

ĐÁNH GIÁ CÁC THAM SỐ CỦA MÔ HÌNH MẠNG NƠ-RON TÍCH CHẬP VÀ ỨNG DỤNG VÀO THIẾT KẾ PHẦN MỀM NHẬN DẠNG CHỮ SỐ VIẾT TAY TRÊN NỀN TẢNG DI ĐỘNG ANDROID

Nguyễn Ngọc Tuấn Anh^{1*}, Vương Quang Phước¹, Phan Hải Phong¹

¹Khoa Điện, Điện tử và Công nghệ vật liệu, Trường Đại học Khoa học, Đại học Huế

*Email: nntuananh97@gmail.com

Ngày nhận bài: 22/6/2020; ngày hoàn thành phản biện: 3/7/2020; ngày duyệt đăng: 02/6/2021

TÓM TẮT

Trong thập kỉ vừa qua, trí nhân tạo nói chung hay kĩ thuật học sâu nói riêng đã có những bước phát triển đáng chú ý. Bài báo trình bày một mô hình mạng trí tuệ nhân tạo, nhận diện chữ số viết tay bằng mạng neuron tích chập (Convolutional neural network - CNN) [1]. Qua đó làm rõ các khái niệm tham số, đánh giá tầm quan trọng các tham số trong mô hình, trình bày kết quả mô phỏng đạt được khi sử dụng mạng neuron nhân tạo để nhận diện các ảnh chữ số viết tay dựa trên tập dữ liệu MNIST (Modified National Institute of Standards and Technology) [2] và đưa mô hình mạng CNN ứng dụng vào bài toán nhận dạng chữ số viết tay trên nền tảng Android. Hiệu năng của mô hình được đánh giá qua tỉ lệ nhận dạng đúng và được khảo sát với nhiều trường hợp khác nhau.

Từ khóa: trí tuệ nhân tạo, mạng nơ-ron tích chập, CNN, MNIST.

1. MỞ ĐẦU

Có thể nói, với sự thỏa mãn về cả ba yếu tố: nguồn dữ liệu đủ lớn, phần cứng hỗ trợ mạnh và các thuật toán tiên tiến, trí tuệ nhân tạo (Artificial Intelligent - AI) đã tạo nên một phong trào công nghệ mới trong kỉ nguyên số hóa hiện tại. Trong đó, việc thu thập thông tin từ hệ thống dữ liệu hình ảnh khổng lồ trên toàn thế giới đang là một lĩnh vực được nhiều nhà khoa học trên toàn thế giới quan tâm và nghiên cứu [3]. Đây là cơ hội và cũng là một thách thức hàng đầu, việc ứng dụng trí tuệ nhân tạo nói chung hay kĩ thuật học sâu (Deep Learning - DL) nói riêng đang là một lĩnh vực đầy tính cạnh tranh, mục tiêu hướng đến là tăng tốc độ xử lý, khả năng trích xuất và thu thập thông tin từ nguồn dữ liệu nói trên cho các mục đích sử dụng khác nhau.

Nghiên cứu trình bày về mạng nơ-ron tích chập (Convolutional Neural Network - CNN), và khả năng ứng dụng của nó trên nền tảng di động android. Trong phạm vi

bài báo, chúng tôi sẽ cố gắng làm rõ mô hình CNN và đánh giá các khối chức năng, tác động của các tham số đến kết quả nhận dạng hình ảnh. Đối tượng cụ thể ở đây là tập cơ sở dữ liệu chữ số viết tay MNIST [2]. Cuối cùng, bài báo cũng đề cập đến tính khả dụng của việc xây dựng mô hình trên nền tảng di động Android cho việc nhận dạng chữ số.

2. KỸ THUẬT HỌC SÂU VÀ MẠNG NƠ-RON TÍCH CHẬP

2.1. Kỹ thuật học sâu

Học sâu (Deep Learning - DL) là một nhánh nhỏ trong Machine Learning hay học máy cho phép máy tính tự huấn luyện để thực hiện việc học hỏi từ một lượng lớn dữ liệu được cung cấp để giải quyết những vấn đề cụ thể.

