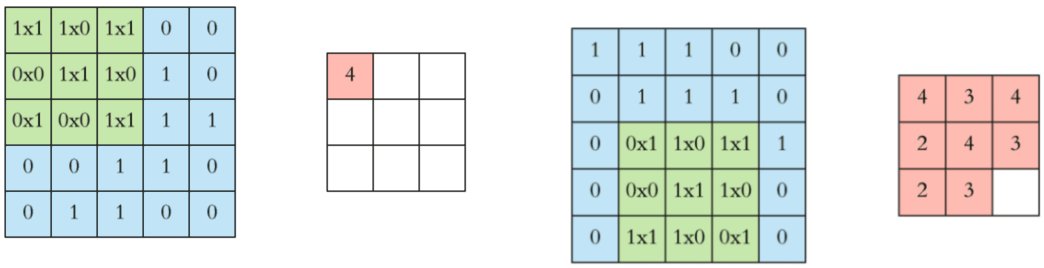
c. Feature map

Tích chập là một khối quan trọng trong CNN. Thuật ngữ tích chập được dựa trên một phép hợp nhất toán học của hai hàm tạo thành hàm thứ ba. Phép toán này kết hợp hai tập thông tin khác nhau.

Trong trường hợp CNN, tích chập được thực hiện trên giá trị đầu vào của dữ liệu và bộ lọc (Kernel/ filter thuật ngữ này được sử dụng khác nhau tùy tình huống) để tạo ra một bản đồ đặc trưng (feature map).

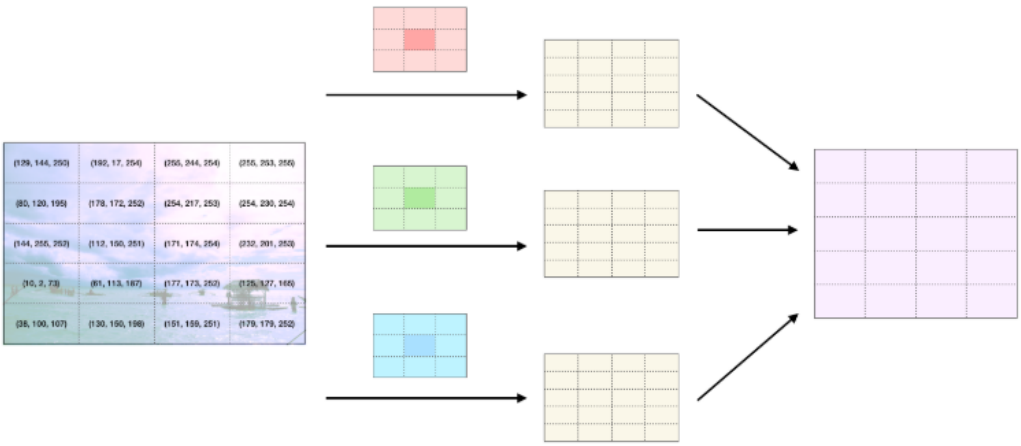
Ta thực hiện phép tích chập bằng cách trượt bộ lọc theo dữ liệu đầu vào. Tại mỗi vị trí, ta tiến hành phép nhân ma trận và tính tổng các giá trị để đưa vào bản đồ đặc trưng.

Trong hình dưới đây, thành phần bộ lọc (màu xanh lá) trượt trên đầu vào (màu xanh dương) và kết quả được trả về bản đồ đặc trưng (màu đỏ). Bộ lọc có kích thước là 3×3 trong ví dụ này.



Hình 2.10 Phép tích chập trên hình ảnh với một giải màu

Đấy là trong trường hợp hình ảnh với một giải màu hoặc là ảnh xám, còn trường hợp quan trọng cần xem xét là cách mà phép tích chập được thực hiện trên hình ảnh màu. Điểm ảnh trong ảnh màu có ba giá trị tương ứng với ba giải màu - giá trị đỏ, lục và lam. Do đó, nếu chúng ta muốn chạy một phép tích chập trên một hình ảnh màu, trước tiên nó phải chia thành các thành phần màu đỏ, xanh lục và xanh lam và thực hiện chạy một bộ lọc trên từng giải dữ liệu đỏ, một trên màu xanh lục và một trên màu xanh lam và tổng hợp tất cả các kết quả.



Hình 2.11 Phép tích chập trên hình ảnh màu

Chúng ta thực hiện phép tích chập trên đầu vào nhiều lần khác nhau. Mỗi lần sử dụng một bộ lọc khác nhau. Kết quả ta sẽ thu được những bản đồ đặc trưng khác nhau. Cuối cùng, ta kết hợp toàn bộ bản đồ đặc trưng này thành kết quả cuối cùng của tầng tích chập. Từ đó phát hiện ra bộ lọc nào cho ra kết quả tương ứng với lớp phân loại hiệu quả nhất. Đối với bài toán tương tự chúng ta thường gọi kết quả của quá trình tích chập là feature map, trọng số xác định các đặc trưng là shared weight và độ lệch xác định một feature map là shared bias.

