Investigate Small Vision Encoders in Multimodal Transformers

CSE641: Computer Vision: Modern Methods And Application

Dhyey Patel
AU2240054

Malav Modi
AU2240214

Prem Patel
AU2240010

***Abstract*—Multimodal transformers like CLIP have demonstrated high-quality overall performance in text-primarily based person search by way of efficaciously aligning textual descriptions with visual functions. However, the huge and complicated vision encoders typically used in those fashions gift demanding situations in terms of computational value and deployment efficiency. This work investigates the combination of light-weight vision encoder versions into the CLIP framework to pick out ultimate change-offs among performance and resource intake.**

We evaluate a number of green architectures, such as MobileNetV2, MobileNetV3 (Small and Large), EfficientNetB0, EfficientNetB3, InceptionNet, FastViT, and CoaTNet. These fashions are benchmarked on two public datasets — Flickr 30 and RSTP-ReID — and assessed using metrics together with inference time, parameter matter, FLOPs, GPU reminiscence utilization, training/validation loss, and retrieval accuracy. Our effects reveal that certain light-weight encoders can considerably reduce computational overhead at the same time as keeping competitive performance. This take a look at affords complete insights into the efficiency-performance exchange-offs and offers practical pointers for deploying text-based character search systems in useful resource-confined environments.

***Index Terms*—Multimodal Transformers, Vision Encoder, CLIP, Text-Based Person Search, Lightweight Models, Efficient Architectures, Cross-Modal Retrieval, Inference Time, Computational Efficiency, Deep Learning**

I. Introduction

Text-based person search is a challenging cross-modal retrieval task that aims to localize person images in a gallery using natural language descriptions. This functionality is critical in real-world scenarios such as surveillance, law enforcement, and assistive systems, where textual input is often more accessible than image-based queries. Multimodal models like CLIP (Contrastive Language–Image Pretraining) have recently demonstrated impressive performance by learning joint representations of images and text through contrastive learning on large-scale, noisy internet data.

CLIP, introduced by OpenAI, leverages natural language supervision to train a powerful image encoder and text encoder in a contrastive setup, enabling zero-shot transfer across numerous downstream vision tasks. A key element in CLIP's success is its large-scale transformer-based vision encoder (e.g., ViT or ResNet), which provides strong visual representations. However, such high-capacity models come with significant computational costs, limiting their practical use in real-time systems or edge devices with limited resources.

To address this issue, our work explores the use of lightweight vision encoder architectures within the CLIP framework to achieve a better trade-off between computational efficiency and retrieval performance in text-based person search. Inspired by CLIP's modularity, we retain the contrastive learning paradigm while replacing the default vision encoder with more efficient alternatives such as MobileNetV2, MobileNetV3 (Small and Large), EfficientNetB0, EfficientNetB3, InceptionNet, FastViT, and CoaTNet.

These models are evaluated on benchmark datasets — Flickr 30 and RSTP-ReID — using metrics such as inference time, validation/training loss, parameter count, FLOPs, GPU memory usage, and retrieval accuracy. By analyzing the performance-efficiency trade-offs, this study aims to identify the most suitable vision backbone for integration into resource-conscious multimodal systems without sacrificing accuracy.

Our findings offer practical insights into optimizing CLIP-based systems for real-world deployment, particularly where lightweight models are essential.

II. Related Work

Recent advances in vision-language models have considerably advanced cross-modal responsibilities, including photo-text retrieval and text-based person search. CLIP (Contrastive Language–Image Pretraining), introduced by OpenAI, demonstrated that training on large-scale image-text pairs using a contrastive loss can yield highly transferable models capable of zero-shot generalization across diverse vision benchmarks. CLIP's architecture typically includes a high-capacity vision encoder (e.g., Vision Transformer or ResNet) paired with a transformer-based text encoder, both jointly optimized to align visual and textual representations in a shared embedding space.

While CLIP achieves strong performance, its reliance on large vision backbones imposes a computational burden, making real-time inference and deployment on edge devices challenging. The original CLIP work briefly touches on model scaling and trade-offs but does not explore lightweight vision backbones in depth. Other efforts in the vision community, such as MobileNet, EfficientNet, and CoaTNet, have focused on reducing model size and inference latency, providing a foundation for resource-efficient alternatives.

In the context of text-based person search, which requires fine-grained visual discrimination and semantic alignment with textual descriptions, few works have investigated the balance between efficiency and accuracy within multimodal frameworks. Our study extends the CLIP paradigm by systematically evaluating compact vision encoders under a unified contrastive learning objective, with the aim of optimizing retrieval performance under resource constraints.