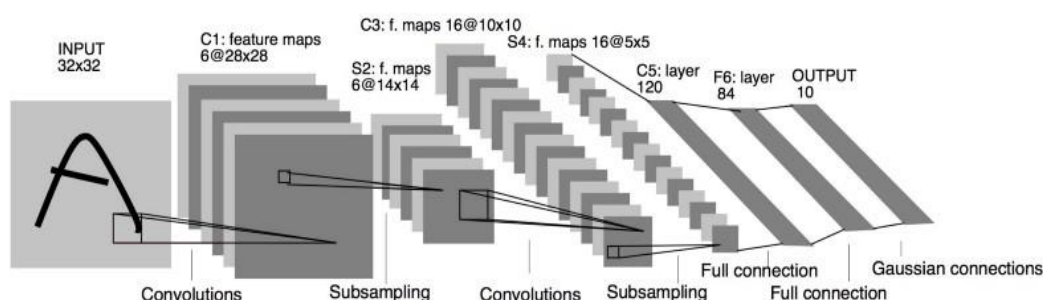
Quá trình nghiên cứu và đánh giá sẽ được thực hiện trên tập dữ liệu chữ số viết tay thông dụng MNIST theo phương thức học có giám sát. Có nghĩa là, các dữ liệu đầu vào đã được xác định sẵn, việc học sẽ thực hiện dự đoán và so sánh kết quả đầu ra với đầu vào, sau đó mô hình mạng sẽ thực hiện “học” và điều chỉnh tham số mạng cho phù hợp với tập dữ liệu đầu vào.

Trong bài báo này, chúng tôi lựa chọn mô hình mạng nơ-ron tích chập thay vì mô hình mạng MLP truyền thống vì ưu điểm xử lý phân tích theo phân vùng không gian, đây là một ưu điểm nổi trội khi đánh giá các tập dữ liệu nhiều chiều, trong trường hợp này là dữ liệu hình ảnh.

2.2. Mạng nơ-ron tích chập CNN

2.2.1. Mạng nơ-ron tích chập

Trong các mạng MLP truyền thống, mỗi nơ-ron trong lớp phía trước sẽ được kết nối đến tất cả các nơ-ron ở lớp phía sau, điều này khiến cho khối lượng tính toán trong mạng tăng mạnh khi tăng độ sâu của mô hình (tăng số lượng lớp) cho mô hình.



Hình 2.1. Sơ đồ khối mô tả mô hình mạng lenet-5 (1998) [4].

Sự ra đời của mạng CNN đã giúp giải quyết vấn đề trên bằng cách sử dụng các vùng tiếp nhận cục bộ, tập trọng số chia sẻ và phương pháp lấy tích chập để trích xuất

thông tin thay cho phương pháp cổ điển. Hình 2.1 phía trên biểu diễn mô hình mạng Lenet-5, mô hình CNN được Yan Lecun công bố đầu tiên vào năm 1998.

Sau đây, chúng tôi sẽ đi vào chi tiết hơn về mạng nơ-ron tích chập và các tham số sẽ được nghiên cứu đánh giá.

2.2.2. Lớp tích chập (convolutional layer)

Lớp tích chập là một thành phần cốt lõi của mạng nơ-ron tích chập (CNN), sử dụng để trích xuất các thông tin đặc tính của hình ảnh (feature map) hỗ trợ cho quá trình “học” của mạng CNN [5]. Phương thức hoạt động của lớp này được thực hiện thông qua quá trình trượt và lấy tích chập của bộ lọc (filter/kernel) trên toàn bộ ảnh. Kết quả đầu ra là đặc tính của ảnh tương ứng với bộ lọc đã sử dụng, với càng nhiều bộ lọc được sử dụng, chúng tôi sẽ thu được càng nhiều đặc tính của ảnh tương ứng.

Trong phạm vi bài báo này, chúng tôi sẽ xét đến các tham số có tác động đến lớp tích chập gồm: kích thước bộ lọc, bước trượt (stride) [5].

2.2.3. Lớp lấy mẫu xuống (pooling layer)

Lớp lấy mẫu xuống có tác dụng giảm kích thước của dữ liệu hình ảnh từ đó giúp cho mạng có thể học được các thông tin có tính chất khái quát hơn, đồng thời quá trình này giảm số lượng các thông số trong mạng [6].

Các phương pháp lấy mẫu xuống thường được sử dụng là Max Pooling và Average Pooling [7].

2.2.4. Lớp dropout

Lớp dropout là một kỹ thuật được sử dụng để hạn chế hiện tượng overfitting (hiện tượng mạng nơ-ron quá bám sát vào tập dữ liệu huấn luyện và không đáp ứng được với các tập dữ liệu mới), thường gặp ở mạng CNN và giúp mô hình tính toán nhanh hơn. Dropout sử dụng phương pháp loại bỏ một số nơ-ron ngẫu nhiên trong mạng với một xác suất cho trước bằng cách thiết lập tất cả trọng số nơ-ron đó bằng 0, đồng nghĩa với các liên kết tới nơ-ron đó đều không có giá trị, khi đó mô hình sẽ phải cố gắng nhận dạng đúng trong khi thiếu thông tin từ các nơ-ron bị loại bỏ. Điều này sẽ giúp tăng tỉ lệ nhận dạng của mô hình nhưng không quá phụ thuộc vào dữ liệu huấn luyện [5].

