MVLIP: MambaVision as an Image Encoder in CLIP

Jaeseo Lee

1. Introduction

Contrastive Language-Image Pretraining, CLIP(Radford et al., 2021) has set a new standard for vision-language alignment by jointly training image and text encoders. However, The ViT(Dosovitskiy et al., 2020)-based image encoder in CLIP is computationally intensive and has less scalability and efficiency, particularly when processing high-resolution images. To be specific, transformer based models essentially have quadratic time complexity due to its self-attention mechanisms, leading to performance degradation when processing high-resolution images. Recently, MambaVision(Hatamizadeh & Kautz, 2024), a hybrid architecture combining Mamba and transformer layers, has demonstrated state-of-the-art accuracy and remarkable inference speed by utilizing both selective state-space based layers and self-attention layers. This work makes the following key contributions: (1) We propose MVLIP, the first integration of MambaVision's hybrid Mamba-Transformer architecture into CLIP's vision-language framework; (2) We design a methodologically sound experimental protocol using CLIP(Swin(Liu et al., 2021)) as an intermediate baseline to enable fair comparison under equivalent pretraining conditions; (3) We demonstrate significant performance improvements (9-10% across all metrics) and computational efficiency gains (33-41% throughput increase) over baselines; (4) We provide empirical evidence that hybrid architectures can effectively address the quadratic complexity limitations of Vision Transformers. The implementation code and experimental details are available at `https://github.com/Tosaaa/MVLIP`.

2. Related Works

• CLIP: Introduced large-scale contrastive pretraining with ViT and ResNet image encoders.

• MambaVision: Proposed a hybrid SSM-Transformer model, outperforming ViT in accuracy and efficiency.

• Swin: Introduced windowed self-attention mechanism, which produces more efficient and accurate results compared to traditional attention models.

3. Methods

In CLIP, ViT was used as the image encoder to extract image features. However, its performance can be increased by replacing ViT with state-of-the-art vision encoders, such as MambaVision.

MVLIP Training Algorithm

1. Initialize MambaVision backbone with ImageNet-21K(Deng et al., 2009) pretrained weights
2. Replace CLIP's ViT image encoder with MambaVision
3. Modify classification head dimensions to match text encoder embedding size (512)
4. Freeze text encoder parameters from CLIP(ViT-B/32)
5. Fine-tune only image encoder and projection layers using contrastive loss with temperature 0.07

4. Experiments

To establish a fair comparison between MVLIP and the original CLIP model, we conducted a controlled experimental setup that addresses the inherent challenges of comparing models with different pretraining backgrounds. Since the original CLIP model was extensively pretrained on OpenAI's proprietary large-scale dataset, directly comparing a newly initialized MVLIP model would not provide meaningful insights due to the substantial difference in pretraining data and computational resources.

To overcome this limitation, we designed a comparative framework using CLIP(Swin) as an intermediate baseline. We replaced CLIP's original ViT image encoder with a Swin Transformer, creating CLIP(Swin), which serves as a more appropriate comparison target for MVLIP. This approach is methodologically sound for several reasons: First, Swin Transformers have demonstrated superior performance compared to standard ViTs across various computer vision benchmarks, making CLIP(Swin) a stronger baseline than the original CLIP(ViT). Second, both the MambaVision backbone used in MVLIP and the Swin Transformer used in CLIP(Swin) were pretrained on ImageNet-21K and fine-tuned on ImageNet-1K, ensuring equivalent pretraining conditions.

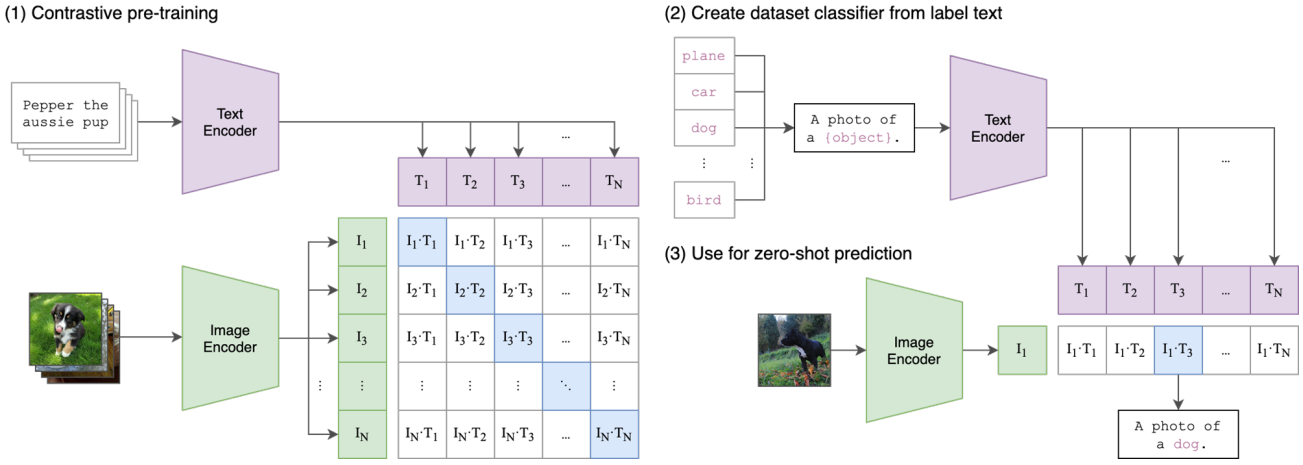

Figure 1. CLIP architecture borrowed from the original CLIP paper. MVLIP has the same architecture as CLIP, but the image encoder is replaced with MambaVision.

