Nhận biết ngôn ngữ ký hiệu sử dụng mạng thần kinh tích chập Inception V3

Nguyễn Đam Trường - 18021333
Nguyễn Hoàng Vũ - 18021435

Tóm tắt nội dung

Giao tiếp với người khiếm thính là một thử thách khó khăn. Người khiếm thính sử dụng ngôn ngữ ký hiệu để giao tiếp, người bình thường muốn hiểu được người khiếm thính nói gì cần phải đối mặt với việc học ngôn ngữ ký hiệu. Điều này gây cản trở trong giao tiếp. Ngôn ngữ ký hiệu sử dụng cử chỉ tay và biểu cảm khuôn mặt. Trong đó cử chỉ tay để nhận diện các chữ cái và các từ. Sử dụng mô hình mạng khoa học thần kinh học sâu để dự đoán dấu hiệu từ ngữ.

1 Mở đầu

Theo Tổ chức Y tế Thế giới (WHO) mới đây cảnh báo: Gần 2,5 tỷ người trên toàn thế giới (hoặc cứ 4 người thì có 1 người) sẽ sống với tình trạng khiếm thính ở một mức độ nào đó vào năm 2050.Ở Việt Nam, theo một số báo cáo nghiên cứu, tỷ lệ trẻ em bị điếc chiếm từ 1/1000 đến 5/1000. Như vậy, ước tính mỗi năm sẽ có 5000 trẻ bị điếc mới trên khoảng 1 triệu trẻ được sinh ra. Ngoài ra theo báo VOV đề cập, gần 1,3 triệu người khuyết tật câm điếc và khiếm thính (theo thống kê chưa đầy đủ ở nước ta) và con số thực tế có thể còn lớn hơn. Những người khiếm thính sẽ bị mất thính lực và rất khó để giao tiếp với người bình thường và ngược lại. Khi muốn giao tiếp một người bình thường sẽ cần học bảng ngôn ngữ ký hiệu để hiểu được ý của người khiếm thính. Ngôn ngữ ký hiệu là biểu diễn trực quan của cử chỉ tay, chuyển động của ngón tay, biểu hiện của khuôn mặt, chuyển động của cơ thể,v.v... Các quốc gia khác nhau có giao tiếp cử chỉ ký hiệu gần giống nhau. Những ngôn ngữ có liên quan trong cùng một họ ngôn ngữ có thể được mong đợi chia sẻ từ 36% đến 79% từ vựng cơ bản (ngôn ngữ ký hiệu của Mỹ và Pháp, những ngôn ngữ được xem là có liên quan trong cùng một họ ngôn ngữ chia sẻ khoảng từ 61% từ vựng cơ bản. Tuy nhiên, ngôn ngữ ký hiệu của Mỹ và Anh không có quan hệ gần nhau vì không chung một họ ngôn ngữ giống nhau, chúng chỉ có 31% cùng nguồn gốc trong từ vựng cơ bản. Cấu trúc ngôn ngữ ký hiệu thay đổi theo không gian và thời gian. Có các phương pháp dựa trên cảm biến và phương pháp dựa trên Vision.

Trong công nghệ nhận dạng cử chỉ dựa trên vision, một máy ảnh sẽ ghi lại chuyển động của cơ thể người, đặc biệt là cử chỉ tay để đưa ra ngôn ngữ ký hiệu.

Trong công nghệ dựa trên cảm biến, các chuyển động của bàn tay và ngón tay trong thời gian thực có thể được theo dõi bằng cách sử dụng cảm biến chuyển động Leap.

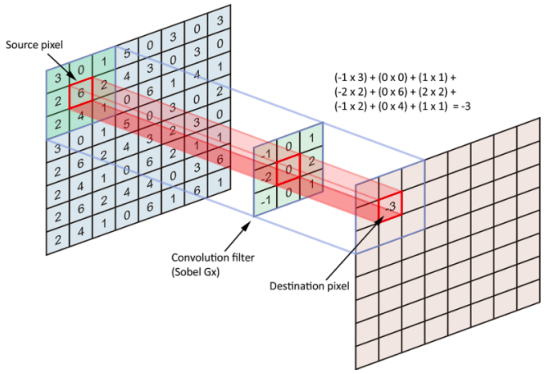
Dựa theo công nghệ nhận dạng cử chỉ dựa trên vision, nhóm đã sử dụng mạng thần kinh tích chập Ĩnception V3 để huấn luyện tạo ra một kiến trúc có thể nhận diện cử chỉ của tay để đưa ra ngôn ngữ.

