Investigating Small Vision Encoders in Multimodal Transformers for Text-Based Person Search

CSE641 Computer Vision: Modern Methods And Application

Dhyey Patel
AU2240054

Malav Modi
AU2240208

Prem Patel
AU2240010

***Abstract*—Multimodal models like CLIP play a vital role in textual content-based totally individual seek, however their huge vision encoders incur high computational expenses. This study investigates replacing ResNet50 with MobileNet, EfficientNet, and Inception to balance performance and overall performance. We conduct experiments at the Flickr30k dataset and examine models based on education/validation loss and computational performance (education and validation time). Our findings highlight the alternate-offs among model complexity and performance, supplying insights into choosing light-weight encoders for green multimodal retrieval.**

***Index Terms*—CLIP model, MobileNet, EfficientNet, Inception, Text Encoder, Image Encoder, Cosine Similarity.**

I. Introduction

Text-based person search requires retrieving images based on natural language descriptions. Large-scale vision encoders such as ResNet50 are commonly used in models like CLIP, but their high computational demand limits deployment on resource-constrained devices.

Objective:

Replace ResNet50 with MobileNet, EfficientNet, and Inception within the CLIP model. Analyze training and validation loss to measure retrieval effectiveness. Evaluate computational efficiency by comparing training and validation time.

II. Related Work

Recent research have explored lightweight CNN architectures for efficient photo processing. MobileNet is known for its depthwise separable convolutions, EfficientNet optimizes scaling, and Inception complements multi-scale function extraction. However, their effect on text-to-picture retrieval stays underexplored. This look at bridges the gap by using evaluating these fashions in a multimodal context.

III. Methodology

A. Dataset Discussion

Dataset: Flickr30k, a huge-scale dataset containing 30,000 pix with more than one textual descriptions per picture. Metrics Used: Performance Metric: Training and validation loss. Efficiency Metric: Training and validation time.

B. Model Architecture

We modify the CLIP version's vision encoder by means of replacing ResNet50 with:

MobileNet: Optimized for cell devices, presenting minimum computation. EfficientNet: Uses compound scaling for better accuracy with decreased complexity. Inception: Extracts multi-scale functions via inception modules.

TABLE I
Hyperparameters Used in Training

Hyperparameter	Value
Debug Mode	False
Batch Size	16
Number of Workers	4
Head Learning Rate	1×10^{-3}
Image Encoder LR	1×10^{-4}
Text Encoder LR	1×10^{-5}
Weight Decay	1×10^{-3}
Patience	1
Factor	0.8
Epochs	2
Text Encoder Model	DistilBERT (base, uncased)
Text Tokenizer	DistilBERT (base, uncased)
Max Token Length	200
Pretrained Models	True
Trainable Models	True
Temperature Parameter	1.0
Image Size	224x224
Projection Layers	1
Projection Dimension	256
Dropout	0.1

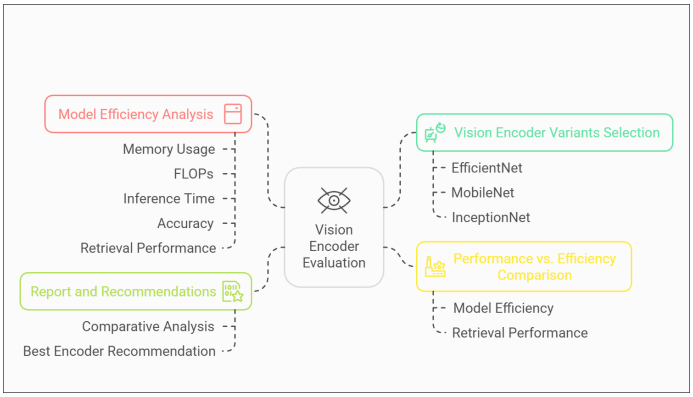

Fig. 1. Methodology

IV. Results and Discussion

Table III presents the training and validation loss along with computational efficiency metrics (training and validation time) for different vision encoders integrated into the CLIP model.

TABLE II
Training and Validation Performance of Different Vision Encoders

Model	Train Loss	Val Loss	Train Time (min)	Val Time (min)
ResNet50	1.99	2.37	36:05	3:09
MobileNet	2.3	2.48	**23:40**	**2:34**
EfficientNet	0.745	2.41	34:14	3:00
Inception	2.05	2.35	45:40	3:54

A. Observations

- **MobileNet** achieves the fastest training and validation time because of its lightweight architecture however indicates a slight increase in loss.
- **EfficientNet** offers a balanced exchange-off among schooling time and loss.
- **Inception** maintains reasonable accuracy but is computationally heavier than MobileNet.

These imply that lightweight encoders can optimize multimodal models for textual content-based individual search at the same time as lowering useful resource utilization. Future paintings will awareness on extending the evaluation to extra datasets and incorporating retrieval-specific accuracy metrics.

TABLE III
Retrieval Matrices

Model	Rank@1	Rank@5	Rank@10
MobileNet	0.0148	0.0726	0.1171
EfficientNet	0.0107	0.0493	0.0867

B. Observations

- MobileNet plays higher than EfficientNet across all ranks. It achieves better accuracy at Rank@1 (0.0148 vs. 0.0107), Rank@five (0.0726 vs. 0.0493), and Rank@10 (0.1171 vs. 0.0867).
- Efficiency vs. Accuracy Trade-off – MobileNet, being lightweight, seems to extract better retrieval features, while EfficientNet, notwithstanding its strong category potential, lags in retrieval.
- MobileNet is most popular for text-primarily based retrieval in this case, however EfficientNet may additionally improve with tuning.

Conclusion

This have a look at highlights the impact of light-weight imaginative and prescient encoders at the performance of multimodal retrieval. Our effects suggest that MobileNet drastically reduces computational overhead, while EfficientNet keeps strong performance. Future paintings includes increasing the evaluation to additional datasets and comparing retrieval accuracy using metrics like Mean Average Precision (mAP).

References

[1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021a, February 26). Learning transferable visual models from Natural Language Supervision. arXiv.org.

[2] Tan, M., Le, Q. V. (2020, September 11). EfficientNet: Rethinking model scaling for Convolutional Neural Networks. arXiv.org. https://arxiv.org/abs/1905.11946v5

[3] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C. (2019, March 21). MobileNetV2: Inverted residuals and linear bottlenecks. arXiv.org. https://arxiv.org/abs/1801.04381

[4] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2014, September 17). Going deeper with convolutions. arXiv.org. https://arxiv.org/abs/1409.4842