NGHIÊN CỨU TỐI ƯU HÓA TẬP TRỌNG SỐ CỦA MẠNG NƠRON CHO CÁC PHẦN CỨNG CẤU HÌNH THẤP

WEIGHT OPTIMIZATION RESEARCH FOR LOW COST HARDWARE

Nguyen Kim Thach

Trường đại học Sư phạm Kỹ thuật TP.HCM

TÓM TẮT

Hiện nay, với sự phát triển mạnh mẽ của thị giác máy tính và các hệ thống IOT, vấn đề tiết kiệm chi phí, năng lượng và tài nguyên máy tính trong các ứng dụng chi phí thấp đang được rất nhiều dự án quan tâm và chú trọng. Trong bài báo này, đề tài đề xuất một giải pháp tối ưu hóa tập trọng số của mạng nơ-ron, đó là lượng tử hóa hầu hết các trọng số về dạng trọng số nhị phân có dấu 1 và -1 để áp dụng cho phần cứng giá rẻ là Raspberry Pi Zero. Trọng số nhị phân sẽ giúp cho việc tính toán trở nên nhanh hơn, tiết kiệm năng lượng trong quá trình huấn luyện so với các trọng số 32bits. Với bộ dữ liệu biển số xe, GTSRB được huấn luyện với thư viện mạng thần kinh nhân tạo TensorFlow, đề tài đã đạt độ chính xác 96,01%. Ngoài ra, đề tài cũng đã huấn luyện mạng BCNN với các bộ dữ liệu khác và đạt độ chính xác là Mnist (99,81%), Cifa-10 (93,59%), Mnist-fashion (96,35) và bộ dữ liệu mã thùng sản phẩm – CNPB (99.89%). Khi huấn luyện trên các tập dữ liệu Mnist, Mnist-fashion, Cifa-10, GTSRB, CNPB thì tập trọng số của mạng nơron BCNN có dung lượng lưu trữ giảm từ 6.16 đến 7.59 lần so với mô hình mạng CNN thông thường cùng cấu trúc. Tốc độ xử lý của mô hình BCNN nhanh gấp 5,2 lần so với mạng CNN thông thường có cấu trúc tương tự khi được triển khai lên phần cứng cấu hình thấp là Raspberry Pi Zero với tập dữ liệu CNPB.

Từ khóa: *Raspberry Pi Zero; BCNN; CNN; GTSRB; CNPB.*

ABSTRACT

Currently, with the strong development of computer vision and IOT systems, the problem of saving costs, energy, and computing resources in low-cost applications is being pursued by many projects. care and attention. In this paper, the topic proposes a solution to optimize the weight set of the neural network, which is to quantize most of the weights in the form of signed binary weights 1 and -1 to apply to cheap hardware such as the Raspberry Pi Zero. Faster, more energy-efficient binary weighting thanks to bitwise operations When training with the license plate dataset, the GTSRB is trained with the TensorFlow artificial neural network library. The topic has achieved 96.01% accuracy. In addition, the project also trained the BCNN network with other data sets and achieved the accuracy of Mnist (99.81%), Cifa-10 (93.59%), Mnist-fashion (96.35) and the product box code dataset - CNPB (99.89%). When training on the data sets Mnist, Mnist-fashion, Cifa-10, GTSRB, CNPB, the inner set of the BCNN neural network, the storage capacity of the BCNN binary neural network weights decreases from 6.16 to 7.59 times compared to the conventional CNN model with the same structure. The processing speed of the BCNN model is 5.2 times faster than a conventional CNN network with a similar structure when deployed to low-profile hardware, the Raspberry Pi Zero with the CNPB dataset.

Keywords: *Raspberry Pi Zero; BCNN; CNN; GTSRB; CNPB.*

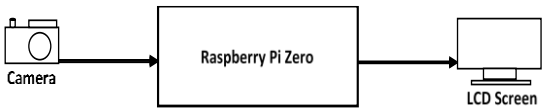
1. GIỚI THIỆU

Mạng nơ-ron nhân tạo (ANN) với độ chính xác cao đã và đang ngày càng đạt được nhiều thành tựu trong nhiều lĩnh vực như nhận dạng hình ảnh, nhận dạng giọng nói và dự đoán [1–7]. Tuy nhiên, mạng ANN rất tốn kém về mặt tính toán. Lý do là ANN được xây dựng trên hệ thống 32 bit và chúng bao gồm một số lượng rất lớn các tác vụ tính toán và tham số bên trong.

