SO SÁNH CÔNG CỤ PHÂN LOẠI HOG-SVM VÀ CNN SỬ DỤNG TRONG MÔ HÌNH NHẬN DẠNG GIỌNG NÓI

Nguyễn Huy Thế, Nguyễn Tuấn Anh
Trường Đại học Thủy lợi, email: nguyenhuythe@tlu.edu.vn

1. GIỚI THIỆU CHUNG

Với sự tiện lợi và linh hoạt, việc áp dụng bộ công cụ nhận dạng giọng nói đang dần trở thành tính năng không thể thiếu trong thiết bị bị thông minh hiện nay. Các công cụ này thường được phát triển dựa trên việc trích xuất các đặc trưng của giọng nói và xây dựng các mô hình nhận dạng dựa trên các đặc trưng đó. Việc lựa chọn mô hình phân loại là bước quan trọng bởi các đặc điểm của mô hình như độ phức tạp, phương pháp huấn luyện có ảnh hưởng lớn đến kết quả nhận dạng và nền tảng phần cứng triển khai.

Hiện nay, có rất nhiều các mô hình nhận dạng đã được phát triển. Nghiên cứu này tập trung vào việc nhận dạng giọng nói bằng hai bộ công cụ: Histogram of Oriented Gradient (HOG) kết hợp với Support Vector Machine (SVM) và mạng nơ-ron tích chập (Convolution Neural Network - CNN). Sau khi thu được bộ dữ liệu đặc trưng của âm thanh Mel Frequency Celtral Coefficient (MFCC), các dữ liệu này sẽ được sử dụng để huấn luyện các mô hình phân loại. Việc tính toán và huấn luyện cho các mô hình này đều được thực hiện bởi ngôn ngữ lập trình mã nguồn mở Python.

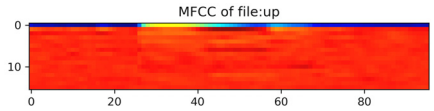
2. PHƯƠNG PHÁP NGHIÊN CỨU

Quá trình xây dựng mô hình nhận dạng giọng nói bao gồm hai bước: thu thập đặc trưng âm thanh và huấn luyện mô hình phân loại.

2.1. Thu thập bộ dữ liệu đặc trưng MFCC

Bộ dữ liệu âm thanh được sử dụng trong nghiên cứu này thuộc tập dữ liệu Google Speech Command datasets [1]. Tập dữ liệu chứa các tệp thu âm ở định dạng .wav của hơn 30 từ tiếng Anh với thời gian khoảng một giây. Nghiên cứu này chỉ sử dụng các bộ dữ liệu của chín từ khóa để tiến hành huấn luyện các mô hình nhận dạng.

Các đặc trưng của các dữ liệu âm thanh nêu trên sẽ được tính toán thông qua kỹ thuật MFCC. Điểm đáng chú ý của kỹ thuật này là việc xây dựng thang đo Mel tương tự với cách tai người cảm nhận âm thanh, ở đó các bộ lọc tần số được bố trí đều nhau tại tần số thấp và được bố trí theo thang logarit đối với các tần số cao, khi đó sẽ thu được các đặc tính quan trọng của tín hiệu giọng nói [2]. Bộ dữ liệu MFCC có cấu trúc mảng hai chiều được minh họa trong Hình 1. Trong nghiên cứu này, bộ dữ liệu đặc trưng của âm thanh MFCC được tính toán bằng cách sử dụng bộ thư viện python-speech-feature trong ngôn ngữ Python.



Hình 1. *Bộ dữ liệu MFCC.*

2.2. Xây dựng mô hình sử dụng bộ công cụ HOG kết hợp SVM

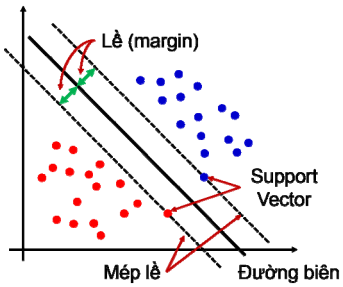
2.2.1. Histogram of Oriented Gradient

Bộ mô tả HOG là một kỹ thuật phổ biến trong lĩnh vực thị giác máy tính trong các bài toán phát hiện và nhận diện đối tượng. Kỹ thuật này tính toán các đặc trưng cục bộ của bức ảnh dựa trên thông tin về độ lớn và hướng của gradient tại mỗi điểm ảnh. Quá trình tính toán véctơ đặc trưng này diễn ra

