Đánh Giá Model Phân Loại Hình Ảnh

Phan Đức Huy

Tháng 9 năm 2024

1 Giới thiệu

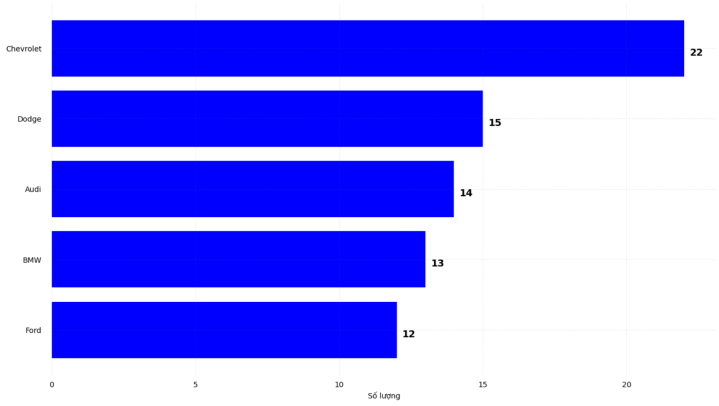
Từ những năm 90, mạng Convolutional Neural Networks (CNN) được coi là giải pháp phổ biến và tối ưu nhất cho bài toán phân loại hình ảnh. Từ đó đến nay, đã có rất nhiều cải tiến về cấu trúc của mạng CNN giúp các model đọc được thông tin của ảnh chính xác hơn, giảm thiểu chi phí tính toán cũng như độ trễ của model. Trong đa số những bài toán phân loại ảnh hiện nay, các model CNN vẫn đưa ra hiệu suất tốt.

Trong vài năm gần đây, dựa theo bảng đánh giá các model trên trang web Papers With Code[1] về độ chính xác trên dataset ImageNet, các model Vision Transformer (ViT) đang thể hiện tốt hơn so với các model CNN truyền thống. Bài báo cáo này sẽ đánh giá các model CNN đã được cải tiến và các model ViT mới trên tập dữ liệu Stanford Cars[2] để phân loại các loại xe hơi.

2 Datasets

2.1 Thông tin về dữ liệu

Tập dữ liệu Stanford Cars được sử dụng bao gồm 16,185 hình ảnh bao gồm 196 loại xe khác nhau. Mỗi loại xe sẽ được phân chia thành các tệp nhỏ bao gồm 70-80 hình ảnh. Các loại xe sẽ được chia ra theo cấu trúc "Hãng sản xuất- Mẫu - Năm sản xuất"



Hình 1: 5 hãng xe có nhiều mẫu nhất

Được cung cấp bởi Stanford University AI Lab vào năm 2013[3], tập dữ liệu này phục vụ chính cho bài toán Fine-Grained Image Classification. Bài toán này đòi hỏi model phải xác định được những thay đổi vô cùng nhỏ về ngoại hình giữa các loại xe để có thể gắn nhãn chính xác.

Hình 2: Hình ảnh được chụp từ nhiều hướng khác nhau

2.2 Xử lý dữ liệu

Thư viện torchvision được sử dụng để xử lý dữ liệu ảnh. Sau khi dữ liệu ảnh được nạp vào dưới dạng PIL Image, ta sẽ thay đổi kích cỡ ảnh về dạng [224,224,3], xoay ảnh ngẫu nhiên sau đó chuyển chúng về dạng torch tensor với kích thước [3,224,224]. Các giá trị pixel của ảnh sẽ được chuyển về trong khoảng 0 đến 1.

Các tập train/val/test sẽ được chia theo tỉ lệ 70/15/15 và được nạp vào model với batch bằng 16. Với 11,000 ảnh để train thì các model CNN đã có thể cho ra kết quả với độ chính xác cao tuy nhiên các model áp dụng Transformers yêu cầu lượng dữ liệu lớn hơn vậy rất nhiều.

