GIẤY • TRUY CẬP MỞ

Nhận dạng giọng nói bằng cách sử dụng Convolution Deep Mạng lưới thần kinh

Cách trích dẫn bài viết này: Ayad Alsobhani et al 2021 J. Vật lý: Hội thảo. Ser. 1973 012166

Tạp chí Vật lý: Chuỗi hội thảo doi:10.1088/1742-6596/1973/1/012166

Nhận dạng giọng nói bằng cách sử dụng Mạng thần kinh sâu Convolution

Ayad Alsobhani1 *, Hanaa MA ALabboodi2 [2,] Haider Mahdi 3 [3]

[1] Khoa Kỹ thuật, Khoa Điện, Đại học Babylon, Iraq.
2Phó giáo sư Tiến sĩ, Khoa Kỹ thuật, Khoa Điện, Đại học Babylon, Iraq.

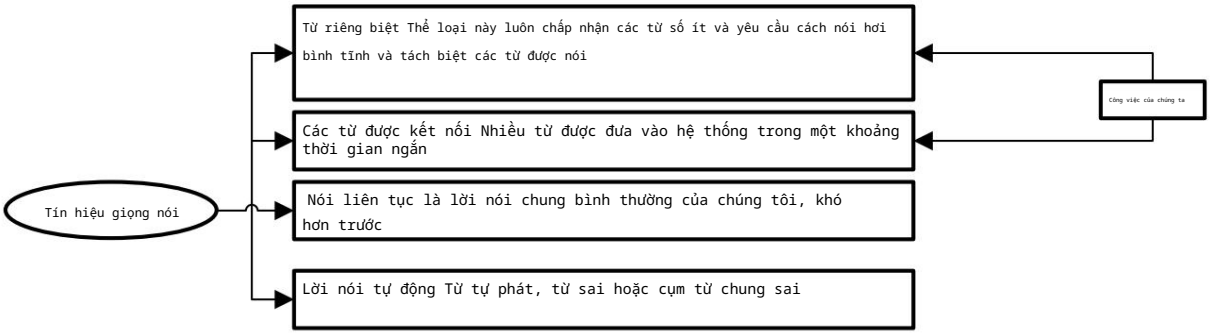
3Phó giáo sư Tiến sĩ, Khoa Kỹ thuật, Khoa Điện, Đại học Babylon, Iraq.

*Email tương ứng của tác giả: ayadsobhan1992@gmail.com

Trừu tượng. Việc sử dụng mô hình nhận dạng giọng nói đã trở nên cực kỳ quan trọng. Kiểm soát lời nói đã trở thành một loại hình quan trọng; Dự án của chúng tôi đã nghiên cứu thiết kế một mô hình theo dõi từ bằng cách áp dụng các tính năng nhận dạng giọng nói với khả năng học tập thần kinh tích chập sâu. Sáu từ điều khiển được sử dụng (bắt đầu, dừng, tiến, lùi, phải, trái). Lời nói từ những người ở các độ tuổi khác nhau. Hai phần bằng nhau, nam và nữ, đóng góp vào tập dữ liệu giọng nói của chúng tôi, được sử dụng để đào tạo và kiểm tra các mạng lưới thần kinh sâu được đề xuất. Thu thập dữ liệu ở những nơi khác nhau trên đường phố, công viên, phòng thí nghiệm và chợ. Các từ có độ dài từ 1 đến 1,30 giây đối với 30 người. Mạng thần kinh chuyển đổi (CNN) được áp dụng như mạng thần kinh sâu tiên tiến để phân loại từng từ từ tập dữ liệu gộp của chúng tôi dưới dạng nhiệm vụ phân loại nhiều lớp. Mạng lưới thần kinh sâu được đề xuất đã trả về 97,06% độ chính xác khi phân loại từ với mẫu giọng nói hoàn toàn không xác định. CNN được sử dụng để đào tạo và kiểm tra dữ liệu của chúng tôi. Công trình của chúng tôi khác biệt so với nhiều bài báo khác thường sử dụng dữ liệu làm sẵn và khá nhất quán của loại từ biệt lập. Mặc dù dữ liệu của chúng tôi được thu thập trong các môi trường ồn ào khác nhau trong các điều kiện khác nhau và từ hai loại giọng nói, từ đơn lẻ và từ liên tục.

1. Giới thiệu

Nhận dạng giọng nói tự động là phương pháp dịch tín hiệu giọng nói thành một chuỗi từ bằng chương trình máy tính và các thuật toán của nó. Mục tiêu chính của nhận dạng giọng nói là cho phép máy móc nhận dạng âm thanh và xử lý chúng. Khả năng máy tính xác định lời nói "nhận và giải thích" và dịch nó thành dạng hoặc văn bản có thể đọc được được gọi là nhận dạng giọng nói tự động. Nhận dạng giọng nói tự động là khả năng máy tính hiểu được lời nói cũng như thực hiện hành động dựa trên hướng dẫn của con người [1]. Âm vị có ba phần xử lý, nhận biết lời nói và nhận biết người nói và phần thứ ba là nhận biết cảm xúc. Các từ được hiển thị bằng văn bản hoặc bằng thiết bị. Chúng được đọc dưới dạng một lệnh cụ thể như những gì đã được thực hiện trong công việc của chúng tôi, khả năng nhận dạng tín hiệu giọng nói có thể được chia thành ba loại dựa trên loại tín hiệu giọng nói và độ dài của nó [2] như trong hình (1)



Hình 1. Các kiểu giọng nói

Nhận dạng âm thanh có vai trò quan trọng trong việc phát triển mạch điều khiển ô tô hoặc ứng dụng robot trong các thiết bị bảo vệ hộ gia đình. Sóng âm có thể được định nghĩa là sóng dọc truyền qua giai đoạn nén và giải nén đoạn nhiệt trong hầu hết các trường hợp. Sóng dọc dao động cùng phương khi truyền đi. Trong xử lý âm thanh, biểu đồ phổ là các mẫu hai chiều hiển thị tần số trên trục tung và thời gian trên trục hoành, biểu thị năng lượng tín hiệu. Nói chung, sự chuyển động của không khí trong đường phát âm tạo ra phụ âm, từ đó phát sinh ra nhiều âm thanh khác nhau. Lộ trình chung để hoàn thành nhận dạng giọng nói được giải thích trong hình (2).

