Ahmedabad University

Investigate Small Vision Encoders in Multimodal Transformers

CSE641 Computer Vision: Modern Methods And Application

Dhyey Patel	AU2240054
Malav Modi	AU2240214
Prem Patel	AU2240010

Ahmedabad University

Problem Statement

- Multimodal transformers like CLIP rely heavily on large vision encoders, leading to high computational costs and memory usage

- This project investigates the use of small vision encoders to reduce model complexity while maintaining competitive performance in image-text tasks.

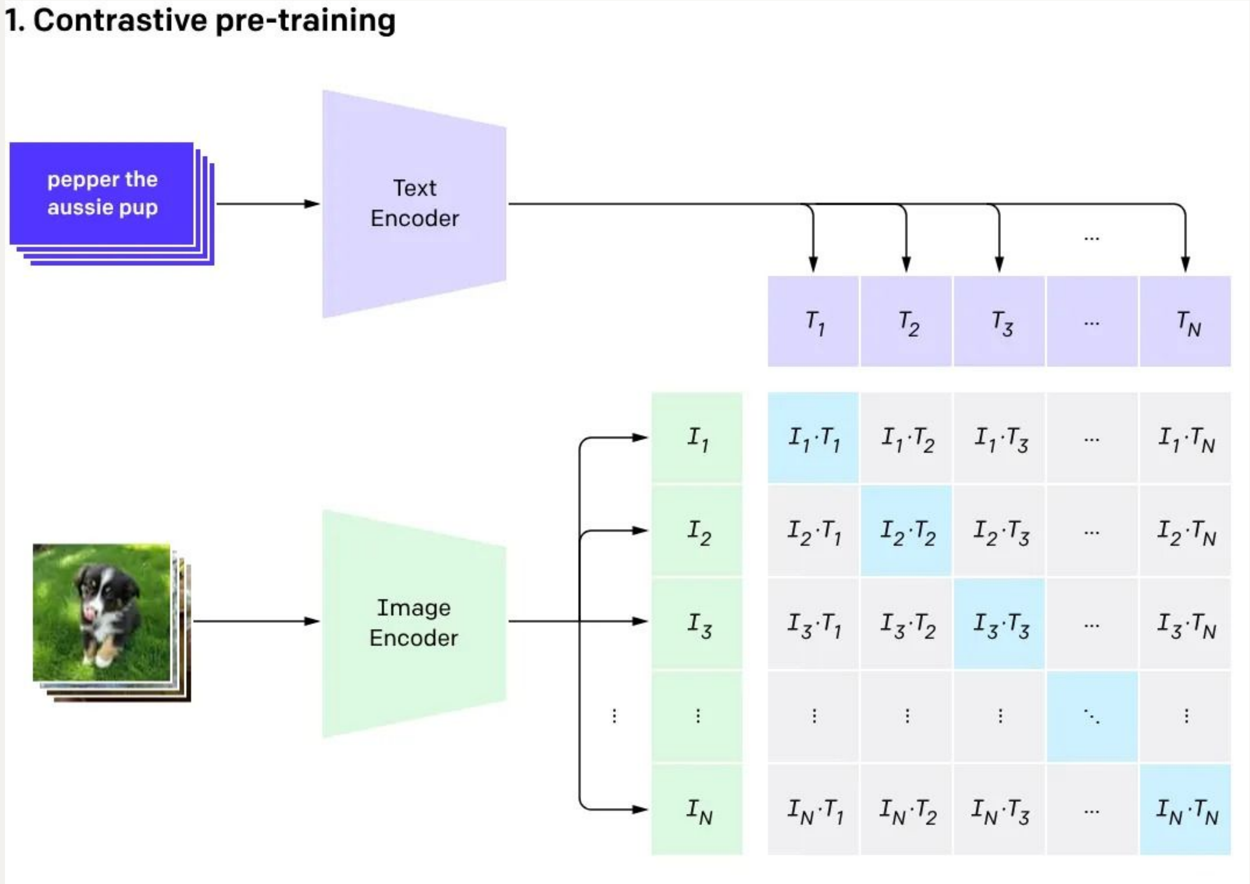

Fig. 1 : CLIP architecture

Ahmedabad University

Pi, "Evaluation of OpenAI's CLIP Model - Neural Engineer - Medium," *Medium*, Nov. 28, 2024. [Online]. Available: https://medium.com/neural-engineer/evaluation-of-openais-clip-model-956b3475a456