Hình 3: Hình ảnh trong một batch sau khi được xử lý

2.3 Môi trường đánh giá

Quá trình tải, xử lý và huấn luyện model đều được chạy trên Kaggle Notebook GPU P100 với 16GiB tối đa. Các package đều sử dụng phiên bản mới nhất được cung cấp bởi Kaggle.

3 Model

3.1 CNN-based Models

Các model CNN được cải tiến gần đây như MobileNetV3[4] hay EfficientNetV2[5] đều đưa ra kết quả khá ổn định cũng như được tối ưu để giảm thiểu chi phí và tiết kiệm thời gian training. Do giới hạn về GPU, EarlyStopping được sử dụng trong quá trình training với patience = 10. Tiêu chí được sử dụng trong bài đánh giá này là Top 1, Top 5 Accuracy và số lượng parameters và thời gian chạy mỗi epoch. Dưới đây là kết quả fine tuning các model này trên tập dữ liệu :

Model Name	Top 1 Acc	Top 5 Acc	Parameters	Runtime (per epoch)	Year
ResNet50	76.5%	95.3%	24M	1m15s	2015
MobileNetV3-S	82.5%	96.8%	3.5M	1m	2019
EfficientNetV2-S	90.4%	98.6%	21M	1m35s	2021
ConvNeXt-Tiny	87.7%	98%	28M	2m	2022
ConVNeXtV2-Tiny	86.5%	97.4%	28M	1m	2023
RDNet-T	86%	97%	23M	4m30s	2024

Bảng 1: Đánh giá các Model CNN

Bảng trên bao gồm kết quả tốt nhất các model đạt được khi train dữ liệu trên các model cỡ nhỏ đã được train sẵn trên tập dữ liệu 'IMAGENET-1k'. Model MobileNetV3 sở hữu lượng tham số vô cùng nhỏ và thời gian train mỗi epoch thấp nhưng vẫn đưa ra độ chính xác khá ổn nhờ cấu trúc Depthwise Convolution và khối MBConv[6] được giới thiệu ở bản tiền nhiệm.

Trong khi đó, model EfficientNetV2 không chỉ có độ chính xác vô cùng cao lên đến 90% mà lượng tham số và thời gian training cũng được tối ưu. Ngoài việc sử dụng NAS Search[7] để tìm ra kiến trúc model tối ưu nhất, điểm khác biệt của EfficientNetV2 nằm ở cơ chế Progressive Learning. Cụ thể EfficientNetV2 sẽ train trên ảnh có kích thước nhỏ với regularization yếu ở những step đầu, sau đó kích thước ảnh và mức độ regularization sẽ được tăng dần ở các bước kế. Điều này giúp model học các feature tốt hơn và giảm thiểu overfitting.

Model ConvNeXt được xây dựng vào năm 2022 với ý tưởng "hiện đại hoá" cấu trúc CNN truyền thống qua những thiết kế Vision Transformers được nghiên cứu trước đó. Thiết kế này giúp ConvNet có thể cạnh tranh với cái các model ViT trên tập dữ liệu lớn và phức tạp.

Vào tháng 3 năm 2024, bài báo nghiên cứu DenseNets Reloaded (RDNet)[8] được công bố về việc tối ưu hoá tiềm năng của cấu trúc DenseNets. Ở DenseNets, skip connection được hình thành khi nối đầu ra của các lớp lại với nhau thay vì cộng lại như ResNet[9]. Điều này dẫn đến những vấn đề về bộ nhớ, nhất là khi cấu trúc các mạng trở nên sâu hơn. Những thay đổi về cấu trúc mạng được đưa ra đã giải quyết các vấn đề này, giúp RDNet có thể cạnh tranh được với các model hiện đại khác mà vẫn giữ kết cấu cốt lõi từ DenseNets.

