Leveraging pretrained unimodal models for efficient image-text retrieval

Anonymous authors
Paper under double-blind review

Abstract

Multimodal models, especially vision-language models, have gained increasing popularity due to their wide range of applications, and show impressive performance especially on retrieval tasks. However, existing approaches often require large-scale models, extensive data, and substantial computational resources, limiting their accessibility for smaller research groups and individuals. We address this issue by introducing an efficient self-supervised vision-language model for image-text retrieval that is significantly cheaper to train and smaller in size. We leverage pretrained unimodal encoders and introduce a randomly initialized shared encoder to align representations using a contrastive loss function. A self-supervised image model is employed for simultaneous knowledge distillation, guiding the alignment through high-level image representations. While not reaching SOTA performance, our approach demonstrates competitive performance with popular vision-language models like CLIP and FLAVA on retrieval tasks, outperforming them on certain metrics while using only 0.75% of the data used by CLIP and 4.3% by FLAVA. These findings underscore the potential for designing efficient multimodal retrieval systems, and therefore lay the foundation for future research on financially accessible models, promoting broader participation in multimodal learning. To promote transparency and facilitate further research, we have made our code for training and evaluating our model publicly available.

1 Introduction

Existing vision-language models have seen a significant increase in parameter count and training data, namely image-text pairs. Along with emerging pretraining objectives like a large-scale contrastive loss (cite clip, coca, vlmo) and especially masked vision-language modeling (cite beit-3, flava) those models have reached near perfect score on the widely used benchmarks MSCOCO (cite) and Flickr30K (cite) for image-text retrieval. While there is the risk that samples from these benchmarks may end up in the training data of those approaches, due to their large-scale training datasets, their ability to connect real-world concepts across image and text remains remarkable.

However, with an increase in parameters and training data, the resources (mainly costs through accelerators) to train these models can only be covered by large companies. For example CLIP (cite) has been trained on 400 million image-text pairs, and the largest model has 428 (cite huggingface) million parameters. Based on our estimate (footnote) a reproduction of this model would cost more than 77 thousand dollars to train. For approaches where we are able to estimate the costs based on the information published by the authors, we observe a similar trend: VLMo (cite) costs more than 9 thousand dollars to train, and CoCa (cite) even more than 350 thousand dollars.

In this paper, we propose a method similar to that of Aytar et al. (2017), and leverage pretrained unimodal models to

Our contributions are as follows:

- We show that using pretrained image and text components can reduce the training costs for image-text retrieval models dramatically.

- We demonstrate that a contrastive loss with a low batch size yields a surprising good performance.

- Using a self-supervised vision model as the teacher for knowledge distillation leads to better performance than a supervised vision model.

- An approach characterised by a fully end-to-end self-supervised training on uncurated image-text data.

2 Related work

Knowledge Distillation for guidance. This paper is motivated by the work of Aytar et al. (2017), which train a multimodal model for the alignment of image, text, and audio. The authors use a supervised vision model as a teacher, which provides a probability distribution over the ImageNet-1K (Russakovsky et al., 2015) classes. Because Aytar et al. (2017) use image-text and image-audio pairs, the multimodal (student) model can predict the probability distribution over the ImageNet-1K (Russakovsky et al., 2015) classes when receiving the same image as the teacher, and most importantly the text and audio of the image-text and image-audio pair respectively. The intuition is that since image and text (or image and audio) contain the same semantic content, the ImageNet-1K (Russakovsky et al., 2015) classes of the image should also describe the content of the corresponding text (audio). An example of this (for the paper relevant) image-text pairs can be seen in TODO in the Appendix. Predicting the probability distribution for an image, and an additional ranking loss, leads to an alignment between image, text, and audio, which can be exploited to perform cross-modal retrieval.

Contrastive learning for image-text alignment. OpenAI's CLIP (Radford et al., 2021) was the first model which exclusively relied on a large scale contrastive loss to align image and text. The authors showed that with sufficient amount of data and a large batch size that the contrastive loss leads to a strong alignment between image and text. This lead to a wide spread adoption of contrastive learning with large batch sizes in vision-language pretraining, and has become the de-facto standard to align image and text (Yu et al. (2022); Bao et al. (2022); Singh et al. (2021); Yao et al. (2022)), and has only recently been shown to not be essential for models upwards of a billion parameters (Wang et al., 2023).

Bootstrapping by pretrained initialization. A well-known practice is to use the weights of models (pre-)trained on tasks similar to the target tasks to reduce data requirements and speed up convergence. Since vision-language models usually have parameters exclusively responsible for image and text, it makes sense to initialize these parts of the model with weights from pretrained image and text models, respectively. This is a practice adopted by Bao et al. (2022) and Singh et al. (2021), and both approaches showed significant improvements compared to a random initialization. While the selection for pretrained language models to initialize the text components of the vision-language model naturally falls to self-supervised trained language models like BERT (Devlin et al., 2019), since masked language modeling leads to a strong understanding of text, one should proceed with care when selecting the right vision model for initialization. It is tempting to use supervised vision models, as they still lead to superior performance compared to self-supervised vision models. However, when using a vision model trained with labeled data the end-to-end process is not fully self-supervised anymore, and can therefore considered as "cheating". It is because of this that using only self-supervised components for the initialization is essential.

References

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *CoRR*, abs/1706.00932, 2017. URL `http://arxiv.org/abs/1706.00932`.

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. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 32897–32912. Curran Associates, Inc., 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pp. 4171–4186, Minneapolis, Minnesota, June 2019.

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. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15617–15629, 2021.

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

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=cpDhcsEDC2`.

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. ISSN 2835-8856. URL `https://openreview.net/forum?id=Ee277P3AYC`.

A Appendix

You may include other additional sections here.