Hình 2. Lộ trình chung để hoàn thành nhận dạng giọng nói.

Con đường chung để hoàn thiện lý thuyết nhận dạng giọng nói là trình bày, ADC (bộ chuyển đổi tương tự sang số) (định hình quang phổ) Chọn sóng âm thanh và chuyển đổi nó thành dạng kỹ thuật số sau đó lọc trước nhấn mạnh theo giai đoạn trích xuất đặc trưng. Các tính năng được trích xuất sẽ gửi đến phân loại dưới dạng giai đoạn được nhắm mục tiêu. Một số kỹ thuật liên quan đến chuyển đổi tham số, đòi hỏi phải chuyển đổi các đặc tính thu được thành tham số tín hiệu bằng cách sử dụng quy trình tách và ghép. Chuyển đổi các tham số trong vectơ quan sát tín hiệu là một phần của mô hình thống kê [3]. Việc nhận dạng âm thanh đóng một vai trò quan trọng trong hệ thống kiểm soát truy cập và bảo mật [4]. Sóng âm là sự rung động hình sin đặc trưng của các âm lớn, rung nhanh và có tần số cao hơn các âm trầm. Năng lượng âm thanh rung được micro chuyển thành năng lượng điện. Hình dạng chung của sóng âm mang lại ấn tượng về lượng năng lượng theo biên độ của tín hiệu. Bản chất của sóng âm thanh là các tần số thay đổi ở chỗ nó bám chặt vào nhau, các thành phần tần số tạo nên sóng âm thanh được phân tích bằng FFT, làm nổi bật nó bằng sơ đồ gọi là sơ đồ phổ. [5]. Sóng hình sin trong không khí được tạo ra bởi giọng nói.

Các âm cao hơn dao động ở tần số cao hơn các âm thấp hơn, do đó chúng rung nhanh hơn. Một micrô có thể cảm nhận được những âm thanh này, sau đó chúng được chuyển đổi từ năng lượng âm thanh mang trong sóng âm thành năng lượng điện và được ghi lại dưới dạng tín hiệu âm thanh. Biên độ của tín hiệu giọng nói cho biết mức năng lượng âm thanh có trong âm thanh và do đó mức độ ồn của nó. Đồng thời, giọng nói của chúng ta được tạo thành từ nhiều tần số khác nhau. Tín hiệu cuối cùng là sản phẩm của việc cộng tất cả các tần số đó lại với nhau. Các tần số thành phần được sử dụng làm đặc điểm để diễn giải tín hiệu tốt hơn. Để phân tách tín hiệu thành các thành phần này, biến đổi Fourier đã được áp dụng. Đối với nhiệm vụ này, thuật toán FFT (Biến đổi Fourier nhanh) thường có sẵn. Âm thanh biến đổi thành biểu đồ phổ bằng cách sử dụng kỹ thuật phân tách này. tín hiệu được chia thành các khung thời gian để tạo ra biểu đồ phổ. Sau đó, bằng cách sử dụng FFT, mỗi khung sẽ chia từng khung thành các thành phần tần số. Một vectơ biên độ ở mỗi tần số đã được sử dụng để mô tả từng khung thời gian. Biểu đồ phổ có thể được căn chỉnh theo thời gian với tín hiệu âm thanh gốc để có được biểu diễn đồ họa của các thành phần âm thanh [5].

Liên quan đến công việc của chúng tôi, sáu lời nói của ba mươi người ở những nơi khác nhau đã được thu thập. Mẫu bài phát biểu đóng góp của chúng tôi có độ tuổi và giới tính khác nhau. Độ dài của các từ được ghi rất khác nhau khiến chúng tôi gặp rất nhiều khó khăn. Một số chương trình đã được sử dụng để giải quyết những khác biệt này (Audacity và Adobe Audition). Các đặc điểm giọng nói đã được trích xuất và huấn luyện bằng cách sử dụng Mạng thần kinh chuyển đổi (CNN). CNN, phương pháp học sâu tiên tiến nhất, đã được áp dụng và hiệu suất nhận dạng tín hiệu giọng nói như một quy trình phân loại đa lớp đã được nghiên cứu. Các loại tham số có thể học CNN khác nhau đã được thử nghiệm và cập nhật để tìm ra cấu trúc CNN tốt nhất có thể giải quyết vấn đề phân loại nhiều lớp của chúng tôi. Động lực của việc sử dụng mô hình deep learning là tìm ra mô hình tốt nhất phù hợp với điều kiện dữ liệu chúng ta thu thập để hoàn thiện quá trình kiểm soát. Như đã đề cập trước đó, dữ liệu ghi âm có nhiều vấn đề về cách phát âm, vị trí phát âm của các từ và lượng tiếng ồn trong đó cũng như tiếng ồn nền và tiếng ồn của thiết bị ghi âm. Nói chung, dữ liệu được thu thập trong những trường hợp như vậy do công việc của chúng tôi nhằm kiểm soát chuyển động của máy thông qua thứ tự tín hiệu giọng nói cho các ứng dụng thời gian thực.