2.2.5. Lớp kết nối đầy đủ (Fully-connected layer - FC)

Đầu vào của lớp kết nối đầy đủ là đầu ra từ lớp lấy mẫu xuống hoặc lớp tích chập cuối cùng, nó được làm phẳng và sau đó được đưa vào lớp kết nối đầy đủ để chuyển tiếp. Lớp FC có nhiệm vụ tổng hợp thông tin đưa ra lớp quyết định (output) cho ra kết quả chính xác nhất.

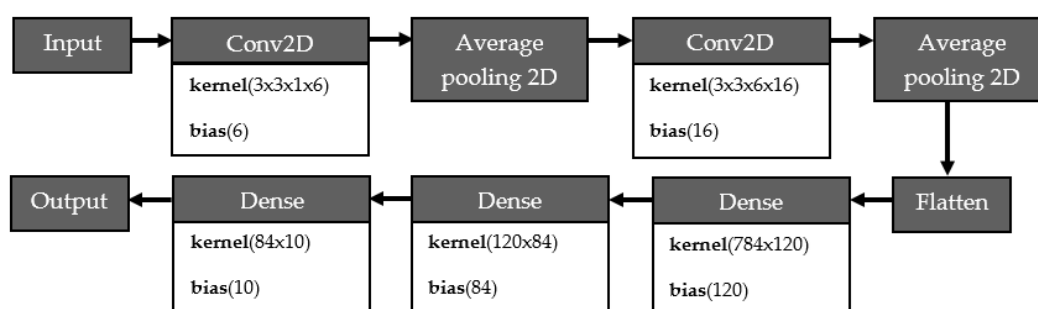
3. KHẢO SÁT VÀ ĐÁNH GIÁ HOẠT ĐỘNG

3.1. Mô hình và cơ sở dữ liệu

3.1.1. Mô hình mạng CNN

Với mục tiêu đánh giá được vai trò của các tham số tác động đến kết quả đầu ra, nhóm nghiên cứu thực hiện đánh giá dựa trên mô hình CNN cơ bản (mô hình Lenet-5) và thực hiện thay đổi các thông số, thay đổi mô hình bằng cách chèn các khối theo yêu cầu của từng khảo sát cụ thể.

Mô hình được mô tả như trong hình 3.1:

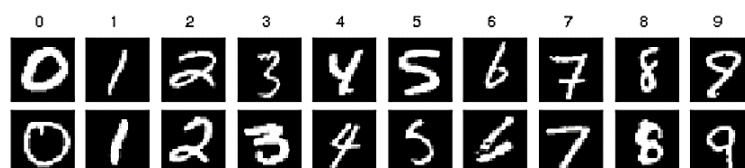


Hình 3.1. Mô hình mạng CNN sử dụng trong nghiên cứu và đánh giá.

3.1.2. Bộ cơ sở dữ liệu MNIST

Tập cơ sở dữ liệu MNIST là một cơ sở dữ liệu lớn gồm các chữ số viết tay, thường được sử dụng để đào tạo các hệ thống học máy nhằm xử lý hình ảnh khác nhau. Cơ sở dữ liệu cũng được sử dụng rộng rãi để đào tạo và thử nghiệm trong lĩnh vực học máy.

Tập MNIST gồm 70000 dữ liệu được chia thành 2 phần: 60000 hình ảnh sử dụng cho huấn luyện và 10000 hình ảnh dùng cho mục đích kiểm tra.



Hình 3.2. Một số hình ảnh ví dụ trích xuất từ cơ sở dữ liệu MNIST [8].

