COMPACT VERSION OF EFFICIENTNET FOR MOBILE DEVICES

MỘT PHIÊN BẢN NHỎ GỌN CỦA MÔ HÌNH EFFICIENTNET DÀNH CHO THIẾT BỊ DI ĐỘNG

Hoàng Văn Thành
Trường Đại học Quảng Bình

ABSTRACT: *EfficientNet is a convolutional neural network architecture generated by performing a neural architecture search using the AutoML MNAS framework, which optimizes both accuracy and efficiency. It is based on the inverted bottleneck residual blocks of MobileNetV2, in addition to squeeze-and-excitation blocks. On the ImageNet challenge, with a much fewer parameter calculation load, EfficientNet could take its place among the state-of-the-art. This paper introduces a compact version of EfficientNet for mobile devices which has similar accuracy on ImageNet dataset but can run approximately 2 times faster.*

Keywords: *CNN, EfficientNet, Mobile device.*

TÓM TẮT: *EfficientNet là một kiến trúc mạng nơ-ron tích chập được tạo ra bằng cách thực hiện tìm kiếm kiến trúc nơ-ron sử dụng nền tảng AutoML MNAS, giúp tối ưu hóa cả độ chính xác và hiệu quả. Nó dựa trên kiến trúc khối nút cổ chai đảo ngược kèm cấu trúc phần dư của MobileNetV2, bên cạnh các khối nén và kích thích. Khi áp dụng vào bài toán phân loại ảnh trên tập dữ liệu ImageNet, với tải tính toán tham số ít hơn nhiều, EfficientNet đạt được một vị trí cao trong các mô hình hiệu quả nhất. Bài viết này giới thiệu một phiên bản nhỏ gọn của EfficientNet dành cho các thiết bị di động. Nó có độ chính xác trên bộ dữ liệu ImageNet tương tự như phiên bản gốc nhưng có thể chạy nhanh hơn khoảng 2 lần.*

Từ khóa: *CNN, EfficientNet, thiết bị di động.*

1. GIỚI THIỆU

Mạng nơ-ron tích chập sâu (Deep convolutional neural network - CNN) đã cho thấy hiệu suất đáng kể trong nhiều tác vụ thị giác máy tính trong những năm gần đây. Xu hướng chính để giải quyết các nhiệm vụ chính là xây dựng các CNN sâu hơn và lớn hơn [2, 12]. Các mô hình CNN chính xác nhất thường có hàng trăm lớp và hàng nghìn kênh [2, 5, 13, 16]. Nhiều ứng dụng trong thế giới thực cần được thực hiện trong thời gian thực và/hoặc trên các thiết bị di động có tài nguyên hạn chế. Do đó, mô hình phải gọn nhẹ và chi phí tính toán thấp. Việc giảm kích thước mô hình CNN đang là sự đánh đổi giữa hiệu quả và độ chính xác.

Gần đây, nhiều công trình nghiên cứu tập trung vào lĩnh vực thiết kế kiến trúc hiệu quả. Lấy cảm hứng từ kiến trúc được đề xuất trong [7], mô-đun Inception được đề xuất trong GoogLeNet[12] để xây dựng các mạng sâu hơn mà không làm tăng kích thước mô hình và chi phí tính toán. Sau đó, nó được cải thiện hơn nữa trong [13] thông qua tích chập nhân tử. Tích chập có thể phân tách theo chiều sâu (Depthwise

Separable Convolution) đã khái quát hóa ý tưởng phân tích thừa số và thực hiện phân tách phép tích chập tiêu chuẩn thành một phép tích chập theo chiều sâu (depthwise convolution) và theo sau là một phép tích chập 1x1 theo từng điểm. EfficientNets [15], một nhóm các mạng nơ-ron tích chập được chia theo một tỷ lệ hiệu quả, gần đây đã nổi lên như một trong các mô hình hiệu quả nhất cho các tác vụ phân loại hình ảnh. EfficientNets tối ưu hóa độ chính xác cũng như hiệu quả bằng cách giảm kích thước mô hình và các phép toán dấu phẩy động được thực hiện trong khi vẫn duy trì tính hiệu quả của mô hình. Kiến trúc mạng cơ bản của mô hình được thiết kế bằng cách sử dụng tìm kiếm kiến trúc thần kinh (neural architecture search - NAS). Bằng cách mở rộng mô hình cơ bản của EfficientNets, các tác giả đã có được 1 sê-ri các biến thể của mô hình EfficientNets. Loạt mô hình này đã đánh bại tất cả các mô hình mạng thần kinh tích chập trước đó về hiệu quả và độ chính xác. Cụ thể, EfficientNet-B7 đạt được độ chính xác top 1 là 84,4% và top 5 chính xác là 97,1% trên bộ dữ liệu ImageNet. Và so với các mô hình khác có độ chính xác cao nhất vào thời điểm đó, kích thước đã giảm 8,4 lần và tính hiệu quả tăng 6,1 lần. Thông qua kĩ thuật học chuyển giao (transfer learning), EfficientNets đã đạt được những kết quả tốt trên nhiều bộ dữ liệu nổi tiếng.