Một số ứng dụng thực tế chi phí thấp, chẳng hạn như nhận dạng biển báo giao thông, nhận dạng cử chỉ tay, nhận dạng mã sản phẩm, v.v., không cần triển khai trên các hệ thống có phần cứng lớn và sử dụng nhiều năng lượng để hoạt động. Một giải pháp để đạt được thiết kế này là mạng thần kinh nhị phân (BNN), hạn chế trọng số và kích hoạt của mô hình thành +1 hoặc -1 [8–11]. Do đó, trong cùng một cấu trúc liên kết mạng, dung lượng lưu trữ mà BNN yêu cầu giảm hơn nhiều lần so với của ANN 32 bit.

So với các phép toán số học DNN, các phép toán thao tác bit BNN cải thiện đáng kể hiệu quả năng lượng [12]. Mục tiêu của luận án trong công việc này là phát triển khung mạng thần kinh nhị phân dựa trên phần cứng chi phí thấp cho Raspberry Pi Zero để nhận dạng mã trên hộp sản phẩm. Các tính toán toán học (cộng, trừ, nhân và chia) của mạng nhị phân có thể được thực hiện bằng phép toán nhân được thay thế bằng phép toán AND logic để giảm dung lượng lưu trữ cần thiết và tăng tốc mạng hiệu suất.

Mạng thần kinh nhị phân được đề xuất triển khai trên bo mạch Raspberry Pi Zero nhận dạng mã để phân loại sản phẩm nhằm tránh nhầm lẫn.



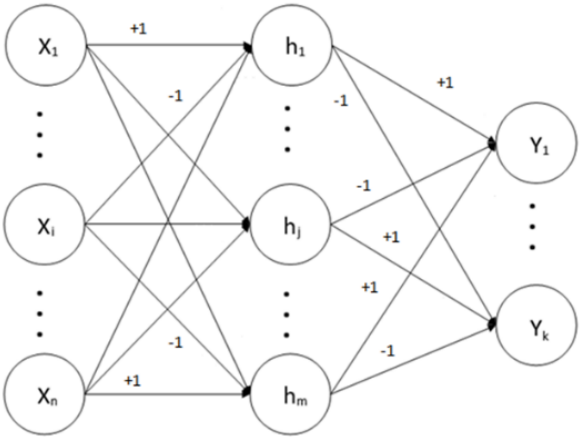
Hình 1.Sơ đồ khối của bộ điều khiển nhận dạng bằng Raspberry Pi. phần cứng

Hình 1 cho thấy sơ đồ mô tả hoạt động điều khiển cho ứng dụng nhận dạng. Bộ điều khiển được triển khai trên bo mạch Raspberry Pi Zero 2W. Hệ thống thực hiện các nhiệm vụ nhận dạng mẫu, bao gồm nhận dạng hình ảnh và nhận dạng đối tượng.

2. CÔNG TRÌNH CÓ LIÊN QUAN

2.1 Mạng thần kinh nhị phân

Để huấn luyện trọng số mạng thần kinh sử dụng số dấu phẩy động 32 bit. Với mục đích tiết kiệm tài nguyên cho cả phần cứng và quá trình tính toán, [12] đã đề xuất một phương pháp sử dụng trọng số giá trị 1 và -1 để huấn luyện mạng thần kinh giúp giảm mức tiêu thụ bộ nhớ và các lần truy cập khi triển khai phần cứng chi phí thấp. Trọng số nhị phân và kích hoạt có thể thay thế các số dấu phẩy động để huấn luyện mạng nơ-ron, điều này có thể làm giảm đáng kể sự dư thừa của mạng nơ-ron tích chập và cũng đạt được độ chính xác phân loại cao hơn so với [22]. [23] cho thấy rằng các CNN với trọng số nhị phân có thể tăng tốc độ tính toán một cách hiệu quả và tốc độ xử lí của chúng được cải thiện hơn.

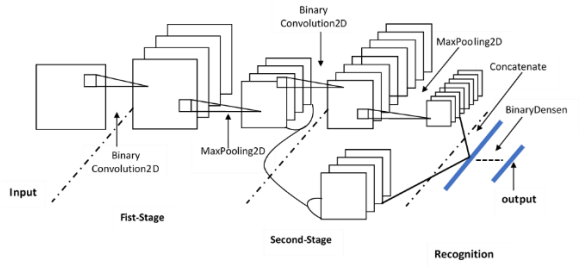


Hình 2.Sơ đồ mô hình mạng thần kinh bậc ba , trong đó các trọng số khớp thần kinh là - 1, + 1.

