Enhancing CLIP with a Third Modality

Efthymios Tsaprazlis
National Technical University of Athens
Electrical and Computer Engineering
etsaprazlis@gmail.com

Georgios Smyrnis
University of Texas at Austin
Electrical and Computer Engineering
gsmyrnis@utexas.edu

Alexandros G. Dimakis
University of Texas at Austin
Electrical and Computer Engineering
dimakis@austin.utexas.edu

Petros Maragos
National Technical University of Athens
Electrical and Computer Engineering
maragos@cs.ntua.gr

Abstract

We study the problem of training a third tower for a new modality given a pre-trained CLIP model. This extra part of the architecture can be used to incorporate other modalities in the model pipeline. In our setting, we consider the use of a model such as BLIP-2, which provides us with a dialogue centered around the image. We evaluate our model in the setting of image and text retrieval, and compare it against the regular image and text based one.

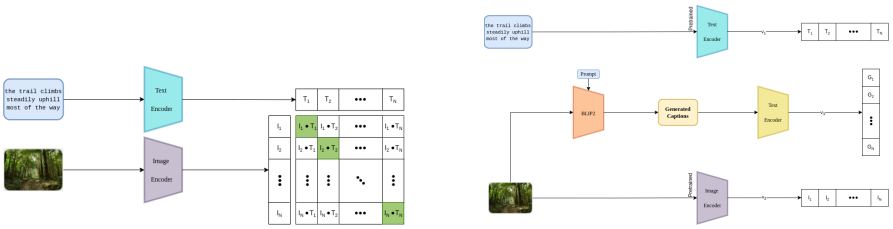
1 Introduction

Image-text models have become fundamental in machine learning, giving rise to several state-of-the-art architectures, such as CLIP, DALL-E and Stable Diffusion, among others [9, 10, 11, 12]. These foundational models can be used in a variety of tasks, usually different than the one they were trained on. This is because they use the each modality to infer knowledge for the former, therefore allowing them to operate without having seen data for the specific task at hand. At the same time, these architectures are also very costly to train. Training is often done on millions of image-text pairs. As such, using these large foundational models often becomes a task of **finetuning** them, rather than training them from scratch. These models can also be used simply for inference, and use them as the starting point around which a larger pipeline can be constructed.

The motivation behind our work is derived from computer vision and the use of additional views [7, 16] of the same object in order to obtain more information about its characteristics. This idea was adapted on contrastive learning by CMC [15], where additional sensory views of the same image where used in order to enhance contrastive training. In the same way, in the case of image-text pairs we find an additional view in dialogues generated by generative models. This extra synthetic data will still be text but of a different nature compared to the captions that were used for the text encoder. We thus consider auxilliary textual datasets like metadata, dialogues about an image or product details obtained from a database to be additional modalities.

Within this context, we examine the use of a **third tower** in these image-text architectures. The addition of the third tower can be seen in Figure 1. We focus our study on a CLIP model trained by OpenCLIP [4], which we augment via an additional encoder, resulting in our **CLIP-3Modal** architecture. This encoder serves to add additional modalities to the input, in a way that is consistent with the paradigm of reusing elements from foundation models. In contrast to previous works that examine the use of a third tower in the architecture [5], we explicitly consider the third tower as operating on a different modality, aside from the usual image and text ones. In our setting, we consider the additional modality to be the dialogue of a user with an image-captioning model such as

37th Conference on Neural Information Processing Systems (NeurIPS 2023).



(a) CLIP (b) CLIP-3Modal

Figure 1: **Our proposed CLIP-3Modal architecture.** We propose incorporating a third tower in the CLIP model, which extends the existing image and text ones. This extra tower can be used during the evaluation of the model, along with the existing ones.

BLIP-2 [6]. With this, we aim to augment the information from our input data with the outputs of a foundation model. We evaluate the performance of this addition to the CLIP architecture via retrieval performance on an image-text retrieval task.

2 Related Work

2.1 Self Supervision and Contrastive Learning

Contrastive learning, as a subset of self supervision, has been a major technique in recent state-of-the-art machine learning models, such as CLIP [9]. It relies on the use of multiple semantically related samples, whose representations are then made as similar as possible. These are called positive samples in this context, and are often derived from the same underlying sample, under different transformations. Aside from views of the same sample, contrastive methods also make use of negative samples, which are considered to be semantically different from the original one. These are often obtained from random samples within the same batch [2] or from a past history of samples from the model [3]. By using these negative samples, the model is able to improve its representations, by learning to distinguish between samples which are different to each other. This form of representation learning has become immensely popular in the context of image-text models, being an integral element of several state-of-the-art works in this setting [9, 17]. An extension to contrastive learning that is related to our work is that of [5], which proposes the use of an additional locked third tower, which contains pretrained image embeddings. While related, this work is different from ours in that we also augment the samples with an additional modality, which serves as input to the third tower of the architecture.