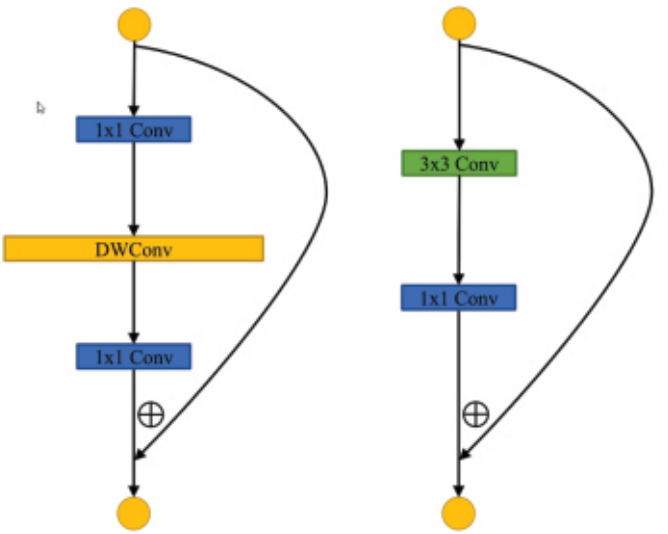
Bài báo này nghiên cứu mô hình EfficientNet-B0, mô hình mạng cơ bản trong chuỗi EfficientNets. Cấu trúc cốt lõi của mạng là mô-đun tích chập nút cổ chai đảo ngược di động (Mobile Inverted Bottleneck Convolution - MBConv), mô-đun này cũng đưa ra ý tưởng chú ý về mạng nén và kích thích (Squeeze-and-Excitation Network- SENet) [4]. Thông qua nghiên cứu này, chúng tôi đã tạo ra một biến thể nhẹ hơn của EfficientNet-B0 để làm cho nó thân thiện hơn với các thiết bị di động.

2. CÁC NGHIÊN CỨU LIÊN QUAN

2.1. Kiến trúc nhỏ gọn cho mô hình CNNs

Gần đây, có nhiều nghiên cứu tập trung vào phương pháp thiết kế kiến trúc nhỏ gọn cho các mô hình CNN [3, 5, 11, 17, 18]. Họ đã đề xuất các mô hình CNN hiệu quả có thể được đào tạo từ đầu đến cuối.

MobileNetV1 [3] đã giới thiệu phép *tích chập có thể phân tách theo chiều sâu* (Depthwise Separable Convolution) như một giải pháp thay thế hiệu quả cho các phép tích chập truyền thống. Phép tích chập có thể phân tách theo chiều sâu, có thể tương đương với phép tích chập truyền thống bằng cách tách tích chập theo không gian ra khỏi quá trình diễn sinh bản đồ đặc trưng. Nó bao gồm hai lớp riêng biệt. Lớp đầu tiên sử dụng toán tử tích chập theo chiều sâu với kích thước rất nhẹ. Nó áp dụng một bộ lọc tích chập duy nhất cho mỗi kênh đầu vào để nắm bắt thông tin không gian trong mỗi kênh. Sau đó, lớp thứ hai sử dụng tích chập theo chiều điểm, nghĩa là tích chập 1x1 có kích thước lớn hơn, để nắm bắt thông tin giữa các kênh nhằm tạo ra được bản đồ đặc trưng hiệu quả.



a) b)

Hình 1. Kiến trúc của các khối được sử dụng trong thực nghiệm của nghiên cứu. a) Khối tích chập nút cổ chai đảo ngược trên thiết bị di động không có mô-đun SE (MBConv) sử dụng trong Efficient-B0-lite. b) Khối tích chập với phép tích chập 3 x 3, và 1 x 1 (sau đây sẽ gọi là khối Conv31) sử dụng trong EfficientNet-B0-compact được đề xuất. Quá trình lấy mẫu giảm độ phân giải (downsampling) sẽ nằm trong phép DWConv hoặc phép tích chập 3 x 3. Kết nối phần dư sẽ không tồn tại nếu kích thước của đầu vào và đầu ra khác nhau. **DWCon**: Lớp tích chập theo chiều sâu. **Conv**: Lớp tích chập.

