Multimodal Distillation of CLIP Models

Georgios Smyrnis **Sriram Ravula** **Sujay Sanghavi** **Alexandros G. Dimakis**
Chandra Family Department of Electrical and Computer Engineering
University of Texas at Austin

Abstract

CLIP-style models are powerful tools for zero-shot image-text tasks, but contain a very large number of parameters, making them expensive to deploy in hardware-constrained settings. We introduce a novel way to distill these large CLIP-based models into significantly smaller ones. Our method is called multimodal distillation because we jointly train two student networks (operating on image and text) from two teacher networks. Our loss tries to preserve the structure of the embeddings of the dataset, as provided by the image and text teacher networks. We are thus able to extract information from the interaction of the teacher embeddings, improving performance on downstream classification tasks.

1 Introduction

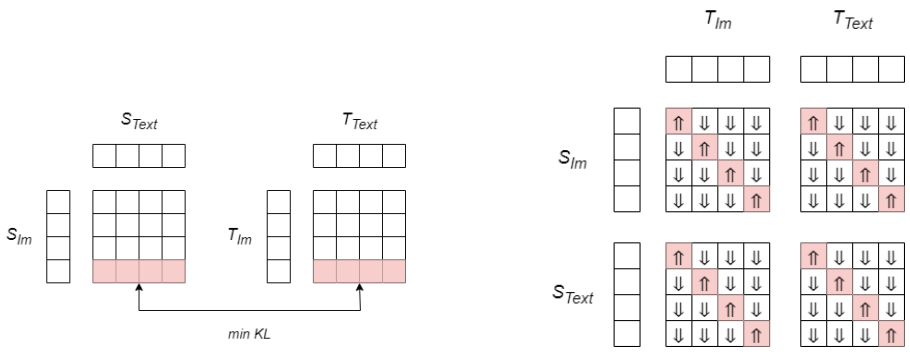
CLIP-style models are among the most prominent image-text models currently in use. These models operate on paired input data, images as well as associated text. Each modality is encoded separately, and the model is trained in a **contrastive** fashion: embeddings that correspond to the same image-text pair are encouraged to have a large inner product, while embeddings that correspond to different image-text pairs are encouraged to have as small of an inner product as possible. This allows for **zero-shot** evaluation, without the use of downstream training images. This allows the model to achieve state of the art performance in various machine learning tasks, without ever having access to the data they are evaluated on. For these models, it is often the case that they require a very large number of parameters to achieve good results, which might make them difficult to use in practice.

Knowledge Distillation (KD) works by distilling information from a pretrained, well-performing teacher model to a (usually smaller) student network. This technique aims to improve the quality of the resulting student, yielding better performance compared to only training on a dataset without the extra information provided by the teacher. Smaller student networks also reduce inference compute and memory requirements and can be deployed on smaller edge devices [2, 3, 9, 25].

In our work, we propose a method for distillation of CLIP style models that aims to circumvent two assumptions commonly made by knowledge distillation techniques:

- The first is that the student model operates on one modality (for example, a vision-based student model learning from a similar vision-based teacher). In our setting, we examine a way to distill information from a model that operates on multiple modalities to a similarly multimodal student. Our method distills both modalities at the same time, therefore informing each modality from the other one.
- The second is that the student model is assumed to have knowledge of the downstream task, most commonly by using distillation for a supervised task. In this work, we shall examine the use of distillation for a setting where the student model is also used for a variety of tasks via zero-shot evaluation. This way, the student aims to be equally diverse to the teacher.

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



(a) Baseline distillation method. (b) Multimodal distillation method.

Figure 1: **Comparison between baseline distillation method and our proposed method.** The baseline distillation method minimizes the KL divergence of the distributions created by the embeddings of the student and the teacher. In contrast, our method performs contrastive learning on each of the student and teacher modality pairs, in order to obtain information from the embeddings directly. Inner products of embeddings on the diagonals are maximized, while the rest are minimized.