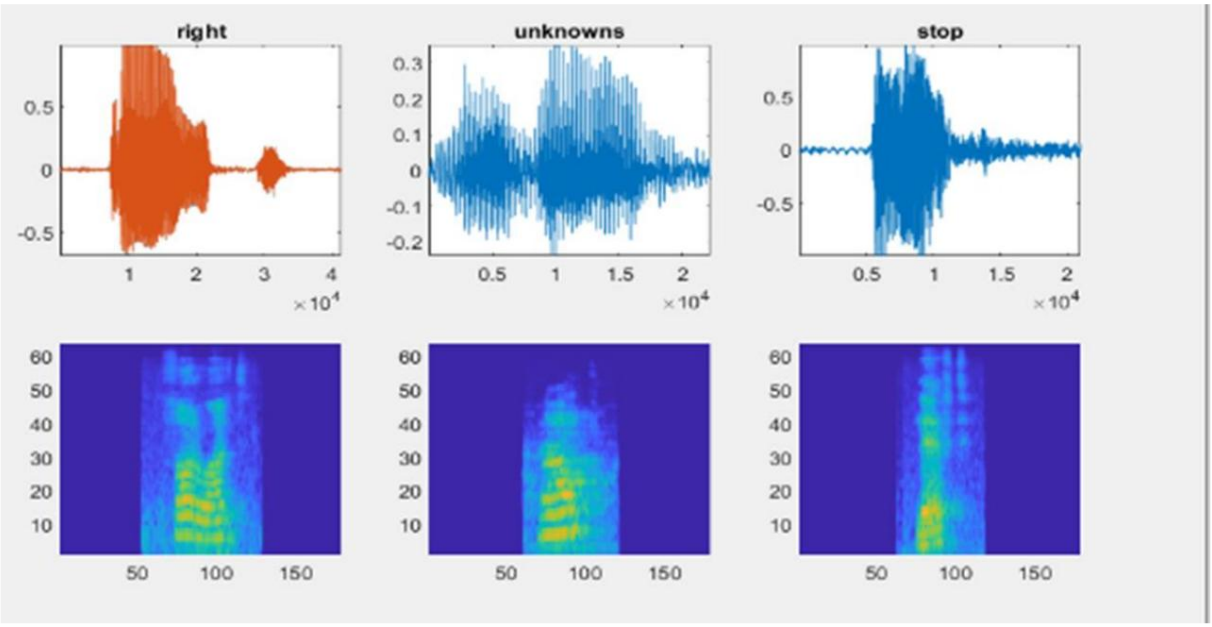
Điều khác biệt giữa công việc của chúng tôi với các công việc khác trong lĩnh vực này là dữ liệu mà chúng tôi tự thu thập ở các khu vực ghi âm khác nhau có tiếng ồn cao, trung bình và ít, chẳng hạn như ở chợ, nhà ở, vườn và phòng thí nghiệm. Đối với nghiên cứu khác, nó chủ yếu sử dụng dữ liệu sẵn sàng, đơn giản và rõ ràng, được ghi lại ở những khu vực không có tiếng ồn và thứ hai chúng tôi sử dụng hai loại từ là các từ riêng biệt như (dừng và bắt đầu)
Và các từ được kết nối (hình 1) như (ngược lại) và các nghiên cứu khác phần lớn chỉ sử dụng các từ riêng biệt. Ưu điểm thứ ba trong công việc của chúng tôi là chúng tôi có thể đạt được hiệu quả tốt so với các nghiên cứu khác bằng cách thực hiện thuật toán xử lý nhiều dữ liệu khác nhau và lọc chúng theo cách cho phép bộ phân loại phân biệt các từ với độ chính xác cao [6,7,8].

2. Công việc liên

quan Vào những năm 50 của thế kỷ trước, nghiên cứu về nhận dạng giọng nói đã bắt đầu từ Đại học Carnegie cho các hệ thống nhận dạng kỹ thuật số riêng biệt cho 10 Nhãn Bell và hành động phát triển đã diễn ra vào những năm 1980. [9]. Vào năm 2017, Vishal Passricha et, đã giới thiệu một phương pháp tổng hợp với phân loại CNN và SVM không đồng nhất trong đó một lớp được thay thế softmax bằng SVM [10]. Y. Yorozu, M.
Hirano, đã chuyển tiếp một mô hình mạng nơ-ron tích chập rất sâu không có các lớp được kết nối và cho thấy rằng VDCNN hoạt động tốt hơn CNN, khi điều tương tự đã được thử nghiệm với MGB-3 [8].
Năm 2020, Yang Xuebin et, hệ thống nhận dạng giọng nói được thiết kế cho một nhóm từ và sử dụng ba phương pháp phân loại, trong đó có CNN và thu được kết quả khoảng 92,88% [11].

3. Tập dữ liệu

Trong dự án của chúng tôi. sáu từ điều khiển (bắt đầu, trái, phải, lùi, tiến và dừng) cho ba mươi người được thu thập. Mỗi người thốt ra sáu từ này một lần bằng tiếng Anh. Người dân là một nửa nam, nửa còn lại là nữ và độ tuổi của họ bắt đầu từ 16 tuổi trở lên. Sáu từ được ghi lại ở những vị trí khác nhau. Trong phòng thí nghiệm, trên đường phố, trong vườn, trong chợ, ở những nơi không có tiếng ồn và những khu vực ồn ào khác. Các từ được ghi có độ dài khác nhau tùy thuộc vào chính từ đó. Ngoài ra, một số từ có độ dài khác nhau ở mỗi người tùy thuộc vào bản thân người nói và cách phát âm của họ. Những điều kiện này làm cho công việc trở nên phức tạp hơn, đặc biệt là đối với quá trình đào tạo và phân loại. Audacity và Adobe Audition được sử dụng để làm sạch và xử lý trước dữ liệu đầu vào nhằm chuẩn bị cho giai đoạn phân loại. Các từ đầu vào có độ dài trong khoảng từ (1 giây đến 1,35 giây). Dữ liệu đã được phân loại thành bảy loại theo các từ cần thiết để dẫn đến mạch điều khiển. Các lớp từ được ghi là lớp tiến, lớp lùi, lớp bắt đầu, lớp dừng, lớp trái, lớp phải và lớp chưa biết. Lớp chưa biết bao gồm các từ nhằm mục đích hoàn thành quá trình đào tạo và những từ này là (vâng, không, là, là, bạn, xin chào, anh ấy và cô ấy). Hình (3) hiển thị ba âm thanh với biểu đồ phổ của chúng từ dữ liệu của chúng tôi