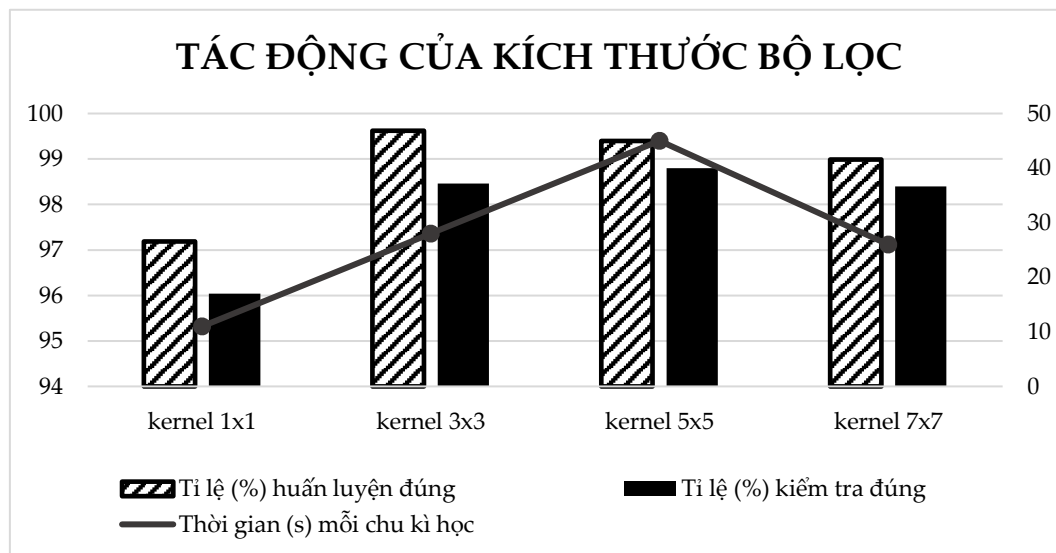
3.2. Đánh giá vai trò của các tham số của mô hình

Tỉ lệ nhận dạng đúng trên 2 tập huấn luyện và kiểm tra sẽ được sử dụng làm tiêu chí để đánh giá mức độ hiệu quả chính xác của mô hình.

3.2.1. Kích thước bộ lọc

Bộ lọc (kernel) chính là cửa sổ có kích thước $n \times n$, có nhiệm vụ thực hiện tích chập với ảnh để trích xuất các đặc tính của ảnh. Các bộ lọc này sử dụng tập trọng số và

ngưỡng chung (shared weights and bias) trên 1 bộ lọc trong suốt quá trình bộ lọc đó trượt trên ảnh, đặc điểm này giúp giảm thiểu được số lượng các tham số của mạng so với mạng no-ron truyền thống (mạng multi layer perceptron – MLP).



Hình 3.3. Sơ đồ biểu diễn kích thước của bộ lọc ảnh hưởng tới tỉ lệ nhận dạng đúng và thời gian mỗi chu kì học trong mô hình.

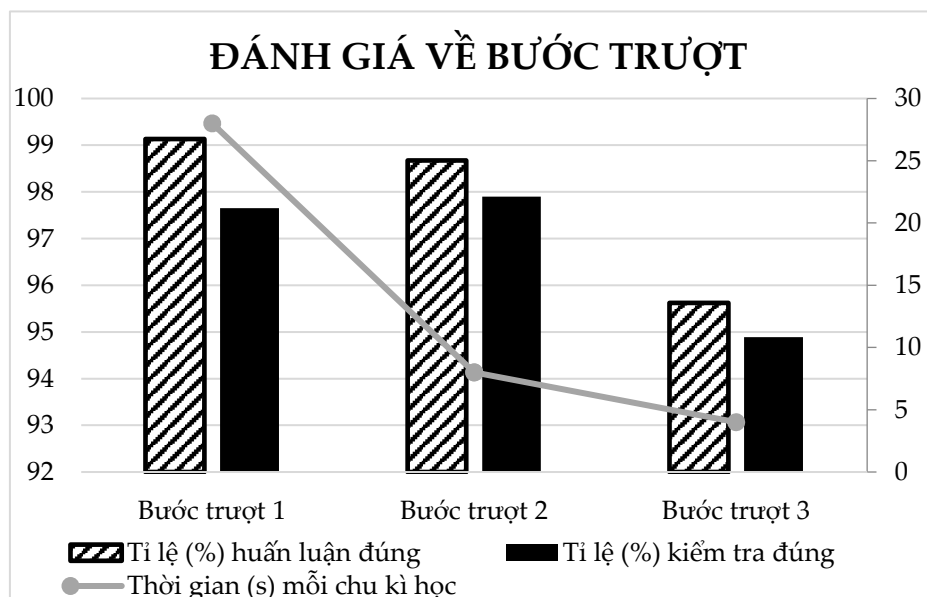
Nhìn vào biểu đồ trên, với 4 kích thước của bộ lọc được xét là: 1x1, 3x3, 5x5, 7x7. Ngoài ra, kích thước bộ lọc 3x3 chính là kích thước mặc định của mô hình Lenet-5 ở hình 2.1. Nhìn chung, tỉ lệ nhận dạng đúng trên tập kiểm tra của các trường hợp 3x3, 5x5, 7x7 khá tốt, đều đạt mức trên 98%; điều ngược lại đúng với trường hợp 1x1, đạt 96%. Bên cạnh đó, thời gian mỗi chu kì học cũng tăng theo kích thước của bộ lọc.

