Self-Distillation using image-language representation for image classification

Pasit Tiwawongrut
Asian Institute of Technology
Klong Luang Pathumthani 12120, Thailand
Pasit.Tiwawongrut@ait.asia

Dr. Chaklam Silpasuwanchai
Asian Institute of Technology
Klong Luang Pathumthani 12120, Thailand
chaklam@ait.asia

Abstract

The ABSTRACT is to be in fully-justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. Leave two blank lines after the Abstract, then begin the main text. Look at previous ICCV abstracts to get a feel for style and length.

1. Introduction

Vision language pre-trained models have show effective performance both in-domain and down-stream task by utilizing both text and image information. CLIP [13] and ALIGN [8] training two-stream encoder with constrastive learning to align vision and language modalities, which result in competitive performance in many vision language task *e.g*. image-text retrieval, visual question answer and zero shot image classification. ALBEF [12], CoCa [17] and mPLUG [9] added cross-attention layers over two-stream encoder to providing better alignment over vision and language modalities with multiple traning objective *e.g*. image-text captioning, image-text contrastive loss, image-text matching and masked-language-modeling loss. As a result these models achieved state-of-the-art multiple vision-language and image classification task

However, the gap of using self-distillation to improve vision language models performance was still underexplored. By using moving average teacher [15, 3], the teacher model weight is updated with average of the student model gradient. As a result, the teacher model output representation is consistent. MixMatch [2], Mixup [18] and Fixmatch [14] is an image input interpolation method for improving output consistency within teacher-student framework by image augmentation and input interpolation between each samples. DINO [4] utilize both moving average teacher and image interpolation to train the teacher-student image encoder model without using any label, which result in competitive down-stream task performance.

In this paper, we proposed a method to improve performance of two-stream encoder vision language model using self-distllation technique. By using moving average teacher, we can remove noise from the encoder model, which is trained using noisy internet image-text pairs. The image interpolation technique were applied to create more consistent encoding. We provide result by applied our method over baseline model including CLIP and CoCa over image classification, image-text retrieval and image captioning task.

2. Related work

2.1. Vision-Language model

In the past few years, many works have shown the ability to utilize textual information with the image task by training with image text pair *e.g*. CLIP [13], UNITER [5], Blip [11, 10], BEiT [16] and CoCa [17]. By training with a large amount of the image-text pair dataset, the ALIGN model could make up for the noisy image description and surpass the model, which was trained with the benchmark dataset in the zero shot image classification task. Recently **Co**ntrastive **Ca**ptioner (CoCa) [17] proposed a vision-language encoder-decoder model which was trained with image-text contrastive loss and captioning loss Cross attention layers were added to join image-text modality. The CoCa model performed linear probing image classification on ImageNet with top-1% 90.6% accuracy. In this research, we adopted the two stream encoder method same as CLIP, and we also used a cross attention layer to create image-text representation for classification. Another methods [5, 16, 5, 1] is to concatenate both image and text embedding and utilize multi-head self-attention to joined vision and language modalities.

2.2. Knowledge Distillation and Self-Distillation

Knowledge Distillation was firstly proposed by [7] to compress the model size while maintaining the model performance as much as possible. The method contained a smaller student model and a single or multiple larger

teacher model. The knowledge was transferred by optimizing the student model output to match the teacher's output. [6] investigated knowledge distillation using a student model size the same as the teacher model, showing improvement in the student model. Such a method is called self-distillation. The self-distillation has widely adopted in semi-supervised image classification tasks, such as Mean Teacher [15], EMAN [3] and FixMatch [14]. DINO [4] proposed self-distillation pre-training without using any label, which resulted in performance improvement. In this paper, we extended the self-distillation by creating representation which was image-text combined representation, and we trained the student model to match teacher softmax outputs.

3. Methodology

References

[1] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

[3] Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Zhuowen Tu, and Stefano Soatto. Exponential moving average normalization for self-supervised and semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 194–203, 2021.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, page 104–120, Berlin, Heidelberg, 2020. Springer-Verlag.

[6] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018.

[7] Geoffrey Hinton, Jeff Dean, and Oriol Vinyals. Distilling the knowledge in a neural network. pages 1–9, 03 2014.

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

[9] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022.

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[12] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

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

[14] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[15] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[16] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023.

[17] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

[18] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.