2.2. Cấu trúc mạng nơ ron tích chập cùng một số mô hình mạng thông dụng trên thực tế

Cấu trúc của một mạng nơ ron tích chập thường sẽ bao gồm các thành phần như:

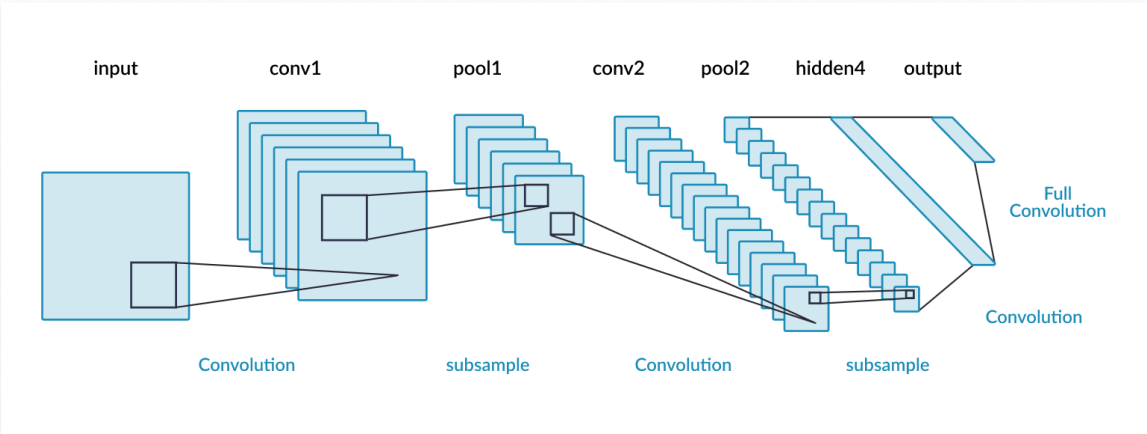
- Convolution layer
- Activaion layer
- Poolling layer
- Flatten layer
- Fully connected layer

Nếu chia theo các loại tầng thì CNN gồm hai thành phần:

Phần tầng ẩn hay phần rút trích đặc trưng: trong phần này, mạng sẽ tiến hành tính toán hàng loạt phép **tích chập** và phép **hợp nhất** (pooling) để phát hiện các đặc trưng. Ví dụ: nếu ta có hình ảnh con ngựa vằn, thì trong phần này mạng sẽ nhận diện các sọc vằn, hai tai, và bốn chân của nó.

Mỗi tầng trong các tầng ẩn tăng cường độ chi tiết và độ phức tạp trong quá trình nhận diện đặc trưng của hình ảnh ví dụ như tầng đầu tiên huấn luyện để phát hiện biên hoặc cạnh của hình ảnh và tầng cuối cùng huấn luyện để phát hiện hình dạng phức tạp hơn như hình tam giác, hình tròn, đôi mắt, mũi, lốp xe. v.v. Các nơ ron trong tầng cuối cùng của tầng ẩn kết nối đến tất cả các nơ ron của tầng đầu ra.

Phần phân lớp: tại phần này, một lớp với các liên kết đầy đủ sẽ đóng vai trò như một bộ phân lớp các đặc trưng đã rút trích được trước đó. Tầng này sẽ đưa ra xác suất của một đối tượng trong hình.



Hình 2.12 Mô phỏng cấu trúc mạng nơ ron tích chập

Cấu trúc mạng CNN là một tập hợp các lớp tích chập (Convolution) chồng lên nhau và sử dụng các hàm kích hoạt như ReLU hoặc tanh để kích hoạt các trọng số trong các nơ ron. Mỗi một lớp sau khi thông qua các hàm kích hoạt sẽ tạo ra các thông tin trừu tượng hơn

cho các lớp tiếp theo. Mô hình CNN thì các tầng liên kết được với nhau thông qua cơ chế gọi là tầng tích chập. Lớp tiếp theo là kết quả tích chập từ tầng trước đó, nhờ vậy mà ta có được các kết nối cục bộ. Như vậy mỗi neuron ở lớp kế tiếp sinh ra từ kết quả tính toán của Kernel hoặc Filter áp đặt lên một vùng ảnh đầu vào của nơ ron trước đó.

Trong mô hình CNN thì ngược lại. Các layer liên kết được với nhau thông qua cơ chế convolution. Layer tiếp theo là kết quả convolution từ layer trước đó, nhờ vậy mà ta có đuược các kết nối cục bộ. Nghĩa là mỗi nơ-ron ở layer tiếp theo sinh ra từ filter áp đặt lên một vùng ảnh cục bộ của nơ-ron layer trước đó.

2.2.1. Convolutional

Đây thường là tầng đầu tiên của mạng nơ ron tích chập, giống như các lớp ẩn khác, lớp tích chập lấy dữ liệu đầu vào, thực hiện các phép chuyển đổi để tạo ra dữ liệu đầu vào cho lớp kế tiếp (đầu ra của lớp này là đầu vào của lớp sau). Phép biến đổi được sử dụng trong lớp tích chập này là phép tính tích chập. Mỗi lớp tích chập chứa một hoặc nhiều bộ lọc - bộ phát hiện đặc trưng (Kernel/Filter) cho phép phát hiện và trích xuất những đặc trưng khác nhau của ảnh.