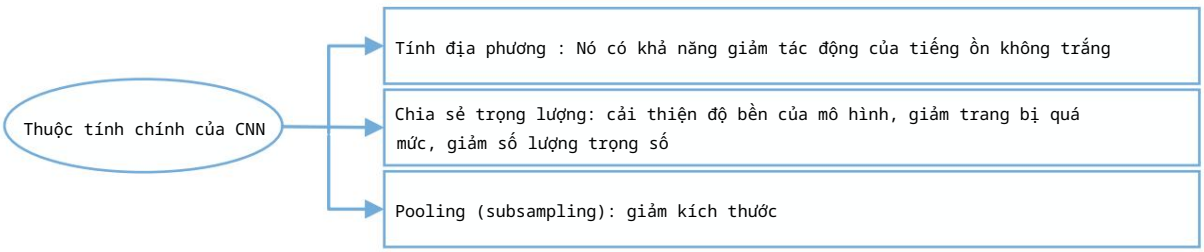
Hình 3. Ba âm học với sơ đồ quang phổ từ dữ liệu của chúng tôi.

4. Mạng nơ ron tích chập (CNN)

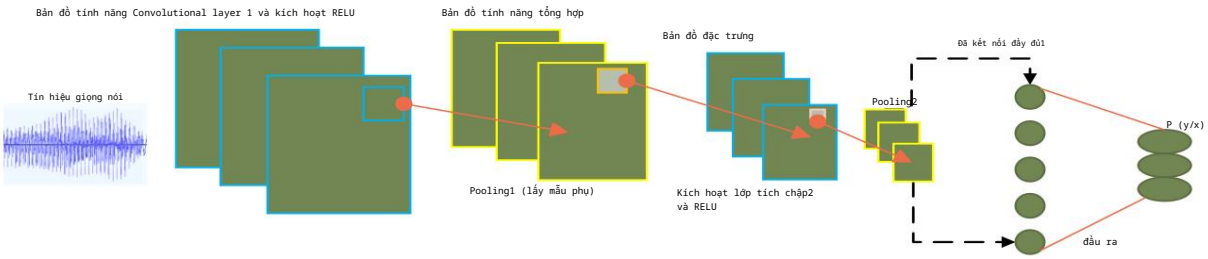
Mạng tích chập là sự khởi đầu mà Hubel và Wiesel đã phát hiện ra rằng một kiến trúc mạng duy nhất có thể làm giảm độ phức tạp trong mạng nơ-ron phản hồi khi nghiên cứu các nơ-ron được sử dụng để lựa chọn độ nhạy và định hướng cục bộ trong vỏ não của mèo.

CNN thường được sử dụng trong xử lý ảnh yêu cầu ma trận hai chiều chứa các đặc điểm và có thể là ba chiều, các giá trị pixel nằm trong các chỉ báo tọa độ ngang và dọc.

CNN là một mô hình mạng lưới thần kinh. Kiến trúc của nó có ba ý tưởng chính, như được giải thích trong hình (4). Mỗi người trong số họ đều có khả năng cải thiện hiệu suất nhận dạng giọng nói [12]. hình (5) giải thích các lớp CNN và cách thức hoạt động của CNN.



