HCMUTE FME
KHOA CƠ KHÍ CHẾ TẠO MÁY
FACULTY OF MECHANICAL ENGINEERING

NHẬN DIỆN CÁC ĐIỂM LANDMARK TRÊN KHUÔN MẶT ỨNG DỤNG CNN

Nguyễn Trọng Nhân ***19146065***

Thông tin

Trường đại học Sư phạm Kỹ thuật thành phố Hồ Chí Minh, khoa Cơ khí Chế tạo máy.

19/06/2022

Từ khóa

Facial Landmark Detection

Face Keypoint Detect

CNN

Tóm tắt

Facial Landmark đươc định nghĩa là các điểm mốc trên khuôn mặt người mà kết quả của bài toán này chính là dự đoán ra các điểm chính tạo nên hình dạng của một đối tượng, cụ thể là khuôn mặt người trong một bức ảnh. Facial Landmark là đầu vào hay còn gọi là bài toán tiên quyết cần phải giải quyết của nhiều bài toán khác như dự đoán tư tế đầu, hoán đổi khuôn mặt, phát hiện nháy mắt, phát hiện cảm xúc, xoay chỉnh cấu trúc khuôn mặt hay dễ hình dung nhất là việc xây dựng FaceID, các Filter trong các phần mềm chụp và chỉnh sửa ảnh,…Quá trình nghiên cứu và tiến hành thực nghiệm nhận dạng các điểm mốc trên khuôn mặt nhằm nâng cao chất lượng, cải tiến chất lượng của các đề tài có quy mô hay yêu cầu lớn hơn. Facial Landmark Detection là bài toán đầu tiên và tiên quyết, nếu muốn tiếp cận được với các đề tài cao hơn, bài toán phức tạp hơn, việc nắm vững bài toán này sẽ thúc đẩy khả năng làm việc với các mô hình bài toán lớn hơn. Hướng tới việc nghiên cứu chuyên sâu về xử lý ảnh được dễ dàng hơn. Quá trình giải quyết bài toán này được thực hiện bởi trí tuệ nhân tạo thông qua các thư viện hỗ trợ Deeplearning, có cơ hội thực hành chuyên sâu với thuật toán Mạng Nơron tích tụ (CNN), được thực hiện thử nghiệm trên Google Colaboratory nhằm tiết kiệm tài nguyên máy tính, làm cho quá trình học tập và nghiên cứu về AI được dễ dàng hơn. Với 3462 hình ảnh và bằng phương pháp huấn luyện, các thuật toán điều chỉnh dataset. Kết thúc quá trình huấn luyện cho ra kết quả khá chính xác với 98% so với thực tế. Do đó, nghiên cứu này bước đầu thành công do với mong đợi và thõa mãn phần lớn các yếu tố ban đầu đề ra.

1. Giới thiệu

Facial Landmark Detection được mô tả dễ hiểu bằng cách cho một hình ảnh với đối tượng chính là khuôn mặt người nhìn trực diện hoặc nghiên vào ống kính để tiến hành nhận dạng. Mục tiêu chính là dự đoán hình dạng hoặc cố gắng định vị được điểm quan trọng theo chính hình dạng đó. Từ đó có thể phát hiện được điểm tạo nên cấu trúc khuôn mặt dựa trên các phương pháp nhận dạng. Sau đó sẽ tiến hành đánh dấu lại các điểm đó, lưu và tạo thành hình ảnh mới với tỷ lệ trùng khớp phụ thuộc vào khả năng làm việc của phương pháp nhận dạng đó.

Việc khai thác và sử dụng Facial Landmark hiện nay đang dần trở nên vô cùng phổ biến và ứng dụng rộng rãi trong nhiều lĩnh vực có sử dụng đến yếu tố xử lý hình ảnh. Quá trình nghiên cứu và tiến hành thực nghiệm nhận dạng các điểm mốc trên khuôn mặt nhằm nâng cao chất lượng, cải tiến chất lượng của các đề tài có quy mô hay yêu cầu lớn hơn.