Với mô hình mạng CNN, lớp tích chập cũng chính là lớp ẩn (Hidden layer), khác ở chỗ lớp tích chập là một tập các bản đồ đặc trưng, và mỗi bản đồ đặc trưng là một bản scan của dữ liệu đầu vào ban đầu, như được trích xuất ra các đặc tính (Feature) cụ thể. Trong tầng này ta sẽ có một ma trận gọi là convolution filter hay kernel thực hiện quét hoặc dịch qua ma trận đầu vào, từ trái qua phải, từ trên xuống dưới và nhân tương ứng với từng giá trị của ma trận đầu vào, rồi cộng lại đưa qua hàm kích hoạt (Sigmoid, ReLU, Elu…), kết quả nhận được là một con số cụ thể và tập hợp lại thành một ma trận đầu ra của tầng này, ma trận này chính là bản đồ đặc trưng.

Giả sử ma trận đầu vào là ***I***, ma trận của bộ lọc là ***K*** có kích thước là ***h* x *w,*** ta có ma trận ***I* x *K*** sẽ được tính bởi công thức dưới :

$$\left(I * K\right)_{xy} = \sum_{i=1}^{h} \sum_{j=1}^{w} K_{ij} \times I_{x+i-1, y+j-1}$$

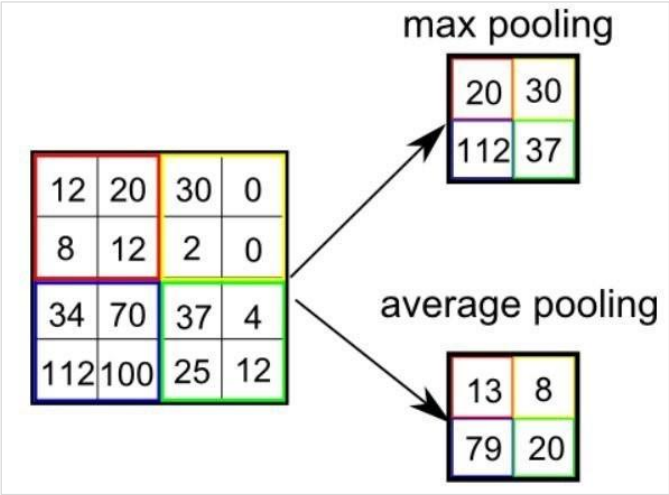
2.2.2. Poolling

Là một lớp được thêm vào giữa các lớp tích chập với mục đích là giảm kích thước của dữ liệu thông qua việc lấy mẫu (sampling), để đơn giản hóa thông tin đầu ra để giảm bớt số

lượng neuron. Việc lấy mẫu này thực hiện bằng cách lấy giá trị lớn nhất hoặc giá trị trung bình của tất cả các giá trị trong cửa sổ pooling được chọn. Pooling là một cách là giảm kích thước ma trận mà vẫn giữ được những thông tin quan trọng nhất của ma trận. Nhiệm vụ của nó là duyệt một ô cửa sổ nhỏ dọc trên một ma trận hình ảnh và lấy giá trị đặc trưng của cửa sổ từ mỗi bước. Trên thực tế cửa sổ được dùng thường có kích thước 2x2 hoặc 3x3. Pooling được xem là một trong những kĩ thuật giúp giảm hiện tượng overfitting trong CNN Chúng ta có thê hình dung hoạt động của nó trong hình sau (Hình 2.13).

Lớp Pooling được sử dụng trong CNN để giảm kích thước đầu vào, tăng tốc độ tính toán và hiệu năng trong việc phát hiện các đặc trưng. Có nhiều hướng Pooling được sử dung, trong đó phổ biến nhất là pooling theo giá trị cực đại (max pooling) và pooling theo giá trị trung bình (average pooling).

- Max Pooling trả về giá trị tối đa từ cửa sổ trượt được bao phủ bởi bộ lọc (Kernel/feature).
- Average Pooling trung bình trả về mức trung bình của tất cả các giá trị từ cửa sổ trượt được bao phủ bởi bộ lọc.



Hình 2.13 Việc thực hiện lấy mẫu trong tầng Pooling

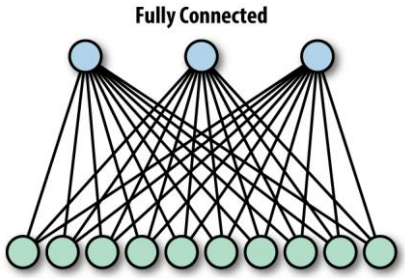
Lớp pooling thường được sử dụng ngay sau lớp tích chập (convulational), với tính chất của lớp pooling, nó làm giảm đáng kể tính chất của ma trận, giúp giảm chi phí tính toán đi đáng kể.

Lớp tích chập và Lớp Pooling, cùng nhau tạo thành lớp thứ i của mạng. Tùy thuộc vào độ phức tạp trong ảnh, số lượng các lớp như vậy có thể được tăng lên để lấy ra được chi tiết ở mức độ sâu hơn, nhưng nó yêu cầu về hiệu năng tính toán của máy tính nhiều hơn. Sau khi

trải qua các bước tích chập thì mô hình có thể hiểu và phân lớp được dữ liệu. Bước tiếp theo là đưa dữu liệu đầu ra của lớp tích chập vào một mảng nơ ron bình thường để thực hiejn quá trình phân lớp.

