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 characterized 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 the contrastive loss leads to a strong alignment between image and text. This lead to a widespread 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 be considered as "cheating". It is because of this that using only self-supervised components for the initialization is essential.

3 Method

Our method is characterized by three main concepts: self-supervised knowledge distillation, contrastive learning, and the initialization. All three concepts are, in order, inspired by the related works presented in the previous section. Before we present the details of our method, we first establish criteria our approach fulfills and why we believe these criteria are important.

3.1 Criteria

End-to-end self-supervised. We believe that a fully self-supervised training process is essential, as this (1) allows to scale up our method if desired (even though this is not the focus of this paper), because we do not rely on labeled data, and (2) we can perform a fair comparison with existing approaches to image-text retrieval. The latter is important because, as already mentioned in the previous section, using supervised models for initialization can be considered as cheating. Even if our training process is self-supervised, the use of (pre-)trained supervised components for initialization turns the whole (end-to-end) process into a supervised one. Whether image-text pairs can be considered as labeled data is a matter of debate, and we discuss this in Appendix A.1.

Independence to pretrained vision-language components. Perhaps the most important criterion is that our method is independent of pretrained vision-language components. We build our method as if the paradigm of vision-language models does not exist, and only rely on pretrained unimodal models. This is important because it allows for a fair comparison with existing approaches, and our results would otherwise most likely be the result of the pretrained vision-language components. Again, this can be considered as cheating and would drastically reduce the significance of our results.

Efficiency. The primary goal of our method is to reduce the costs of (pre-)training image-text retrieval models. Therefore, an obvious criterion is that our method is efficient in terms of parameter count, training data, and computational resources. It follows that our method should be significantly cheaper to train than existing approaches.

Performance. While the primary goal of our method is to reduce the costs of training image-text retrieval models, we still aim to achieve competitive performance with existing approaches, for example CLIP (Radford et al., 2021). However, since this work is **fully self-funded** and not backed by a large enterprise or research institution, we do not aim to reach state-of-the-art performance. Instead, the goal is to demonstrate that it is possible to achieve somewhat competitive performance with a fraction of the costs. What will come apparent when we present our results is that we neither reach the state-of-the-art performance on MSCOCO (Lin et al., 2014) and Flickr30K (Young et al., 2014) retrieval, as currently[1] held by BEiT-3 (Wang et al., 2023), nor do we aim to do so.

3.2 Architecture and Initialization

Our vision-language model consists of three components: a pretrained image encoder, a pretrained text encoder, and a randomly initialized shared encoder. The latter has to be randomly initialized to fulfill the criterion of independence to pretrained vision-language components. As their name suggests, the image encoder is responsible for encoding images, and the text encoder is responsible for encoding text. Therefore, they are specific to their respective modality, and we can therefore initialize them with pretrained unimodal models. For the image encoder, we use the pretrained Data2Vec2 (Baevski et al., 2023) image model, and for the text encoder, we use the pretrained BERT base (Devlin et al., 2019) model. Since each of these models is a 12-layer Transformer (Vaswani et al., 2017), which already has 86 million parameters, we only take the first 6 layers of each model to reduce the parameter count. Using other strategies like every second layer (e.g. 1, 3, 5, ...) leads to a worse performance in preliminary experiments. Note that both models have been trained self-supervised on image and text data, respectively, and therefore fulfill our criteria defined in the previous section. To keep the parameter count manageable, the shared encoder is merely a single Transformer layer and follows the ViT (Dosovitskiy et al., 2021) architecture. An overview of the architecture can be seen in Figure TODO.

Text Representation. Each caption/text is tokenized according to the BERT base uncased tokenizer (Devlin et al., 2019) and token ids are converted to embeddings using the BERT base model. The input $\mathbf{H}^{s}_{0,w} \in \mathbb{R}^{M+2 \times D}$ to the cropped BERT base model (our text encoder) is the sequence of token embeddings summed element-wise with the positional embeddings $\mathbf{T}^{pos}_{w} \in \mathbb{R}^{M+2 \times D}$ of BERT.

[1]As of September 2024.

$$\mathbf{H}_{0,w}^{s} = [\mathbf{h}_{0,w,\texttt{[T\_CLS]}}^{s}, \mathbf{h}_{0,w,1}^{s}, ..., \mathbf{h}_{0,w,M}^{s}, \mathbf{h}_{0,w,\texttt{[T\_SEP]}}^{s}] + \mathbf{T}_{w}^{pos}$$

Here the superscript s denotes that the representation stems from the student model, which will later be important for knowledge distillation. The subscript for a single token is of the form <layer, modality, token>, where 0 denotes the input to the first layer of the BERT base model. Correspondingly 1 denotes the output of the first layer and therefore the input to the second layer, and so on. The subscript w denotes the text modality. Inspired by Wang et al. (2023), we set the maximum sequence length M to 64 for efficiency, which means that we only utilize the first 64 pretraining positional embeddings of the BERT model. The special tokens `[T_CLS]` and `[T_SEP]` are taken directly from the BERT base model and are also pretrained. The notation is inspired by Bao et al. (2022).