MobileNetV2 [11] đã giới thiệu kiểu thiết kế nút cổ chai đảo ngược kèm cấu trúc phần dư để tạo ra các cấu trúc lớp hiệu quả hơn nữa. Cấu trúc này được thể hiện trong Hình 1.a). Nó bao gồm một phép tích chập mở rộng 1 x 1, theo sau đó là phép tích chập theo chiều sâu và lớp phép chiếu 1 x 1. Đầu vào và đầu ra được kết nối bằng một kết nối phần dư khi và chỉ khi chúng có cùng số lượng kênh và độ phân giải. Cấu trúc này duy trì một biểu diễn nhỏ gọn ở đầu vào và đầu ra trong khi mở rộng sang không gian đặc trưng có nhiều chiều hơn bên trong để tăng tính đặc trưng của các phép biến đổi phi tuyến tính trên mỗi kênh.

EfficientNets [15], một nhóm các mạng nơ-ron tích chập được mở rộng theo một tỉ lệ hiệu quả, gần đây đã nổi lên như một mô hình tiên tiến nhất cho các tác vụ phân loại hình ảnh. EfficientNets tối ưu hóa độ chính xác cũng như hiệu quả bằng cách giảm kích thước mô hình và các phép toán dấu phẩy động được thực hiện trong khi vẫn duy trì chất lượng mô hình.

Kiến trúc của EfficientNets được thiết kế bằng cách thực hiện tìm kiếm kiến trúc mạng nơ-ron, một kỹ thuật để tự động thiết kế mạng nơ-ron. Nó tối ưu hóa cả độ chính xác và hiệu quả như được đo trên cơ sở phép toán dấu phẩy động mỗi giây (FLoating-point Operations Per Second - FLOPS). Kiến trúc này sử dụng khối tích chập nút cổ chai đảo ngược trên thiết bị di động (Mobile Inverted Bottleneck Convolution -

MBConv). Sau đó, các nhà nghiên cứu đã mở rộng mạng cơ sở này để có được một họ các mô hình học sâu, được gọi là chung EfficientNets. Dựa vào đó, các tác giả cũng đề xuất họ mô hình EfficientNet-lite với khác biệt là không có mô-đun nén và kích thích (SE). Kiến trúc của EfficientNet-B0-lite, mô hình mạng cơ bản trong chuỗi EfficientNets-lite, được đưa ra trong Bảng 1.

Mức độ hiệu quả là vấn đề quan trọng nhất đối với các mô hình tích chập cho thiết bị di động. Các kiến trúc tích chập hiệu quả đã được nghiên cứu rộng rãi trong dòng MobileNet [3, 11]. Phép tích chập có thể phân tách theo chiều sâu và khối tích chập nút cổ chai đảo ngược là những ý tưởng cốt lõi cho một mạng có kích thước di động hiệu quả. MnasNet [14] và EfficientNets [15] tiến thêm một bước để đưa ra khối MBConv dựa trên nút cổ chai đảo ngược trên thiết bị di động trong [11]. EfficientNet cho thấy rằng các mô hình có sử dụng khối MBConv không chỉ đạt được kết quả tốt trên bộ dữ liệu ImageNet mà còn rất hiệu quả.

2.2. Khối tích chập nút cổ chai đảo ngược

Khối tích chập nút cổ chai đảo ngược thu được thông qua kĩ thuật tự động tìm kiếm kiến trúc mạng nơ-ron. Cấu trúc mô-đun tương tự như tích chập có thể phân tách theo chiều sâu, như có thể thấy trong Hình 1a.

Trước tiên, khối tích chập nút cổ chai đảo ngược thực hiện phép tích chập 1 x 1 từng điểm trên đầu vào và dựa trên hệ số mở rộng (expand ratio), thay đổi kích thước của kênh đầu ra (ví dụ: khi hệ số mở rộng là 3, kích thước của kênh sẽ tăng lên 3 lần). Sau đó tiến hành phép tích chập theo chiều sâu với kích thước nhân k x k Sau đó, khôi phục kích thước kênh ban đầu khi kết thúc bằng phép tích chập từng điểm 1 x 1. Cuối cùng, quá trình kết nối phần dư với đầu vào được thực hiện nếu cần thiết.