Độ chính xác cao nhất được báo cáo trong báo cáo này khi tiến hành xác định các điểm mốc trên khuôn mặt chính xác là 98%.

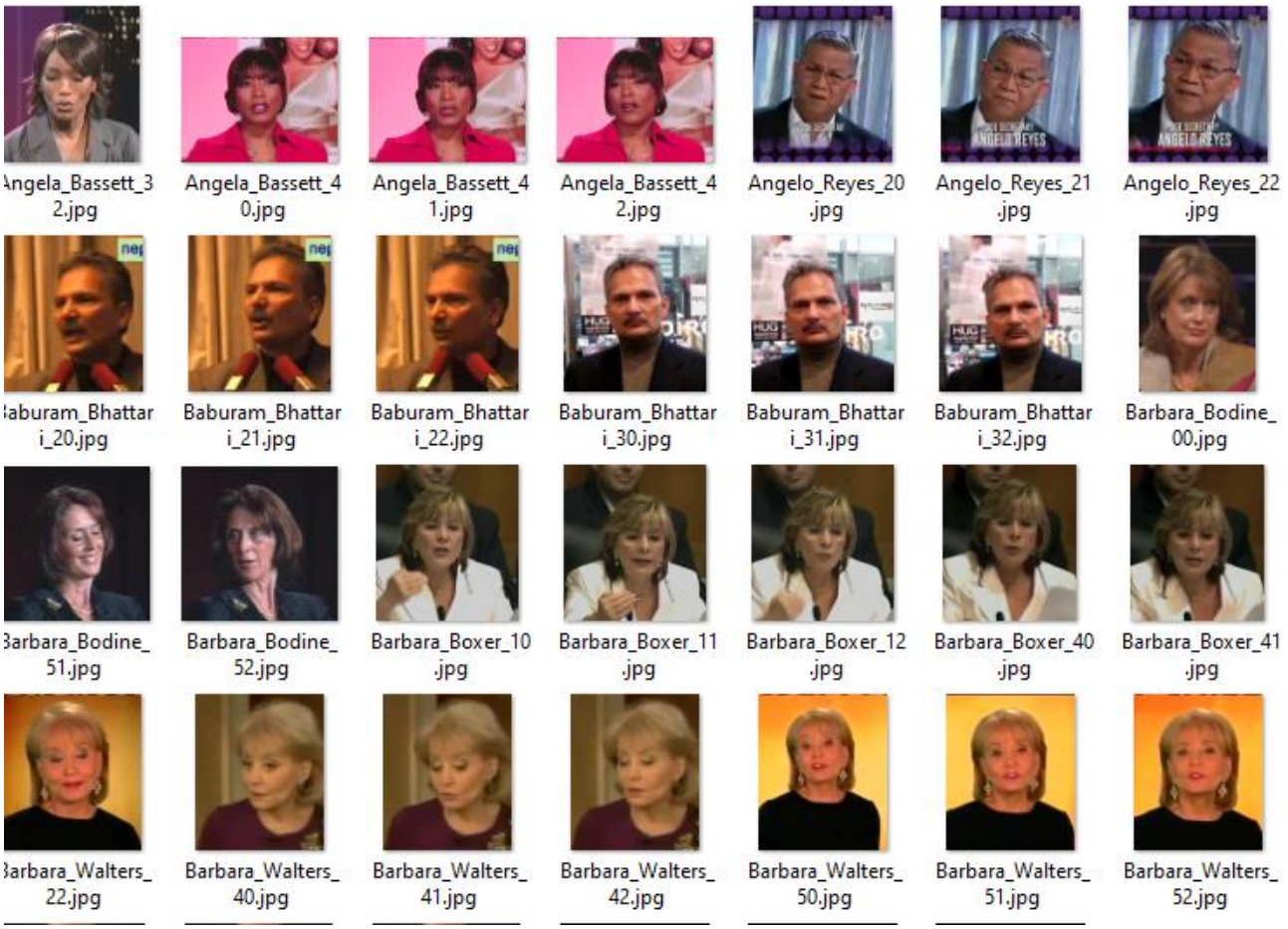
2. Phương pháp và dữ liệu

Trong nghiên cứu này, sử dụng thuật toán CNN nh Sử dụng CNN như một phương pháp nghiên cứu tìm ra các điểm Facial Landmarks.