Trường hợp sử dụng của số kích thước 1x1 có phương thức hoạt động khá tương đồng với cơ chế mạng MLP, một điểm ở lớp này sẽ kết nối đến tất cả các bản đồ đặc tính ở phía sau nên tỉ lệ nhận dạng đúng trên tập kiểm tra không cao. Ở các kích thước của số còn lại hiện tượng overfitting ít hơn 1x1 nhưng kích thước 5x5 vẫn cho tỉ lệ nhận dạng trên tập kiểm tra cao nhất trong các kích thước còn lại. Theo khuyến nghị của chúng tôi khi sử dụng hình ảnh đầu vào có kích cỡ nhỏ (nhỏ hơn 128x128 điểm ảnh) thì nên dùng bộ lọc 3x3 thích hợp cho mô hình.

3.2.2. Bước trượt

Bước trượt (stride) là số điểm ảnh mà bộ lọc sẽ dịch chuyển so với vị trí trước đó. Giả sử chọn bộ lọc với giá trị bước trượt bằng 1, bộ lọc sẽ dịch chuyển toàn bộ cửa sổ tích chập từ trái sang phải và từ trên xuống 1 điểm ảnh. Giá trị bước trượt càng lớn, tốc độ tính toán càng nhanh (số lượng phép tính ít), tuy nhiên kích thước của thông tin đặc tính của hình ảnh càng nhỏ, lượng thông tin bị mất càng nhiều.

Dựa vào biểu đồ ở hình 3.4, với 3 kích thước bước trượt được xét là: 1, 2, 3. Có thể thấy rằng tỉ lệ nhận dạng đúng trên tập kiểm tra ở bước trượt bằng 1, 2 đạt trên 97,5% và ngược lại với bước trượt bằng 3 thấp hơn 95% bên cạnh đó thời gian mỗi chu kì học cũng giảm theo kích thước của bước trượt.



Hình 3.4. Sơ đồ thể hiện kích thước của các bước trượt ảnh hưởng tới tỉ lệ nhận dạng đúng và thời gian mỗi chu kì học trong mô hình mạng.

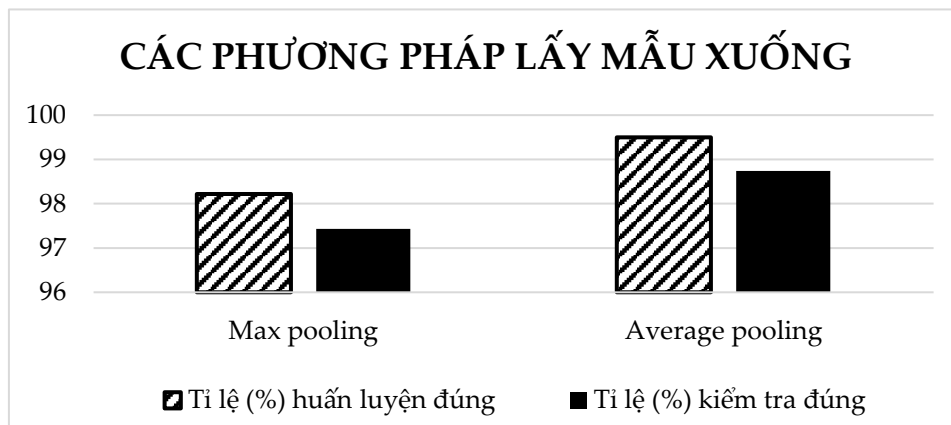
Nhận thấy rằng ở bước trượt 1 xảy ra tình trạng overfitting dẫn đến tỉ lệ nhận dạng thấp hơn bước trượt 2. Ở số bước trượt 2 tỉ lệ nhận dạng đúng trên tập kiểm tra cao nhất trong 3 trường hợp được xét. Đối với trường hợp bước trượt bằng 3 thì tỉ lệ nhận dạng trên tập kiểm tra thấp do khối lượng thông tin bị mất mát quá nhiều. Xét chu kì học thì tại mỗi trường hợp thì bước trượt bằng 1 cho thời gian lâu nhất do sự dịch chuyển tại bước này nhỏ nên thời gian để bộ lọc trượt hết toàn bộ ảnh càng lâu và điều này ngược lại so với các trường hợp khác. Với kết quả trong hình 3.4 thì nên sử dụng bước trượt bằng 2 vì tại bước trượt này cho kết quả nhận dạng trên tập kiểm tra cao với chu kì học chấp nhận được.