Hình 2 cho thấy một mạng thần kinh nhị phân trong đó các trọng số khớp thần kinh là -1 hoặc +1. Để biểu thị các giá trị có trọng số -1 hoặc +1, có thể sử dụng bộ nhớ 8-bit thay vì bộ nhớ động 32-bit được sử dụng trong các mạng thần kinh có độ chính xác đầy đủ để lượng bộ nhớ cần thiết cho các tham số mô hình của mạng thần kinh nhị phân nhỏ hơn nhiều so với mạng thần kinh có độ chính xác cao (giảm hơn 6-7.59 lần). Do dung lượng lưu trữ cần thiết cho bộ đã giảm nên có thể áp dụng mạng thần kinh này cho hệ thống nhúng chi phí thấp như bo mạch Raspberry Pi Zero.

Để nâng cao độ chính xác, luận án thay thế các CNN thông thường và mạng nơ-ron liên thông đầy đủ bằng các CNN nhị phân và

mạng liên thông đầy đủ. Chúng tôi cũng sử dụng chức năng kích hoạt nhị phân, chức năng ký hiệu. Do đó, BCNN sẽ được hiển thị trong hình bên dưới. Kiến trúc mạng chi tiết được hiển thị trong hình 3.



Hình 3. Kiến trúc. BCNN có 9 lớp tích chập (trong đó tầng 1 có 3 lớp và tầng 2 có 6 lớp) tiếp theo là 3 tầng liên thông đầy đủ với tập số được lượng tử hóa thành hệ nhị phân.

2.2 Lượng tử hóa nhị phân

Khi đào tạo BNN, chúng tôi sẽ lượng tử hóa trọng số thành +1 hoặc -1. Hai giá trị đó rất thuận tiện từ góc độ phần cứng. Để biến đổi các biến có giá trị thực thành hai giá trị đó, chúng ta sẽ phải sử dụng hai hàm khác nhau [19–21].

$$x^b = \text{Sign}(x) = \begin{cases} +1 & if \ x \geq 0 \\ -1 & if \ x < 0 \end{cases} \qquad (1)$$

Với x^b là biến nhị phân (trọng số hoặc kích hoạt) và x có giá trị thực.

2.3 Thuật toán tính toán và lan truyền gradient

Vì đạo hàm của hàm (1) luôn bằng 0 nên các trọng số sẽ không phù hợp với phương pháp lan truyền ngược. Để giải quyết vấn đề trên trong nghiên cứu này, chúng tôi sử dụng phương pháp ước lượng lan truyền tuyến tính cho lan truyền gradient:

$$q = \text{sign}(x), \qquad (2)$$

$$g_x = g_q 1_{|x|\leq 1}, \qquad (3)$$

Trong đó g_q đại diện cho giá trị độ dốc của q và g_x đại diện cho giá trị độ dốc của x. Nếu giá trị tuyệt đối của x nhỏ hơn hoặc bằng 1 thì g_x bằng g_q; ngược lại, g_x bằng 0.

Thuật toán lan truyền của chúng tôi dựa trên [24], vì vậy chúng tôi sẽ không đi sâu vào chi tiết ở đây.

3. TRIỂN KHAI MẠNG BCNN TRÊN CÁC BỘ DỮ LIỆU HUẤN LUYỆN

3.1 Bộ dữ liệu GTSRB

Đây là bộ dữ liệu bao gồm hơn 50.000 hình ảnh về biển báo giao thông. Các hình ảnh trong tập dữ liệu này có kích thước từ 15x15 pixel đến 222x193 pixel. Tỷ lệ giữa tập dữ liệu huấn luyện và tập dữ liệu đánh giá là 3:1, cụ thể là 12630 cho tập huấn luyện và 3909 hình ảnh cho tập dữ liệu đánh giá.

Hình 4.Các mẫu ngẫu nhiên trong bộ GTSRB

3.2 Bộ dữ liệu MNIST

Cơ sở dữ liệu MNIST (tiếng Anh: MNIST database, viết tắt từ Modified National Institute of Standards and Technology database) là một cơ sở dữ liệu lớn chứa các chữ số viết tay thường được sử dụng trong đào tạo các hệ thống xử lý ảnh cho các bức ảnh khác nhau. Bộ dữ liệu này có khoảng 60000 ảnh huấn luyện và 10000 ảnh đánh giá , bao gồm 10 loại số. Cơ sở dữ liệu này cũng được sử dụng rộng rãi để đào tạo và thử nghiệm trong lĩnh vực học máy.

