Supplementary for AlignCLIP: Align Multi Domains of Texts Input for CLIP models with Object-IoU Loss

Anonymous Authors

1 DETAILED RELATED WORK

Here, we present detailed related work, including vision-language pre-training, long-tail data learning and soft-label CLIP.

1.1 Vision-Language Pre-training

In the realm of Vision-Language Pre-training (VLP), the endeavor to synergize visual and textual modalities has been operationalized through extensive training on image-text pairs. Architecturally, VLP models bifurcate into two predominant streams: single-stream and dual-stream frameworks. Single-stream architectures integrate image and text inputs early in the process, utilizing a unified transformer to process the amalgamated embeddings, typified by models such as VisualBERT [10], OSCAR [12], UNITER [4], UNICODER [8], UNIMO [11] and HAMMER [20]. This architecture facilitates direct interaction between modalities within a shared semantic space. Conversely, dual-stream architectures advocate for a modular approach, encoding images and texts through distinct pathways before convergence. Models like CLIP [19], ALIGN [9], DeCLIP [13], SoftCLIP [6], PyramidCLIP [7] and LaCLIP [5] exemplify this approach, underscoring the advantage of discrete yet complementary processing of modal information. Most of this work is to improve certain shortcomings of CLIP. For example, DeCLIP [13] speeds up training through self-supervision. PyramidCLIP [7] uses object detectors for more fine-grained alignment. SoftCLIP [6] uses object detectors to construct many-to-many relationships.

The proposed AlignCLIP belongs to the dual-stream architecture. Differently, AlignCLIP sets out to solve the long-tail distribution in CLIP and misalignment in multiple text domains. Furthermore, we achieve soft label training at low cost based on caption objectf parsing. Compared with previous methods, AlignCLIP training cost is lower and its performance is better.

1.2 Long-tail Data Learning

The long-tail distribution [14], where few categories are common and many are rare, presents a significant challenge in data mining and machine learning. Addressing this, researchers have developed three main strategies: re-sampling, re-weighting, and transfer learning. Re-sampling [2, 21, 22] methods adjust the dataset to balance the distribution between common and rare classes, either by increasing the presence of rare classes or reducing that of common ones. Re-weighting [3, 18, 24] approaches alter the loss function to prioritize rare classes during training, giving them more importance. Transfer learning [16, 17, 25] techniques use the knowledge gained from common classes to improve the learning of rare classes, enriching their feature representation. These strategies, from adjusting data distribution to modifying training emphasis, offer pathways to mitigate the long-tail problem, aiming for a more balanced learning across classes.

However, in multi-modal pre-training, there are relatively few solutions to the long-tail problem, which has been grossly ignored. Although there are some works that use visual descriptions to improve the performance [15, 23], however, they only generate category attributes at the test stage, which leads to the multi-domain misalignment, limiting model performance. We propose to use visual descriptions while solving the distribution shift of multiple domains during the training stage, achieving better results.

1.3 Soft-label CLIP

The original CLIP assumes that images and texts are strictly one-to-one, that is, during training, the labels of matching image and text are 1, and the unmatched ones are 0. However, this assumption does not hold true in many cases, and there is some correlation between different negative samples. Therefore, many works propose the use of soft labels in CLIP training. One of the most intuitive ideas is to use self-distillation, that is, using a trained teacher model to assign soft labels to get a better student model, and continuously iterate, as did in CLIP-PSD [1]. PyramidCLIP [7] proposes to use label smoothing to alleviate strict one-to-one correspondence. However, it gives the same weight to all negative samples, which brings limited improvement. SoftCLIP [6] proposes to use a object detector to obtain the object information contained in the image, so as to obtain the similarity and relationship between negative samples. However, this method brings a lot of additional overhead to training, resulting in increased training costs.

Differently, we propose to use caption object parsing to obtain the objects in the caption, thereby constructing a many-to-many relationship to generate soft labels. It can achieve higher performance with lower training costs.

2 DATASET STATISTICS

Here we display detailed statistical information of the datasets, including the number of image-text pairs in the dataset, the total number of parsed objects, and the number of tail objects.

Table 1: Dataset Statistics

Dataset	Image-text Pairs	Parsed Objects	Tail Objects
CC3M	2,901,344	109,341	32,802
CC12M	10,841,279	276,123	82,836
YFCC15M	13,930,140	330,982	99,294

3 PROMPTS FOR LLMS

In our method, there are two places where LLM may be used:

1) The first one is caption object analysis. Our method includes two solutions, namely part-of-speech tagging (POS)and large language models (LLM). For LLM, the prompt for LLM is “The following is a caption for an image. Please directly indicate the objects contained in it, separated by commas.”

