Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00084

NHẬN DẠNG PHƯƠNG NGỮ TIẾNG VIỆT SỬ DỤNG MẠNG NƠRON TÍCH CHẬP CNN

Nguyễn Hồng Quang[1], Trinh Văn Loan[1], Phạm Ngọc Hưng[2]

[1] Viện Công nghệ Thông tin và Truyền thông, Trường Đại học Bách Khoa Hà Nội

[2] Khoa Công nghệ Thông tin, Trường Đại học Sư phạm Kỹ thuật Hưng Yên

quangnh@soict.hust.edu.vn, loantv@soict.hust.edu.vn, phamngochung@soict.hust.edu.vn

***TÓM TẮT**: Bài báo này trình bày phương pháp nhận dạng phương ngữ tiếng Việt sử dụng mạng nơron tích chập CNN. Các nghiên cứu hiện nay về nhận dạng phương ngữ tiếng Việt mới sử dụng các phương pháp học máy truyền thống như K láng giềng gần nhất KNN, cây quyết định, máy hỗ trợ véctơ SVM. Nghiên cứu này bước đầu áp dụng phương pháp mạng nơron học sâu vào bài toán này. Quá trình trích chọn tham số đã biểu diễn âm thanh tiếng nói ở dạng phổ spectrogram. Kiến trúc mạng được chọn lựa thử nghiệm là mạng nơron tích chập CNN. Kết quả thử nghiệm cho thấy phương pháp này đã đạt kết quả vượt trội so với các phương pháp học máy truyền thống.*

***Từ khóa**: Nhận dạng phương ngữ tiếng Việt, mạng nơron học sâu, phổ tiếng nói spectrogram, mạng nơron tích chập CNN.*

I. GIỚI THIỆU

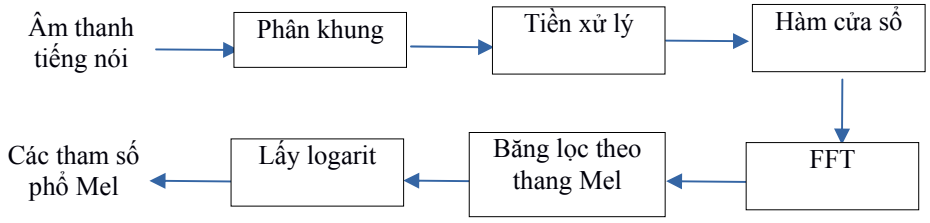
Nhận dạng tự động phương ngữ (Automatic Dialect Identification, viết tắt là ADI) là một trong những bài toán quan trọng trong cả lĩnh vực xử lý ngôn ngữ tự nhiên và xử lý tiếng nói. Các nghiên cứu về nhận dạng phương ngữ đã được thực hiện với ngôn ngữ Ả rập [9], tiếng Thái Lan [10], tiếng Nhật [11], tiếng Anh Mỹ [12],… Tiếng Việt là ngôn ngữ có phương ngữ phong phú và đa dạng [1]. Mặc dù đã có một số nghiên cứu nhận dạng tiếng Việt nói [6] [7] [8], song số lượng nghiên cứu nhận dạng phương ngữ và ảnh hưởng của phương ngữ với hệ thống nhận dạng tiếng Việt nói còn chưa nhiều. Các nghiên cứu hiện nay mới chủ yếu sử dụng các phương pháp học máy truyền thống như K láng giềng gần nhất, máy hỗ trợ véctơ, mô hình Markov ẩn, mô hình hỗn hợp Gauss, cây quyết định,... [1]. Hầu như chưa có nghiên cứu nào sử dụng kỹ thuật học sâu vào vấn đề này. Mạng nơron học sâu hiện nay đang nhận được sự quan tâm của nhiều nhà nghiên cứu và được áp dụng hiệu quả trong nhiều lĩnh vực trong đó có xử lý tiếng nói. Vì vậy trong bài báo này chúng tôi trình bày phương pháp sử dụng một kiến trúc mạng nơron học sâu rất phổ biến hiện nay là mạng nơron tích chập (Convolutional Neural Network CNN) [5] cho vấn đề nhận dạng phương ngữ. Sự khác biệt giữa các phương ngữ thể hiện ở nhiều yếu tố như ngữ âm, từ vựng, ngữ pháp. Trong nghiên cứu này, chúng tôi khai thác sự khác biệt về ngữ âm để nhận dạng phương ngữ tiếng Việt.