Tầng i	Phép toán $\hat{\mathcal{F}}_i$	Độ phân giải $\hat{H}_i$ x $\hat{W}_i$	Số kênh $\hat{C}_i$	Số lặp $\hat{L}_i$
1	Conv3x3	224 x 224	32	1
2	MBConv1, k3x3	112x112	16	1
3	MBConv6, k3x3	112x112	24	2
4	MBConv6, k5x5	56 x 56	40	2
5	MBConv6, k3x3	28 x 28	80	3
6	MBConv6, k5x5	14 x 14	112	3
7	MBConv6, k5x5	14 x 14	192	4
8	MBConv6, k3x3	7 x 7	320	1
9	Conv1x1 & Pooling	7 x 7	1280	1
10	FC	1 x 1	1000	1

Bảng 1. Kiến trúc tổng thể của mạng EfficientNet-B0-lite. **MBConv**: Khối tích chập cổ chai đảo ngược trên thiết bị di động, số theo sau là hệ số mở rộng.

Tầng i	Phép toán $\hat{\mathcal{F}}_i$	Độ phân giải $\hat{H}_i$ x $\hat{W}_i$	Số kênh $\hat{C}_i$	Số lặp $\hat{L}_i$
1	Conv3x3	224 x 224	32	1
2	MBConv1, k3x3	112x112	24	1
3	MBConv6, k3x3	112x112	24	1
4	MBConv6, k5x5	56 x 56	40	2
5	MBConv6, k3x3	28 x 28	80	3
6	MBConv6, k5x5	14 x 14	112	3
7	MBConv6, k5x5	14 x 14	192	4
8	MBConv6, k3x3	7 x 7	320	1
9	Conv1x1 & Pooling	7 x 7	1280	1
10	FC	1 x 1	1000	1

Bảng 2. Kiến trúc tổng thể mô hình EfficientNet-B0-compact được đề xuất. **Conv31**: Khối tích chập có lớp chập 3 x 3 và lớp chập 1 x 1. **MBConv**: Khối Tích chập nút cổ chai đảo ngược trên thiết bị di động, số theo sau là hệ số mở rộng.

3. KIẾN TRÚC MÔ HÌNH

3.1. Kiến trúc mô hình EfficientNet-B0-lite

EfficientNet-B0 là mô hình mạng cơ bản trong chuỗi EfficientNets. Cấu trúc cốt lõi của mạng là mô-đun tích chập nút cổ chai đảo ngược di động (MBConv), mô-đun này cũng đưa ra ý tưởng về sự chú ý dựa vào mô-đun nén và kích thích (SE) [4] Các tác giả cũng đã giới thiệu một phiên bản rút gọn của nó, phù hợp hơn cho các thiết bị di động/IoT. Họ chủ yếu thực hiện các thay đổi sau dựa trên mô hình EfficientNet-B0 ban đầu: 1) Loại bỏ mô-đun nén và kích thích (SE), do mô-đun SE không được hỗ trợ tốt cho các thiết bị di động; và 2) Thay thế tất cả hàm kích hoạt (activation) swish bằng hàm kích hoạt RELU6 để nhằm hậu lượng tử hóa dễ dàng hơn. Kiến trúc của nó được thể hiện trong Bảng 1. Tất cả các lớp được theo sau bởi một lớp định mức hàng loạt (batch normalization) và kích hoạt phi tuyến tính ReLU ngoại trừ lớp được kết nối đầy đủ (fully connected) cuối cùng, không có phi tuyến tính và được đưa qua một lớp softmax để phân loại. Việc lấy mẫu xuống được xử lý với phép tích chập theo chiều sâu của khối đầu tiên trong từng tầng, cũng như trong lớp đầu tiên. Lớp tổng hợp trung bình (global average pooling) cuối cùng làm giảm độ phân giải không gian xuống 1 trước lớp được kết nối đầy đủ.

3.2. Kiến trúc mô hình đề xuất EfficientNet-B0-compact

Chúng tôi sẽ giới thiệu một phiên bản mới có tên là EfficientNet-B0-compact để có một mô hình thân thiện hơn dựa trên EfficientNet-B0-lite. Mô hình được đề xuất có một số khác biệt nhỏ ở giai đoạn đầu so với EfficientNet-B0-lite. Đó là:

Các cài đặt khác là: phân rã trọng lượng là 0,0001, động lượng là 0,9 và kích thước lô là 2048.

Tất cả các mạng đã được huấn luyện trên dịch vụ Google Colab với môi trường TPU. Tốc độ của tất cả các model được đánh giá trên laptop có CPU, RAM 8GB không có thiết bị GPU khi chạy mô hình với kích thước lô là 8.