2.2.3. Lớp kết nối đầy đủ (Fully connected layer)

Tên tiếng viết là Mạng liên kết đầy đủ. Tại lớp mạng này, mỗi một nơ-ron của layer này sẽ liên kết tới mọi nơ-ron của lớp khác. Để đưa ảnh từ các layer trước vào mạng này, buộc phải dàn phẳng bức ảnh ra thành 1 vector thay vì là mảng nhiều chiều như trước. Tại layer cuối cùng sẽ sử dụng 1 hàm kinh điển trong học máy mà bất kì ai cũng từng sử dụng đó là softmax để phân loại đối tượng dựa vào vector đặc trưng đã được tính toán của các lớp trước đó.

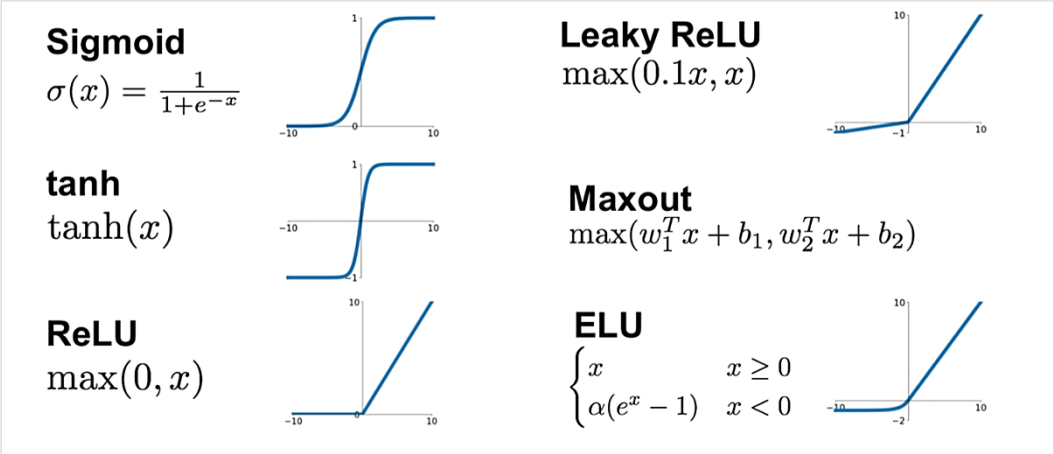


Hình 2.14 Minh họa lớp kết nối đầy đủ

Một lớp được kết nối đầy đủ với đầu ra của lớp trước đó, mỗi nơ ron tại lớp này được kết nối đến tất cả các nơ ron tại lớp tiếp theo được gọi là lớp kết nối đầy đủ. Lớp này sẽ nhận giá trị đầu vào từ lớp pooling và xác định kết quả đầu ra là gì. Đầu ra của lớp này sẽ thực hiện cuộc bầu chọn xem những đặc trưng của mình giống với kết quả hoặc nhãn đầu ra nào nhất, từ đó sẽ xác định được nhãn của dữ liệu đầu vào này là gì.

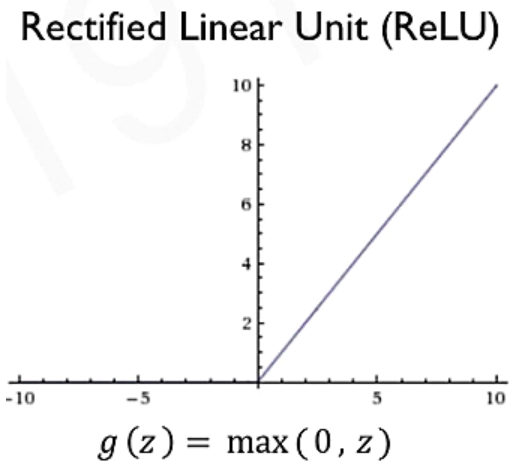
2.2.4. Hàm Kích hoạt (Activation Function)

Hàm kích hoạt là một nút được đặt ở cuối hoặc ở giữa của cấu trúc mạng nơ ron, có rất nhiều loại hàm kích hoạt khác nhau như hàm Sigmoid, Maxout, ReLU… (Hình 2.1). Việc lựa chọn hàm kích hoạt đôi khi là kinh nghiệm của người xây dựng mạng và nó còn phụ thuộc khá nhiều ở bài toán mà chúng ta đang giải quyết. Tuy nhiên hàm ReLU hoạt động khá tốt cho phần lớn các bài toán trong DL.



Hình 2.15 Các hàm kích hoạt phổ biết trong mô hình mạng nơ ron.

- **ReLU** (Rectified Linear Unit) được dựa trên tư tưởng của việc loại bỏ bớt những tham số không quan trọng trong quá trình training và điều đó là cho mạng của chúng ta trở nên nhẹ hơn và việc training cũng nhanh chóng và có hiệu quả hơn. Hàm này thực hiện một việc rất đơn giản như sau: giữ nguyên những giá trị đầu vào lớn hơn 0, nếu giá trị đầu vào nhỏ hơn 0 thì coi là 0. Chúng ta có thể hình dung kĩ hơn trong hình sau (Hình 2.2):



Hình 2.16 Hàm kích hoạt ReLU

- **Softmax** là một loại của hàm kích hoat - activation function. Nó rất hữu ích trong bài toán phân loại đa lớp. Softmax nhận đầu vào là một mảng số thực và đầu ra là một phân phối xác suất với mỗi phần tử nằm trong khoảng [0, 1] và tổng các phần tử là 1 (tương ứng với 100%).