Phần tiếp theo của bài báo sẽ trình bày phương pháp đề xuất, bao gồm quá trình trích chọn tham số phổ spectrogram và kiến trúc mạng CNN sử dụng. Phần III trình bày các thử nghiệm nhận dạng phương ngữ tiếng Việt trên mô hình đề xuất. Phần IV là kết luận và hướng nghiên cứu tiếp theo.

II. PHƯƠNG PHÁP ĐỀ XUẤT

A. Trích chọn tham số

Các tham số sử dụng là phổ mel (mel-spectrogram). Hình 1 mô tả sơ đồ khối quá trình tính các tham số này.



Hình 1. Tính phổ mel của file âm thanh tiếng nói

Các file âm thanh trong cơ sở dữ liệu tiếng nói VDSPEC [1] được thu âm với tần số lấy mẫu Fs = 16 kHz. Mỗi file âm thanh tiếng nói được thực hiện phân khung với độ rộng khung 512 mẫu, độ dịch khung 256 mẫu. Số hệ số Mel cho mỗi khung tín hiệu tiếng nói là 96.

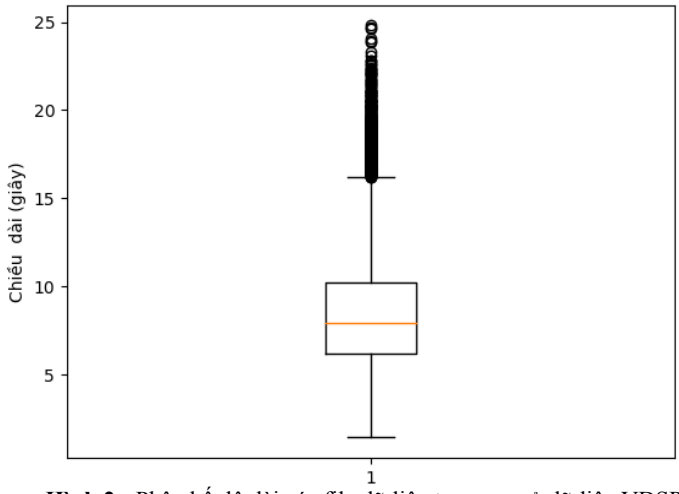
Do đặc thù của tín hiệu âm thanh nên các file âm thanh trong cơ sở dữ liệu có độ dài khác nhau, biến thiên từ 1,52 giây tới 25,09 giây [Hình 2]. Vì vậy chúng tôi đưa ra hai phương pháp biểu diễn tham số cho các file âm thanh:

- Phương pháp 1. Chuẩn hóa chiều dài file âm thanh. Mỗi file âm thanh được chọn độ dài chuẩn 1.52 giây (bao gồm 96 khung). Với các file âm thanh có độ dài lớn hơn thì chỉ chọn đoạn âm thanh 1,52 giây nằm ở giữa file. Như vậy toàn bộ các file âm thanh đều được phân tích tham số phổ mel với kích thước 96 x 96 (96 khung, mỗi khung bao gồm 96 hệ số Mel). Phương pháp này đặt tên là PARAMETER_NORM.

- Phương pháp 2. Sử dụng toàn bộ dữ liệu file âm thanh. Mỗi file âm thanh được chia thành các đoạn 1,52 giây. Như vậy mỗi file âm thanh có số lượng các đoạn khác nhau. Ví dụ như một file âm thanh có N đoạn, khi đó file âm thanh này được phân tích tham số phổ mel với kích thước N x 96 x 96. Phương pháp này được đặt tên là PARAMETER_FULL.