4.3. Đánh giá hiệu năng

Bảng 3 thể hiện so sánh về độ chính xác Top 1 và Top 5 của các mô hình EfficientNet-B0, EfficientNet-B0-lite và EfficientNet-B0-compact trên bộ dữ liệu ImageNet.

Dễ dàng nhận thấy rằng mặc dù phiên bản lite và compact có sai số cao hơn một chút so với phiên bản gốc EfficientNet-B0 nhưng lại có tốc độ chạy nhanh gấp 2 lần. Hơn nữa, phiên bản compact có thể đạt được hiệu suất tương đương với phiên bản rút gọn trong khi chạy nhanh hơn 15%.

5. KẾT LUẬN

Bài báo này đã nghiên cứu kiến trúc của mô hình EfficientNet-B0. Trên cơ sở đó đề xuất mô hình mới EfficientNet-B0-compact có độ chính xác tương đương nhưng có tốc độ nhanh hơn so với mô hình EfficientNet-B0 gốc, và mô hình EfficientNet-B0-lite. Mô hình mới được xây dựng dựa trên EfficientNet-B0-lite với các thay đổi ở giai đoạn đầu. Đó là: 1) Xóa khối MBConv1 đầu tiên có độ phân giải 112 x 112; 2) Thay đổi khối MBConv thứ hai ở tầng có độ phân giải 112 x 112 thành khối Conv31. Kết quả thực nghiệm trên cơ sở dữ liệu ImageNet cho thấy rằng mô hình mới EfficientNet-B0-compact có độ chính xác tương đương nhưng có tốc độ nhanh hơn 2 lần so với mô hình EfficientNet-B0 gốc, cũng như nhanh hơn 15% so với mô hình EfficientNet-B0-lite.

TÀI LIỆU THAM KHẢO

[1] Glorot, X. and Bengio, Y. (2010), Understanding the Difficulty of Training Deep Feedforward Neural Networks, *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 249-256.

[2] He, K., Zhang, X., Ren, S. and Sun, J. (2016), Deep Residual Learning for Image Recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770-778.

[3] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H. (2017), MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

[4] Hu, J., Shen, L. and Sun, G. (2018), Squeeze-and-excitation networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132-7141.

[5] Huang, G., Liu, Z., Maaten, L.V.D. and Weinberger, K.Q. (2017), Densely Connected Convolutional Network, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700-4708.

[6] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W. and Jackel, L.D. (1989), Backpropagation Applied to Handwritten Zip Code Recognition, *Neural Computation*. 1, 4, 541-551.

[7] Lin, M., Chen, Q. and Yan, S. (2014), Network In Network. *Proceedings of the International Conference on Learning Representations*.

[8] Nesterov, Y.E. (1983), A Method for Solving the Convex Programming Problem with Convergence Rate O (1/k 2), *Dokl. Akad. Nauk SSSR*, 543-547.

[9] Robbins, H. and Monro, S. (1951), A Stochastic Approximation Method, *The annals of mathematical statistics*, 400-407.

[10] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C. and Fei-Fei, L. (2015), ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision*. 115, 3, 211-252.

[11] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.-C. (2018), MobileNetV2: Inverted Residuals and Linear Bottlenecks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510-4520.

[12] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015), Going Deeper with Convolutions, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1-9.

[13] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016), Rethinking the Inception Architecture for Computer Vision, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 2818-2826.

[14] Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A. and Le, Q.V. (2019), MnasNet: Platform-Aware Neural Architecture Search for Mobile, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2820-2828.

[15] Tan, M. and Le, Q. (2019, Efficientnet: Rethinking model scaling for convolutional neural networks, *Proceedings of the International Conference on Machine Learning*, 6105-6114.

[16] Zagoruyko, S. and Komodakis, N. (2016), Wide Residual Networks. *Proceedings of the British Machine Vision Conference*, 87.1-87.12.

[17] Zhang, X., Zhou, X., Lin, M. and Sun, J. (2018), ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6848-6856.

[18] Zoph, B., Vasudevan, V., Shlens, J. and Le, Q.V. (2018), Learning Transferable Architectures for Scalable Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8697-8710.

Liên hệ:
TS. Hoàng Văn Thành
Khoa Kỹ thuật - Công nghệ thông tin, Trường Đại học Quảng Bình
Địa chỉ: 312 Lý Thường Kiệt, Đồng Hới, Quảng Bình
Email: thanhhv@qbu.edu.vn

Ngày nhận bài: 02/01/2023
Ngày gửi phản biện: 05/01/2023
Ngày duyệt đăng: 25/02/2023