$$softmax_i(a) = \frac{\exp a_i}{\sum \exp a_i}$$

2.2.5. Một số mô hình mạng thông dụng trên thực tế

Trên thực tế mô hình mạng nơ ron được sử dụng phổ biến với các kiến trúc mạng sử dụng lớp tích chập nhiều tầng với kích thước của feature map của từng lớp tăng dần, nhưng có nhiều mô hình với kiến trúc mạng mới đây đã thiết kế sáng tạo hơn và cho kết quả hiệu quả hơn. Dưới đây là các ví dụ của một số kiến trúc mạng nơ ron tích chập thông dụng:

LeNet

Alexnet

VGGNet

2.3. Xây dựng tập dữ liệu cho bài toán

2.3.1. Giới thệu về bộ dữ liệu sử dụng trong bài toán

Trong bài toán phân loại độ tuổi của người bằng hình ảnh sử dụng mạng nơ ron tích chập chúng ta sử dụng tập dữ liệu khuôn mặt các diễn viên trong phim của Ấn Độ (Idian Movies Face Database) hay viết tắt là IMFDB là một bộ dữ liệu khuôn mặt lớn bao gồm 26742 hình ảnh của 100 diễn viên Ấn Độ được thu thập từ hơn 100 video. Tất cả các hình ảnh được lựa chọn và cắt xén thủ công từ các khung hình video dẫn đến rất nhiều kích thước, tư thế, biểu hiện, độ sáng, độ tuổi và độ phân giải. Bộ dữ liệu IMFDB là cơ sở dữ liệu khuôn mặt đầu tiên cung cấp nhãn chi tiết cho mọi hình ảnh về độ tuổi, tư thế, giới tính và biểu hiện có thể giúp nhiều ứng dụng khác nhau liên quan đến phân tích khuôn mặt.

Ảnh khuôn mặt người trong bộ dữ liệu IMFDB được thu thập từ các video, vì thế khuôn mặt trong các video phim được cho là có rất đa dạng ảnh khác nhau về độ chiếu sáng, góc nhìn, độ phân giải, độ mờ, v.v. Video được thu thập từ hai thập kỷ trước nên nó chứa rất nhiều ảnh khuôn mặt khác nhau về độ tuổi so với hình ảnh được thu thập từ Internet thông qua truy vấn tìm kiếm trong hiện nay[9]. Bộ dữ liệu IMFDB được xây dựng bằng cách chọn thủ công và cắt xén các khung hình video dẫn đến mức độ đa dạng của các biểu cảm của khuôn mặt người và có thể được sử dụng để phát triển các thuật toán để phân tích biểu cảm của mặt người.

Cắt xén khuôn mặt: Khuôn mặt được cắt bằng một khung hình vừa với một khuôn mặt. Để duy trì tính nhất quán trên các hình ảnh, tôi đã theo dõi một cách cắt xén khuôn mặt từ trán đến cằm.

Hình 2.20 Một số hình ảnh ví dụ của bộ dữ liệu IMFDB

Trong bộ dữ liệu kèm theo tệp bảng dữ liệu tương ứng với thông tin mã hình ảnh, các thông tin khác và nhãn phân loại của từng hình ảnh [9]. Bao gồm các trường dữ liệu như :

ID : Image id
Expressions : Anger, Happiness, Sadness, Surprise, Fear, Disgust
Illumination : Bad, Medium, High
Pose : Frontal, Left, Right, Up, Down
Occlusion : Glasses, Beard, Ornaments, Hair, Hand, None, Others
Makeup : Partial makeup, Over-makeup
Gender : Male, Female
Age : Young, Middle, Old

Bảng 2.1 Mẫu bộ dữ liệu IMFDB

ID	Expression	Illumination	Pose	Occlusion	Makeup	Gender	Age
088.jpg	SURPRISE	MEDIUM	FRONTAL	GLASSES	OVER	Male	Young
177.jpg	NEUTRAL	BAD	RIGHT	NONE	PARTIAL	Male	Old
634.jpg	HAPPINESS	MEDIUM	FRONTAL	OTHERS	PARTIAL	Male	Young
671.jpg	SADNESS	BAD	LEFT	OTHERS	PARTIAL	Female	Young
799.jpg	HAPPINESS	MEDIUM	DOWN	OTHERS	PARTIAL	Female	Middle

807.jpg	HAPPINESS	MEDIUM	UP	OTHERS	PARTIAL	Male	Middle
908.jpg	DISGUST	BAD	DOWN	OTHERS	PARTIAL	Male	Middle
938.jpg	SADNESS	BAD	LEFT	NONE	PARTIAL	Male	Young
033.jpg	HAPPINESS	MEDIUM	FRONTAL	NONE	PARTIAL	Male	Young
183.jpg	SURPRISE	MEDIUM	FRONTAL	NONE	OVER	Female	Young

2.3.2. Tiền xử lý và chuẩn bị dữ liệu

a. Loại bỏ đặc trưng không cần thiết

b. Loại bỏ nhiễu trong dữ liệu

c. Chỉnh kích thước ảnh sang kích thước phù hợp

d. Phân chia dữ liệu

2.4. Xây dựng mô hình mạng nơ ron tích chập để giải quyết bài toán phân loại độ tuổi của người bằng hình ảnh.

2.4.1. Cấu trúc mô hình

Việc xây dựng mô hình để đạt được hiệu quả cao phụ thuộc vào các yếu tố như cấu trúc cảu mô hình mạng, lựa chọn thuật toán, xác định các biến dữ liệu phù hợp và điều chỉnh các tham số để cho phù hợp dựa trên bộ dữ liệu sử dụng để huấn luyện mô hình. Đối với bài toán này chúng ta sẽ sử dụng mô hình mạng nơ ron tích chập CNN để phân loại độ tuổi bằng hình ảnh với tập dữ liệu dã giới thiệu ở mục trên.