Nhìn chung các model CNN đưa ra kết quả khá tốt với số lượng ảnh nhỏ và thời gian training ngắn song vẫn gặp khó khăn trong bài toán Fine-grained Image Classification. Tuy vậy, hướng phát triển của các model CNN trong các năm gần đây phần lớn lại tập trung vào việc cải thiện hiệu suất, thu nhỏ model mà vẫn giữ được độ chính xác. Do đó, các model CNN sẽ hợp lý hơn cho các bài toán yêu cầu real-time hay độ trễ thấp.

3.2 ViT-based Models

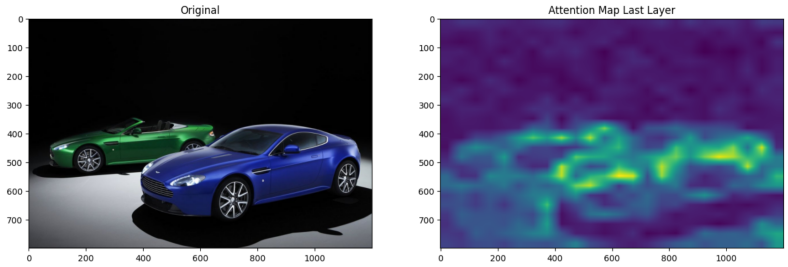
Vision Transformer (ViT) được giới thiệu vào năm 2021 và là giải pháp thay thế phổ biến để cạnh tranh với các model CNN. Các model ViT yêu cầu lượng dữ liệu vô cùng lớn để có thể học được mối quan hệ giữa các mảng của ảnh, tuy vậy rất nhiều pre-trained model đã được cung cấp giúp rút ngắn thời gian training. Sau đây là kết quả fine tuning các model ViT-based trên tập dữ liệu:

Model Name	Top 1 Acc	Top 5 Acc	Parameters	Runtime (per epoch)	Year
ViT-Small	70.2%	90.5%	22M	1m30s	2021
ViT-Base	83.9%	97.2%	86M	3m	2021
DeiT-Base	87.9%	97.9%	86M	3m	2021
SwinV2-Tiny	87.4%	98%	27M	2m25s	2022
EfficientViT_B3	88%	97.6%	21M	2m30s	2023

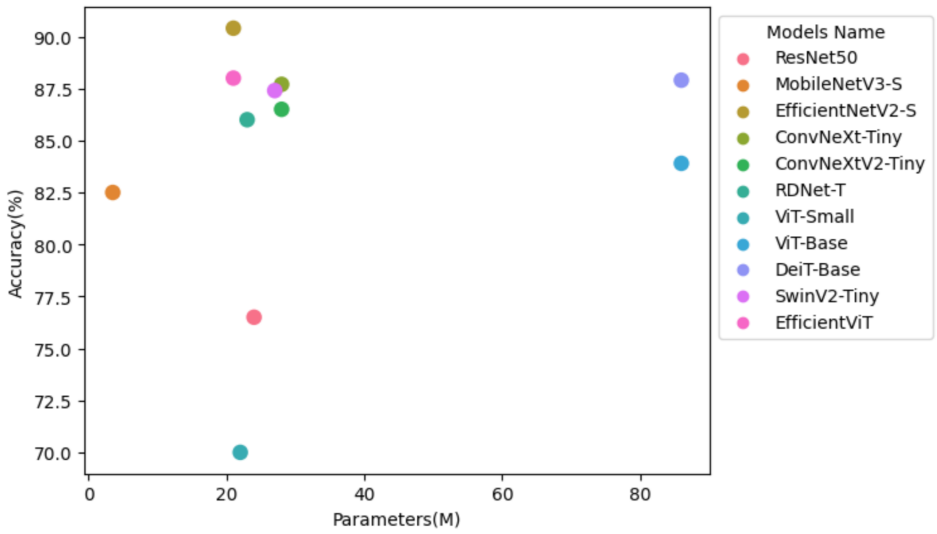
Bảng 2: Đánh giá các Model ViT

Nhìn qua tổng quan, độ chính xác của các model ViT cũng không có khác biết quá lớn so với các model CNN khi train trên tập dữ liệu. So với lượng thời gian chạy mỗi epoch phải đánh đổi, ta có thể thấy đối với tập dữ liệu chỉ hơn 8000 ảnh mà ta sở hữu, việc sử dụng ViT không mang lại hiệu quả quá cao.