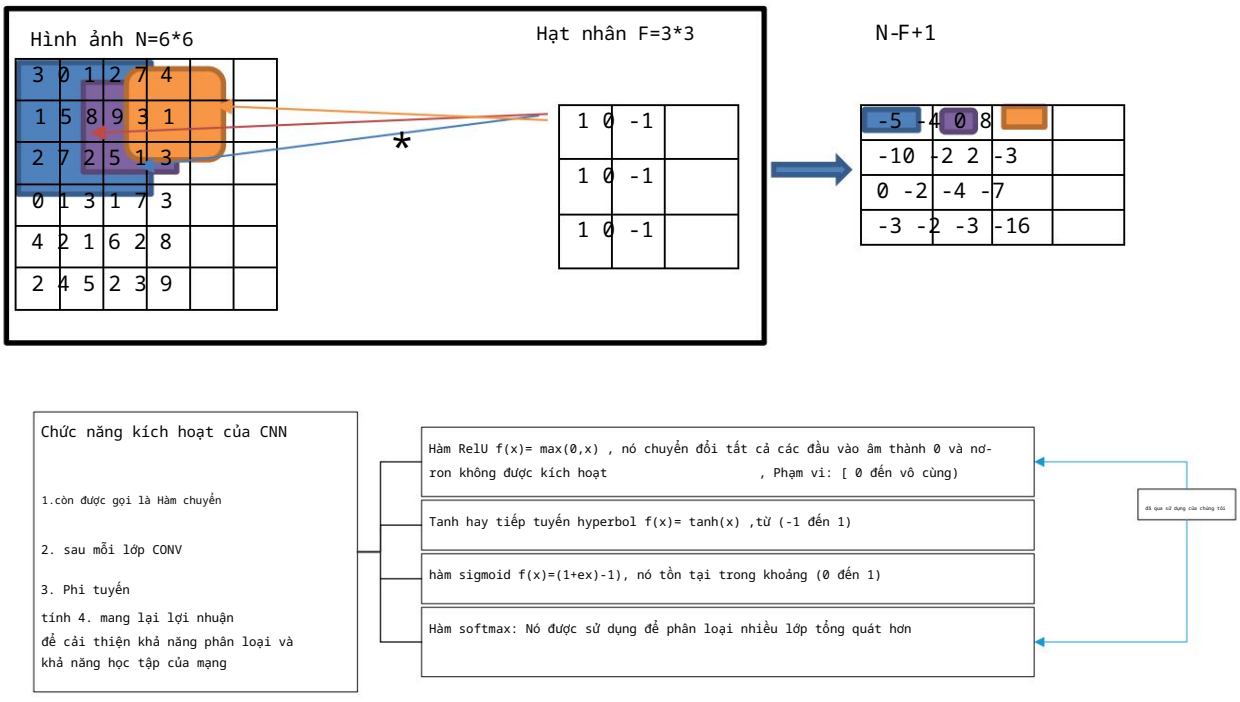
Hình 4. Kiến trúc thuộc tính CNN



Hình 5. Kiến trúc của các lớp CNN

CNN có bộ lọc chuyển qua hình ảnh để tạo ra bản đồ đặc trưng ở các lớp chập, thông qua cửa sổ hoặc bộ lọc này, các trọng số của mạng có thể xác định các đặc điểm khác nhau của hình ảnh đến. Hàm kích hoạt quyết định xem một tính năng cụ thể có xuất hiện ở một vị trí cụ thể trong ảnh hay không. Thường sử dụng rất nhiều bộ lọc trên ảnh để tìm ra những đặc điểm cần thiết [13]. CNN thường được gọi là mạng cục bộ vì các đơn vị riêng lẻ được tính toán ở một vị trí cụ thể của cửa sổ phụ thuộc vào khu vực cục bộ mà cửa sổ hiện đang xem xét. Kiến trúc tích chập được phối hợp bởi ba lớp chính được sắp xếp theo cấu trúc cấp dữ liệu chuyển tiếp. Lớp tích chập để trích xuất tính năng, lớp lấy mẫu phụ, lớp tổng hợp (gộp), để giảm kích thước của dữ liệu đầu vào và đầu ra, lớp được kết nối đầy đủ để dự đoán các lớp cuối cùng [14]. bộ lọc tuyến tính và hàm kích hoạt phi tuyến, Một trong những yếu tố quan trọng nhất [15]. Trong lớp chập, mỗi mặt phẳng được kết nối với một hoặc nhiều bản đồ đặc trưng của lớp trước [16]. Một hàm kích hoạt được áp dụng cho kết quả thu được đầu ra của mặt phẳng. Đầu ra mặt phẳng là ma trận 2 chiều được gọi là bản đồ đặc trưng; Tên này phát sinh vì mỗi đầu ra tích chập cho biết sự hiện diện của một tính năng trực quan tại một vị trí pixel nhất định [16]. Lớp tích chập tạo ra một hoặc nhiều bản đồ đặc trưng. Sau đó, mỗi bản đồ đặc trưng được kết nối với chính xác một mặt phẳng trong lớp lấy mẫu phụ (tổng hợp) tiếp theo [15].
Việc chia sẻ trọng số và vị trí là cần thiết đối với các thuộc tính của nhóm, các giá trị đặc trưng được tính toán ở các vị trí khác nhau được nhóm lại với nhau và được biểu thị bằng một giá trị duy nhất nhằm giảm thiểu sự khác biệt trong các đặc điểm được trích xuất dọc theo chiều tần số khi các mẫu đầu vào bị dịch chuyển. Điều này rất quan trọng khi xử lý những thay đổi tần số nhỏ thường gặp trong lời nói do giọng hát có độ dài đường truyền khác nhau. CNN đã sử dụng phe kích hoạt [16] như hình (6) và bảng (1) giải thích một số thuộc tính

của lớp CNN Sau khi chuyển đổi âm thanh thành Spectrogram dưới dạng hình ảnh, giả sử chúng ta có một hình ảnh có kích thước N = 6 * 6 và bộ lọc là bộ lọc hạt nhân F = 3 * 3, ví dụ bên dưới và sau khi thực hiện tích chập (*) kết quả là 4 * 4 theo công thức Out =N-F+1.
N=6, F=3 với phần đệm P=0 và tích chập Strided S=1, out =4=n-f+1



Hình 6. Chức năng kích hoạt và hoạt động [16]

Bảng 1. Minh họa các thuộc tính của lớp CNN [13,14,15].

Lớp tích chập Lớp tổng hợp Các bộ lọc được bao gồm để tìm các tính năng Giảm kích thước của hình ảnh		Các lớp được kết nối đầy đủ Thông tin tổng hợp từ tính năng cuối cùng
Bộ lọc bao gồm các hạt nhân nhỏ (số lượng hạt nhân) Một thành kiến cho mỗi bộ lọc	Diện tích tối đa hoặc trung bình được trích xuất	Phân loại cuối cùng chung
Đối với mọi giá trị của bản đồ đặc trưng phải áp dụng hàm kích hoạt	Phương pháp cửa sổ trượt	Các tham số được kết nối đầy đủ, (số nút, chức năng kích hoạt; thường thay đổi tùy theo vai trò của các lớp. RELU được sử dụng để tổng hợp thông tin và SOFTMAX để tạo ra nhiều phân loại cuối cùng)
các tham số của các lớp CONV, (kích thước của hạt nhân, chức năng kích hoạt, bước tiến, phần đệm và loại và giá trị chính quy hóa)	Các tham số của tổng hợp, (sải chân và kích thước của cửa sổ). [16]	