Cấu trúc mô hình mạng nơ ron tích chập được sử dụng trong bài toán dựa vào một mô hình mạng LeNet, có cấu trúc bao gồm ba lớp tích chập (Convolution) với mỗi lớp sẽ có các lớp Pooling ở giữa của từng lớp, tiếp theo đến ba lớp kết nối đầy đủ (Fullyconnected) với lớp kết nối đầy đủ cuối cùng là lớp giá trị đầu ra với sô nơ ron bằng sô nhãn phân loại. Mô hình bao gồm các chi tiết cụ thể như sau:

❖ **Lớp tích chập (Convolution)**

Ba lớp đầu tiên của mô hình mạng chúng ta đều là lớp tích chập, với lớp tích chập đầu tiên nhận dữ liệu đàu vào là mảng cẩu các hình ảnh có kích thước 128 x 128 pixel. Tại lớp này chúng ta khai báo với số bộ lọc (Kernel) sử dụng là 25 với kích thước của từng bộ lọc mà 3 x 3. Tiếp theo là khai báo hàm kích hoạt cho lớp này, chúng ta sử dụng hàm “ReLu” đã giới thiệu ở mục trên.

Tương tự với lớp tích chập đầu tiên, lớp tích chập thứ hai và thứ ba chúng ta sẽ khai báo tương ứng nhưng với số dặc trưng sử dụng khác nhau là 50 và 75 theo lần lượt. Còn hàm kích hoạt thì chúng ta vẫn khai báo hàm "ReLu" tương tự như trên.

Giữa các lớp kích hoạt chúng ta sẽ khai báo một lớp hợp nhất (Pooling), ở đây chúng ta sử dụng phép hợp nhất tối đa (Max pooling) với giá trị kích thước là 2 x 2.

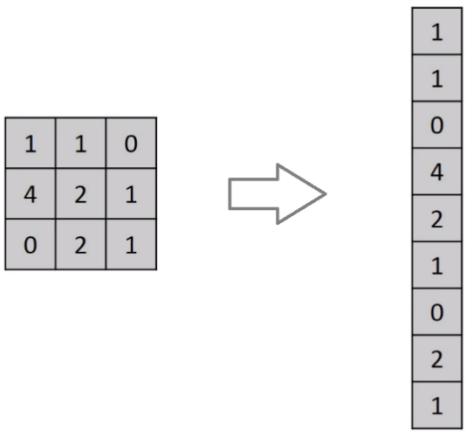
❖ **Hàm kích hoạt sử dụng (Activation function)**

Hàm kích hoạt sử dụng trong bài toán này gồm có hai hàm là "ReLu" (Rectified Linear Unit) và hàm "Softmax".

Tại các lớp tích chập chúng ta sử dụng hàm "ReLu" là hàm kích hoạt. Hàm này có công thức dễ thực hiện tính toán và hiệu quả với nhiều loại bài toán, với tốc độ thữ hiện nhanh đãn đến thời gian huấn luyện mô hình tương đối nhanh so vưới hàm kích hoạt khác. Tại tầng liên kết đầy đủ cuối cùng, chung ta sử dụng hàm "Softmax". Hàm "Softmax" thường được sử dụng ở tầng đầu ra, nhằm đánh giá sắc xuất nhãn phân loại của dữ liệu đầu vào của tầng đấy.

❖ **Lớp làm phẳng (Flatten)**

Lớp này có nhiệm vụ chuyển đổi kết quả đầu ra từ lớp tích chập là mảng nhiều chiều và chuyển đổi thành vec tơ một chiều trước khi được vào tầng kết nối đầy đủ để thực hiện quá trình phân loại.



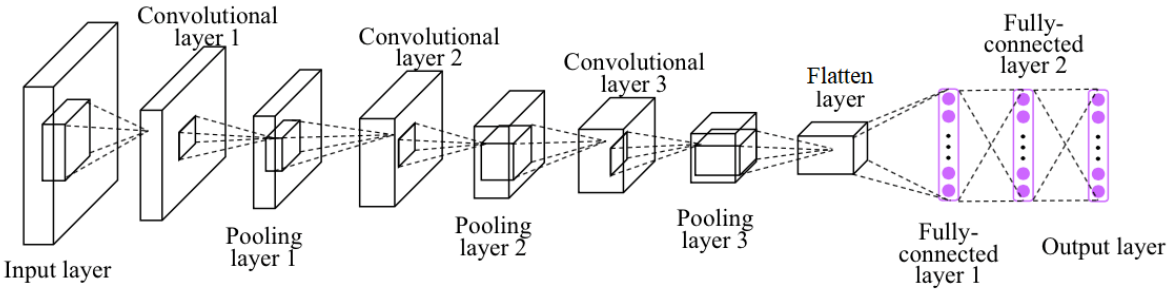
Hình 2.26 Minh họa phương thức làm phẳng (Flatten)

❖ **Lớp kết nối đầy đủ (Fully connected layer)**

Tại lớp này chúng ta sử dụng tất cả 3 lớp, hai lớp đầu tiên là một lớp kết nối đầy đủ với số nơ ron bằng 32 và sử dụng hàm "ReLU" là hàm kích hoạt.

