FINE-GRAINED TEXT UNDERSTANDING IN CLIP-STYLE ENCODER MODELS

Yiqing Zhu
(ID: 260948489)
McGill University
yiqing.zhu2@mail.mcgill.ca

Claire Yang
(ID: 260898597)
McGill University
youyou.yang@mail.mcgill.ca

Shawn Hu
(ID: 260901823)
McGill University
xiding.hu@mail.mcgill.ca

Motivation CLIP (Radford et al., 2021) is widely used for image-text alignment. It is constrained by a 77-token text encoder, limiting its ability to capture semantic relations, such as attributes, spatial relations, and logical modifiers. Prior works such as LongCLIP (Zhang et al., 2024) extend input length, and CLIP-Adapter (Gao et al., 2021) improves generalization. To our knowledge, no prior work addresses fine-grained semantic comprehension in a text-side-only method while preserving the original CLIP alignment.

Our research question is: How can CLIP better capture fine-grained semantic structures and relations through lightweight text-side adaptation?

Methods We designed a four-step pipeline: (i) Process input text into semantically coherent chunks. (ii) Inject lightweight adapters at the text encoder, may reference Tip-Adapter (Zhang et al., 2021). (iii) Aggregate chunk-level embeddings. We will start with simple baselines such as mean/max pooling. Next, we will introduce a learnable attention-based pooling, where each chunk embedding receives adaptive weights according to its semantic contribution. (iv) Training objective combines the original CLIP contrastive loss with a distillation term that constrains the new representations close to the frozen part.

Hypothesis and Expected Results We expect that text-side adaptation with semantic aggregation will improve CLIP's performance on fine-grained retrieval and reasoning tasks (e.g., color, spatial, and logical relations), while maintaining the short-prompt accuracy.

Experimental Design Our baseline is the original CLIP model, followed by our four-step pipeline. Training will be limited to the text encoder using parameter-efficient fine-tuning, with the image tower frozen. Experiments will run on Compute Canada HPC resources.

We will use public image-text datasets with a focus on captions that contain rich modifiers and relations, with caption length and semantic statistics recorded. For evaluation, we can select CLEVR-Text for controlled compositional reasoning and Flickr30k Entities for natural image grounding of attributes and relations (Vongala et al., 2025). Evaluation includes Recall@K for standard retrieval. In addition, we can test on contrastive sentence pairs (inspired by Winoground, Thrush et al. (2022)) that differ by a single token (e.g., red vs blue, left vs right) to quantify semantic consistency.

Project Timeline and Roles The project will proceed in four phases: (i) pipeline implementation and code setup (Nov 1–7); (ii) baseline model training and validation (Nov 7–14); (iii) model variants, ablations, and evaluation (Nov 14–20); (iv) documentation and final presentation (Nov 20–25).

All members will collaborate on the main pipeline and report writing. Yiqing will lead the training phase, Claire will coordinate evaluation and data analysis, and Shawn will oversee model variants and ablation studies.

REFERENCES

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. URL https://arxiv.org/abs/2110.04544.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022. URL https://arxiv.org/abs/2204.03162.

Madhukar Reddy Vongala, Saurabh Srivastava, and Jana Košecká. Compositional image-text matching and retrieval by grounding entities, 2025. URL https://arxiv.org/abs/2505.02278.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*, 2024.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *CoRR*, abs/2111.03930, 2021. URL https://arxiv.org/abs/2111.03930.