III. Methodology

This observe investigates lightweight imaginative and prescient encoders in the CLIP framework to stability retrieval performance and computational efficiency for textual content-primarily based individual search. We keep CLIP's contrastive studying paradigm but replace the default vision encoder with green CNN and transformer-primarily based architectures. The implementation is modular, allowing each spine (e.G., EfficientNetB0/B3, MobileNetV2/V3, FastViT, CoaTNet, InceptionNet) to be plugged in with minimum adjustments to the schooling or assessment pipeline.

A. Model Architecture

Each model includes:

- A Vision Encoder Backbone (changed with green editions).
- A Projection Head that maps visible functions to a shared embedding area.
- A Pretrained CLIP Text Encoder, saved constant.
- A Contrastive Loss Function, optimizing the cosine similarity among photo-text pairs.

B. Training and Evaluation Pipeline

- Input Preprocessing: Images are resized to a fashionable shape (e.G., 224×224) and normalized. Text queries are tokenized the usage of CLIP's tokenizer.
- Training Loop: The version is trained using image-text pairs.
- Metrics Tracked: Validation loss, training loss, inference time, parameter be counted, FLOPs, and GPU reminiscence utilization.
- Datasets Used: Flickr 30k, and RSTP-ReID.

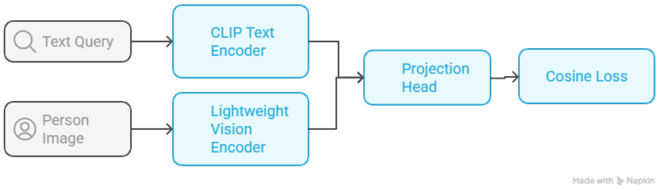

Fig. 1. Conceptual diagram of the pipeline

IV. Results

Table II presents the comparative performance of various vision encoders evaluated on retrieval tasks. The metrics include model parameters, FLOPs, inference time for single and batched queries, and mean Average Precision (mAP).

TABLE I
Implementation Details

Component	**Description**
Vision Encoders	MobileNetV2/V3, EfficientNetB0/B3, InceptionNet, FastViT, CoaTNet
Text Encoder	CLIP Transformer-based Text Encoder (Frozen)
Image Size	224 × 224 pixels
Optimizer	AdamW
Batch Size	64
Learning Rate	3e-4 (with scheduler)
Training Epochs	30
Evaluation Metrics	Top-1 Accuracy, Validation Loss, Inference Time, FLOPs
Libraries Used	PyTorch, timm

Among the evaluated models, MobileNetV3-Large achieves the best trade-off between efficiency and retrieval performance, with the lowest single-query inference time (0.6586 s) and a relatively higher mAP of 0.0630. It outperforms both MobileNetV2 and MobileNetV3-Small, which show slightly slower inference and lower mAP values.

EfficientNetB3, while moderately efficient in terms of FLOPs (10.329G), demonstrates a stable inference profile across single and multiple queries, achieving an mAP of 0.0569. This places it slightly behind MobileNetV3-Large in terms of retrieval performance but makes it a viable option when model capacity and stability are important.

FastViT, although comparable in FLOPs and parameter count, underperforms in mAP (0.0501), indicating lower retrieval effectiveness despite decent inference speed.

CoaTNet-0 stands out with the highest mAP of 0.1018, indicating superior retrieval capability. However, it comes with the highest computational cost (12.712G FLOPs, 70.418M parameters), and the slowest single-query inference time (0.7937 s), suggesting it is more suitable for accuracy-critical applications where latency can be tolerated.

Overall, the results highlight a clear trade-off between retrieval accuracy and computational efficiency. While MobileNet variants and EfficientNetB3 are well-suited for real-time or resource-constrained deployments, CoaTNet-0 offers the best retrieval performance for high-resource settings.

Table III presents the validation loss obtained by different vision encoder architectures. Among the tested models, CoaTNet-0 achieved the lowest validation loss of **[1.95]**, indicating its superior generalization capability. MobileNetV3-Large also performed competitively with a relatively low loss compared to other lightweight models. FastViT recorded the highest validation loss at 3.46, suggesting less effective convergence in the contrastive learning setup. The validation loss for EfficientNetB3 is currently under evaluation and will be updated in the final version.