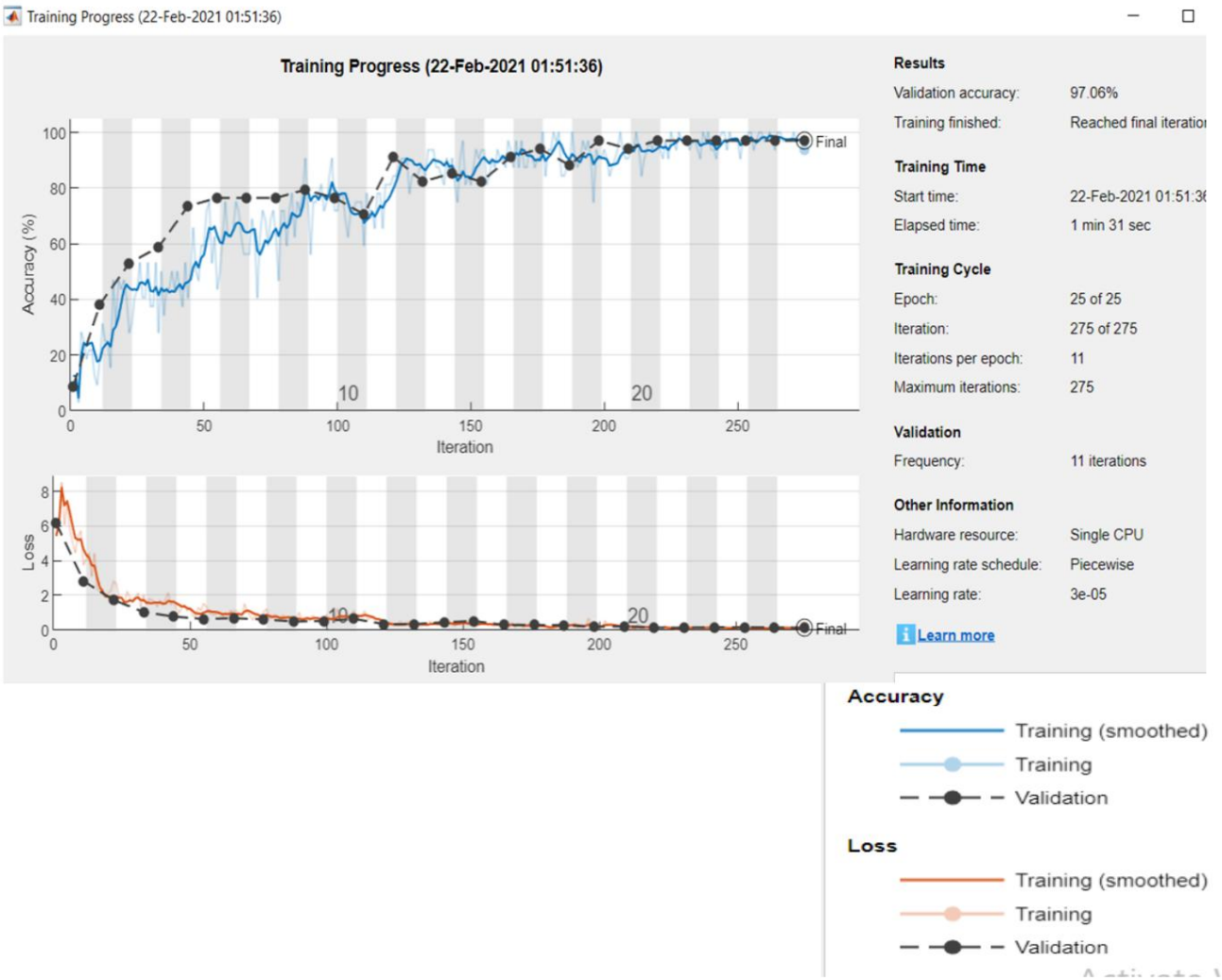
5. Kết quả thí nghiệm và thảo luận

Học có giám sát với mô hình nơ-ron sâu được đề xuất đã được sử dụng để huấn luyện và kiểm tra mạng nơ-ron tích chập CNN. Cấu trúc CNN được đề xuất có 13 lớp. Trong CNN, tệp âm thanh đã được nhập trực tiếp vào cấu trúc mạng được thiết kế chứa quy trình học đa cấp để đạt được nhiệm vụ phân loại đa mô hình nhằm phân loại sáu nhãn cho các từ (bắt đầu, dừng, phải, trái, tiến, lùi). Kích thước của dữ liệu đã được mở rộng bằng cách tăng số lượng dữ liệu đào tạo thông qua việc tăng cường quy trình. Số lượng tăng thêm là ba cho mỗi hình ảnh (phổ) của âm thanh. Phổ thính giác trong lấy mẫu tần số 48K đã được tính toán với thời lượng phân đoạn bằng 1,8 và thời lượng khung hình 0,02Ms, độ dài FFT 1024 và số băng tần 64 trên melspectum đã được áp dụng. Bảng (2) giải thích tất cả chi tiết về các lớp CNN của chúng tôi