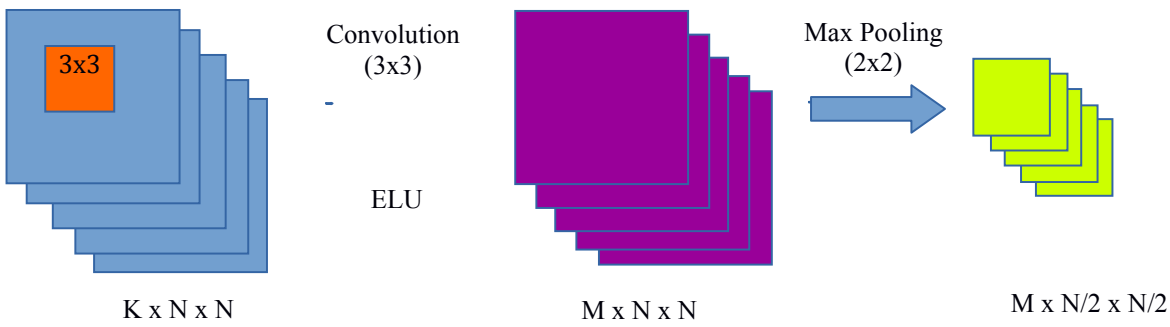
B. Mạng nơron tích chập CNN

Để huấn luyện mạng nơron tích chập CNN chúng tôi sử dụng phương pháp biểu diễn tham số PARAMETER_NORM. Mỗi file âm thanh tiếng nói trong tập huấn luyện được biểu diễn bởi một ma trận 96 x 96. Đồng thời file âm thanh này cũng đã xác định được phương ngữ tương ứng. Như vậy đây là bài toán huấn luyện có giám sát.



Hình 2. Phân bố độ dài các file dữ liệu trong cơ sở dữ liệu VDSPEC

Mạng nơron đề xuất sử dụng trong bài báo này là mạng nơron tích chập CNN. Kiến trúc môđun cơ bản của mạng CNN được mô tả ở Hình 3, môđun này được chúng tôi đặt tên là BM_CNN (Basic Module of CNN). Trong hình này ta thấy mạng CNN gồm hai thao tác cơ bản là tích chập (Convolution) và thao tác chọn lớn nhất (Max Pooling).



Hình 3. Kiến trúc môđun cơ bản trong mô hình CNN (BM_CNN)

Thao tác đầu tiên là tích chập (Convolution):

- Đầu vào là một tensor N x N gồm K thành phần. Ví dụ với ảnh đầu vào thì N = 96 và K = 1.
- Thao tác tổng chập sử dụng cửa sổ có kích thước 3 x 3. Cửa sổ này cho phép khối tổng chập ghi nhận được các mẫu cơ bản trong ảnh phổ. Thao tác tổng chập được thực hiện quét lần lượt từ trái sang phải, từ trên xuống dưới. Công thức (1) mô tả thao tác tính tổng chập tại một vị trí cụ thể.

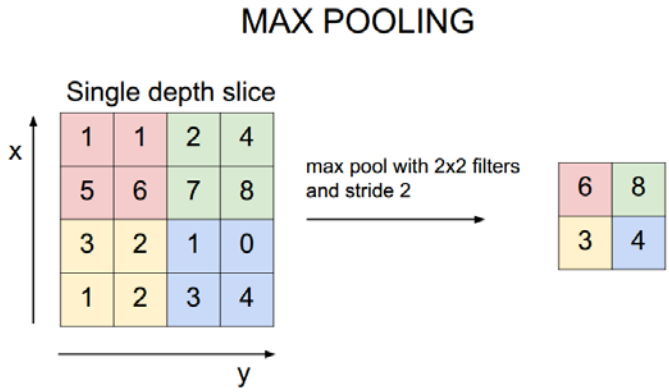
$$y = \sum_i w_i x_i + b \tag{1}$$

Trong công thức (1), x_i bao gồm các điểm ảnh phổ nằm trong phạm vi cửa sổ đang quét, nghĩa là bao gồm 3 x 3 điểm ảnh phổ. Số giá trị phải quét nằm tại cùng vị trí cho tất cả các thành phần, do vậy sẽ có K x 3 x 3 giá trị x_i và w_i tương ứng. Ngoài ra còn thêm 1 hệ số độ lệch b trong công thức này. Do vậy số tham số cần thiết cho thao tác tổng chập bao gồm w_i và b là K x 3 x 3 + 1 tham số. Do số bộ lọc sử dụng là M bộ lọc, do đó số tham số bao gồm Mx (K x 3 x 3 + 1) tham số. Sau khi tính tổng chập, giá trị y được biến đổi phi tuyến bởi hàm ELU (Exponential Linear Unit). Hàm này sẽ giúp làm