2.1. Datasets

Trong bài báo cáo này, sử dụng nguồn dữ liệu được tổng hợp nhiều nguồn trên Internet, trong đó sử dụng bộ hình ảnh và các điểm mốc dữ liệu chứa 68 điểm trên mỗi ảnh. Tổng cộng bao gồm:

Số lượng ảnh dùng cho File Test	**2308**
Số lượng ảnh dùng cho File Train	3462



Hình 1. Hình ảnh từ dataset

2.2. Convolutional Neural Network là gì

Convolutional Neural Network (CNN hoặc ConvNet) được tạm dịch ra thành "Mạng nơ ron tích chập" là một trong những mô hình Deep Learning mà trong đó sử dụng nhiều thuật toán thu được mô hình dữ liệu trừu tượng hóa ở mức nâng cao bằng việc sử dụng nhiều lớp xử lý cấu trúc, trong đó đối tượng làm việc chủ yếu là hình ảnh. Bằng cách phân tích hình ảnh, CNN là một lớp của mạng nơ ron sâu. Nó giúp cho chúng ta xây dựng được những hệ thống thông minh với độ chính xác cao và trực quan hóa kết quả.

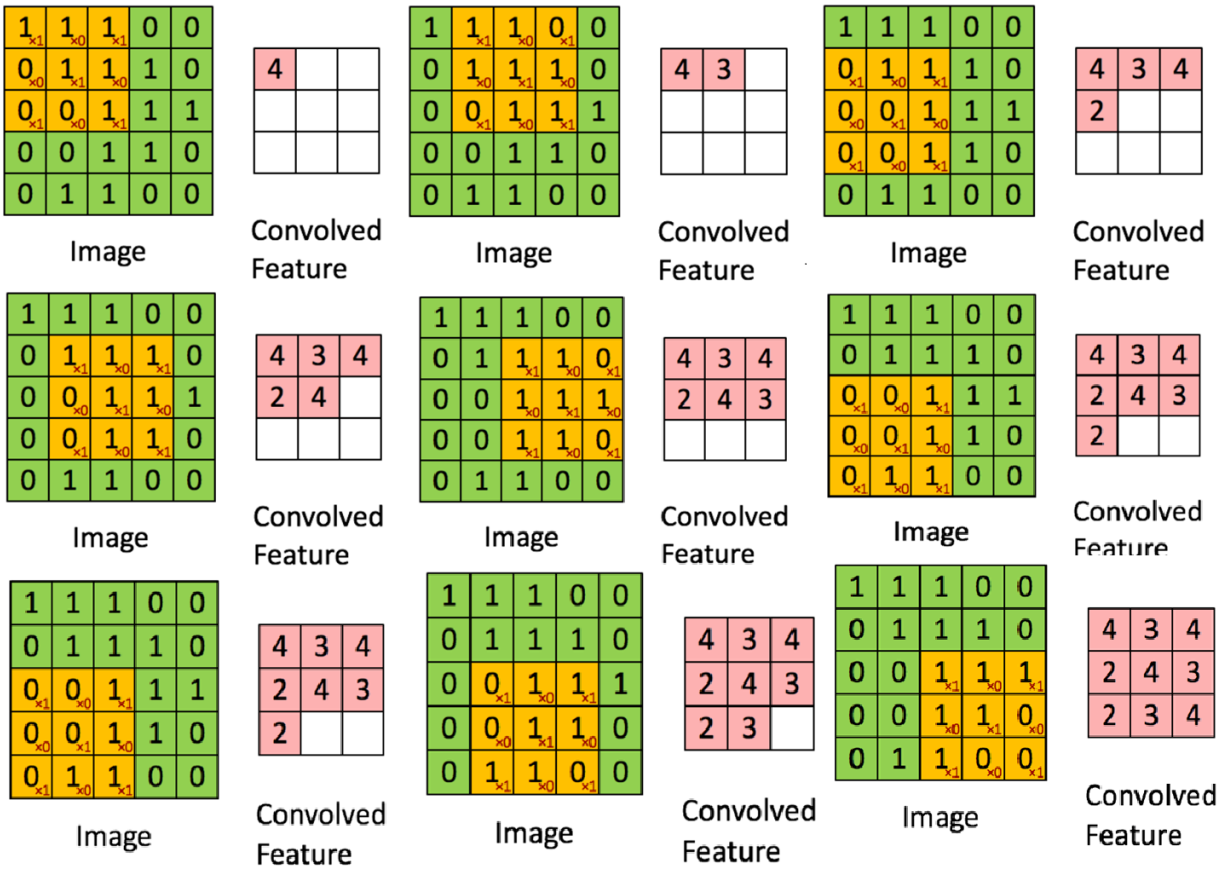
Lấy ví dụ cơ bản, ta có thể sử dụng lớp mạng này trong các ứng dụng như nhận dạng hình ảnh, nhận dạng khuôn mặt, hay thuật toán được ứng dụng nhiều vào quảng cáo tự động trên các nền tảng mạng xã hội như Facebook, Instagram, Google, Youtube,…