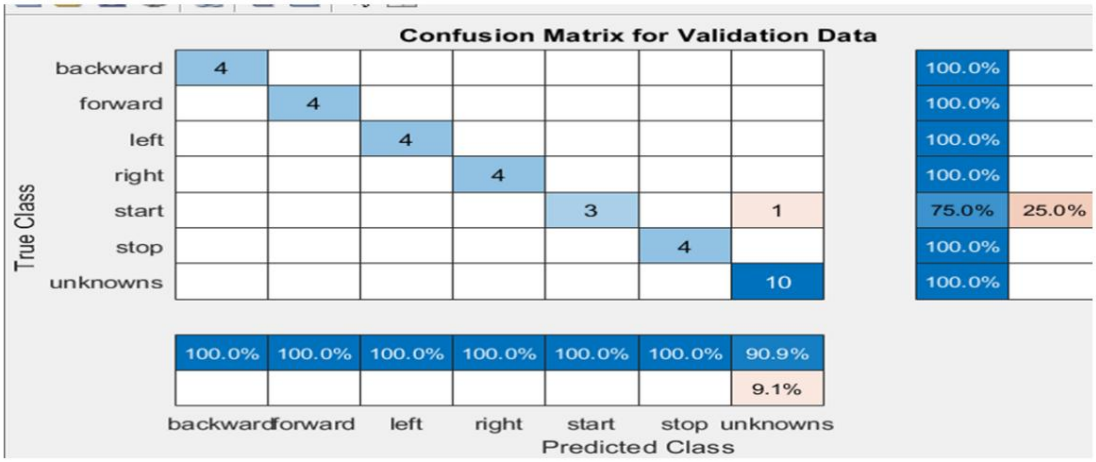
Bảng 2. Các lớp CNN cho các lớp Cnn mô hình được đề xuất của chúng

tôi (công việc của chúng tôi)	Lớp mô tả Lớp
thứ 1 đầu vào hình ảnh để điều chỉnh kích thước hình ảnh theo (số lượng hy vọng và số lượng dải)	
2 thứ lớp	Để thêm kích thước bộ lọc của pixel (phần đệm bằng 3)
thứ 3 lớp	Để cân bằng dữ liệu và đặt giá trị trung bình và độ lệch chuẩn bằng 0, đồng thời thực hiện độ dốc mượt mà hơn, đào tạo nhanh hơn và độ chính xác tổng quát hóa tốt hơn [chuẩn hóa] với lớp RelU
4 lớp	Thêm tính năng gộp nhóm để giảm kích thước với 3 bước và 2 phần
5 lớp	đệm 2*số bộ lọc (numF) cho phần đệm [số bộ lọc =10]
6 lớp	Lớp chuẩn hóa hàng loạt với lớp RelU và maxpooling2dlayer với bước 3 và phần đệm 2
7 lớp	Lớp 2D tích chập (3,4*numF, 'Đệm', 'giống nhau'), numf=10
số 8 lớp	Lớp chuẩn hóa hàng loạt (lớp RelU)
9 lớp	Lớp 2D gộp tối đa([timePoolSize,1])
thứ 10 lớp	Dropout Layer (cướp dropout), dropout, chống overfitting
thứ 11 lớp	Lớp được kết nối đầy đủ (Số lớp)
thứ 12 lớp	Lớp Softmax, tính xác suất của từng nhãn
thứ 13	Lớp phân loại, phân loại dựa trên softmax, chi phí sẽ là x-entropy

Hình (7) giải thích các giai đoạn đào tạo và xác nhận của CNN và nó hiển thị rõ ràng tất cả các tham số đào tạo. Các mẫu giọng nói của chúng tôi đã được đào tạo và xác thực bằng cách áp dụng mạng lưới thần kinh sâu được đề xuất. Quá trình học mô hình bắt đầu từ việc thu thập các kiểu học tương ứng trong phổ giọng nói đầu vào. Các tính năng học tập đi qua các lớp mạng được đề xuất của chúng tôi và các thông số đào tạo đã được cập nhật trong quá trình học tập. Hình (7) cho thấy độ chính xác phân loại từ cho cả dữ liệu huấn luyện và xác thực đã được nâng cao bằng cách tăng số lần lặp để đạt độ chính xác tốt hơn với 275 lần lặp. Mặt khác, dữ liệu phân loại sai sẽ giảm trong suốt quá trình đào tạo và xác nhận. Mô hình bao gồm 13 lớp, lớp đầu tiên có 179 nơ-ron tương ứng với các mẫu giọng nói trong ma trận đầu vào và lớp cuối cùng là lớp phân loại có 7 nơ-ron tham chiếu đến bảy lớp của chúng ta (nhãn nhắm mục tiêu). Tất cả các chi tiết mô hình được giải thích rõ ràng trong hình (7). Trong khi hình (8) hiển thị ma trận nhầm lẫn của các kết quả xác thực, hiển thị vị trí của lỗi và chỉ thấy một lỗi từ quá trình xác thực của chúng tôi
data, trong đó lỗi trong từ 'bắt đầu' xuất hiện trong lớp không xác định. Hình (9) hiển thị phân phối nhãn đào tạo so với phân phối nhãn được nhắm mục tiêu.



Hình 7. Kết quả huấn luyện và kiểm định phương pháp CNN



Hình 8. Ma trận nhầm lẫn cho dữ liệu xác thực của CNN

Hình 9. Phân bổ nhãn huấn luyện

6. Kết luận Dữ

liệu của chúng tôi bao gồm sáu từ dành cho 25 người, mỗi từ đại diện cho một nhãn trong số sáu nhãn cho các từ của chúng tôi (bắt đầu, dừng, phải, trái, tiến, lùi) áp dụng cho việc học có giám sát mô hình thần kinh sâu được đề xuất của chúng tôi. Cấu trúc CNN được đề xuất có 13 lớp. Trong CNN, tệp âm thanh đã được nhập trực tiếp vào cấu trúc mạng được thiết kế có chứa quy trình học tập đa cấp để đạt được nhiệm vụ phân loại đa mô hình. Kết quả thử nghiệm cho thấy mạng lưới thần kinh sâu có khả năng giải quyết các thách thức nhận dạng giọng nói liên quan đến tín hiệu âm thanh có độ phân giải thấp và nhiễu. CNN đã trả về hiệu suất có thể chấp nhận được cho tác vụ nhận dạng từ bằng cách sử dụng tín hiệu giọng nói ồn ào của chúng tôi. Hiệu suất của mô hình có thể cao hơn và mang lại độ chính xác phân loại tốt hơn khi có đủ dữ liệu đào tạo cho mô hình học sâu. Vì mục đích này, dữ liệu đã được tăng lên và một lớp khác đã được thêm vào sáu lớp từ của chúng tôi để trở thành 7 lớp. Lớp thứ bảy là lớp chưa biết có khoảng 50 từ, hiện nay, tổng lượng dữ liệu và từ ngữ mà các lớp đóng góp vào