2) The second place is the generation of appearance description. Here, the prompt is “Please briefly introduce the appearance of CLASS in ordinary language”.

4 ADDITIONAL ABLATION EXPERIMENTS

4.1 Object-IoU Loss v.s. Label Smoothing

Here, we compare the object-IoU loss in this paper with the ordinary label smoothing [7] loss function. The experiments were pre-trained on CC3M and evaluated with the zero-shot classification accuracy of ImageNet 1K. For label smoothing, we follow the PyramidCLIP [7] and take the smoothing parameter 0.2. For hard-label, we use the original cross-entropy loss, leaving everything else unchanged. The results are shown in Tab. 2. We can see that label smoothing can bring a 1.5 points of improvement in Top-1 accuracy, and our method can greatly exceed label smoothing.

Table 2: Object-IoU Loss v.s. Label Smoothing

Method	Top-1 ACC	Top-5 ACC
Hard Label	23.0	45.8
Label Smoothing	24.5	48.7
Object-IoU Loss	**27.0**	**52.2**

REFERENCES

[1] Alex Andonian, Shixing Chen, and Raffay Hamid. 2022. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16430–16441.

[2] Nadine Chang, Zhiding Yu, Yu-Xiong Wang, Animashree Anandkumar, Sanja Fidler, and Jose M Alvarez. 2021. Image-level or object-level? a tale of two resampling strategies for long-tailed detection. In *International conference on machine learning*. PMLR, 1463–1472.

[3] Xiaohua Chen, Yucan Zhou, Dayan Wu, Chule Yang, Bo Li, Qinghua Hu, and Weiping Wang. 2023. Area: adaptive reweighting via effective area for long-tailed classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19277–19287.

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.

[5] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2024. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems* 36 (2024).

[6] Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, and Xing Sun. 2024. Softclip: Softer cross-modal alignment makes clip stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 1860–1868.

[7] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. 2022. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems* 35 (2022), 35959–35970.

[8] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. *arXiv preprint arXiv:1909.00964* (2019).

[9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.

[10] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[11] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2592–2607.

[12] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 121–137.

[13] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208* (2021).

[14] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2537–2546.

[15] Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E O’Connor. 2023. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 262–271.

[16] Aakanksha Naik, Jill Lehman, and Carolyn Rosé. 2022. Adapting to the long tail: A meta-analysis of transfer learning research for language understanding tasks. *Transactions of the Association for Computational Linguistics* 10 (2022), 956–980.

[17] Sarah Parisot, Pedro M Esperança, Steven McDonagh, Tamas J Madarasz, Yongxin Yang, and Zhenguo Li. 2022. Long-tail recognition via compositional knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6939–6948.

[18] Hanyu Peng, Weiguo Pian, Mingming Sun, and Ping Li. 2023. Dynamic re-weighting for long-tailed semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 6464–6474.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[20] Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and Grounding Multi-Modal Media Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6904–6913.

[21] Jiang-Xin Shi, Tong Wei, Yuke Xiang, and Yu-Feng Li. 2023. How Re-sampling Helps for Long-Tail Learning? *Advances in Neural Information Processing Systems* 36 (2023).

[22] Hongxin Wei, Lue Tao, Renchunzi Xie, Lei Feng, and Bo An. 2022. Open-sampling: Exploring out-of-distribution data for re-balancing long-tailed datasets. In *International Conference on Machine Learning*. PMLR, 23615–23630.

[23] Wenhao Wu, Huanjin Yao, Mengxi Zhang, Yuxin Song, Wanli Ouyang, and Jingdong Wang. 2023. GPT4Vis: What Can GPT-4 Do for Zero-shot Visual Recognition? *arXiv preprint arXiv:2311.15732* (2023).

[24] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. 2021. Distribution Alignment: A Unified Framework for Long-Tail Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2361–2370.

[25] Yin Zhang, Derek Zhiyuan Cheng, Tiansheng Yao, Xinyang Yi, Lichan Hong, and Ed H Chi. 2021. A model of two tales: Dual transfer learning framework for improved long-tail item recommendation. In *Proceedings of the web conference 2021*. 2220–2231.