Our experimental protocol involved replacing the image encoders in both models while maintaining the identical text encoder from CLIP(ViT-B/32), then fine-tuning both models on the ImageNet-1000 (mini) dataset under identical training conditions. This controlled setup enables us to isolate the impact of the image encoder architecture while maintaining all other variables constant.

Used Models

• CLIP(ViT-B/32)

• MambaVision: MambaVision-B-21K (97.7M Params)

• Swin: swin_base_patch4_window7_224.ms_in22k_ft_in1k (87.8M Params)

Dataset

Experiments were conducted on ImageNet-1000 (mini), a subset of original ImageNet dataset containing 1,000 classes.

Implementation Details

• Input image resolution: 224×224 pixels

• Training/Evalutaion Batch Size: 32

• Image preprocessing: All input images were preprocessed following CLIP's original transformation pipeline.

• Gradient clipping: max norm 1.0

• Data type: FP32

• Total Epochs: 20

• Learning Rate: 1e-4

• Optimzer: AdamW with weight decay 0.01

• Scheduler: CosineAnnealing

Hardware Specifications

• Operating System: Windows 10 with WSL2

• Python(3.10.12), PyTorch(2.6.0+cu118)

• GPU: NVIDIA GeForce RTX 4070 Ti Super (16GB VRAM)

5. Results and Discussion

Training and Validation Loss

Training and validation loss of MVLIP and CLIP(Swin) over 20 epochs is shown in Figure 2 and Figure 3. MVLIP achieves substantially lower validation loss compared to CLIP(Swin). This substantial gap indicates that MVLIP's hybrid Mamba-Transformer architecture provides superior feature representations that generalize better to unseen data.

Figure 2. Training Loss Comparison

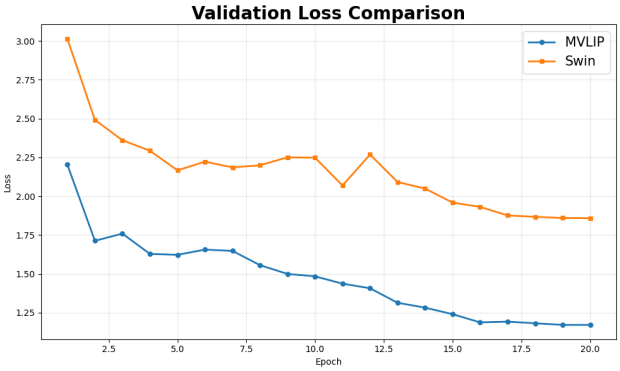

Figure 3. Validation Loss Comparison

Zero-Shot Classification Performance Metrics

We conducted a zero-shot classification evaluation on validation set of ImageNet-1000 (mini). The final epoch comparison demonstrates MVLIP's clear superiority across all evaluation metrics. The result is shown in Table 1. These consistent improvements suggest that MVLIP provides more balanced and robust classification performance.

Table 1. Final Epoch (Best) Performance Comparison

Metric	MVLIP	CLIP(Swin)	Improvement (%)
Accuracy	0.7466	0.6798	9.83
Precision	0.7803	0.7131	9.42
Recall	0.7466	0.6798	9.83
F1-Score	0.7390	0.6717	10.02

Computational Efficiency

The throughput analysis reveals MVLIP's significant computational advantages. MVLIP achieves 33.49% higher throughput than standard CLIP(ViT). The substantial efficiency gain validates the effectiveness of the Mamba architectures' linear complexity compared to the quadratic complexity of Vision Transformers.

Table 2. Throughput Comparison

Model	Throughput (samples/s)	vs CLIP(ViT) (%)
MVLIP	**626.35**	+33.49
CLIP(Swin)	443.51	-5.48
CLIP(ViT)	469.22	-

Experiment Summary

The superior performance of MVLIP over CLIP(Swin) in this controlled comparison provides strong evidence that MVLIP would likely outperform the original CLIP(ViT) if subjected to the same large-scale pretraining process employed by OpenAI.

6. Conclusion

This study presents MVLIP, a novel vision-language model that replaces CLIP's Vision Transformer image encoder with MambaVision, a hybrid Mamba-Transformer architecture. Through comprehensive experimental evaluation, we demonstrate that MVLIP achieves significant improvements over strong baselines across multiple performance dimensions.

Our controlled comparison with CLIP(Swin) reveals that MVLIP consistently outperforms the Swin Transformer-based variant by substantial margins. These consistent improvements across all evaluation metrics, combined with superior training convergence characteristics and better generalization performance (37% lower validation loss), provide compelling evidence for the effectiveness of the hybrid Mamba-Transformer architecture in vision-language tasks.

Beyond accuracy improvements, MVLIP demonstrates remarkable computational efficiency gains, achieving 33.5% higher throughput compared to original CLIP(ViT). This substantial efficiency improvement addresses one of the primary limitations of Vision Transformers—their quadratic computational complexity—while maintaining or improving performance quality.

The experimental design, which ensures fair comparison through equivalent pretraining conditions and controlled fine-tuning, suggests that MVLIP would achieve superior performance compared to the original CLIP model if subjected to the same large-scale pretraining process. The combination of better accuracy, superior computational efficiency, and more stable training dynamics positions MVLIP as a promising advancement for practical vision-language applications.

7. Future Work

Future work should explore scaling MVLIP to larger datasets and investigating its performance across diverse downstream tasks to fully realize its potential as a more efficient and effective alternative to current vision-language models. Although the current state-of-the-art models may achieve higher overall performance than CLIP, integrating MambaVision into CLIP is still meaningful as it may bring a paradigm shift toward hybrid architectures and open new possibilities for vision backbone designs.

References

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

Hatamizadeh, A. and Kautz, J. Mambavision: A hybrid mamba-transformer vision backbone. *arXiv preprint arXiv:2407.08083*, 2024.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.