giảm thiểu ảnh hưởng của hiện tượng suy giảm gradient trong quá trình huấn luyện mạng CNN [2]. Ý nghĩa của thao tác tích chập là xác định khả năng xuất hiện các mẫu tại các vị trí nhất định trong ảnh. Mỗi mẫu này được biểu diễn bằng trọng số của cửa sổ tương ứng với một bộ lọc (3 x 3 trong trường hợp này). Tổng số mẫu mà mạng cần học chính là số bộ lọc M sử dụng.

Trong nghiên cứu này thao tác tổng chập được thực hiện kỹ thuật bổ sung các điểm ảnh. Kỹ thuật này được gọi là padding. Mục đích của kỹ thuật này nhằm đảm bảo thành phần đầu ra sẽ có kích thước giống với thành phần đầu vào. Vì ở đây sử dụng cửa sổ 3 x 3, do đó chúng tôi đã sử dụng padding 2 điểm ảnh.

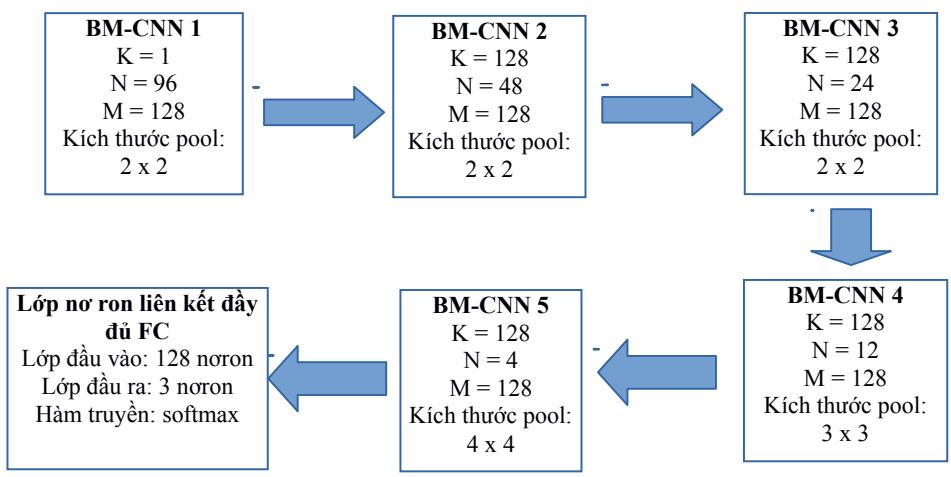
Thao tác thứ 2 là thao tác "Max Pooling". Hình 4 mô tả thao tác này sử dụng bộ lọc 2 x 2 với độ dịch 2. Với mỗi vùng ảnh kích thước 2 x 2 (tương ứng với kích thước của bộ lọc), thao tác này sẽ trả về giá trị điểm ảnh lớn nhất. Như vậy qua thao tác này, ta sẽ thu được một ảnh mới có kích thước mỗi chiều giảm một nửa. Ý nghĩa của thao tác này là cho phép nổi bật lên đặc trưng có trong ảnh đồng thời làm giảm kích thước ảnh.



Hình 4. Thao tác "Max Pooling" sử dụng trong mạng CNN

Kiến trúc đầy đủ của mạng nơron tích chập cho bài toán nhận dạng phương ngữ tiếng Việt được mô tả ở Hình 5:

- Khối BM-CNN1: lớp đầu vào nhận ảnh spectrogram có kích thước 96 x 96 (thu được từ phân tích phổ của file âm thanh, xem phần II.A). Tất cả các khối tích chập được thực hiện thống nhất trong mạng với 128 bộ lọc kích thước 3 x 3 được khởi tạo ban đầu là giá trị ngẫu nhiên từ 0 đến 1 với kỹ thuật padding. Trong khối này, lớp "Max-Pooling" sử dụng bộ lọc kích thước 2 x 2. Do đó đầu ra của lớp này bao gồm 128 thành phần với kích thước 48 x 48. Khối dữ liệu này sử dụng làm đầu vào cho khối BM-CNN2.
- Các khối BM-CNN2 có tính năng tương tự nhau: lớp vào bao gồm 128 thành phần với kích thước 48 x 48. Đầu ra của khối là 128 thành phần kích thước 24 x 24. Tương tự cho các khối BM-CNN3, BM-CNN4 và BM-CNN5. Đầu ra của khối BM-CNN5 là dữ liệu 1 x 1 x 128, nghĩa là có 128 giá trị.