theo các bước như sau: chia bức ảnh thành các ô có kích thước xác định, sau đó thực hiện tính toán đạo hàm cho từng điểm ảnh theo phương x và y, thu được độ lớn và hướng của gradient cho từng điểm ảnh trong mỗi ô. Véctơ đặc trưng cho mỗi ô này được tính bằng cách xác định độ lớn gradient cho hướng được lựa chọn. Các véctơ đặc trưng cho các ô trong một khối sẽ được ghép với nhau và được chuẩn hóa. Kết hợp tất cả các véctơ của các khối trong bức ảnh sẽ thu được véctơ đặc trưng cho cả bức ảnh. Trong nghiên cứu này, véctơ đặc trưng HOG được tính toán bằng cách sử dụng bộ thư viện skimage trong ngôn ngữ Python.

2.2.2. Support Vector Machine

SVM là môt thuật toán của lớp bài toán học giám sát, được sử dụng chủ yếu trong các bài toán phân loại, ở đó dữ liệu đầu vào có n chiều. Nhiệm vụ của SVM là tìm ra siêu phẳng (hyper-plane) để phân chia các bộ dữ liệu đầu vào này theo các nhãn đã được định sẵn. Bài toán SVM được minh họa như trong Hình 2.



Hình 2. *Bài toán SVM phân loại nhị phân.*

Siêu phẳng sẽ tạo ra biên giới phân chia các bộ dữ liệu. Các điểm dữ liệu gần với đường biên này nhất được gọi là các véctơ hỗ trợ (support vector), khoảng cách từ các điểm này đến đường biên được gọi là lề (margin). Có rất nhiều đường có thể được sử dụng làm biên giới để phân chia bộ dữ liệu. Thuật toán SVM tập trung vào tìm đường biên có lề rộng nhất, tương ứng với khả năng phân loại dữ liệu tốt nhất [3]. Trong nghiên cứu này, mô hình SVM được huấn luyện bằng cách sử dụng bộ thư viện scikit-learn trong ngôn ngữ Python.

2.3. Mô hình sử dụng mạng nơ-ron tích chập

Mạng nơ-ron tích chập CNN là một mạng nơ-ron nhân tạo có cấu trúc đặc biệt, được sử dụng phổ biến trong các bài toán phân loại hình ảnh, nhận diện đối tượng. Cấu trúc của mô hình CNN gồm hai lớp chính: lớp trích xuất thông tin và lớp phân loại. Lớp trích xuất lấy thông tin đặc trưng của bộ dữ liệu đầu vào thông qua việc sử dụng các phép tính tích chập (convolution), phép gộp dữ liệu (pooling). Dữ liệu đặc trưng này sẽ được dàn phẳng thành một véctơ và đưa đến lớp phân loại, là lớp mạng nơ-ron kết nối đầy đủ (fully connected). Bên cạnh các lớp nơ-ron thông thường, lớp đầu ra sử dụng hàm kích hoạt là softmax để đưa ra xác suất xuất hiện của các lớp đầu ra.

Trong nghiên cứu này, mô hình mạng CNN được huấn luyện bằng cách sử dụng bộ thư viện tensorflow trong ngôn ngữ Python.

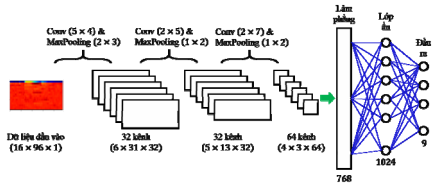
3. KẾT QUẢ NGHIÊN CỨU

Các từ tiếng Anh được sử dụng trong nghiên cứu này bao gồm ‘down’, ‘go’, ‘left’, ‘no’, ‘off’, ‘right’, ‘stop’, ‘up’, ‘yes’. Để thuận tiện cho việc huấn luyện mô hình, các từ này sẽ được đánh nhãn tương ứng từ 0 đến 8. Dữ liệu âm thanh của các từ này sau khi đưa qua phương pháp trích xuất MFCC có kích thước là 16×96 điểm ảnh.