Với lớp phân nhãn cuối cùng với số nơ ron bằng 3 tương ứng với số nhãn phân loại của tập dữ liệu là (“Middle”, “Old”, “Young”).

Cuối cùng chúng ta nhận được mô hình mạng được minh họa ở hình dưới (Hình 2.27):



Hình 2.27 Minh họa mô hình mạng sử dụng trong bài toán

2.4.2. Các hàm và kỹ thuật sử dụng

a. Thuật toán tối ưu (Optimizer)
b. Hàm lỗi (Loss function)
c. Kỹ thuật Drop-out để giảm Over-fitting
d. Kỹ thuật tăng cường dữ liệu (Data augmentation)
e. Cân bằng trọng số (weight balancing)

2.5. Kết chương

Trong chương II, Luận văn trình bày giới thiệu về mạng nơ ron tích chập, cấu trúc mạng và các ứng dụng trên thực tế sử dụng mạng nơ ron tích chập. Giới thiệu về bộ dữ liệu sủa dụng trong luận văn tiền xử lý dữ liệu và chuẩn bị cho mô hình mạng, sau đó xây dựng mộ mô hình mạng để giải quyết bài toán phân loại độ tuổi người và thực hiện huấn luyện mô hình. Trong chương tiếp theo tôi sẽ đánh giá kết quả huấn luyện và kiểm chứng của mô hình.

CHƯƠNG 3: CÀI ĐẶT VÀ THỬ NGHIỆM

3.1. Cài đặt môi trường thực hiện huấn luyện và thử nghiệm mạng nơ ron tích chập áp dụng trên bộ dữ liệu thực tế.

Trong quá trình triển khai và xây dựng mô hình của luận văn này tôi đã áp dụng các giải pháp phần mềm sau:

1. Ngôn ngữ lập trình Python
2. Công cụ và môi trường tích hợp mã nguồn Python sử dụng là Jupyter Notebook
3. Các thư viện hỗ trợ Python:
   - Keras
   - Tensorflow
   - Pandas
   - OpenCV
   - Mathplotlib
   - Numpy
   - Scikit-learn
4. Mối trường cài đặt:

3.2. Phương pháp đánh giá

Để đánh giá hiệu suất của bài toán phân loại văn bản chúng ta sử dụng các độ đo như: Accuracy, Precision, Recall, và bảng Confusion matrix. Được định nghĩa ở phần dưới. Để ước lượng các độ đo này, có thể dựa vào bảng sau:

Bảng 3.1 Bảng Confusion matrix

		Dự đoán	
Nhãn		0	1
	0	TN	FP
	1	FN	TP

Một số tiêu chí mô tả độ hiệu quả của mô hình phân loại bao gồm có:

- Accuracy : Khả năng mô hình phân loại dự báo chính xác, phân loại chính xác hay xác định đúng nhãn hoặc lớp đối với dữ liệu cần phân loại. Được tính bằng công thức:

$$Accuracy = \frac{TP + TN}{(TP + TN) + (FP + FN)}$$

- Precision : Được định nghĩa như là xác suất mà một dữ liệu phân loại là 1 là một phân loại đúng (độ chính xác của mỗi lần dự đoán). Được tính toán ước lượng như sau:

$$Precision = \frac{TP}{TP + FP}$$

- Recall : Được định nghĩa như là xác suất mà một dữ liệu với nhãn là 1 đã được phân loại đúng (độ chính xác của dự đoán cho từng nhãn). Được tính toán ước lượng như sau:

$$Recall = \frac{TP}{TP + FN}$$

3.3. Đánh giá kết quả

Từ kết quả của quá trình huấn luyện trên, chúng ta thực hiện kiểm chứng mô hình sau đã huấn luyện trên với dữ liệu kiểm chứng đã chuẩn bị trước đó như sau:

Kết quả nhận được được thể hiện dưới đây:

```
Score = model.evaluate(test_x, test_y, verbose=0)
print("Accuracy: %.4f%%" % (Score[1]*100))
```

```
Accuracy: 86.3667%
```

Từ kết quả kiểm chứng với chỉ số đánh giá là độ chính xác phân lớp của mô hình chúng ta nhận được ở mức tương đương 86.37% với bộ dữ liệu kiểm chứng. Chúng ta sẽ xem xét với độ đo khác như Precision và Recall được thể hiện dưới bảng sau:

Bảng 3.2 Kết quả phân loại của mô hình

Nhãn	Precision	Recall
Middle	0.92	0.84
Old	0.78	0.82
Young	0.82	0.91
Trung bình	0.84	0.86

Với số đo độ Precision cho thấy trong dữ liệu kiểm chứng thì 84% dữ liệu được phân loại đúng nhãn. Độ đo Recall cho chúng ta thấy là mẫu dữ liệu có nhãn phân loại được phân loại đúng nhãn của nó với xác suất là 86%.

