

Adaptive Representation Construction from CLIP Embeddings for Test-Time Adaptation

Hyewook Kim¹, Yoonsuk Hyun¹

1) *Department of Mathematics, Inha University, Incheon 22212, KOREA*

hyewook@inha.edu

ABSTRACT

CLIP[1], pretrained on a large-scale collection of image-text pairs, demonstrates strong zero-shot classification performance across diverse domains. However, its accuracy can still degrade significantly under domain shift at test time. To address this issue, recent studies have explored test-time adaptation (TTA) techniques that allow models to adjust to incoming data distributions without source access. Our work builds upon the existing cache-based TTA approach, Efficient Test-Time Adaptation of Vision-Language Models (TDA)[2], and proposes several improvements to its representation extraction and usage mechanism. Specifically, we introduce a memory-efficient method that incrementally stores informative features from test samples in a cache model and utilizes them to enhance future predictions. Operating without backpropagation, our method achieves gradual performance improvements with reduced computational cost. Experimental results demonstrate that our approach significantly outperforms TDA in diverse scenarios.

REFERENCES

1. Radford, A., et al. *Learning transferable visual models from natural language supervision*. in *International conference on machine learning*. 2021. PmLR.
2. Karmanov, A., et al. *Efficient test-time adaptation of vision-language models*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.