tăng cường mạng lưới thần kinh tích chập trong nhiệm vụ nhận dạng giọng nói từ. Và độ chính xác phân loại mô hình tăng lên đạt 97,06%. CNN yêu cầu một lượng lớn dữ liệu cho mục đích học tập, càng đưa ra nhiều dữ liệu, CNN sẽ trả về độ chính xác phân loại tốt và chính xác hơn. Kết quả phân loại của chúng tôi dựa trên tỷ lệ phân chia dữ liệu và tỷ lệ đào tạo là 85% và xác thực là 15% dữ liệu đầu vào.

7. Tài liệu tham khảo

[1] Fadlilah AF, Djamal EC (2019) Nhận dạng giọng nói và người nói bằng cách sử dụng máy vectơ hỗ trợ phân cấp và truyền ngược. Tại: Hội nghị quốc tế lần thứ 6 về Kỹ thuật điện, Khoa học máy tính và Tin học (EECSI) lần thứ 6 năm 2019. IEEE, tr. 404-409

[2] Shaikh Naziya S., Deshmukh RR (2016) Hệ thống nhận dạng giọng nói—đánh giá. IOSR J. Máy tính. Tiếng Anh, 8.4: 3-8. [3] Aderhold J, Davydov V Yu, Fedler F, Klausing H, Mistele D, Rotter T, Semchinova O, Stemmer J và Graul J 2001 J. Cryst. Tăng trưởng 222 701

[3] Kesarkar MP, Rao, P. (2003) Trích xuất tính năng để nhận dạng giọng nói. Hệ thống điện tử, EE. Phòng, IIT Bombay.

[4] Alías F., Socoró JC, Sevillano X. (2016) Đánh giá các kỹ thuật trích xuất đặc điểm vật lý và nhận thức đối với lời nói, âm nhạc và âm thanh môi trường. Khoa học Ứng dụng, 6(5), 143.

[5] Sinh viên PG (2016) Lựa chọn và trích xuất tính năng của tín hiệu âm thanh. thuật toán, 5(3).

[6] Abdel-Hamid O., Mohamed AR, Jiang, H., Deng L., Penn, G., Yu D. (2014) Mạng lưới thần kinh tích chập để nhận dạng giọng nói. Giao dịch IEEE/ACM về xử lý âm thanh, giọng nói và ngôn ngữ, 22(10), 1533-1545.

[7] Li X., Chu Z. (2017) Nhận dạng lệnh giọng nói bằng mạng nơ-ron tích chập. CS229 giáo dục Stanford

[8] Poudel S., Anuradha, R. (2020) Nhận dạng lệnh giọng nói bằng cách sử dụng Mạng thần kinh nhân tạo. JOIV: Tạp chí quốc tế về trực quan hóa tin học, 4(2), 73-75.

[9] Song Z. (2020) Nhận dạng giọng nói tiếng Anh dựa trên deep learning với nhiều tính năng. Máy tính, 102(3), 663-682

[10] Passricha V., Aggarwal RK (2019) Máy vectơ hỗ trợ tích chập để nhận dạng giọng nói. Tạp chí Quốc tế về Công nghệ Giọng nói, 22(3), 601-609

[11] Yang X., Yu H., Jia L. (2020) Nhận dạng giọng nói của các từ lệnh dựa trên mạng nơ ron tích chập. Trong: Hội nghị quốc tế về thông tin máy tính và ứng dụng dữ liệu lớn (CIBDA) năm 2020 (trang 465-469).

[12] Saitoh T., Chu Z., Zhao, G., Pietikäinen M. (2016) CNN dựa trên hình ảnh khung nối để nhận dạng giọng nói trực quan. Trong: Hội nghị Châu Á về Thị giác Máy tính. P. 277-289.15- Phùng, Sơn Lâm và Abdesselam Bouzerdoum. "Phòng thí nghiệm xử lý tín hiệu hình ảnh và âm thanh của Đại học Wollongong", 2009.

[13] Nanni L., Costa YM, Aguiar RL, Mangolin, RB, Brahnam S., Silla CN (2020) Tập hợp các mạng thần kinh tích chập để cải thiện khả năng phân loại âm thanh của động vật. Tạp chí EURASIP về Xử lý Âm thanh, Lời nói và Âm nhạc, 1-14.

[14] Patel S. (2020) Phân tích toàn diện về các mô hình mạng nơ-ron tích chập.
Tạp chí Quốc tế Khoa học và Công nghệ Tiên tiến, 29(4), 771-777.

[15] Kubanek M., Bobulski J., Kulawik, J. (2019) Một phương pháp mã hóa giọng nói để nhận dạng giọng nói bằng cách sử dụng mạng thần kinh tích chập. Đối xứng, 11(9), 1185.

[16] Nwankpa C., Ijomah W., Gachagan, A., Marshall, S. (2018) Chức năng kích hoạt: So sánh về các xu hướng trong thực hành và nghiên cứu về học sâu. bản in trước arXiv arXiv:1811.03378.

[17] Phùng SL, Bouzerdoum A. (2009) Phòng thí nghiệm xử lý tín hiệu hình ảnh và âm thanh Đại học Wollongong