Các nghiên cứu như SwinV2[10] hay EffcientViT[11] tập trung vào cải thiện nhược điểm của Transformers trên dữ liệu có độ phân giải lớn do khác với word token trong NLP, các phần tử visual có kích thước, độ phân giải khác nhau. Do một số hạn chế về bộ nhớ, hình ảnh đầu vào đã được chuyển về kích thước 224x224, phần nào không thể tận dụng hết khả năng của các model này.



Hình 4: Attention Map của model ViT-Base



Hình 5: Độ chính xác và kích cỡ các model được sử dụng

4 Kết luận

Qua quá trình training, model EfficientNetV2 đưa ra kết quả tốt nhất với 90,4% cùng với tốc độ train không quá cao. Chỉ số Top 5 Accuracy của đa số các model lên đến 97-98% cho thấy các model đang gặp khó khăn trong việc phân loại các mẫu xe có nhiều điểm tương đồng. Tuy vậy, vấn đề này có thể được cải thiện bằng việc gia tăng kích cỡ model, chất lượng ảnh đầu vào để nắm bắt kĩ hơn mối quan hệ giữa các pixel. Các model phân loại hình ảnh ra mắt gần đây đưa ra nhiều cải tiến trong cấu trúc của mạng truyền thống thống giúp tối ưu hoá chi phí và thời gian tính toán mà vẫn giữ được độ chính xác cao. Với những tài nguyên và dữ liệu ta có, sử dụng các model trong họ EfficientNetV2 sẽ là lựa chọn hợp lý nhất.

References

[1] Papers With Codes. *Image Classification on ImageNet*. URL: https://paperswithcode.com/sota/image-classification-on-imagenet.

[2] JessicaLi. *Stanford Cars Dataset*. URL: https://www.kaggle.com/datasets/jessicali9530/stanford-cars-dataset.

[3] Jonathan Krause Michael Stark Jia Deng Li Fei-Fei. "3D Object Representations for Fine-Grained Categorization". In: *2013 IEEE International Conference on Computer Vision Workshops* (2013).

[4] Andrew Howard et al. *Searching for MobileNetV3*. 2019. arXiv: 1905.02244 [cs.CV]. URL: https://arxiv.org/abs/1905.02244.

[5] Mingxing Tan and Quoc V. Le. "EfficientNetV2: Smaller Models and Faster Training". In: *CoRR* abs/2104.00298 (2021). arXiv: 2104.00298. URL: https://arxiv.org/abs/2104.00298.

[6] Mark Sandler et al. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. 2019. arXiv: 1801.04381 [cs.CV]. URL: https://arxiv.org/abs/1801.04381.

[7] Barret Zoph and Quoc V. Le. *Neural Architecture Search with Reinforcement Learning*. 2017. arXiv: 1611.01578 [cs.LG]. URL: https://arxiv.org/abs/1611.01578.

[8] Donghyun Kim, Byeongho Heo, and Dongyoon Han. *DenseNets Reloaded: Paradigm Shift Beyond ResNets and ViTs*. 2024. arXiv: 2403.19588 [cs.CV]. URL: https://arxiv.org/abs/2403.19588.

[9] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV]. URL: https://arxiv.org/abs/1512.03385.

[10] Ze Liu et al. *Swin Transformer V2: Scaling Up Capacity and Resolution*. 2022. arXiv: 2111.09883 [cs.CV]. URL: https://arxiv.org/abs/2111.09883.

[11] Xinyu Liu et al. *EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention*. 2023. arXiv: 2305.07027 [cs.CV]. URL: https://arxiv.org/abs/2305.07027.