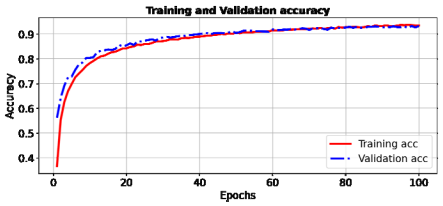
Trong phần tính toán véctơ HOG, các ô có kích thước 8×8 điểm ảnh, mỗi véctơ đặc trưng cho các ô có 9 phần tử đại diện cho các góc hướng 0°, 20°, ..., 160°. Các khối có kích thước là 2×2 ô = 16×16 điểm ảnh. Như vậy, mỗi khối sẽ có véctơ đặc trưng có độ dài là 2×2×9 = 36 phần tử. Bộ dữ liệu MFCC có kích thước là 16×96 điểm ảnh nên sẽ có 11 khối. Do đó, véctơ HOG đặc trưng cho mỗi bộ dữ liệu MFCC có kích thước là 36×11 = 396 phần tử.

Cấu trúc của mạng CNN được thể hiển trong Hình 3. Trong đó, lớp trích xuất gồm ba lớp tích chập. Dữ liệu sau các lớp tích chập được dàn phẳng thành véctơ có kích thước 768 phần tử và được đưa vào lớp phân loại có một lớp ẩn và một lớp đầu ra. Quá trình huấn

luyện mô hình CNN được biểu diễn trong Hình 4. Độ chính xác của mô hình với tập dữ liệu huấn luyện (training) và tập dữ liệu xác nhận (validation) qua các giai đoạn đều tăng lên và đạt kết quả rất tốt (trên 90%).



Hình 3. *Cấu trúc mạng CNN*



Hình 4. *Quá trình huấn luyện mô hình CNN*

Áp dụng các mô hình nhận dạng trên với tập dữ liệu kiểm tra và tính toán hiệu suất của mô hình thông qua các thông số Precision, Recall và F1-score [4], thu được kết quả như trong Bảng 1 và Bảng 2. Giá trị của các thông số đánh giá tương ứng với phương pháp HOG-SVM cho thấy cách tiếp cận này có hiệu suất trung bình, mô hình vẫn có khả năng phân loại dữ liệu. Đối với mô hình sử dụng CNN, các giá trị trên đều lớn hơn 0.8, chứng tỏ phương pháp có khả năng phân loại rất tốt với ít sai sót.

Bảng 1. Kết quả nhận dạng sử dụng HOG-SVM

Từ khóa	Precision	Recall	F1-score
‘down’	0.53	0.63	0.57
‘go’	0.50	0.48	0.49
‘left’	0.69	0.72	0.70
‘no’	0.54	0.72	0.53
‘off’	0.67	0.65	0.66
‘right’	0.79	0.78	0.78
‘stop’	0.69	0.64	0.66
‘up’	0.59	0.6	0.59
‘yes’	0.76	0.72	0.74
Trung bình	0.64	0.64	0.64

Bảng 2. Kết quả nhận dạng sử dụng CNN

Từ khóa	Precision	Recall	F1-score
‘down’	0.91	0.91	0.91
‘go’	0.92	0.87	0.90
‘left’	0.93	0.89	0.91
‘no’	0.92	0.93	0.92
‘off’	0.91	0.96	0.93
‘right’	0.95	0.95	0.95
‘stop’	0.96	0.96	0.96
‘up’	0.83	0.87	0.85
‘yes’	0.97	0.98	0.97
Trung bình	0.92	0.92	0.92

4. KẾT LUẬN

Bài báo trình bày về hai bộ công cụ phân loại sử dụng trong mô hình nhận dạng giọng nói bao gồm HOG-SVM và CNN. Trong cả hai mô hình, bộ trích xuất đặc trưng lấy dữ liệu từ bộ MFCC của âm thanh và đưa qua bộ phân loại. Việc áp dụng mạng CNN rất hiệu quả do quá trình trích xuất đặc trưng dữ liệu và phân loại trong mạng CNN được huấn luyện đồng thời với các tham số được điều chỉnh phù hợp. Tuy nhiên, mạng CNN có cấu trúc phức tạp nên quá trình huấn luyện cần nhiều thời gian hơn. Do đó, hướng nghiên cứu khai thác cách mô tả đặc trưng dữ liệu HOG và cải tiến cấu trúc SVM cũng rất cần thiết để có thể triển khai mô hình nhận dạng gọn nhẹ trên nền tảng phần cứng.

5. TÀI LIỆU THAM KHẢO

[1] P. Warden. (2018). A dataset for limited-vocabulary speech recognition. arXiv: 1804.03209.

[2] Alim, S. A., & Rashid, N. K. A. (2018). Some commonly used speech feature extraction algorithms (pp. 2-19). London, UK:: IntechOpen.

[3] Chang, C. C. (2001). A library for support vector machines.

[4] Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correla-tion. arXiv:2010.16061.