3.4. Max Pooling và Average Pooling

Như đã chỉ ra ở mục 2.2.3, ta tiến hành khảo sát với 2 phương pháp lấy mẫu là Max Pooling và Average Pooling.

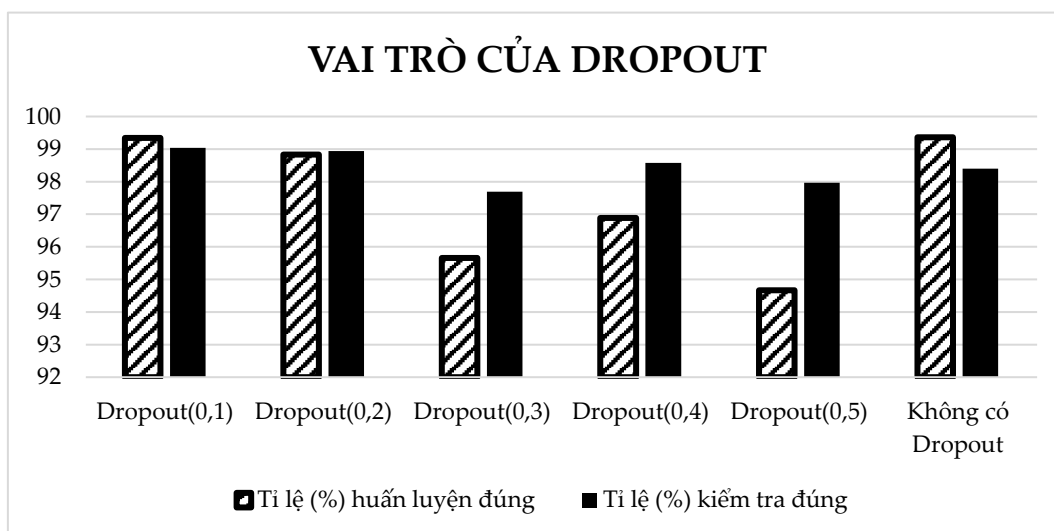
Hình 3.5 thể hiện kết quả sau 50 chu kì học, mô hình được thiết lập với bước trượt bằng 2. Dễ dàng nhận thấy được, tỉ lệ nhận dạng đúng của Average Pooling (~99% trên tập kiểm tra) cao hơn Max Pooling (~97% trên tập kiểm tra) và hiện tượng overfitting của hai phương pháp ngang nhau (~1% chênh lệch giữa 2 tập dữ liệu).

Trong thực tế, Average Pooling thường được sử dụng để trích xuất các nội dung mang tính chất tổng hợp, còn Max Pooling được sử dụng cho các trường hợp muốn tách các thông tin nổi trội trong dữ liệu được học. Tùy vào từng điều kiện cụ thể mà có thể sử dụng từng loại hoặc kết hợp cả 2 phương pháp để tăng tính linh hoạt cho mô hình.



Hình 3.5. Phương pháp lấy mẫu xuống ảnh hưởng tới tỉ lệ nhận dạng đúng trong mô hình mạng.

3.5. Tác động của lớp dropout



Hình 3.6. Sơ đồ biểu diễn sự ảnh hưởng của dropout lên mô hình mạng.