2 Related Work

2.1 Contrastive Learning

Contrastive learning is a form of self-supervision that works by contrasting samples in the input data. The aim of techniques under this categorization is to learn representations that are aligned for similar inputs. This is often done by performing augmentations on the data, creating multiple views from a single sample whose representations the model wants to align. These also need to be as varied as possible, so that views from different images have different representations. This dual operation of alignment and uniformity of representations learned using contrastive learning is a known result of one of its standard training objectives [19]. Overall, contrastive learning has become very popular in recent years, with multiple works demonstrating various learning objectives which exhibit these desired properties [4, 5, 11, 24]. The usage of multiple views is not restricted to contrastive learning, as there also exist methods that make use of the positive samples without relying on negatives [1, 10].

Contrastive learning is an excellent choice in the multimodal setting, where models operate on paired datasets. Each sample in the data has one view per modality, all of them containing different but related information about the sample in question. This has led to many state-of-the-art image-text models that use contrastive learning or architectures trained with it as part of their pipeline [14, 15, 16]. In what follows, we shall focus on CLIP in particular, an architecture trained to learn both image and text representations from unlabeled data. Recently, via OpenCLIP [13] it has become possible to reproduce this CLIP training [7], allowing further exploration into CLIP-style models.

2.2 Knowledge Distillation

Knowledge distillation [12] is a technique that allows the distillation of the information from a pretrained teacher model to a student one. Traditionally, this technique works by training the student to match the soft labels assigned by the teacher. Recently, contrastive learning has proven to be useful in this setting as well, as it can be used in the form of contrastive representation distillation [18], to obtain results which are better than regular knowledge distillation in the unimodal setting.

In the setting of image-text models, the study of distillation of large teacher models to smaller ones has been limited. While there are a few studies that examine this technique [8, 20, 21], they rely on cross-modal transformers and architectures which provide only a single representation per sample, instead of one per modality. More importantly, they also incorporate the downstream tasks as part of their complete distillation pipelines. This is somewhat limiting, given that we want to distill an all-purpose teacher model into an equally all-purpose student one. In contrast to that, there exist a pair of works [6, 22] that use distillation in a way that is model and downstream task agnostic (for the special case of self distillation), by considering the logits of the teacher in the pretraining stage

(given distributions of images and texts within the batch). A concurrent work to ours [23] also uses this loss function as well as weight inheritance to distill CLIP-style models. Our work will examine a similarly general approach, where two teachers are jointly distilled into two students.

3 Method

Let us assume that we have a CLIP-style teacher model which consists of two encoders, $T_{im} : \mathcal{X} \to \mathbb{R}^{d_T}$ and $T_{text} : \mathcal{Y} \to \mathbb{R}^{d_T}$. These two encoders map an image $x \in \mathcal{X}$ and a text $y \in \mathcal{Y}$ to a common representation space. Our goal is to learn a smaller student model, consisting of encoders $S_{im} : \mathcal{X} \to \mathbb{R}^{d_S}$ and $S_{text} : \mathcal{Y} \to \mathbb{R}^{d_S}$, by leveraging information from the teacher model during training. Normally, CLIP training on the student is done via the following loss:

$$\mathcal{L}_{CLIP} = -\frac{1}{2N}\sum_{i=1}^{N}\big(\log\sigma(S_{im}(x), S_{text}(y); i, \tau) + \log\sigma(S_{text}(y), S_{im}(x); i, \tau)\big), \tag{1}$$

$$\sigma(f(a), g(b); i, \tau) = \frac{\exp(f(a_i)^T g(b_i)/\tau)}{\sum_{j=1}^{N}\exp(f(a_i)^T g(b_j)/\tau)} \tag{2}$$

Minimizing this loss function enables the student image and text encoders to learn similar representations for the correct image and text pairs, and different ones for the rest. The loss function can be extended by including an additional term which allows for distillation between the teacher and the student. In other words, our loss becomes $\mathcal{L} = \mathcal{L}_{CLIP} + \mathcal{L}_{Distill}$. The key question then becomes how to formulate $\mathcal{L}_{Distill}$ properly, so that the we obtain relevant information from the teacher.