Convolutional Neural Network là hệ thống mạng nhận đầu vào là một mảng hai chiều và hoạt động trực tiếp trên hình ảnh thay vì tập trung trích xuất tính năng thường thấy ở các mạng nơ ron khác. Do đó, để tìm hiểu CNN là gì cần tập trung vào một số thuật ngữ như sau:

Feature được dịch theo nghĩa tiếng Việt là *đặc điểm*. Khi sử dụng thuật toán CNN so sánh hình ảnh theo từng mảnh, mỗi mảnh đó được gọi là Feature.

Mỗi Feature được xem như một hình ảnh mini hay gọi là những mảng hai chiều nhỏ. Các Feature được khớp với những khía cạnh chung trong bức ảnh đó. Nghĩa là Feature sẽ tương ứng với khía cạnh nào đó của hình ảnh và chúng sẽ khớp lại với nhau.

Convolutional được hiểu là *tích chập*. Xét về cơ bản, khi xem một hình ảnh mới, thuật toán CNN sẽ không nhận biết được nó ở vị trí nào, các Feature sẽ khớp với nhau ở đâu? Chính vì vậy, Convolutional sẽ thử chúng với tất cả các vị trí khác nhau và tạo thành một bộ lọc gọi là Filter. Quá trình này được thực hiện thông qua phần toán nơ ron tích chập. Convolutional được hiểu như một cửa sổ trượt (Sliding Windows) trên một ma trận



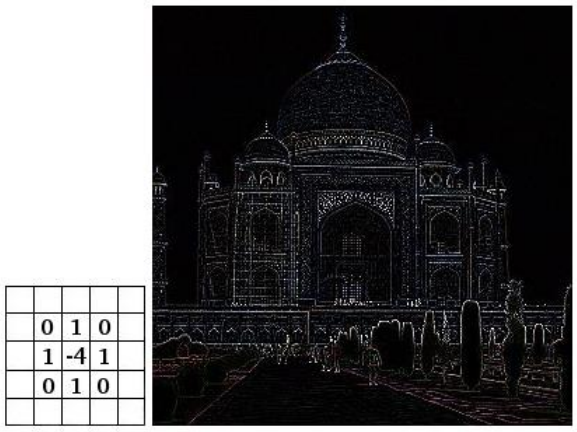
Hình 2. Ví dụ về Convolutional

Các convolutional layer có các parameter(kernel) đã được học để tự điều chỉnh lấy ra những thông tin chính xác nhất mà không cần chọn các feature.

Trong hình ảnh ví dụ Hình 2.2, ma trận bên trái là một hình ảnh trắng đen được số hóa. Ma trận có kích thước 5×5 và mỗi điểm ảnh có giá trị 1 hoặc 0 là giao điểm của dòng và cột.