Hình 5.Đồ thị của các mẫu ngẫu nhiên trong tập hợp MNIST

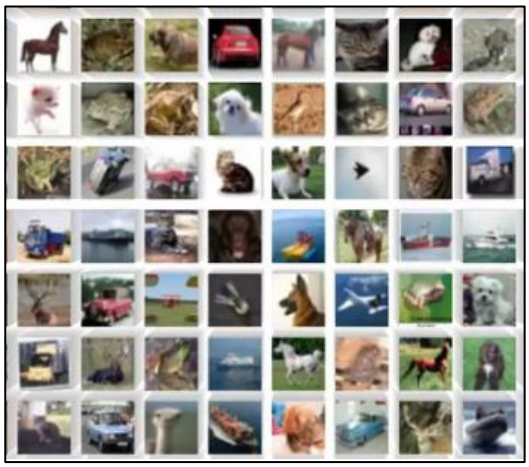
3.3 Bộ dữ liệu thời trang MNIST

Là bộ dữ liệu thời trang được lấy từ các bài viết của thương hiệu thời trang Zalando. Bộ dữ liệu này có khoảng 60000 hình ảnh đào tạo và 10000 hình ảnh đánh giá, bao gồm 10 loại mặt hàng thời trang và phụ kiện, cụ thể là áo phông, áo polo, áo sơ mi, quần, váy, áo khoác, xăng đan, giày thể thao, bốt ngắn và túi xách. Mỗi hình ảnh trong bộ Thời trang MNIST có kích thước 28x28, với mỗi pixel có giá trị từ 0 đến 255.

Hình 6.Các mẫu ngẫu nhiên trong bộ dữ liệu MNIST Fashion

3.4 Bộ dữ liệu CIFAR-10

Bộ dữ liệu bao gồm 60000 ảnh màu 32x32 chia thành 10 lớp, tương đương 6000 ảnh cho mỗi lớp. Tỷ lệ ảnh huấn luyện và đánh giá là 5:1, cụ thể là 50.000 ảnh huấn luyện và 10.000 ảnh đánh giá. Tập dữ liệu này được chia thành năm phiên đào tạo và một bài kiểm tra, mỗi phiên có 10.000 hình ảnh.



Hình 7.Ảnh các ảnh ngẫu nhiên trong bộ CIFAR-10

3.5 Bộ dữ liệu CNPB

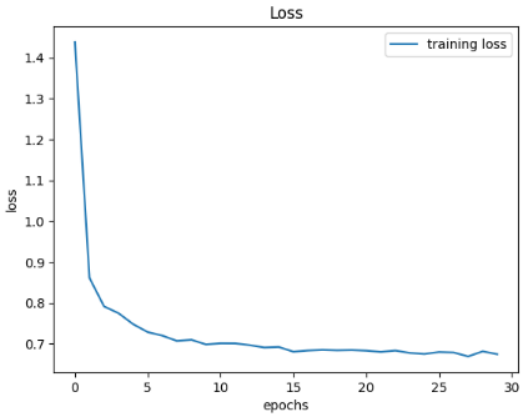
Bộ dữ liệu CNPB ((tiếng Anh: CNPB dataset, viết tắt từ Code Number on Product Boxes) do chúng tôi tạo ra. Nó bao gồm 20000 hình ảnh với kích thước 100 x 100 pixel. Trong đó có 16.000 hình ảnh để huấn luyện mô hình và 4000 hình ảnh để thử nghiệm mô hình. Trong bộ dữ liệu này chúng tôi có các mã là ' F06 ', ' K11 ', ' S19 ', ' V22 '. Các mã này có phần số tương ứng với số thứ tự của các dây chuyền sản xuất chuyên biệt và phần chữ có vị trí trong bảng 24 chữ cái tương ứng với số thứ tự của dây chuyền sản xuất.

