Facial Keypoints Detection

Ngô Thành Phát

Ngày 1 tháng 1 năm 2023

Tóm tắt nội dung

Báo cáo này mô tả một cách tiếp cận để dự đoán các vị trí điểm chính của khuôn mặt, như một phần của cuộc thi Kaggle Nhận diện khuôn mặt (2016).Các điểm chính trên khuôn mặt bao gồm trung tâm và các góc của mắt, lông mày, mũi và miệng, và các vị trí khác. Phương pháp của chúng tôi bao gồm ba bước để tạo ra các dự đoán đầu ra của bài toán. Bước đầu tiên, chúng ta sẽ xem xét đánh giá tổng quan bộ dữ liệu. Tiếp theo, chúng ta sẽ thảo luận về mạng thần kinh tích chập mà chúng ta sẽ đào tạo. Cuối cùng, chúng tôi sẽ khái quát hóa bài toán để áp dụng cho nhiều trường hợp.

1 Giới thiệu

1.1 Lý do chọn đề tài

Ngày nay, phát hiện các điểm chính trên khuôn mặt đã trở thành một chủ đề phổ biến và các ứng dụng của nó bao gồm Face Filter (Những bức ảnh phủ lên khuôn mặt của mọi người với những điều thú vị, đã trở nên phổ biến. Chúng thường được tìm thấy trên các nền tảng truyền thông xã hội như Facebook, Instargram,), Snap Pay(Thanh toán bằng nhận diện khuôn mặt) , . . . đã thu hút một lượng lớn người dùng. Mục tiêu của phát hiện các điểm chính trên khuôn mặt là tìm ra khuôn mặt và các điểm chính trong một khuôn mặt nhất định, điều này rất khó khăn do đặc điểm khuôn mặt rất khác nhau từ người này sang người khác. Các ý tưởng về học sâu đã được áp dụng cho vấn đề này, chẳng hạn như mạng nơ-ron và mạng nơ-ron xếp tầng.Chúng tôi đặc biệt chọn cuộc thi Kaggle phát hiện điểm chính trên khuôn mặt vì nó cho chúng tôi nhiều cơ hội để thử nghiệm nhiều cách tiếp cận và mô hình mạng nơ-ron khác nhau để giải quyết vấn đề. Yếu tố cạnh tranh cũng cho phép chúng tôi so sánh kết quả của mình với cộng đồng lớn hơn và so sánh hiệu quả của các phương pháp của chúng tôi với các phương pháp khác.Phát hiện các điểm chính trên khuôn mặt là một vấn đề đầy thách thức với các biến thể về cả đặc điểm khuôn mặt cũng như điều kiện hình ảnh. Các đặc điểm khuôn mặt khác nhau tùy theo kích thước, vị trí, tư thế và biểu cảm, trong khi điều kiện hình ảnh thay đổi theo độ sáng và góc nhìn. Trong đồ án này, chúng ta sẽ xác định vị trí các điểm chính trong một hình ảnh nhất định bằng cách sử dụng kiến trúc học sâu để không chỉ có được tổn thất thấp hơn cho nhiệm vụ phát hiện mà còn tăng tốc quá trình đào tạo và thử nghiệm cho các ứng dụng trong thế giới thực.Dưới đây là ví dụ đầu vào(hàng đầu tiên) và đầu ra (hàng thứ hai).

1.2 Đặt vấn đề

Với sự phát triển nhanh chóng trong lĩnh vực thị giác máy tính, ngày càng có nhiều và nhiều công trình nghiên cứu và ứng dụng trong ngành tập trung vào phát hiện điểm chính trên khuôn mặt. Trong đồ án của chúng ta, chúng ta sẽ sử dụng các cấu trúc học sâu để phát hiện các điểm chính trên khuôn mặt, có thể học tốt từ các điểm khác nhau trên khuôn mặt và khắc phục sự khác biệt giữa các khuôn mặt của những người khác nhau hoặc của các điều kiện khác nhau ở mức độ lớn. Hai mô hình được sử dụng rộng rãi, Mạng nơ-ron một lớp ẩn(One Hidden Layer Neural Network) và Convolutional Neural Network(Mạng thần kinh tích chập). Mất mát được tính bằng sai số bình phương trung bình(MSE), thước đo hiệu quả của độ lệch về khoảng cách giữa 15 tọa độ điểm trên khuôn mặt thực và dự đoán.