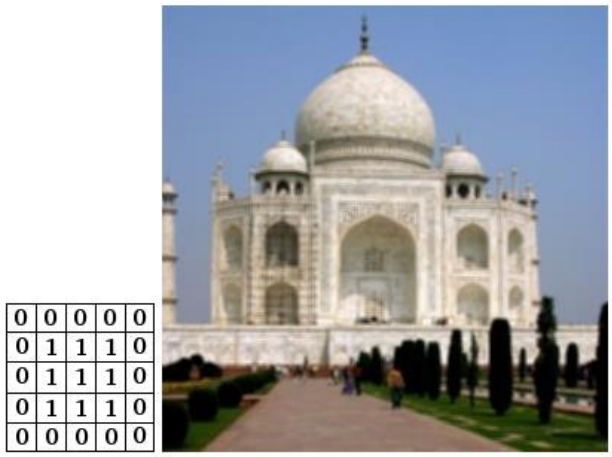
Convolution hay tích chập là nhân từng phần tử trong ma trận 3. Sliding Window hay còn gọi là kernel, filter hoặc feature detect là một ma trận có kích thước nhỏ như trong ví dụ trên là 3×3.

Convolution hay tích chập là nhân từng phần tử bên trong ma trận 3×3 với ma trận bên trái. Kết quả được một ma trận gọi là Convoled feature được sinh ra từ việc nhận ma trận Filter với ma trận ảnh 5×5 bên trái.



	0	1	0	
	1	-4	1	
	0	1	0	

Hình 3. Hình ảnh trắng đen được số hóa



0	0	0	0	0
0	1	1	1	0
0	1	1	1	0
0	1	1	1	0
0	0	0	0	0

Hình 4. Hình ảnh được Convoled feature

2.3. Cấu trúc của mạng CNN

Mạng CNN là một tập hợp các lớp Convolution chồng lên nhau và sử dụng các hàm nonlinear activation như ReLU và tanh để kích hoạt các trọng số trong các node. 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.

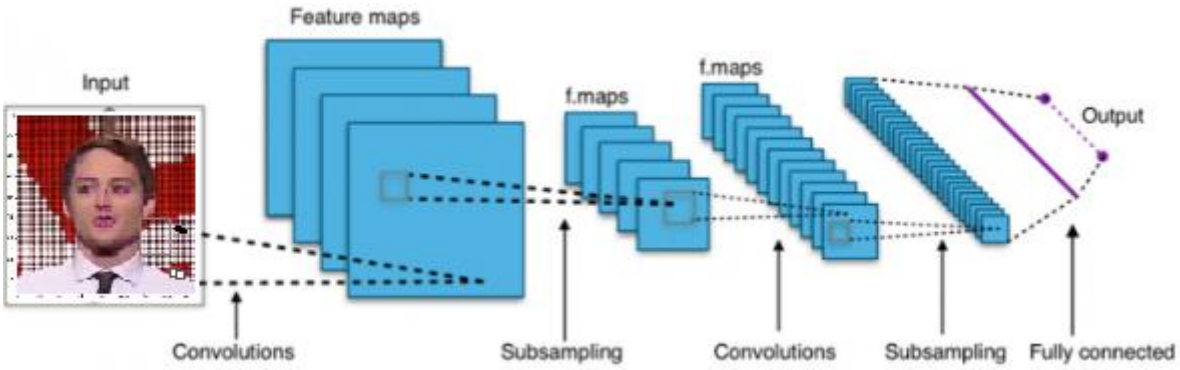
Mỗi một lớp sau khi thông qua các hàm kích hoạt sẽ tạo ra các thông tin trừu tượng hơn cho các lớp tiếp theo. Trong mô hình mạng truyền ngược (feedforward neural network) thì mỗi neural đầu vào (input node) cho mỗi neural đầu ra trong các lớp tiếp theo.