Hình 8.Các mẫu trong bộ dữ liệu mã số thùng sản phẩm

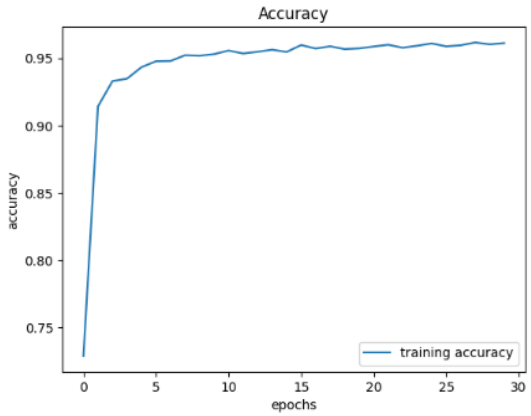
4. KẾT QUẢ THỰC NGHIỆM

4.1 Bộ dữ liệu GTSRB

Trình nhận dạng biển báo giao thông dựa trên BCNN được đánh giá bằng cách sử dụng bộ thử nghiệm GTSRB, bao gồm 12630 hình ảnh. Độ chính xác của mô hình là 96,01%. Chúng tôi cũng vẽ sơ đồ các thay đổi về độ chính xác của các bộ xác thực và đào tạo của mô hình trong quá trình đào tạo, như thể hiện trong hình 10.



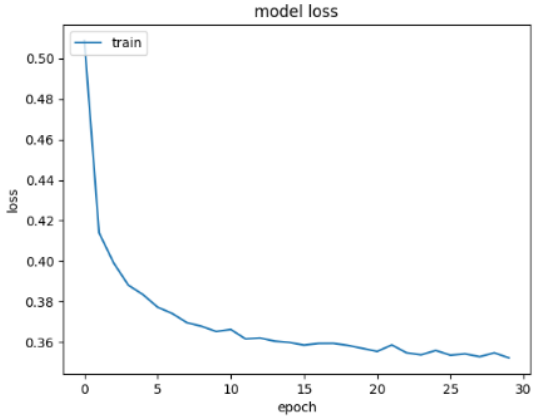
Hình 9. Độ mất mát của mô hình mạng nơron 1bit với tập GTSRB



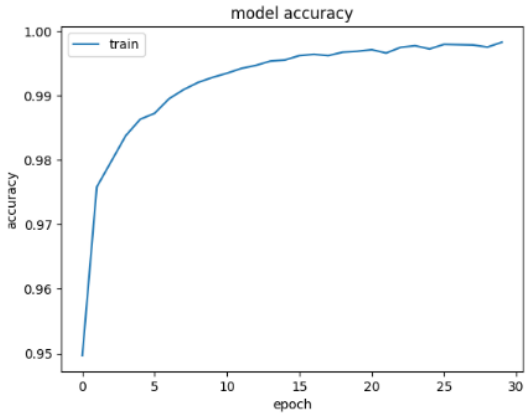
Hình 10. Độ chính xác mô hình mạng nơron 1bit với tập GTSRB

4.2 Tập dữ liệu MNIST

Trình nhận dạng chữ viết tay dựa trên BCNN được đánh giá bằng cách sử dụng bộ thử nghiệm GTSRB, bao gồm 12630 hình ảnh. Độ chính xác của mô hình là 99.81%. Chúng tôi cũng vẽ sơ đồ các thay đổi về độ chính xác của các bộ xác thực và đào tạo của mô hình trong quá trình đào tạo, như thể hiện trong hình 12.



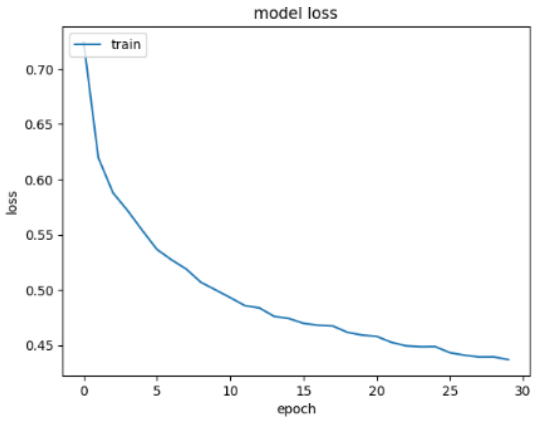
Hình 11.Độ mất mát của mô hình mạng nơroṇ 1bit với tập MNIST



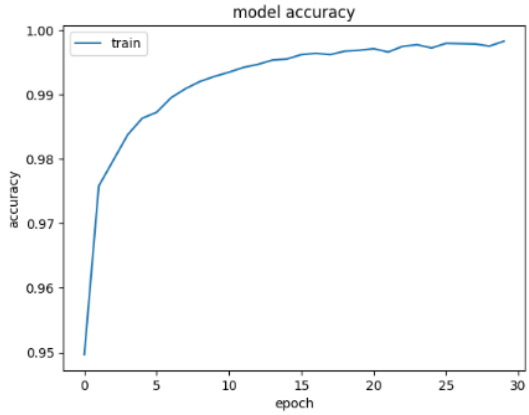
Hình 12. Độ chính xác mô hình mạng nơron 1bit với tập MNIST