Hình 5. Kiến trúc mạng nơron tích chập CNN sử dụng cho bài toán nhận dạng phương ngữ tiếng Việt

- Lớp liên kết nơron đầy đủ (FC: Fully Connected Layer) nhận vào 128 giá trị kết xuất ra bởi khối BM-CNN5 và được kết nối đầy đủ với 3 nơron trong lớp (đại diện cho 3 phương ngữ). Như vậy lớp này có 3x128 + 3 = 387 tham số (mỗi nơron có một giá trị độ lệch b bên cạnh 128 trọng số kết nối đến 128 giá trị đầu vào). Lớp này sử dụng hàm softmax để biểu diễn phân bố xác suất cho từng phương ngữ.

Như vậy với kiến trúc mạng nơron CNN được đề xuất ở hình 5, mạng này đã chuyển đổi từ âm thanh tiếng nói đầu vào thành 3 giá trị biểu diễn phân bố xác suất ứng với 3 phương ngữ. Tổng số tham số của mạng là 594947 tham số.

C. Sử dụng mạng nơron tích chập CNN nhận dạng phương ngữ tiếng Việt

Do có hai phương pháp tham số hóa file âm thanh, do vậy chúng tôi cũng đề xuất phương pháp sử dụng mạng nơron tích chập ở trên cho từng phương pháp này.

- Phương pháp 1 (PARAMETER_NORM): do với phương pháp này, mỗi file âm thanh được biểu diễn bởi ma trận với kích thước 96 x 96, do vậy dữ liệu này được đưa trực tiếp vào mạng nơron tích chập trên, và đầu ra của mạng nơron tích chập chính là phương ngữ nhận dạng được.
- Phương pháp 2 (PARAMETER_FULL): với phương pháp này, mỗi file âm thanh được biểu diễn bởi một tập các ma trận kích thước 96 x 96. Chúng tôi sử dụng phương pháp bình chọn (voting): từng ma trận này được đưa vào mạng nơron tích chập để xác định được phương ngữ nhận dạng, phương ngữ nào xuất hiện nhiều nhất thì chính là phương ngữ nhận dạng được.

III. THỬ NGHIỆM VÀ ĐÁNH GIÁ

A. Cơ sở dữ liệu tiếng nói

Bộ ngữ liệu VDSPEC được ghi âm trực tiếp từ người nói thông qua việc đọc các đoạn văn bản đã được chuẩn bị sẵn. Văn bản này được tổ chức theo 5 chủ đề khác nhau bao gồm: khoa học, đời sống, kinh doanh, pháp luật, ô tô - xe máy. Nội dung văn bản được thu thập tự động từ báo điện tử vnexpress.net. Tiếp theo, văn bản được lựa chọn để đảm bảo sự cân bằng về thanh điệu với số lượng các từ cho mỗi thanh điệu là xấp xỉ như nhau, khoảng 717 từ. Tiếng nói được ghi âm với tần số lấy mẫu là 16000 Hz, 16 bit cho mỗi mẫu. Độ tuổi của người nói trung bình là 21 tuổi. Ở độ tuổi này, tiếng nói đã ổn định và thể hiện rõ được tiếng địa phương. Mỗi phương ngữ có 50 người nói bao gồm 25 nữ và 25 nam. Giọng Hà Nội được chọn đại diện cho phương ngữ Bắc, giọng Huế cho phương ngữ Trung và giọng Thành phố Hồ Chí Minh đại diện cho phương ngữ Nam. Với mỗi chủ đề, người nói đọc 25 câu, mỗi câu có độ dài ghi âm khoảng 10 giây. Tổng thời gian tiếng nói đã ghi âm của VDSPEC là 45,11 giờ, chiếm dung lượng 4,84 GB bộ nhớ.