Mô hình này gọi là mạng kết nối đầy đủ (fully connected layer) hay mạng toàn vẹn (affine layer). Còn trong mô hình CNNs 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 được các kết nối cục bộ. Như vậy mỗi neuron ở lớp kế tiếp sinh ra từ kết quả của filter áp đặt lên một vùng ảnh cục bộ của neuron trước đó.

Mỗi một lớp được sử dụng các filter khác nhau thông thường có hàng trăm hàng nghìn filter như vậy và kết hợp kết quả của chúng lại. Ngoài ra có một số layer khác như pooling/subsampling layer dùng để chắt lọc lại các thông tin hữu ích hơn (loại bỏ các thông tin nhiễu).

Trong quá trình huấn luyện mạng (traning) CNN tự động học các giá trị qua các lớp filter dựa vào cách thức mà bạn thực hiện. Ví dụ trong tác vụ phân lớp ảnh, CNNs sẽ cố gắng tìm ra thông số tối ưu cho các filter tương ứng theo thứ tự raw pixel > edges > shapes > facial > high-level features. Layer cuối cùng được dùng để phân lớp ảnh.



Hình 5. Cấu trúc mạng CNN

Trong mô hình CNN có 2 khía cạnh cần quan tâm là tính bất biến (Location Invariance) và tính kết hợp (Compositionality). Với cùng một đối tượng, nếu đối tượng này được chiếu theo các gốc độ khác nhau (translation, rotation, scaling) thì độ chính xác của thuật toán sẽ bị ảnh hưởng đáng kể.

Pooling layer sẽ cho bạn tính bất biến đối với phép dịch chuyển (translation), phép quay (rotation) và phép co giãn (scaling). Tính kết hợp cục bộ cho ta các cấp độ biểu diễn thông tin từ mức độ thấp đến mức độ cao và trừu tượng hơn thông qua convolution từ các filter. Đó là lý do tại sao CNNs cho ra mô hình với độ chính xác rất cao. Cũng giống như cách con người nhận biết các vật thể trong tự nhiên.

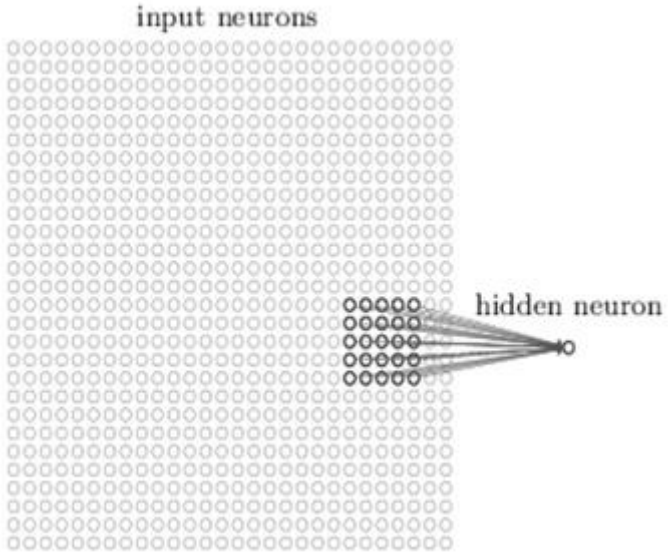
Mạng CNN sử dụng 3 ý tưởng cơ bản:

- **Các trường tiếp nhận cục bộ** (local receptive field)
- **Trọng số chia sẻ** (shared weights)
- **Tổng hợp** (pooling).

2.3.1 Trường tiếp nhận cục bộ (local receptive field)

Đầu vào của mạng CNN là một ảnh. Ví dụ như ảnh có kích thước 28×28 thì tương ứng đầu vào là một ma trận có 28×28 và giá trị mỗi điểm ảnh là một ô trong ma trận. Trong mô hình mạng ANN truyền thống thì chúng ta sẽ kết nối các neuron đầu vào vào tầng ảnh.

Tuy nhiên trong CNN chúng ta không làm như vậy mà chúng ta chỉ kết nối trong một vùng nhỏ của các neuron đầu vào như một filter có kích thước 5×5 tương ứng (28- 5 + 1) 24 điểm ảnh đầu vào. Mỗi một kết nối sẽ học một trọng số và mỗi neuron ẩn sẽ học một bias. Mỗi một vùng 5×5 đấy gọi là một trường tiếp nhận cục bộ.