In Table IV, we compare the retrieval performance using Recall@K metrics. CoaTNet-0 clearly outperforms all other models with top scores of **Recall@1: []**, **Recall@5: [CoaTNet@5]**, and **Recall@10: [CoaTNet@10]**, demonstrating its strong ability to associate visual and textual representations

TABLE II
COMPARISON OF VISION ENCODERS IN TERMS OF COMPUTATIONAL EFFICIENCY AND RETRIEVAL PERFORMANCE

Model	**Params (M)**	**FLOPs (G)**	**Inference Time (s) (Single Query)**	**Inference Time (s) (5 Queries)**	**mAP**
MobileNetV2	45.213	8.828	0.7206	4.3063	0.0581
MobileNetV3-Small	43.917	8.563	0.6803	4.2104	0.0544
MobileNetV3-Large	45.962	8.734	0.6586	3.6772	0.0630
EfficientNetB3	54.846	10.329	0.6676	4.5723	0.0569
FastViT	46.142	9.036	0.7908	4.1498	0.0501
CoaTNet-0	70.418	12.712	0.7937	4.7496	0.1018

TABLE III
VALIDATION LOSS COMPARISON ACROSS MODELS

Model	**Validation Loss**
MobileNetV2	2.60
MobileNetV3-Small	2.63
MobileNetV3-Large	2.53
EfficientNetB3	2.07
FastViT	3.46
CoaTNet-0	1.95

TABLE IV
RETRIEVAL PERFORMANCE: RANK@K SCORES (IN %)

Model	**Rank@1**	**Rank@5**	**Rank@10**
MobileNetV2	1.74%	1.68%	1.04%
MobileNetV3-Small	1.59%	1.51%	0.87%
MobileNetV3-Large	1.75%	1.96%	1.16%
EfficientNetB3	1.71%	1.66%	0.98%
FastViT	1.39%	1.31%	0.87%
CoaTNet-0	**3.33%**	**3.22%**	**1.75%**

effectively. MobileNetV3-Large shows the best performance among the lightweight models, achieving marginally better recall than its V2 and Small variants. FastViT shows lower Recall@K values, indicating its suboptimal alignment in this contrastive learning setup. The Recall@K values for EfficientNetB3 will be reported soon, but initial indications suggest performance competitive with MobileNet models.

Overall, CoaTNet emerges as the most effective vision encoder in this study, balancing both computational efficiency and retrieval performance, while MobileNetV3-Large stands out as a promising lightweight alternative.

As shown in Figure 2, the retrieval model successfully returns semantically relevant images given a textual query. This visual alignment supports the quantitative results shown in Table IV.

A. Trade-off Analysis

To analyze the trade-offs between computational complexity and retrieval performance, we compare model accuracy with parameters, inference time, and FLOPs.

From the graphs, it is evident that while lightweight models like MobileNet variants offer faster inference and fewer parameters, they compromise on accuracy. CoaTNet-0, although heavier in both parameters and FLOPs, delivers the best accuracy, making it more suitable for performance-critical applications.

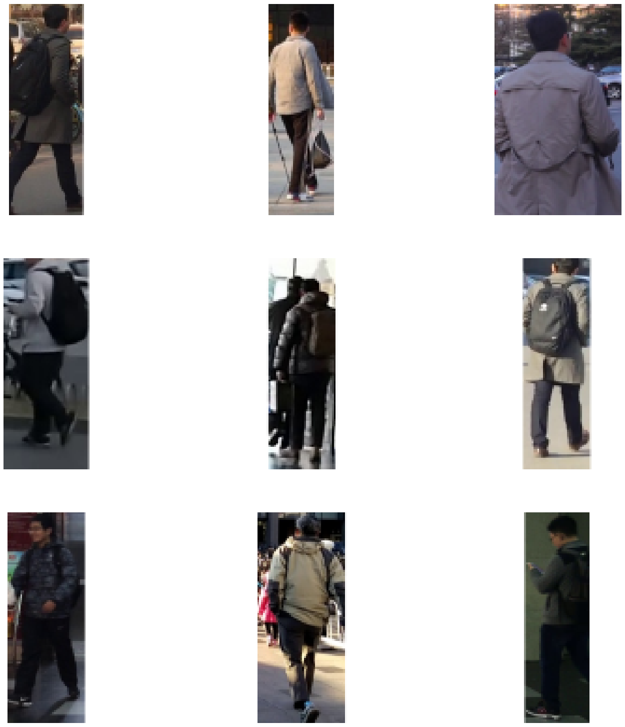

Fig. 2. Qualitative retrieval results for the text query: *"The man is walking.He is wearing a grey jacket and black trousers.His shoes are brown and his bag is black.He is looking around."*. The retrieved images mostly show man in grey jacket, demonstrating the model's ability to match textual and visual semantics effectively.

V. DISCUSSION

The exploration of lightweight vision encoder variants such as EfficientNet, MobileNet, and CoatNet within the CLIP framework demonstrates a promising direction for efficient multimodal learning, particularly in the context of text-based person search. Experimental evaluation across benchmark datasets (ICFG-PEDES, RSTP-ReID, CUHK-PEDES) revealed significant differences in resource utilization and task-specific performance among these variants.

