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

In computer vision, self-distillation [4, 11, 8] is a technique for improving deep learning models without increasing model size. This paradigm involves training a student model whose parameter size is equal to the teacher model with new parameter initialization. One method from this paradigm can work without any label called Self-**di**stillation with **no** labels (DINO) [2]. The method has been shown to improve the performance of both ResNet [5] and Vision Transformers (ViT) [3]. According to [1], when using the self-distillation technique, the student model is forced to learn soft-label features, which were extracted from the dataset. Additionally, by training the model with difference parameter initialization, the student model acquires knowledge from multiple views of images. The result shows around 2% improvement by the self-distillation method over multiple ResNet models [10]. test

In another branch of research, a multimodal approach demonstrates that the model's performance can be improved when combining both image and text data into the model. Contrastive Language-Image Pre-training (CLIP) [7] and **A** **L**arge-scale **I**ma**G**e and **N**oisy-text embedding (ALIGN) [6] both achieved performance on par with fully supervised image classification across multiple benchmarks. These models are obtained by training the models with image-text pairs using the contrastive vision language pre-training method. The current state-of-the-art is **Co**ntrastive **Ca**ptioner (CoCa) [9]. This approach used image-text pairs with contrastive language-image loss and image captioning loss. Thus, it is a clear benefit of the training model in utilizing image and text information.

By merging the two paradigms, we proposed a new approach to train an image classification model by distilling knowledge from a multimodal teacher as shown in Figure 1. Multimodal teacher models were constructed by leveraging a pre-trained language model and a pre-trained image encoder. The output of both encoders was combined using cross-attention and a linear classification layer, called "image-text representation head". The detail of the image-text representation head is described in Figure **??**. In this work, the encoded text was used as a query to extract the relevant information from the image encoding. The student model, which had the same architecture as the teacher image encoder model, was trained using teacher output as a target. Thus, the student learned with high-level semantic information.

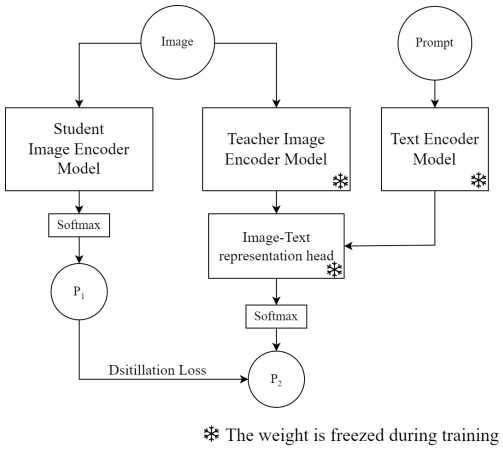

Figure 1. Self-distillation training with image text joined representation.

The result showed that by combining textual information with images, our approach improved accuracy by 3% in both ResNet [5] and ViT [3] model compared to the baseline self-distillation method. The ablation study showed that the

student model achieved 3% higher accuracy by providing detailed descriptions in the training process. This suggested that by using the text encodings with cross-attention, the model extracted higher semantic information and more precise image representations from the images.

To summarize our contribution. Firstly this paper investigated the effectiveness of combining text-image representation by using text as a query to emphasize image representation in the self-distillation method. Secondly, this work proposed a method to efficiently combine textual information and images for the self-distillation method. Lastly, this work also investigated the effect of prompts in our methods to create image descriptions for training.

2. Related work

2.1. Vision-Language model

2.2. Knowledge Distillation and Self-Distillation

3. Methodology

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[2] 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.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[4] 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.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] 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.

[7] 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.

[8] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[9] 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.

[10] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

[11] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019.