4.3 Bộ dữ liệu MNIST- Fashion

Trình nhận dạng thời trang dựa trên BCNN được đánh giá bằng cách sử dụng bộ thử nghiệm Mnist fashion, bao gồm 70000 hình ảnh. Độ chính xác của mô hình là 96,35%. Chúng tôi cũng vẽ sơ đồ các thay đổi về độ chính xác của các bộ huấn luyện và xác thực của mô hình trong quá trình huấn luyện, như thể hiện trong hình 14.



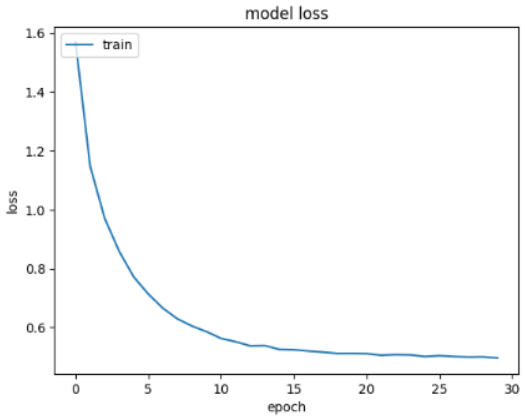
Hình 13. Độ mất mát của mô hình mạng nơroṇ 1bit với tập MNIST- Fashion



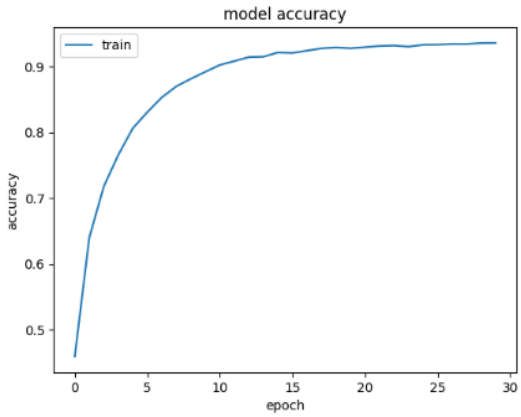
Hình 14. Độ chính xác mô hình mạng nơron 1bit với tập MNIST- Fashion

4.4 Bộ dữ liệu CIFAR-10

Trình nhận dạng hình ảnh trong tập dữ liệu Cifar-10 dựa trên B-MNN được đánh giá bằng cách sử dụng bộ kiểm tra GTSRB bao gồm 60000 hình ảnh. Độ chính xác của mô hình là 93,59%. Chúng tôi cũng vẽ sơ đồ các thay đổi về độ chính xác của các tập xác thực và huấn luyện của mô hình trong quá trình huấn luyện, như thể hiện trong hình 16.



Hình 15. Độ mất mát của mô hình mạng nơron 1bit với tập CIFAR-10

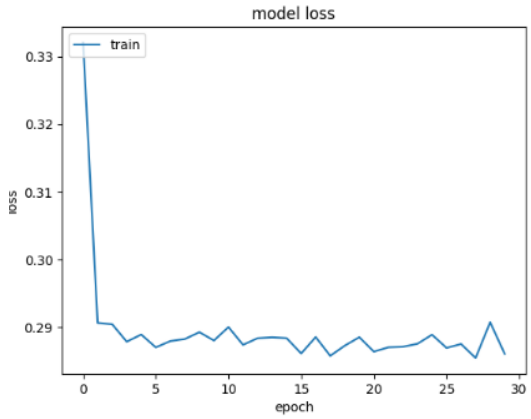


Hình 16. Độ chính xác mô hình mạng nơron 1bit với tập CIFAR-10

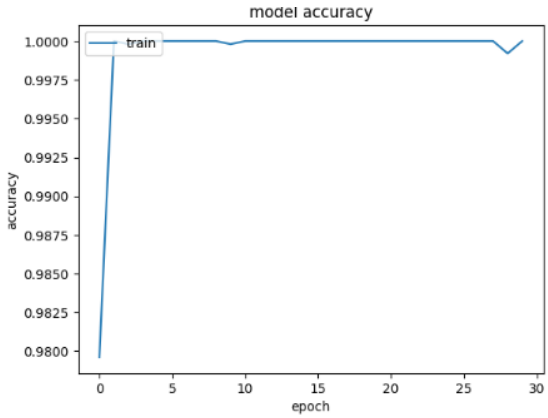
4.5 Bộ dữ liệu CNPB

Trình nhận dạng hình ảnh mã số mục trên hộp dựa trên BCNN được đánh giá bằng cách sử dụng bộ thử nghiệm CNPB, bao gồm 20000 hình ảnh. Độ chính xác của mô hình là 99,89%.