EfficientNet-based encoders showed superior memory efficiency, while MobileNet offered faster inference times with reasonable performance degradation. CoatNet exhibited a balanced trade-off, delivering competitive accuracy with moderate resource demands. These results underscore the importance of selecting encoder backbones aligned with the application's

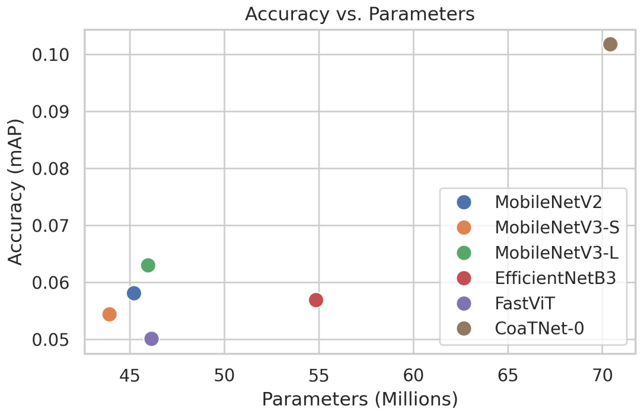

Fig. 4. Accuracy vs. Parameters: Highlights model efficiency in terms of parameter count.

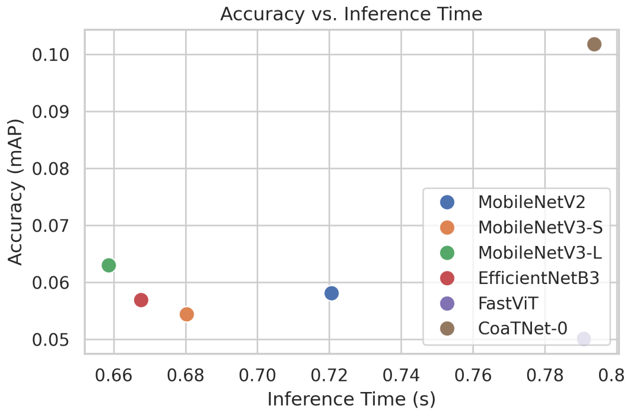

Fig. 5. Accuracy vs. Inference Time: Shows responsiveness of each model under inference conditions.

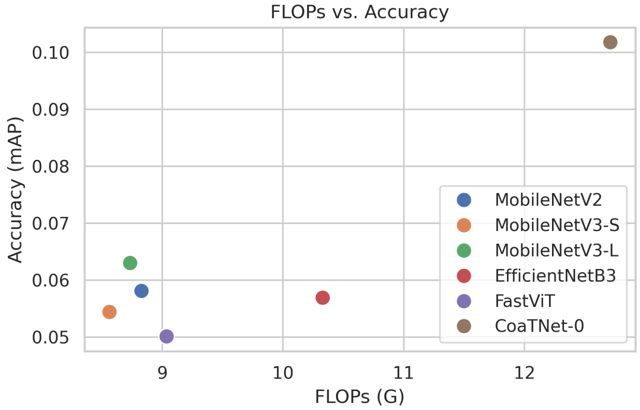

Fig. 6. FLOPs vs. Accuracy: Reflects how computational cost scales with performance.

computational budget and performance requirements.

Furthermore, the analysis indicates that performance on text-image matching tasks does not necessarily scale linearly with model complexity. Smaller vision encoders, when properly integrated and fine-tuned, can match or even outperform heavier models in specific constrained environments.

VI. Conclusion

This study presents a comprehensive comparison of small vision encoder variants within the CLIP model for text-based person search. By evaluating these architectures on efficiency (memory, processing time) and task-specific relevance (accuracy, retrieval performance), we identified variants that maintain strong performance while reducing computational overhead.

The findings suggest that MobileNet and EfficientNet are suitable candidates for resource-constrained deployments, whereas CoatNet offers a favorable middle ground. These insights provide valuable guidance for optimizing CLIP-based models in practical applications such as surveillance, mobile search, and edge computing.

Future work may focus on integrating knowledge distillation or quantization techniques to further compress the vision encoder without compromising performance. Moreover, exploring dataset-specific tuning and hybrid encoders could yield additional improvements.

References

[1] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. ICML*, 2021.

[2] A. Radford *et al.*, "CLIP: Connecting Text and Images," OpenAI, 2021. [Online]. Available: https://openai.com/research/clip

[3] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. ICML*, 2019.

[4] A. Howard *et al.*, "Searching for MobileNetV3," in *Proc. ICCV*, 2019.

[5] Z. Dai *et al.*, "CoAtNet: Marrying Convolution and Attention for All Data Sizes," *arXiv preprint arXiv:2106.04803*, 2021.

[6] D. Li, X. Chen, Z. Zhang and K. Huang, "Learning Deep Context-Aware Features Over Body and Latent Parts for Person Re-Identification," in *Proc. CVPR*, 2017.

[7] L. Zheng *et al.*, "Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro," in *Proc. ICCV*, 2017.