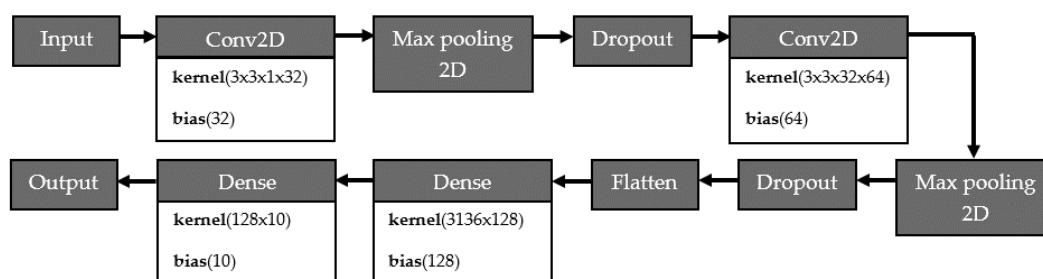
Nhìn vào hình 3.6 thì tỉ lệ nhận dạng đúng trên tập kiểm tra của dropout 0,1 là cao nhất đạt trên 99% và giảm dần từ 0,2 $\rightarrow$ 0,4 $\rightarrow$ “không có dropout” $\rightarrow$ 0,5 $\rightarrow$ 0,3. Qua đó dropout của 0,1; 0,2 và “không có dropout” xảy ra hiện tượng overfitting còn ở 0,3; 0,4; 0,5 xảy ra hiện tượng underfitting (tỉ lệ huấn luyện đúng thấp hơn tỉ lệ kiểm tra đúng). Cụ thể trong bài toán được khảo sát, chọn tỉ lệ dropout 0,1 và 0,2 cho được kết quả khả quan hơn các trường hợp còn lại. Tuy nhiên, đối với các trường hợp và bài toán khác cần có sự khảo sát với nhiều trường hợp để có tỉ lệ tối ưu hơn.

3.7. So sánh hai mô hình lý thuyết và thực tế

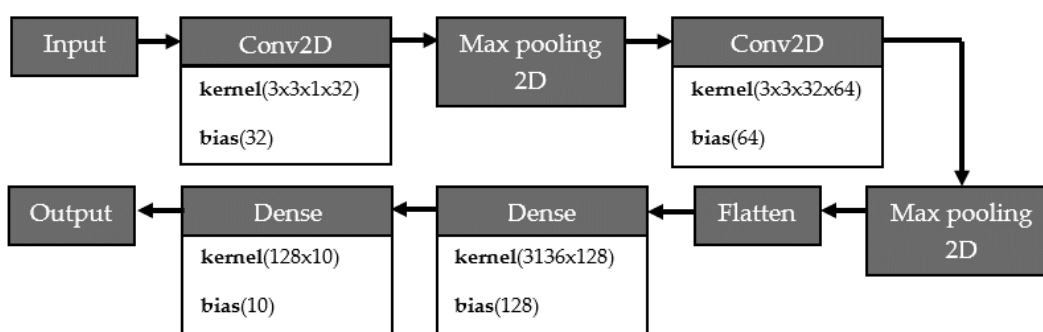
Vì một số vấn đề liên quan đến sự chênh lệch về giới hạn phần cứng và xử lý phần mềm trên máy tính và thiết bị di động, nhóm nghiên cứu đề xuất mô hình CNN với cấu trúc như hình 3.7. Một số ưu điểm của mô hình có thể kể đến: đơn giản, linh hoạt, đáp ứng được bài toán đề xuất.

Tuy nhiên, để có thể ứng dụng mô hình trên vào nền tảng di động cần có sự chuyển đổi giữa Tensorflow (TF) và Tensorflow Lite (TF Lite), quá trình chuyển đổi giữa hai nền tảng thuật toán cũng có sự khác biệt. Dễ dàng nhận thấy, đó chính là việc khuyết thiếu khối chức năng Dropout khi chuyển từ nền tảng TF sang TF Lite, lưu ý, các tham số khác như tập trọng số chung, ngưỡng chung, kích thước bộ lọc, hay bước trượt ... vẫn được giữ nguyên.

Hình 3.7 và 3.8 mô tả sự tương quan giữa hai mô hình hoạt động trên hai nền tảng khác nhau.



Hình 3.7. Mô hình mạng CNN sử dụng thư viện Tensorflow.



Hình 3.8. Mô hình mạng CNN sử dụng thư viện Tensorflow Lite.

Hình 3.9, mô tả một số kết quả khi thực hiện trên ứng dụng di động. Với mô hình được sử dụng như trình bày, kết quả nhận được khá khả quan. Tuy nhiên, với dữ liệu đưa vào là bộ dữ liệu mới (do người sử dụng viết và chưa qua học tập) vẫn còn một số trường hợp nhận dạng chưa chính xác, ví dụ: liên quan đến vấn đề chữ số viết nghiêng,

chữ số được viết trong một phần của giao diện viết hoặc trường hợp thử nghiệm với nhiều hơn 1 chữ số, hình 3.10 thể hiện các kết quả còn chưa tối ưu.

Có thể nhận thấy, ứng dụng chỉ đáp ứng được việc xử lý bài toán đơn giản, trong một số điều kiện cụ thể và còn cần phát triển thêm trong tương lai với nhiều tác vụ phức tạp hơn.



Hình 3.9. Giao diện của ứng dụng nhận dạng chữ số viết tay trên thiết bị di động Android.