Để kiểm tra hiệu suất, chúng tôi đã triển khai mạng BCNN và CNN trên phần cứng Raspberry Pi Zero 2W với 500 MB RAM. Tốc độ xử lý hình ảnh áp dụng BCNN trung bình là 47,198 giây và trên mạng CNN, thông thường mạng cùng kích thước có tốc độ xử lý là 245,428 giây. Ta thấy tốc độ xử lý của mô hình BCNN nhanh gấp 5,2 lần so với CNN thông thường. Chúng tôi cũng vẽ sơ đồ các thay đổi về độ chính xác của các bộ huấn luyện và xác thực của mô hình trong quá trình huấn luyện, như thể hiện trong hình 18.



Hình 17. Độ mất mát của mô hình mạng nơron 1bit với tập CNPB.



Hình 18. Độ chính xác mô hình mạng nơron 1bit với tập CNPB.

5 KẾT LUẬN

Các thử nghiệm trên Raspberry Pi với tập dữ liệu CNPB chứng minh phương pháp đề xuất hiệu quả hơn mô hình mạng nơ-ron thông thường. Khi áp dụng thực tế, phương pháp mới BCNN giảm đáng kể dung lượng và tăng tốc độ xử lí so với mạng CNN, cụ thể dung lượng lưu trữ giảm 6,84 lần và tốc độ nhanh hơn 5,2 lần.

Trong tương lai, đề tài sẽ huấn luyện với dữ liệu lớn hơn, triển khai trên nhiều thiết bị và server để quản lý tốt hơn. Đề tài cũng có thể phát triển ứng dụng di động để quản lý trên nhiều thiết bị, đồng thời đưa hệ thống lên web để liên kết vào mạng IOT cho các dự án lớn.

TÀI LIỆU THAM KHẢO

[1] Son Truong Ngoc"Low cost artificial neural network model for Raspberry Pi", *Engineering, Technology & Applied Science Research*, Vol. 10, Issue 2, 2020, 5466-5469, 2020

[2] A. Krizhevsky, I. Sutskever, GE Hinton, "Imagenet Classification with Deep Convolutional Neural Networks", *Advances in Neural Information Processing Systems*, Lake Tahoe, USA, December 3-8, 2012.

[3] BMZahran, "Using neural networks to predict the hardness of aluminum alloys", *Engineering, Technology & Applied Science Research*, Vol. 5, No. 1, pp. 757-759, 2015

[4] GS Fesghandis, A. Pooya, M. Kazemi, ZN Azimi, "Comparison of perceptual-based and multilayer afferent-based functional neural networks in predicting the success of new product development", *Research in Engineering Science, Technology & Applications*, Vol. 7, No. 1, pp. 1425-1428, 2015

[5] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, IEEE, 2016

[6] Y. Choukroun, E. Kravchik, F. Yang and P. Kisilev, "Low-bit Quantization of Neural Networks for Efficient Inference," *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3009-3018, 2019

[7] R. Timofte, VA Prisacariu, LV Gool and I. Reid, "Combining traffic sign detection with 3d tracking to better assist drivers," *Emerging topics in computer vision and other topics its application*, pp. 425–446, World Science, 2012

[8] SP Rajendran, L. Shine, R. Pradeep and S. Vijayaraghavan, "Real-time Traffic Sign Recognition Using yolov3-based detectors", *2019 10th International Conference on Technology Computer, Communication and Networking Technology (ICCCNT)*, pp. 1–7, IEEE, 2019.