Hình 6. Ma trận 28x28

Như vậy, local receptive field thích hợp cho việc phân tách dữ liệu ảnh, giúp chọn ra những vùng ảnh có giá trị nhất cho việc đánh giá phân lớp.

2.3.2 Trọng số chia sẻ (shared weight and bias)

Đầu tiên, các trọng số cho mỗi filter (kernel) phải giống nhau. Tất cả các nơ-ron trong lớp ẩn đầu sẽ phát hiện chính xác feature tương tự chỉ ở các vị trí khác nhau trong hình ảnh đầu vào. Chúng ta gọi việc map từ input layer sang hidden layer là một feature map. Tóm lại, một convolutional layer bao gồm các feature map khác nhau. Mỗi một feature map giúp detect một vài feature trong bức ảnh. Lợi ích lớn nhất của trọng số chia sẻ là giảm tối đa số lượng tham số trong mạng CNN.

2.3.3 Lớp tổng hợp (pooling layer)

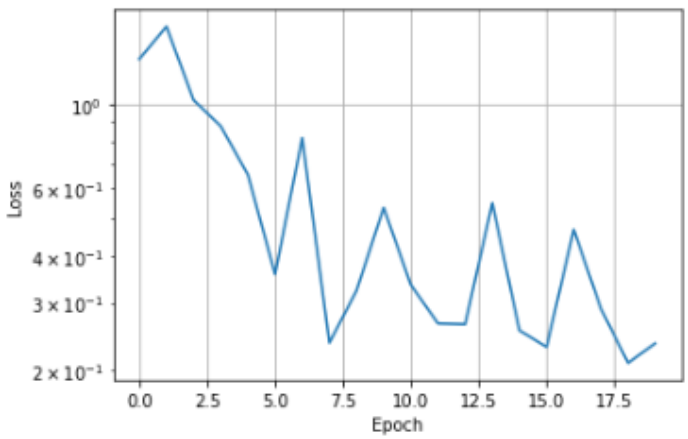
Lớp pooling thường được sử dụng ngay sau lớp convulational để đơn giản hóa thông tin đầu ra để giảm bớt số lượng neuron.

Như vậy qua lớp Max Pooling thì số lượng neuron giảm đi phân nửa. Trong một mạng CNN có nhiều Feature Map nên mỗi Feature Map chúng ta sẽ cho mỗi Max Pooling khác nhau. Chúng ta có thể thấy rằng Max Pooling là cách hỏi xem trong các đặc trưng này thì đặc trưng nào là đặc trưng nhất. Ngoài Max Pooling còn có L2 Pooling.

Cuối cùng ta đặt tất cả các lớp lại với nhau thành một CNN với đầu ra gồm các neuron với số lượng tùy bài toán.