Để thực hiện nghiên cứu nhận dạng phương ngữ tiếng Việt, từ bộ dữ liệu VDSPEC ở trên, chúng tôi tiến hành xây dựng tập TRAIN để sử dụng huấn luyện các mô hình hệ thống, tập VALID để điều chỉnh các tham số của hệ thống trong quá trình huấn luyện và tập TEST để đánh giá. Để đánh giá khách quan, các tập này cần độc lập với nhau về người nói và về chủ đề. Do vậy trong 5 chủ đề, chúng tôi chọn 3 chủ đề cho tập TRAIN, 1 chủ đề cho tập VALID và 1 chủ đề cho tập TEST. Về người nói, mỗi phương ngữ chọn 15 nam và 15 nữ cho tập TRAIN, 5 nam và 5 nữ cho tập VALID và 5 nam và 5 nữ cho tập TEST. Kết quả tập TRAIN có 6750 câu tiếng nói, tập VALID và tập TEST mỗi tập có 750 câu tiếng nói.

B. Các phương pháp truyền thống

Đầu tiên chúng tôi đã thử nghiệm nhận dạng phương ngữ bằng hai phương pháp học máy truyền thống: K láng giềng gần nhất (KNN : K Nearest Neighbor), máy hỗ trợ véctơ (SVM: support vector machine) với sự hỗ trợ của bộ công cụ Weka [3]. Bộ tham số sử dụng bao gồm 384 hệ số do bộ công cụ OpenSMILE [6] thực hiện. Đây là dữ liệu thống kê của mỗi file ghi âm. Với mỗi file tiếng nói được trích chọn đặc trưng, OpenSMILE sẽ cho ra 384 hệ số trong đó có các đặc trưng sau:

- Năng lượng khung;
- Các hệ số MFCC (Mel-/Bark-Frequency-Cepstral Coefficients (MFCC));
- Tỷ lệ biến thiên qua trục không (Zero-Crossing Rate);
- Xác suất âm hữu thanh;
- Tần số cơ bản;
- Một số đặc trưng phổ (năng lượng theo băng tần, dải quá độ, trọng tâm phổ, phương sai…).

Các đặc trưng này lại được biểu diễn theo: các giá trị và vị trí cực biên, trung bình, mômen, thống kê theo phần trăm, xấp xỉ bình phương và tuyến tính, trọng tâm, giá trị đỉnh, phân đoạn, các giá trị mẫu.

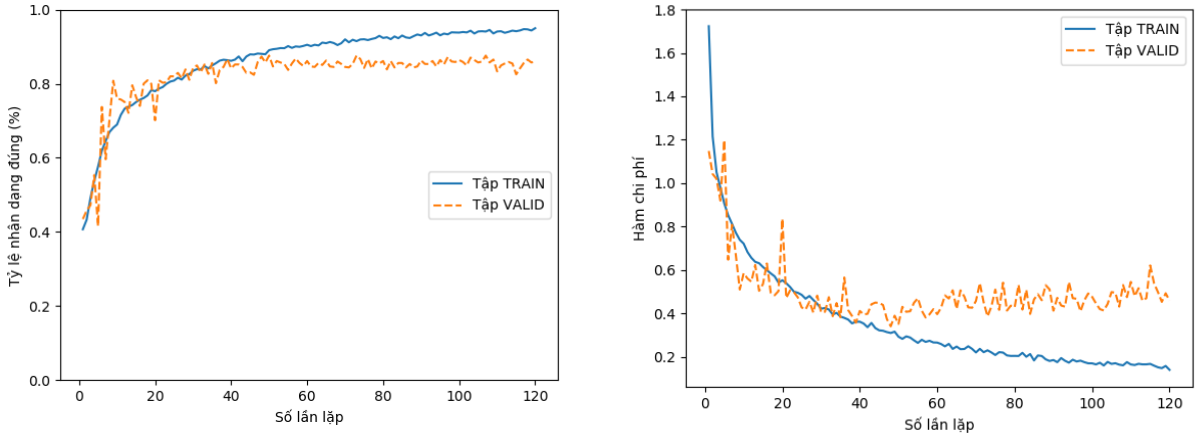
Bảng 1. Kết quả nhận dạng (độ chính xác tính theo tỷ lệ phần trăm) phương ngữ tiếng Việt sử dụng các phương pháp học máy truyền thống