2 Lí thuyết chung

1. Mạng tích chập (CNNs):

Convolutional Neural Network là một trong những mô hình Deep Learning tiên tiến, CNN được sử dụng nhiều trong các bài toán nhận dạng các object trong ảnh. Nó giúp cho chúng ta xây dựng được những hệ thống thông minh với độ chính xác cao như hiện nay.

Phép tích chập là một cửa sổ trượt trên một ma trận như trong hình:

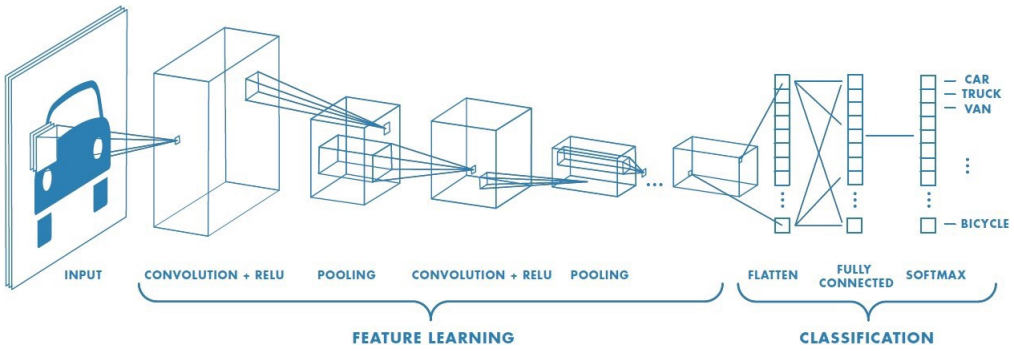


Hình 1: Phép tích chập thực hiện trên một ảnh

Convolution hay tích chập nhân từng phần tử trong filter rồi tính tổng của chúng, thay thế cho giá trị điểm ảnh ở trung tâm của filter. 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. Kết quả là một ma trận gọi là Convoled feature được sinh ra từ việc dịch filter đi toàn bộ ảnh gốc.

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. Trong mô hình mạng truyền ngược (feedforward neural network) thì mỗi neural là đầu vào cho mỗi neural trong các lớp tiếp theo. 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. 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.

Trong quá trình huấn luyện mạng CNN tự động học các giá trị qua các lớp filter dựa vào cách thức 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 2: Mạng tích chập

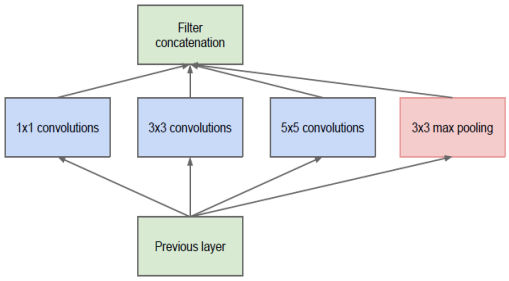
3 Cấu trúc Inception

Inception là một mô hình được công bố bởi Google. Ý tưởng xuất phát từ việc phải quyết định xem kích cỡ của cửa sổ trượt trong mỗi lớp tích chập trong mô hình mạng neural, có thể là 1x1, 3x3 hoặc 5x5. Thay vì phải lựa chọn, mô hình Inception bao gồm tất cả cửa sổ trượt đó và để mô hình tự lựa chọn. Việc đó có thể được thực hiện bằng cách tính tích chập với các kích thước của cửa sổ trượt khác nhau và nối các feature map lại trước khi đưa vào layer tiếp theo. Giả sử lớp tiếp theo cũng là một Inception, mỗi feature map của lớp Inception trước sẽ được truyền qua một hỗn hợp tích

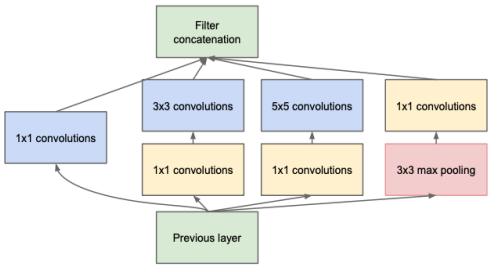
chập của lớp hiện tại. Thú vị ở chỗ, chúng ta không cần quan tâm về thứ tự thực hiện nhân tích chập của các cửa sổ trượt, ví dụ 3x3 sau đó 5x5. Thay vào đó, ta tính tích chập toàn bộ và để mô hình tự chọn những thuộc tính tốt nhất. Thêm nữa, kiến trúc này sẽ cho phép mô hình trích xuất được những thuộc tính khu vực dựa vào những cửa sổ trượt nhỏ và những thuộc tính trừu tượng lớn dựa vào tích chập với những cửa sổ trượt lớn hơn.