Hình 3.10. Các trường hợp cho kết quả chưa chính xác.

4. KẾT LUẬN

Bài báo đã thực hiện giới thiệu về mạng nơ-ron tích chập và khảo sát các tham số cũng như các khối chức năng, đánh giá vai trò tác động của chúng đến tỉ lệ nhận dạng hình ảnh ở đầu ra. Nghiên cứu cũng thực hiện triển khai mô hình mạng CNN nhằm giải quyết bài toán nhận dạng chữ số viết tay trên nền tảng Android sử dụng thư viện Tensorflow và Tensorflow Lite. Tuy kết quả đánh giá trên thực tế khá khả quan, nhưng vẫn còn tồn tại một số trường hợp nhận diện sai do sai khác về tập dữ liệu của mô hình giữa lý thuyết và thực tế. Do vậy, nhóm nghiên cứu có định hướng xây dựng ứng dụng nhận diện hoàn thiện hơn trong tương lai, và mở rộng cơ sở dữ liệu không chỉ chữ số viết tay mà còn thể ứng dụng lên các tập khác như CIFAR-10, CIFAR-100, hay Fashion MNIST, ...

TÀI LIỆU THAM KHẢO

- [1]. Chhavi Yadav, Léon Bottou (2019). arXiv:1905.10498[cs.LG].
- [2]. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE*, 86(11):2278-2324, November 1998.
- [3]. Himadri Sankar Chatterjee (2019) "A Basic Introduction to Convolutional Neural Network".
- [4]. Eugenio Culurciello (19/4/2017) "The History of Neural Networks".
- [5]. Michael A. Nielsen (2015) "Neural Network and Deep Learning".
- [6]. Adrian Rosebrock (31/2018) "Keras Conv2D and Convolutional Layers".
- [7]. Waseem Rawat and Zenghui Wang (9/6/2017) "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review" DOI: 10.1162/NECO_a_00990.
- [8]. Henry Chinner (18/3/2015) "Simple Neural Network Implementation in Ruby".

EVALUATING THE PARAMETERS OF THE CONVOLUTION NEURAL NETWORK MODEL (CNN) AND APPLYING TO DESIGN SOFTWARE IDENTIFICATION OF WRITTEN NUMBERS ON ANDROID MOBILE PLATFORM

Nguyen Ngoc Tuan Anh*, Vuong Quang Phuoc, Phan Hai Phong

University of Sciences, Hue University

*Email: nntuananh97@gmail.com

ABSTRACTS

Over the past decade, artificial intelligence in general or deep learning in particular has made remarkable developments. The report presents a model of artificial intelligence network, handwritten number recognition by convolutional neural network (CNN) [1]. Thereby, the parameter concept is clarified to evaluate the importance of the parameters in the model, presenting simulation results achieved when using artificial neural networks to identify handwritten numerical images based on MNIST data set (Modified National Institute of Standards and Technology) [2] and bring simulation models onto the Android application. The result of the model is appreciated through the rate of right reconization and investigated with many different cases.

Keywords: artificial intelligence, convolutional neural network, CNN, MNIST.



Nguyễn Ngọc Tuấn Anh sinh năm 1997 tại Thừa Thiên Huế. Năm 2020, ông tốt nghiệp kỹ sư chuyên ngành Điện tử viễn thông tại trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: mạng Neuron nhân tạo.



Vương Quang Phước sinh năm 1990 tại Thừa Thiên Huế. Năm 2013, ông tốt nghiệp kỹ sư chuyên ngành Điện tử viễn thông tại trường Đại học Khoa học, Đại học Huế. Năm 2018, ông nhận bằng thạc sĩ chuyên ngành Kỹ thuật Điện tử tại trường Đại học Bách khoa Đà Nẵng. Hiện nay, ông đang công tác tại Khoa Điện, Điện tử và Công nghệ vật liệu, trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: Hệ thống thông tin quang, mạng Neuron nhân tạo.



Phan Hải Phong sinh năm 1982 tại Thừa Thiên Huế. Ông tốt nghiệp Đại học ngành Vật lý năm 2004 tại trường Đại học Khoa học, ĐH Huế; nhận bằng Thạc sĩ (2010) và Tiến sĩ (2019) về chuyên ngành Kỹ thuật điện tử tại trường Đại học Công nghệ, ĐH Quốc Gia Hà Nội. Hiện ông công tác tại Khoa Điện, Điện tử và Công nghệ vật liệu, trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: thiết kế vi mạch, hệ thống nhúng, hệ thống tích hợp trên vi mạch.