Phương pháp	Kết quả thử nghiệm trên tập TEST
K láng giềng gần nhất	61,8
Máy hỗ trợ véctơ	72,4

Kết quả nhận dạng phương ngữ sử dụng KNN và SVM được mô tả ở bảng 1. Bảng 1 cho thấy phương pháp máy hỗ trợ véctơ cho kết quả tốt hơn so với thử nghiệm sử dụng phương pháp K láng giềng gần nhất. Tuy nhiên kết quả nhận dạng đúng cao nhất đạt được mới là 72,4%.

C. Nhận dạng phương ngữ với mạng CNN

Để huấn luyện mạng CNN, đầu tiên mỗi file âm thanh tiếng nói được phân tích phổ theo phương pháp PARAMETER_NORM. Kết quả mỗi file âm thanh được biểu diễn bởi ma trận với kích thước 96 x 96. Chúng tôi phát triển giải thuật huấn luyện và nhận dạng mạng CNN sử dụng bộ công cụ Keras [4]. Hình 6 mô tả sự biến thiên của hàm chi phí và độ chính xác nhận dạng trên tập VALID theo số lần lặp trong quá trình huấn luyện.



Hình 6. Sự biến thiên của hàm chi phí (hình bên phải) và độ chính xác nhận dạng (hình bên trái) trên tập VALID theo số lần lặp trong quá trình huấn luyện

Kết quả ở hình 6 cho thấy khi số lần lặp trong quá trình huấn luyện mạng CNN tăng lên thì tỷ lệ nhận dạng đúng với tập TRAIN tăng dần, tuy nhiên từ bước lặp thứ 40 thì tỷ lệ này với tập VALID lại gần như không đổi. Điều này có thể quan sát tương tự với hàm chi phí: với tập TRAIN thì càng huấn luyện thì hàm chi phí càng giảm nhưng với tập VALID thì từ bước lặp 40 hàm chi phí dao động xung quanh một giá trị nhất định mà không có xu hướng giảm rõ ràng. Vì vậy chúng tôi lấy kết quả mô hình thu được ở bước lặp 40 để tiến hành các thử nghiệm nhận dạng phương ngữ tiếng Việt.

Hai thử nghiệm nhận dạng phương ngữ tiếng Việt đã được thực hiện ứng với hai phương pháp PARAMETER_NORM và PARAMETER_FULL (mô tả ở phần II.C). Kết quả thử nghiệm được mô tả ở Bảng 2.

Bảng 2. Kết quả nhận dạng phương ngữ tiếng Việt (tỷ lệ %) sử dụng mạng CNN

Phương pháp	Kết quả thử nghiệm trên tập TEST
PARAMETER_NORM	85,5
PARAMETER_FULL	88,5

Kết quả ở Bảng 2 cho thấy phương pháp PARAMETER_FULL cho kết quả nhận dạng tốt hơn phương pháp PARAMETER_NORM. Ngoài ra khi so sánh với Bảng 1 cho thấy phương pháp đề xuất CNN cho kết quả nhận dạng trên tập TEST cho kết quả vượt trội so với các phương pháp học máy truyền thống như K láng giềng gần nhất hay máy hỗ trợ véctơ.

IV. KẾT LUẬN

Bài báo này trình bày phương pháp nhận dạng phương ngữ tiếng Việt sử dụng mạng nơron tích chập CNN. Các nghiên cứu hiện nay về nhận dạng phương ngữ tiếng Việt mới sử dụng các phương pháp học máy truyền thống như K láng giềng gần nhất KNN, cây quyết định, máy hỗ trợ véctơ SVM. Nghiên cứu này bước đầu áp dụng phương pháp mạng nơron học sâu vào bài toán này. Quá trình trích chọn tham số đã biểu diễn âm thanh tiếng nói ở dạng phổ spectrogram. Kiến trúc mạng được chọn lựa thử nghiệm là mạng nơron tích chập CNN. Kết quả thử nghiệm cho thấy phương pháp này đã đạt kết quả vượt trội (tỷ lệ nhận dạng đúng đạt 88,5%) so với các phương pháp học máy truyền thống bao gồm K láng giềng gần nhất (tỷ lệ nhận dạng đúng đạt 61,8%) và máy hỗ trợ véctơ (tỷ lệ nhận dạng đúng đạt 72,4%).