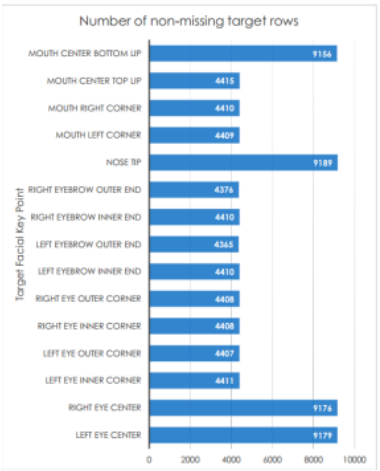
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2.$$

Mean Squared Error (MSE) có lẽ là số liệu phổ biến nhất được sử dụng cho các bài toán hồi quy. Về cơ bản, nó tìm thấy sai số bình phương trung bình giữa các giá trị được dự đoán và thực tế. MSE là thước đo chất lượng của một công cụ ước tính - nó luôn không âm và các giá trị càng gần 0 càng tốt.Trong đó n là số điểm dữ liệu, y là giá trị quan sát và ŷ là giá trị dự đoán. Nó thực hiện điều này bằng cách lấy khoảng cách từ các điểm đến đường hồi quy (những khoảng cách này là "sai số") và bình phương chúng. Bình phương là rất quan trọng để giảm độ phức tạp với các dấu hiệu tiêu cực. Nó cũng tạo ra nhiều trọng lượng hơn cho sự khác biệt lớn hơn.

2 Bộ Dữ liệu

2.1 Mô tả

Bộ dữ liệu của chúng tôi sử dụng được cung cấp bởi University of Montreal Medical , bao gồm dữ liệu đào tạo gồm 7049 hình ảnh 96x96x1 thang độ xám, với tọa độ (x, y) được gắn nhãn cho 15 điểm chính trên khuôn mặt. Các điểm chính trên khuôn mặt bao gồm bên phải, bên trái và trung tâm của mắt, lông mày, mũi, môi trên và môi dưới. Nhiệm vụ của chúng tôi là phát triển một mô hình có thể dự đoán các vị trí cụ thể của các điểm chính trên khuôn mặt này trên các hình ảnh của bộ thử nghiệm 1783, với hàm mất mát MSE (Lỗi bình phương trung bình). Trong số 7049 hình ảnh đào tạo, tồn tại một tập hợp con riêng biệt của 2140 được dán nhãn đầy đủ và chính xác, trong khi những ảnh còn lại đều không được gắn nhãn đầy đủ. Chúng ta sẽ chọn loại bỏ các hình ảnh khác vì chúng thường có chất lượng thấp hơn, có nhiều lỗi và được dán nhãn không đầy đủ.



Hình 1: Thông kê bộ dữ liệu

2.2 Thao tác trên bộ dữ liệu

Để làm sạch bộ dữ liệu có nhiều công cụ mạnh làm tốt công việc này và một trong số đó là Pandas. Pandas là một thư viện phân tích dữ liệu được sử dụng rộng rãi cho Python và chúng ta sẽ giá trị bị thiếu bằng một hàm cung cấp bởi Pandas đó là dropna(). Đơn giản là dropna() được sử dụng để loại bỏ các giá trị bị thiếu.

3 Mô hình

3.1 Sơ lược về Convolutional Neutral Network(CNN)

Mạng thần kinh tích chập là một loại mạng thần kinh nhân tạo chuyên dụng sử dụng một phép toán gọi là tích chập thay cho phép nhân ma trận tổng quát ở ít nhất một trong các lớp của chúng. Chúng được thiết kế đặc biệt để xử lý dữ liệu pixel và được sử dụng trong quá trình nhận dạng và xử lý hình ảnh.