Literature Survey

Question	Pros	Cons	References
How does natural language supervision improve visual models?	CLIP demonstrates that natural language supervision enables robust, transferable visual representations for multimodal tasks.	Requires large-scale datasets and computational resources for pretraining effectively.	Learning Transferable Visual Models From Natural Language Supervision. https://arxiv.org/pdf/2103.00020
How can EfficientNet improve Model Scaling for CNN's?	Balances network width, depth, and resolution using a compound coefficient, leading to better accuracy and efficiency.	Scaling might not always translate directly to improved performance in all multimodal tasks.	EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks https://arxiv.org/pdf/1905.11946v5
How does MobileNetV2 achieve efficiency?	Utilizes inverted residuals and linear bottlenecks to significantly reduce the number of parameters and computational cost.	May have limitations in capturing very fine-grained details due to information bottlenecking.	MobileNetV2: Inverted Residuals and Linear Bottlenecks. https://arxiv.org/pdf/1801.04381
How does the Inception architecture contribute to efficient visual encoding?	Multi-scale feature extraction within a single layer, potentially reducing depth and parameters.	Increased complexity in layer design and potential for vanishing gradients in deeper networks.	Going deeper with convolutions. https://arxiv.org/pdf/1409.4842

Dataset Discussion

- **Total Images:** 31,783
- **Total Captions:** 158,915 (5 captions per image)
- **Source**: Flickr (real-world photographs)
- **Caption Format:** Short descriptive English sentences
- **Common Caption Themes:** Human activities, sports, nature scenes, animals, vehicles, objects, and urban settings.

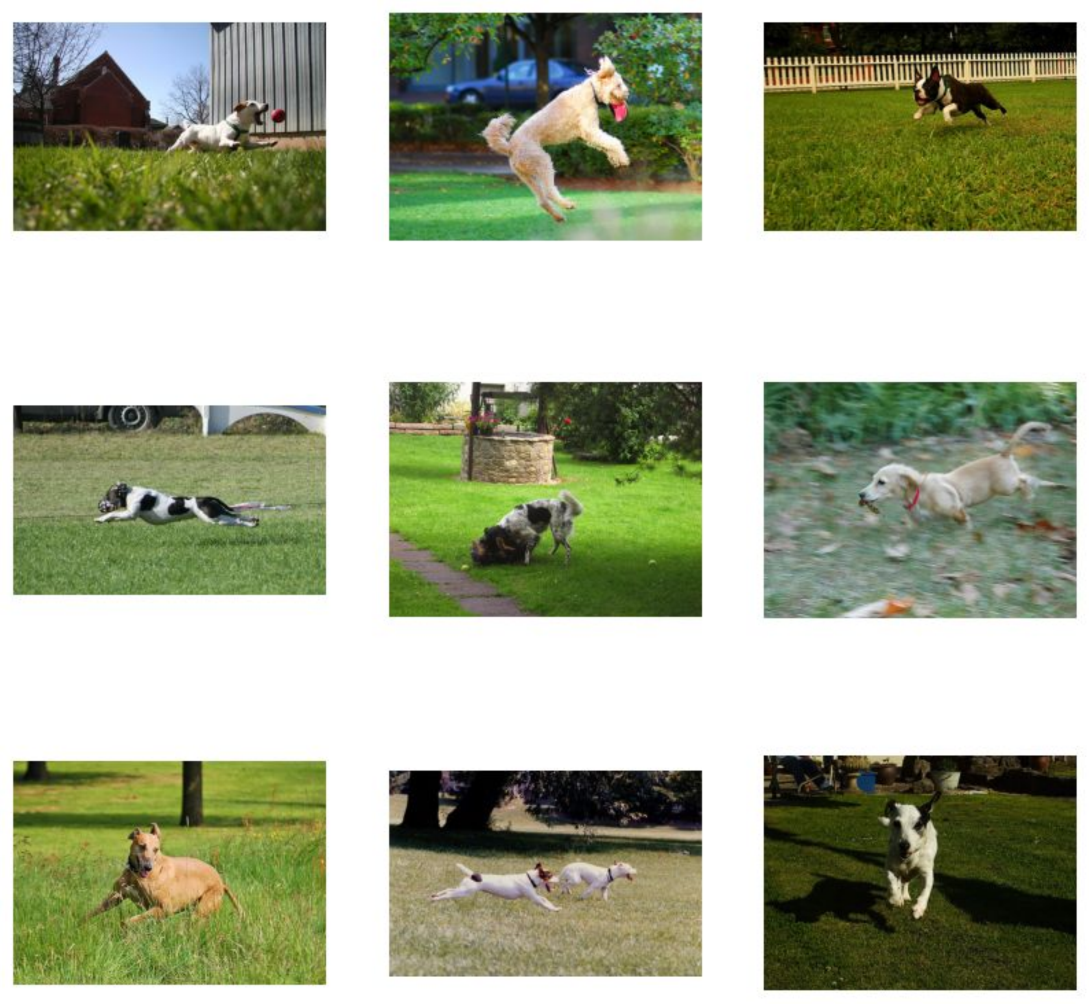

Fig. 2 : Dataset Example

Ahmedabad University

Methodology

Used **CLIP (Contrastive Language-Image Pretraining)** model.