Image Representation. Each image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ is patchified, flattened, and each resulting patch is projected into a D-dimensional embedding according to (Dosovitskiy et al., 2021). The parameter used for the patch projection stem directly from the pretrained Data2Vec2 (Baevski et al., 2023) image model. Since we make use of the ViT-B/16 architecture (Dosovitskiy et al., 2021), D equals 768, which also holds for the BERT base model. In all our experiments the image resolution is set to 224×224 pixels. Similar to the text representation, we define the image representation $\mathbf{H}_{0,v}^{s} \in \mathbb{R}^{N+1 \times D}$ as:

$$\mathbf{H}_{0,v}^{s} = [\mathbf{h}_{0,v,\texttt{[I\_CLS]}}^{s}, \mathbf{h}_{0,v,1}^{s}, ..., \mathbf{h}_{0,v,N}^{s}] + \mathbf{T}_{v}^{pos}$$

Here, the subscript v denotes the image modality, and N is the number of patches. The special token `[I_CLS]` is taken directly from the Data2Vec2 (Baevski et al., 2023) image model and is also pretrained. The positional embeddings $\mathbf{T}_{v}^{pos} \in \mathbb{R}^{N+1 \times D}$ are sinusoidal.

Forward Pass Let our cropped BERT base model be denoted by $f_w(\cdot)$, the cropped Data2Vec2 image model by $f_v(\cdot)$, and the shared encoder by $f_s(\cdot)$. Each image representation $\mathbf{H}_{0,v}^{s}$ and text representation $\mathbf{H}_{0,w}^{s}$ is first passed through the pretrained image and text encoder, respectively.

$$\mathbf{H}_{L,v}^{s} = f_v(\mathbf{H}_{0,v}^{s}) \quad \text{and} \quad \mathbf{H}_{L,w}^{s} = f_w(\mathbf{H}_{0,w}^{s})$$

Since both encoders have 6 layers, it holds that $L = 6$. After being passed through the encoders, the representations are passed separately through the shared encoder.

$$\mathbf{H}_{K,v}^{s} = f_s(\mathbf{H}_{L,v}^{s}) \quad \text{and} \quad \mathbf{H}_{K,w}^{s} = f_s(\mathbf{H}_{L,w}^{s})$$

Again, this is just one Transformer layer, so $K = L + 1 = 7$. The final representations for image and text are the representations of the `[I_CLS]` and `[T_CLS]` tokens. They are denoted by $\mathbf{h}_{K,v,\texttt{[I\_CLS]}}^{s}$ and $\mathbf{h}_{K,w,\texttt{[T\_CLS]}}^{s}$, respectively.

3.3 Contrastive Learning

An approach central, but not unique to our approach is the use of a contrastive loss to align image and text. We use the contrastive loss as presented by Radford et al. (2021), and follow the approach of Bao et al. (2022) by gathering negative examples from all GPUs to increase the effectiveness of the contrastive loss. We use the representations $\mathbf{h}_{K,v,\texttt{[I\_CLS]}}^{s}$ and $\mathbf{h}_{K,w,\texttt{[T\_CLS]}}^{s}$ for image and text, respectively, which are normalized before computing the cosine similarity between all possible pairs of image and text in the current batch. For equations that formulate the contrastive loss, we refer to Bao et al. (2022).

3.4 Self-Supervised Knowledge Distillation

What makes our approach unique is the use of knowledge distillation to guide the alignment between image and text. We use BEiTv2 (Peng et al., 2022) as the teacher model. For each image-text pair in a batch, we pass the image to BEiTv2 and extract the representations of the `[I_CLS]` token, denoted by $\mathbf{h}^t_{K,v,\texttt{[I\_CLS]}}$, from the last layer. Since BEiTv2 acts as the teacher model, we add the subscript t to the representation.

Contrastive Distillation. Unlike Aytar et al. (2017), our teacher is self-supervised, so we cannot use the kl-divergence loss on the probability distributions of the ImageNet-1K (Russakovsky et al., 2015) classes. Instead, we perform a contrastive loss between the student and teacher representations. Let $f_p(\cdot)$ denote a projection head, which is a single linear layer. The image representations and text representations, created by our vision-language, are passed through the projection head and normalized.

$$\mathbf{z}^s_v = ||f_p(\mathbf{h}^s_{K,v,\texttt{[I\_CLS]}})||_2 \quad \text{and} \quad \mathbf{z}^s_w = ||f_p(\mathbf{h}^s_{K,w,\texttt{[T\_CLS]}})||_2$$

We then perform the contrastive loss once between the image representations of the student $\mathbf{z}^s_v$ and teacher $\mathbf{z}^t_v$, and once between the text representations of the student $\mathbf{z}^s_w$ and the image representations of the teacher $\mathbf{z}^t_w$.

Memory Bank.

Note that we use a self-supervised vision teacher model in order to fulfill both the criteria of being self-supervised and independent of pretrained vision-language components.

The final training objective is given by:

$$\min \mathcal{L}_{cl} + \mathcal{L}_{kd}$$

3.5 Pretraining Setup

4 Results

4.1 Image-Text Retrieval

4.2 Image Classification

4.3 Text Classification

4.4 Ablation Studies

5 Limitations and Future Work

- mome arch

6 Conclusion

Broader Impact Statement

Acknowledgments

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`.

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning*, volume 202, pp. 1416–1429, Honolulu, Hawaii, USA, July 2023.

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.

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 *9th International Conference on Learning Representations*. OpenReview.net, May 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, 2014. ISBN 978-3-319-10602-1.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *CoRR*, abs/2208.06366, 2022. URL `https://arxiv.org/abs/2208.06366`.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

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`.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, February 2014.

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

A.1 Discussion on Curated Data