Để đánh giá được mô hình cụ thể hơn chúng ta sẽ xem kết quả phân loại trên từng nhãn, xem mô hình có tỷ lệ phân nhãn của từng nhãn với độ chính xác là bao nhiêu bằng cách sinh bảng Confusion matrix dưới đây:

Bảng 3.3 Confusion matrix

		Dự đoán		
		Middle	Old	Young
Nhãn	Middle	1346	73	178
	Old	49	305	20
	Young	73	13	921

Bảng 3.3 là ma trận phân tích độ chính xác của dự đoán nhãn sau khi chạy mô hình. Từ bảng trên chúng ta có thể thấy được kết quả phân loại đúng, sai của từng nhãn như bảng 3.4 dưới đây:

Bảng 3.4 Kết quả phân loại theo từng nhãn

Nhãn phân loại	**Phân loại đúng**	**Phân loại sai**	**Tổng số mẫu**
Middle	1346	251	1597
Old	305	69	374
Young	921	86	1007

Tại lớp có nhãn phân loại độ tuổi là “Old” thì mô hình phân loại đúng 305 mẫu trên tất cả 374 mẫu, tương đương với 81.55%. Với lớp có nhãn phân loại là Middle, mô hình phân loại đúng 1346 mẫu trên tổng số 1597 mẫu, chiếm 84,28%. Tương tự, xác suất phân loại mẫu chính xác của nhãn “Young” là 91,46%.

Nhận xét: Với kết quả đạt được từ mô hình trên cho thấy độ chính xác phân loại của mô hình tương đối ổn định nhưng chưa đạt được kết quả tốt nhất, cũng như độ chính xác phân lớp của nhãn độ tuổi già (Old) có độ chính xác tương đối thấp so với các lớp khác. Từ đó tôi đã nhận thấy được hai vấn đề chưa giải quyết được trong đo bài toán này là: Vấn đề về xử lý bộ dữ liệu mất cân bằng, do số tổng số mẫu của dữ liệu mang nhãn độ tuổi già chỉ chiếm 12% trong tất cả bộ dữ liệu. Vấn đề thứ hai là mô hình phân loại, chúng ta chưa xây dựng được mô hình tốt nhất để thực hiện bài toán phân loại độ tuổi với bộ dữ liệu này, do hạn chế về tài nguyên máy tính nên không khả năng xử lý mô hình mô hình với độ phức tạp cao hơn với số tham số mô hình cao hơn.

3.4. Kết chương

Trong chương này, tôi đã trình bày về môi trường cài đặt, ngôn ngữ lập trình và thư viện hỗ trợ được sử dụng và đưa ra kết quả, cũng như những phương pháp phân tích đánh giá mô hình từ đó đánh giá được những kết quả đạt được. Ngoài ra, tôi đã đưa ra những vấn đề ảnh hưởng đến độ chính xác của mô hình huấn luyện, từ đó cải thiện độ chính xác của mô hình trong những nghiên cứu sau này được tốt hơn.

KẾT LUẬN

Trong bài báo này, tôi đã đề xuất một mô hình học sâu sử dụng mạng CNN để nhận diện độ tuổi của người dựa vào hình ảnh khuôn mặt. Mô hình mới này cho phép sử dụng một số lượng nhỏ các tham số nhưng đạt hiệu suất tốt hơn các mô hình đã được công bố gần đây, đồng thời góp phần giải quyết vấn đế nhận diện trong thời gian thực.Trong tương lai gần, tôi đang có kế hoạch cải thiện độ chính xác của mô hình, đặc biệt là đối với ước lượng độ tuổi bằng cách thử áp dụng dữ liệu mới tự thu thập. Mặt khác, tôi sẽ áp dụng mô hình của tôi cho các bài toán khác trong lĩnh vực thị giác máy tính và học sâu. Những kết quả hoạt động chính của luận văn:

- Nghiên cứu về phương pháp học sâu so sánh với phương pháp học máy truyền thống.
- Giới thiệu tổng quan về mạng nơ ron tích chập, các thành phần kiến trúc và mô hình mạng, chức năng và ứng dụng thực tế của nó.
- Xây dụng một mô hình mạng nơ ron tích chập cho bài toán phân loại độ tuổi người bằng hình ảnh. Trong đó thực hiện huấn luyện, điều chỉnh các thông số trong mạng và áp dụng các kỹ thuật để đạt được khả năng dự đoán độ tuổi với độ chính xác trên 83%.
- Phân tích đánh giá kết quả sau khi huấn luyện mô hình.

Luận văn còn một số vấn đề tiếp tục phát triển. Do thời gian, kinh nghiệm, ngôn ngữ Tiếng Việt hạn chế nên không tránh khỏi các sai sót, kính mong được sự thông cảm của các Thầy Cô, các Nhà khoa học.

Định hướng phát triển của luận văn

Hướng nghiên cứu tiếp theo của luận văn sẽ tập trung vào phần xây dựng mô hình mạng phân loại độ tuổi của người với độ chính xác cao hơn, có thể sử dụng mô hình áp dụng nhiều tầng tích chập theo kiến trúc mạng VGGNet để phân tích được các đặc trưng chi tiết hơn. Thử nghiệp áp dụng kỹ thuật over-sampling vào dữ liệu với nhãn phân loại là già ”Old”, để tăng thêm độ cân bằng của bộ dữ liệu.