One way to do this is to view the embeddings for each batch as a distribution of the images across the possible texts and vice-versa. This means that, for each image x_i in the batch, we can assign a distribution $\boldsymbol{z}_{S_{im},i} = \underset{j=1,\dots,N}{\text{softmax}}\left(S_{im}(x_i)^T S_{text}(y_j)\right)$ over the texts in the batch. Similarly, for each text y_j we can define a distribution $\boldsymbol{z}_{S_{text},j} = \underset{i=1,\dots,N}{\text{softmax}}\left(S_{text}(y_j)^T S_{im}(x_i)\right)$ over the images. We can define distributions $\boldsymbol{z}_{T_{im},i}$ and $\boldsymbol{z}_{T_{text},j}$ in the same way. The term below is then added to the loss:

$$\mathcal{L}_{Distill} = \mathcal{L}_{KD} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} z_{T_{im},ij}\log z_{S_{im},ij} - \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{N} z_{T_{text},ji}\log z_{S_{text},ji} \tag{3}$$

This method has been presented in [6] and [22]. There, it was used in the context of self-distillation (where the teacher model is a previous version of the student). Visually, this operation can be seen in Figure 1a. This extra term works by minimizing the KL divergence between the distributions corresponding to the rows and columns of the matrix in Figure 1a. This term essentially encourages the student to assign images across the texts in the batch as similarly to the teacher as it can.

In the previous method, distillation is performed at the level of predictions, instead of using the embeddings directly. This is somewhat of an extraneous task, since at this point there is no inherent classification task that we want our student to perform. Rather, the implicit goal is to actually learn the representations of the teacher, in a way that is useful for downstream classification. With that in mind, we propose an alternative loss that directly operates on the embedding space of the teacher and the student. Our **Multimodal Distillation** loss is $\mathcal{L}_{Distill} = \mathcal{L}_{MM}$, where:

$$\begin{aligned}\mathcal{L}_{MM} = -\frac{1}{N}\sum_{i=1}^{N}\big(&\log\sigma(S_{im}(x), W_{im}T_{im}(x); \tau) + \log\sigma(S_{im}(x), W_{text}T_{text}(y); \tau) + \\ &\log\sigma(S_{text}(y), W_{im}T_{im}(x); \tau) + \log\sigma(S_{text}(y), W_{text}T_{text}(y); \tau)\big)\end{aligned} \tag{4}$$

In the above, W_{im} and W_{text} are $d_S \times d_T$ matrices, learnt during training and then discarded. These matrices project the teacher embeddings to the student embedding space if needed, as it is often the case that $d_T > d_S$. The above loss can be seen visually in Figure 1b. This loss function operates in a way similar to Contrastive Distillation [18], applied in the setting of multiple modalities. We consider all possible combinations of modalities for samples in the batch. The above loss encourages similarity between student and teacher embeddings of the same sample across both modalities. At the same time, it decreases the similarity of embeddings of different samples. This is precisely the goal of contrastive learning and general, making our method better related to the paradigm that is used to train these image-text models. This encourages the student to preserve the structure learned by the teacher, which is the key aspect of image-text training.

Table 1: **Results on ViT-S-32, training for 150 million samples.** We compare our method to the baseline knowledge distillation approach [6], as described in Section 3. We can see that we have a benefit in ImageNet-1K Zero-Shot accuracy.

Method	ImageNet-1K Zero-Shot Accuracy
No Distillation	24.35%
Knowledge Distillation	29.50%
Multimodal Distillation	**30.18%**

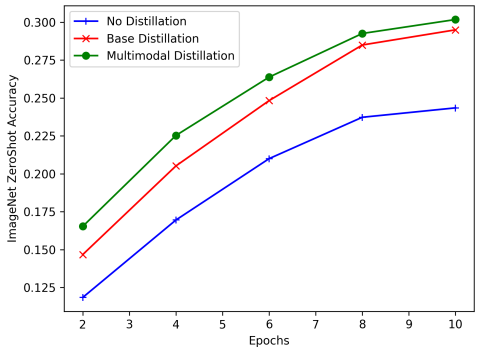


Figure 2: **Performance of methods across training.** We see that our loss outperforms the baseline across training, with the difference being higher at the earlier stages of training.