2.2 Image Captioning

Image captioning is among the key uses of foundation models that employ both image and text models. In recent years, two models have been prominent in the context of image captioning. The first is Flamingo [1], a model which operates by interleaving image and text tokens. This allows the model to perform captioning with a guided prompt, by combining the image tokens with those of the prompt into a single sequence. The second one is BLIP-2 [6], which uses an extra transformer module (called a querying transformer) to combine the image and text modalities of the underlying models, enabling image captioning by passing information from the visual module to the language model. This follows the paradigm of foundation models, in that parts of the large pretrained models can be used as modules in architectures for a variety of tasks, with only a small part added to enable interaction between the two. In what follows, we shall use BLIP-2 to generate our additional modality for training.

3 Method

3.1 Using BLIP-2 for Captions

To incorporate the third tower into our architecture, we need to include a third modality in our input data. To do this, we use BLIP-2 [6] to expand an image text dataset with an additional modality. We

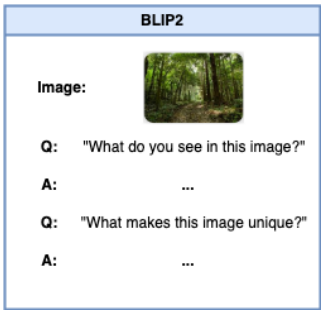


Figure 2: **Use of BLIP-2 model.** The questions with which the model is prompted provide a base form of dialogue for use as our third modality.

choose CC3M [14] as the dataset that we enrich with BLIP-2 generated captions. For each image, we provide it as input to the BLIP-2 model. We then provide the following two questions, in sequence, as our prompt: (a) "What do you see in this image?" and (b) "What makes this image unique?". This can also be seen in Figure 2. This pair of questions provides a basic form of dialogue between the model and a user. Moreover, when providing the second question to the captioning model, we also give it the response to the first question as input. This allows for the response to the second question to be contextualized by the model by its own response to the first one.

We note here that the model that generates the third modality does not take the text from the image-text pairs as input. This means that, despite the third modality being provided to the training process in text format, the actual content is not directly dependent on the existing caption of the image. This allows the BLIP-2 model to provide new information about this particular sample.

3.2 Training the Third Tower

We now aim to train a CLIP architecture that incorporates a third tower in its construction, using BLIP-2 captions as input. To do this, we start from a pretrained CLIP model provided by OpenCLIP. The image and text towers from this model become the image and text towers for our architecture as well. To construct the third tower of our architecture, we start by making a copy of the pretrained text tower. We then freeze the original image and text towers, and train only the third tower on our extended CC3M dataset. Our loss function is similar to the one used in regular CLIP training:

$$\begin{aligned}\mathcal{L} = &-\frac{a}{N}\sum_{i=1}^{N}\log\frac{\exp(I(x_i)^T G(z_i)/\tau)}{\sum_{j=1}^{N}\exp(I(x_j)^T G(z_i)/\tau)} \\ &-\frac{(1-a)}{N}\sum_{i=1}^{N}\log\frac{\exp(T(y_i)^T G(z_i)/\tau)}{\sum_{j=1}^{N}\exp(T(y_j)^T G(z_i)/\tau)},\ a\in[0,1]\end{aligned} \tag{1}$$

where (x_i, y_i, z_i) is one of our samples and I, T, G are our image, text, and generated dialogue towers respectively. In the above, a is a blending hyperparameter between the two losses. Intuitively, we want a to be high enough to encourage proper behavior of the generated caption representations with respect to the corresponding image representation. At the same time, we don't want too high value of a, since this will just lead to simply replacing the pretrained text encoder with the third tower. Careful assignment of parameter a can lead to the third tower taking into account both original modalities.

4 Experimental Results

4.1 Evaluation Method

For the evaluation of CLIP-3Modal we focused on zero-shot retrieval tasks between image and text due to CLIP. We used the ViT-B-32 architecture of CLIP provided by OpenCLIP as the baseline for the evaluation. This is the same model used as the foundation of CLIP-3Modal. The aforementioned pretrained model was trained on 32 billion samples of the LAION-2B dataset [13]. This provides us with a good initial point for our third encoder. For the weighting hyperparameter a in our loss

	Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
OpenCLIP	39.7%	65.4%	75.6%	56.3%	79.8%	87.1%
CLIP-3Modal	**40.2%**	**65.9%**	**75.9%**	**57.0%**	**80.6%**	**87.5%**

Table 1: **CLIP-3Modal improves recall on every zero-shot retrieval task.** Both models are using ViT-B-32 architecture and have been pretrained (their image and text encoders) on 32 billion samples from LAION-2B dataset. Evaluation is done on MSCOCO. We see that our model outperforms the baseline OpenCLIP one.