Hình 3: Ý tưởng của cái tên Inception được bắt nguồn từ hình ảnh trong bộ phim cùng tên

Kiến trúc Inception sử dụng đa dạng các cửa sổ trượt tích chập, đặc biệt là các cửa sổ 1x1, 3x3, và 5x5 convolutions bên cạnh 3x3 max pooling. Việc sử dụng đồng thời nhiều cửa số một lớp đòi hỏi một khối lượng tính toán rất lớn. Do đó bài báo có đề cập tới một phương pháp để giảm khối lượng tính toán, giải pháp đầu tiên thực hiện tích chập 1x1 đóng vai trò như một nút thắt cổ chai để giảm kích thước của các feature map, sau đó truyền các feature map qua hàm relu, sau đó tiếp tục tính tích chập lớn hơn (5x5 hoặc 3x3).



(a) Inception module, phiên bản naive



(b) Inception module, giảm số chiều

Hình 4: Inception module

4 Nhận diện ngôn ngữ kí tự

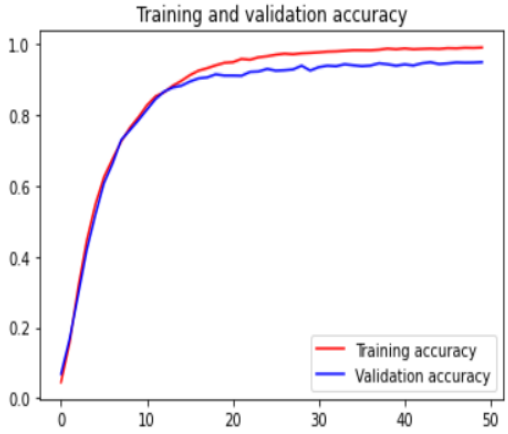
Để thực hiện nhận diện ngôn ngữ kí hiệu, chúng tôi sử dung phương pháp transfer learning, model đã được huấn luyện sẵn trên một dataset cực lớn với hàng trăm giờ huấn luyện trên các GPUs công suất lớn. Để thực hiện nhiệm vụ này, chúng tôi sử dụng mô hình InceptionV3, mô hình đã được huấn luyện với dataset ImageNet gồm 1000 lớp với hơn 1 triệu ảnh huấn luyện.

Nhiêm vụ phân loại ngôn ngữ kí hiệu theo bảng chữ cái ASL(American sign language) bao gồm phân loại 51 kí tự. Mô hình InceptionV3 sau khi được lấy từ keras được xếp thêm một lớp dense với 1024 neuron, và đầu ra gồm 51 neuron sử dụng hàm kích hoạt softmax tương ứng với 51 kí tự. Mạng IncepionV3 bao gồm 314 lớp, trong đó chúng tôi đóng băng 249 lớp đầu tiên và thục hiện huấn luyện

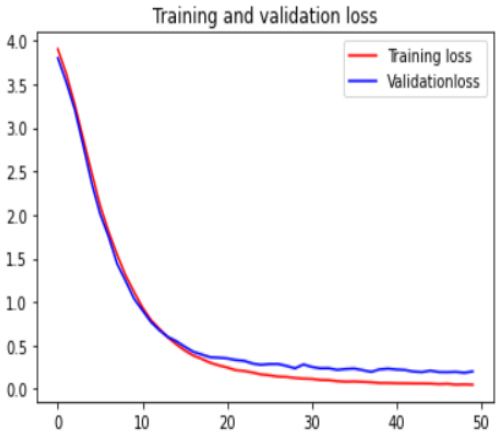
với tất cả các lớp còn lại. Chúng tôi tiếp tục Huấn luyện mô hình phân loại ngôn ngữ kí hiệu với bộ dataset ASL_and_same_words được lấy từ Kaggle, gồm 4000 ảnh cỡ 200x200x3 cho mỗi kí tự.

5 Kết quả

Sau khi mô hình được huấn luyện qua 50 epoch, độ chính xác của mô hình lên tới 0.9899 trên tập huấn luyện, 0.9488 trên tập validation và mất mát trên tập huấn luyện bằng 0.0502, 0.2026 trên tập validation.



(a) Độ chính xác trên tập huấn luyện và tập validation



(b) Mất mát trên tập huấn luyện và tập validation