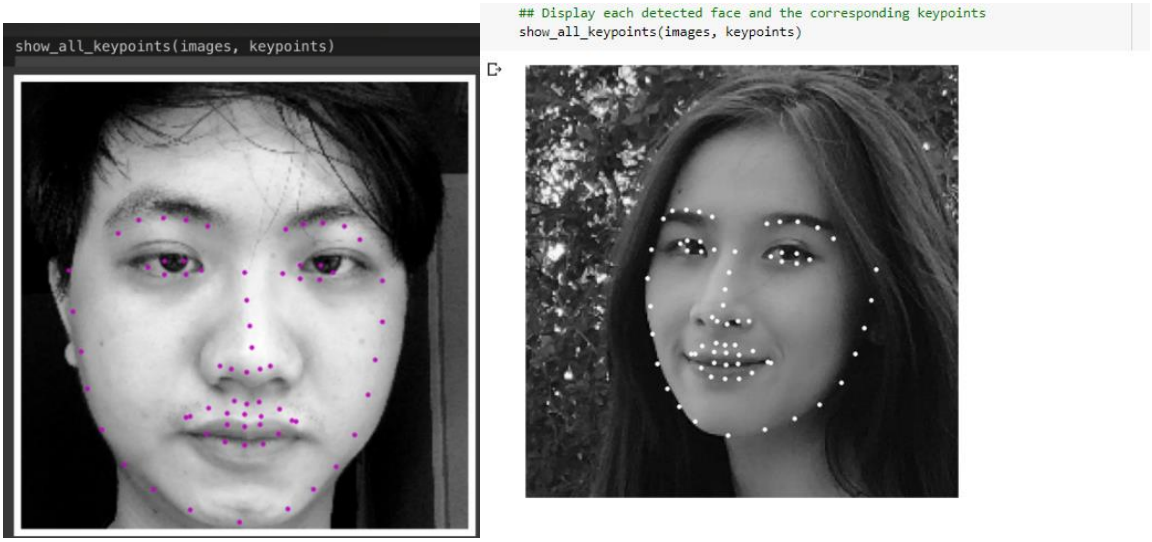
3. Kết quả

Mô hình CNN được xây dựng trong việc xác định các điểm mốc với độ chính xác 98%, cho ra kết quả chuẩn đoán chính xác các vị trí điểm mốc trên hình ảnh một khuôn mặt ngẫu nhiên. Đây là đồ thị giá trị loss sau khi được train:



Hình 7. Đồ thị giá trị mất mát theo số lần học

Các hình ảnh được đưa lên driver, sau khi xử lí dữ liệu sẽ được đưa vào để dự đoán, ta có được kết quả như hình 5:



Hình 8. Kết quả dự đoán trên Google Colab

4. Kết luận

Kết thúc quá trình huấn luyện cho ra kết quả khá chính xác với 98% so với thực tế. Do đó, nghiên cứu này bước đầu thành công do với mong đợi và thõa mãn phần lớn các yếu tố ban đầu đề ra.

Mô hình tạo ra tuy đã nhận diện được khuôn mặt và tạo ra được các điểm mốc cho khuôn mặt khá chính xác nhưng vẫn chưa đạt được giá trị tiệm cận tuyệt đối như những mô hình huấn luyện chuẩn khác

Mô hình chưa kết hợp với chạy thời gian thực và chưa có web app nên tính ứng dụng chưa cao.

Trong tương lai, hướng phát triển trực tiếp sẽ là định hình và nghiên cứu thêm khả năng ứng dụng của việc xác định các điểm Facial Landmark. Tiến hành huấn luyện lại cho model và áp dụng các thuật toán tăng cường nhằm tối ưu tốt nhất có thể cho các dự án sau

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1]. Mì AI, Xây dựng web app, triển khai model, Link truy cập: https://miai.vn/2021/12/29/xay-dung-web-app-trien-khai-model-trong-1-phut-voi-streamlit-mi-ai/

[2]. Trần Văn Huy, Introduction to Streamlit, Link truy cập: https://huytranvan2010.github.io/Introduction-to-streamlit/

[3]. TOPDev, Thuật toán CNN – Convolutional Neural Network, Link truy cập: https://topdev.vn/blog/thuat-toan-cnn-convolutional-neural-network/

Tiếng Anh

[1]. Streamlit, A faster way to build and share data apps, Available: https://streamlit.io/

[2]. DenseNet: Better CNN model than ResNet. Available: http://www.programmersought.com/article/7780717554/

[3]. Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in Proceedings of 2010 IEEE international symposium on circuits and systems, 2010, pp. 253-256.

[4]. ResNet, AlexNet, VGGNet, Inception: Understanding various architectures of Convolutional Networks. Available: https://cv-tricks.com/cnn/understand-resnet-alexnet-vgg-inception/

[5]. H. Jung, B. Kim, I. Lee, J. Lee, and J. Kang, "Classification of lung nodules in CT scans using three-dimensional deep convolutional neural networks with a checkpoint ensemble method," *BMC medical imaging*, vol. 18, p. 48, 2018.

S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, pp. 1345-1359, 2009.