	Image Zero-Shot Retrieval			Text Zero-Shot Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Baseline (β = 1)	39.7%	65.4%	75.6%	56.3%	79.8%	87.1%
β = 0.95	39.9%	65.7%	75.8%	**57.1%**	80.4%	87.3%
β = 0.90	**40.2%**	**65.9%**	**75.9%**	57.0%	**80.6%**	**87.5%**
β = 0.80	40.0%	65.6%	75.8%	56.3%	79.6%	86.2%
β = 0.60	38.7%	63.8%	73.8%	53.7%	75.6%	84.8%

Table 2: **High values of the blending parameter improve performance, while smaller drop the evaluation scores lower than the baseline**. In this figure β denotes the blending weight hyperparameter. Smaller weight on the generated captions encoder benefits the fused embeddings by preserving their initial information and enhancing them with different aspects of the input. The best results were occurred for $\beta = 0.9$.

function, we use $a = 0.65$. We observed that overall our model performs the best when the weight of the loss between image and generated captions is higher than 0.5.

After training our model, we fuse the output embeddings of the text and the generated captions encoders, to obtain a final embedding for the text. For the fusion of the outputs we take a weighted sum of the embeddings provided by the text and the generated captions towers:

$$X_{ensemble} = \beta \cdot T + (1 - \beta) \cdot G,\ \beta \in [0, 1] \tag{2}$$

where T and G are the output embeddings of the original text and the generated captions tower. By weighting the two outputs appropriately, the third tower will incorporate additional information that enhances the predictions learned by the original model. In this case, we use a blending parameter of $\beta = 0.9$. The insight for this high value of β is that we want the embeddings from the third tower to influence the output, but we still want the changes to be small so as to not lose their initial valuable information. An ablation study on the blending parameter is presented in Table 2.

4.2 Results

We trained our third tower as described above, using a ViT-B-32 based CLIP pretrained on LAION-2B (provided by OpenCLIP) as our foundation. For our training, we used batch size 1024, learning rate 10^{-5}, weight decay 0.1 and trained our model for 2 epochs on our custom dataset. The training of the third tower takes approximately 1 hour per epoch on a single GPU, which is significantly less than the training time of the foundation model. We based our evaluation on Microsoft COCO [8] by studying the zero-shot retrieval performance on this specific dataset, using the same evaluation metric provided by OpenCLIP. We managed to outperform the OpenCLIP's model in both image and text zero-shot retrieval, with a margin of 0.3% to 0.8%. More details are presented in Table 1.

5 Conclusion

We can see here that the use of an extra modality is a viable way to improve upon an image and text model. The additional tower which expands the architecture of the model is a way to incorporate additional modalities into the training process, which can be used to extend existing image-text models. In the future, we aim to further analyze the use of a third tower in CLIP style models, as well as examine alternative choices for the third modality.

Acknowledgements

This research has been supported by NSF Grants AF 1901292, CNS 2148141, Tripods CCF 1934932, IFML CCF 2019844 and research gifts by Western Digital, Amazon, WNCG IAP, UT Austin Machine Learning Lab (MLL), Cisco and the Stanly P. Finch Centennial Professorship in Engineering. This work was also supported by the Onassis Foundation - Scholarship ID: F ZS 056-1/2022-2023.

References

[1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[3] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[4] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt. Openclip, July 2021. URL `https://doi.org/10.5281/zenodo.5143773`.

[5] J. Kossen, M. Collier, B. Mustafa, X. Wang, X. Zhai, L. Beyer, A. Steiner, J. Berent, R. Jenatton, and E. Kokiopoulou. Three towers: Flexible contrastive learning with pretrained image models. *arXiv preprint arXiv:2305.16999*, 2023.

[6] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023.

[7] Y. Li, M. Yang, and Z. Zhang. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1863–1883, oct 2019. doi: 10.1109/tkde.2018.2872063. URL `https://doi.org/10.1109%2Ftkde.2018.2872063`.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[10] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[11] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[13] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

[14] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

[15] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 776–794, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58621-8.

[16] C. Xu, D. Tao, and C. Xu. A survey on multi-view learning, 2013.

[17] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. 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.