Về kỹ thuật, mô hình CNN dùng để training và kiểm tra, mỗi hình ảnh đầu vào sẽ chuyển nó qua 1 loạt các lớp tích chập với các bộ lọc, tổng hợp lại các lớp được kết nối đầy đủ (FullConnected). Các CNN có một số bộ lọc/hạt nhân khác nhau bao gồm các tham số có thể huấn luyện có thể kết hợp với một hình ảnh nhất định về mặt không gian để phát hiện các tính năng như các cạnh và hình dạng. Số lượng bộ lọc cao này về cơ bản học cách nắm bắt các tính năng không gian từ hình ảnh dựa trên các trọng số đã học thông qua việc truyền ngược và các lớp lọc xếp chồng lên nhau có thể được sử dụng để phát hiện các hình dạng không gian phức tạp từ các tính năng không gian ở mọi cấp độ tiếp theo. Do đó, nếu các lớp của mô hình càng sâu nó có thể học được những tính năng ở mức trừu tượng cao.

3.2 Kiến trúc

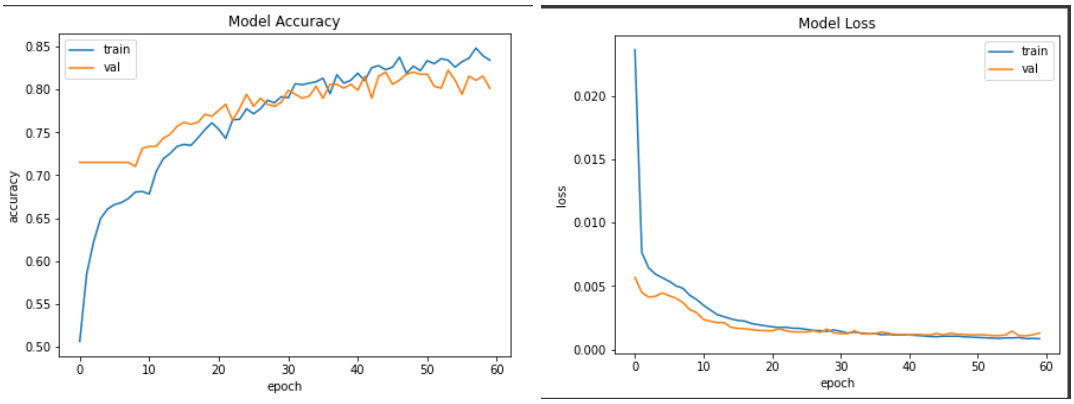
Chúng ta có một kiến trúc CNN chứa 5 lớp chính, và 1 lớp chính được kết nối đầy đủ và một lớp đầu ra mục tiêu. Mỗi lớp chính chứa 1 lớp tích chập(convotional layer) và một lớp tổng hợp(pooling layer) có kích thước 2x2. Lớp tích chấp đầu tiên có 16 bộ lọc, lớp 2 32, lớp 3 64, lớp 4 128,lớp 5 256. Tất cả bộ lọc này có kích thước 3x3 và chúng ta sẽ sử dụng chức năng kích hoạt(activation) 'relu' trong những lớp này. Các giá trị được làm phẳng ở lớp thứ 5(Flatten()) sẽ được đưa đến lớp kết nối đầy đủ với 512 nút(nodes) sau đó các giá trị cuối cùng sẽ được gửi đến lớp đầu ra có 30 nút Chúng ta sẽ đặt số lượng epochs thành 60, batch size thành 64, và sử dụng trình tối ưu hóa 'Adam' với chức năng lỗi bình phương trung bình

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 96, 96, 16)	160
max_pooling2d (MaxPooling2D)	(None, 48, 48, 16)	0
conv2d_1 (Conv2D)	(None, 48, 48, 32)	4640
max_pooling2d_1 (MaxPooling2D)	(None, 24, 24, 32)	0
conv2d_2 (Conv2D)	(None, 24, 24, 64)	18496
max_pooling2d_2 (MaxPooling2D)	(None, 12, 12, 64)	0
conv2d_3 (Conv2D)	(None, 12, 12, 128)	73856
max_pooling2d_3 (MaxPooling2D)	(None, 6, 6, 128)	0
conv2d_4 (Conv2D)	(None, 6, 6, 256)	295168
max_pooling2d_4 (MaxPooling2D)	(None, 3, 3, 256)	0
flatten (Flatten)	(None, 2304)	0
dense (Dense)	(None, 512)	1180160
dropout (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 30)	15390

Hình 2: Kiến trúc

3.3 Kết quả

Kết quả sơ bộ của chúng ta được hiển thị dưới đây:



Hình 3: Biểu đồ quá trình đào tạo

Như biểu đồ trên, chúng ta thấy biểu đồ tổn thất của quá trình đào tạo, sẽ là một đường cong, cả training loss và valid loss sẽ giảm xuống và gần như hội tụ. Điều này thể hiện mô hình không quá phù hợp với dữ liệu đào tạo nhưng thực sự có thể hiểu rõ các điểm trên khuôn mặt

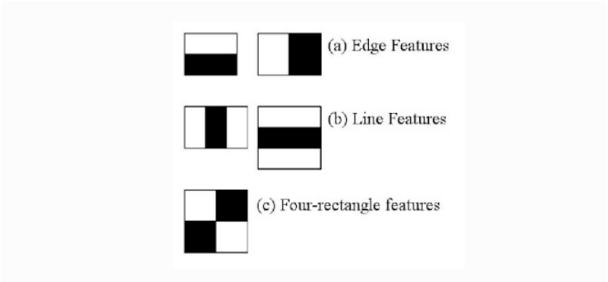
Val accuracy cho thấy độ chính xác của các dự đoán của một xác thực được phân tách ngẫu nhiên sau mỗi thời gian đào tạo. Với các biểu đồ này, chúng tôi đã đạt được mức độ val accuracy gần 83 phần trăm.

4 Khái quát bài toán

Chúng ta cần phải đảm bảo những hình ảnh đầu vào là những hình ảnh khuôn mặt màu xám có kích thước 96x96 chứ không phải một hình ảnh có nhiều khuôn mặt ở những vị trí khác nhau, điều này làm giảm khả năng khái quát của bài toán. Để giải quyết vấn đề này, chúng ta cần một hệ thống có thể giúp chúng ta phát hiện ra những khuôn mặt trong ảnh. Để thực hiện chúng ta có thể tạo ra một model khác hoặc sử dụng thuật toán nhận diện khuôn mặt Haar Cascade

4.1 Haar Cascade

Phát hiện đối tượng Sử dụng các phân loại tầng dựa trên tính năng Haar là một phương pháp phát hiện đối tượng hiệu quả được đề xuất bởi Paul Viola và Michael Jones trong bài báo của họ, "Rapid Object Detection using a Boosted Cascade of Simple Features" vào năm 2001. Tính năng Haar: Độ sáng của các vùng trên khuôn mặt là khác nhau. Ví dụ: Vùng mắt tối hơn vùng má, vùng mũi sáng hơn vùng hai bên. Sử dụng cửa sổ 24x24 để đánh các đặc trưng của ảnh. Tại mỗi pixel, bạn sẽ trích xuất những các đặc điểm,và sẽ phân loại các điểm mà bạn đã trích xuất là khuôn mặt hay không phải khuôn mặt



Hình 4: Haar Feature

Đây là một mô hình đã được đào tạo từ trước và cho phép mọi người sử dụng mà không phải đào tạo lại từ đầu. Tuy nhiên chúng phải xem xét một số yêu cầu khi sử dụng. Bao gồm: ảnh chính diện(khuôn mặt đủ lớn), ánh sáng tốt, không vật cản trước gương mặt, nước da sáng

5 Hướng cải thiện và phát triển

5.1 Dữ liệu

Áp dụng thêm các kỹ thuật Data Augmentation (flip, rotane, alter brightness, ...). Chọn lựa phù hợp các kỹ thuật tăng cường khác nhau để phù hợp với bộ dữ liệu. Tăng cường sự đa dạng của dữ liệu bằng cách thu thập thêm nhiều ảnh về các khuôn mặt ở độ tuổi khác nhau.

5.2 Mô hình

Áp dụng thêm nhiều pre-trained model khác nhau để có thể tìm được mô hình phù hợp nhất với bài toán và bộ dữ liệu.Áp dụng một số kỹ thuật như thay đổi cấu trúc mô hình, tùy chỉnh tham số để có thể cải thiện mô hình hơn

Tài liệu

https://www.youtube.com/watch?v=vC3bTziLRTA

https://www.kaggle.com/code/erdenabaatar/facial-keypoint-resnet50-aug-finetune

http://cs231n.stanford.edu/reports/2016/pdfs/007Report.pdf

https://www.kaggle.com/code/abhinandanreddy/facial-keypoint-detection-getting-started