Trong nghiên cứu tiếp theo, chúng tôi sẽ sử dụng mạng CNN vào nhận dạng tiếng Việt nói, cũng như sẽ tích hợp mô hình CNN cho phương ngữ vào hệ thống này để làm tăng tỷ lệ nhận dạng đúng của hệ thống.

LỜI CẢM ƠN

Bài báo này được thực hiện trong khuôn khổ đề tài nghiên cứu khoa học cấp trường "Nghiên cứu xây dựng hệ thống nhận dạng phương ngữ tiếng Việt sử dụng phương pháp học sâu", mã số T2016-PC-044 của Trường Đại học Bách khoa Hà Nội. Các tác giả chân thành cảm ơn Trường Đại học Bách khoa Hà Nội, Phòng Khoa học Công nghệ, Viện Công nghệ Thông tin và Truyền thông đã hỗ trợ để chúng tôi có thể thực hiện thành công đề tài.

TÀI LIỆU THAM KHẢO

[1] Phạm Ngọc Hưng, Trịnh Văn Loan, Nguyễn Hồng Quang, "Automatic identification of Vietnamese speech", Tạp chí Tin học và Điều khiển học, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, trang 19-30, Tập 32, Số 1 năm 2016. Journal of Computer Science and Cybernetics, Vietnam Academy of Science and Technology, Volume 32, Number 1, 2016, ISSN 1813-9663

[2] Djork-Arné Clevert, Thomas Unterthiner, Sepp Hochreiter, Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs), ICLR 2016, May 2 - 4, 2016, Brazil.

[3] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[4] François Chollet, Keras, https://github.com/fchollet/keras, 2015

[5] Justin Johnson, Andrej Karpathy, "Convolutional Neural Networks for Visual Recognition course", Stanford University, 2016

[6] Nguyen Hong Quang, P. Nocera, E. Castelli, Trinh Van Loan, "A Novel Approach in Continuous Speech Recognition for Vietnamese, an Isolating Tonal Language". Proceedings of the INTERSPEECH, Brisbane, Australia, 2008, pp. 1149-1152.

[7] Quan Vu Hai, Kris Demuynck and Dirk Van Compernolle, "Vietnamese Automatic Speech Recognition: the FLaVoR Approach", International Symposium on Chinese Spoken Language Processing, Singapore, 2006.

[8] Thang Tat Vu, Dung Tien Nguyen, Mai Chi Luong and John-Paul Hosom, "Vietnamese Large Vocabulary Continuous Speech Recognition", Proceedings of Eurospeech, Lisboa, 2006

[9] Lei, Y. & Hansen, J. H. (2009). Factor analysis-based information for arabic dialect identification. In Acoustics, speech and signal processing, 2009, ICASSP 2009, pp 4337-4340.

[10] Sittichok Aunkaew, "Development of a Corpus for Southern Thai Dialect Speech Recognition: Design and Text Preparation", The 10th International Symposium on Natural Language Processing, pp 28-30, 2013.

[11] Ikuo Kudo, Takao Nakama, Tomoko Watanabe and Reiko Kameyama, "Data Collection of Japanese Dialects and Its Influence into Speech Recognition", The Fourth International Conference on Spoken Language Processing ICSLP 1996, October 3-6, Philadenphia, USA.

[12] Wendy Baker, David Eddington and Lyndsey Nay, Dialect recognition: The Effects of Region of Origin and Amount of Experience, American Speech 84:48-71, January 2009.

VIETNAMESE DIALECT RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK

Nguyen Hong Quang, Trinh Van Loan, Pham Ngoc Hung

***ABSTRACT**: This paper presents the method of Vietnamese dialect identification using convolutional neural network. Current studies on Vietnamese dialect r utilize traditional machine learning methods such as the K nearest neighbor, decision tree, support vector machine. This study initially applied neural network methodology to this problem. The parametric extraction process performed speech spectrogram. The chosen network architecture is the convolutional neural network. The results show that this method has outperformed traditional machine learning methods.*