[9] M. Coubarariaux, Y. Bengio, and J.-P. David, "Binary Connections: Training Deep Neural Networks with Binary Weights During Propagation," *in Advances in Neural Information Processing Systems*, pp. 3123–3131, 2015

[10] Sajad Darabi, Mouloud Belbahri, Matthieu Courbariaux, Vahid Partovi Nia, "Regular Binary Network Training", available at: https://doi.org/10.48550/arXiv.1812.11800, 2020

[11] Zhou, Shuchang & Ni, Zekun & Zhou, Xinyu & Wen, He & Wu, Yuxin & Zou, Yuheng, "DoReFa-Net: Training low bitband convolutional neural networks with low bandwidth gradients", 2016

[12] M. Courbaraux, I. Hubara, D. Soudry, R. El-Yaniv, Y. Bengio, "BinaryNet: Training Deep Neural Networks with +1 or −1 bound activations and weights", available at: https://arxiv.org/abs/1602.02830, 2016

[13] P. Sermanet and Y. LeCun, "Traffic Sign Recognition with Multi-Scale Convolutional Networks." in IJCNN, pp. 2809–2813, 2011. Y. Yuan, Z. Xiong and Q. Wang, "Incremental Framework for Video-Based Traffic Sign Detection, Tracking, and Recognition," *IEEE Transaction on Intelligent Transportation Systems*, vol. 18, no. 7, pp. 1918–1929, 2016

[14] L. Deng, P. Jiao, J. Pei, Z. Wu, G. Li "GXNOR-Net: Training deep neural networks with weights and cubic activations without full precision memory in the die unified discriminant framework", *Neural Networks*, Vol. 100, pages 49-58, 2018

[15] H. Luo, Y. Yang, B. Tong, F. Wu and B. Fan, "Traffic Sign Recognition Using Multitasking Convolutional Neural Networks," *IEEE Transactions on Traffic Systems Smart, practice. 19*, 4, pp. 1100-1111, 2017.

[16] Y. Akhauri, "HadaNets: Flexible Quantization Strategies for Neural Networks," *in IEEE/CVF Computer Vision and Pattern Recognition Workshop (CVPRW), Long Beach, CA, USA* , pp. 526-534, 2019.

[17] Y. Akhauri, "HadaNets: Flexible Quantization Strategies for Neural Networks," *IEEE/CVF 2019 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW), 2019*, pp. 526-534, XNOR-Net++: Advanced Binary Neural Networks, Adrian Bulat, Georgios Tzimiropoulos, 2019.

[18] R. Girshick, "Fast R-cnn," *in Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and AC Berg, "Ssd: Single shot multibox detector," *in European conference on computer vision*, pp. 21–37, Springer, 2016.

[20] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-cnn: Towards Real-Time Object Detection with Area Recommendation Networks," *in Advances in Processing Systems neuroinformation processing*, pp. 91–99, 2015 .

[21] Y. Umuroglu, NJ Fraser, G. Gambardella, M. Blott, P. Leong, M. Jahre and K. Vissers, "Finn: A Framework for Scalable, Fast Binary Neural Network Inference," *in Proceedings of ACM/2017 SIGDA International Symposium on Field Programmable Gate Arrays*, pp. 65–74, ACM, 2017.

[22] C. Ma, Y. Guo, Y. Lei and W. An, "Binary Volumetric Convolutional Neural Networks for 3-D Object Recognition," *IEEE Transactions on Instrumentation and Measurement, vol. 68,* 1, pp. 38–48, 2018.

[23] X. Song, H. You, S. Zhou and W. Xie, "Traffic Sign Recognition with Binary Multi-Scale Neural Networks," *the Association's 35th Youth Academic Annual Meeting 2020 China automation (YAC), Zhanjiang, China*, pp. 116-121, doi: 10.1109/YAC51587.2020.9337571, 2020

[24] H. Qin et al., "Forward and backward information retention for precise binary neural networks," *IEEE/CVF 2020 Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2247-2256, 2020.

[25] M. Coubarariaux, Y. Bengio, and J.-P. David, "Binary Connections: Training Deep Neural Networks with Binary Weights During Propagation," *in Advances in Neural Information Processing Systems*, pp. 3123–3131, 2015.

[26] J. Chen, L. Liu, Y. Liu and X. Zeng, "Learning Framework for n-Bit Quantum Neural Networks Towards FPGAs," *in IEEE Transactions on Neural Networks and Systems study, practice. 32*, 3, pp. 1067-1081, March 2021.

[27] K. Hwang, W. Sung, "Designing a fixed-point relay deep neural network using +1, 0 and −1 weights", *in IEEE Workshop on Signal Processing Systems, Belfast, Kingdom England*, October 20–22, 2014.

Tác giả chịu trách nhiệm bài viết:

Họ tên: Nguyễn Kim Thạch

Đơn vị: Khoa Điện-Điện tử, Trường Đại học Sư phạm Kỹ thuật TP.HCM

Điện thoại: 0966-579-617

Email: nguyenkimthach.a@ gmail.com