Trained the model on different variants of vision encoders within the CLIP model.

- **EfficientNet:** High accuracy with efficiency
- **MobileNet:** Lightweight and fast
- **Inception:** Multi-scale feature extraction

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

Methodology

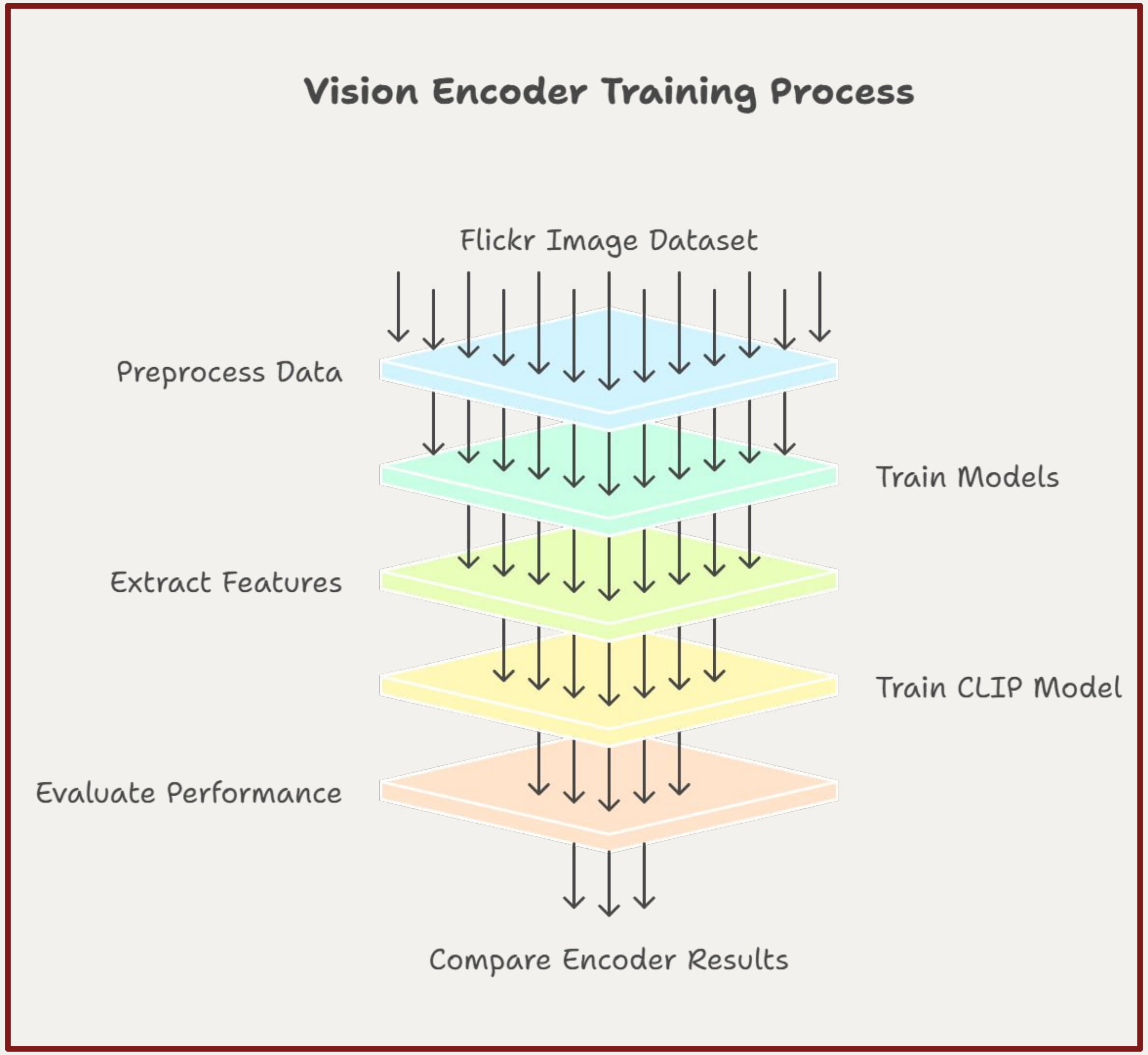

Fig. 3 : Methodology

Results

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

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

Ahmedabad University

Future Work

- Assess the computational cost and resource efficiency of each vision encoder variant.
- Evaluate the performance of each variant in terms of accuracy and relevance for text-based person search tasks.
- Compare the trade-offs between model efficiency and performance, identifying the most efficient vision encoder variant that still performs well for the text-to-image matching task.

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. https://arxiv.org/abs/2103.00020

[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