4 Experiments

For our experiments, we trained a CLIP model using the above distillation losses. We used the OpenCLIP implementation of CLIP training [13]. For the teacher model, we used a ViT-L/14 based CLIP model, trained by OpenAI. This is a well performing teacher model, achieving 75.3% accuracy on ImageNet. Note that to achieve this accuracy, the model was trained on a very large amount of data, a setting that we don't match in this work due to compute limitations. For the student model, we used a ViT-S/32 based model, as defined in the OpenCLIP repository. Our teacher model has 428 million parameters while the student has 63 million, which correspond to a major reduction in the theoretical expressive power of the network. Training was done on an HPC cluster, using a single node with 3 A100s GPUs. Training is done for 10 epochs on the YFCC15M dataset, a 15 million subset of YFCC100M [17] defined in the CLIP repository. We use a batch size of 1024 per GPU, and a learning rate of 0.0005.

We evaluate our models on their ImageNet-1K accuracy. Evaluation is done in a zero-shot way (so we do not train a linear classifier on top of the representations). Results can be seen in Table 1. We can see that all distillation methods outperform pure CLIP training, and that our method performs better than using the KD based variant or distilling each modality separately.

Moreover, in Figure 2 we see that our method outperfoms the baseline Knowledge Distillation approach across training, with the benefit being higher earlier in the training process. This shows the benefit of our method, when used for shorter training, if fewer resources are available.

5 Conclusion

We have presented a method to distill a large CLIP model to a smaller one, in a way that prioritizes similarity between the embeddings of the teacher and the student. We see that there is a benefit over using the logits of the teacher within the batch. In the future, we aim to expand on these techniques and further examine the capabilities of our method to create all-purpose students, without access to downstream tasks.

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. This work was also supported by NSF Tripods Grant 2217069 and a research award from Amazon.

References

[1] A. Bardes, J. Ponce, and Y. LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=xm6YD62D1Ub`.

[2] K. Bhardwaj, N. Suda, and R. Marculescu. Dream distillation: A data-independent model compression framework. *arXiv preprint arXiv:1905.07072*, 2019.

[3] K. Bhardwaj, N. Suda, and R. Marculescu. Edgeal: A vision for deep learning in the iot era. *IEEE Design & Test*, 38(4):37–43, 2019.

[4] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[5] 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.

[6] R. Cheng, B. Wu, P. Zhang, P. Vajda, and J. E. Gonzalez. Data-efficient language-supervised zero-shot learning with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3119–3124, June 2021.

[7] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[8] Z. Fang, J. Wang, X. Hu, L. Wang, Y. Yang, and Z. Liu. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1428–1438, 2021.

[9] J. Gou, B. Yu, S. J. Maybank, and D. Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

[10] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33: 21271–21284, 2020.

[11] 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.

[12] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[13] 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`.

[14] 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.

[15] 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.

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

[17] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Commun. ACM*, 59(2):64–73, jan 2016. ISSN 0001-0782. doi: 10.1145/2812802. URL https://doi.org/10.1145/2812802.

[18] Y. Tian, D. Krishnan, and P. Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkgpBJrtvS.

[19] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*. PMLR, 2020.

[20] T. Wang, W. Zhou, Y. Zeng, and X. Zhang. Efficientvlm: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning. *arXiv preprint arXiv:2210.07795*, 2022.

[21] Z. Wang, W. Wang, H. Zhu, M. Liu, B. Qin, and F. Wei. Distilled dual-encoder model for vision-language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8901–8913, 2022.

[22] B. Wu, R. Cheng, P. Zhang, T. Gao, J. E. Gonzalez, and P. Vajda. Data efficient language-supervised zero-shot recognition with optimal transport distillation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=G89-1yZLFHk.

[23] K. Wu, H. Peng, Z. Zhou, B. Xiao, M. Liu, L. Yuan, H. Xuan, M. Valenzuela, X. S. Chen, X. Wang, et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21970–21980, 2023.

[24] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[25] Y. Zhang, H. Chen, X. Chen, Y. Deng, C. Xu, and Y. Wang. Data-free knowledge distillation for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7852–7861, 2021.