CLIP-Embed-KD: Computationally Efficient Knowledge Distillation Using Embeddings as Teachers

Lakshmi Nair
Georgia Institute of Technology, Atlanta, GA 30332, USA

***Abstract*—This extended abstract investigates the application of *Contrastive Language-Image Pre-training* (*CLIP*) for efficient knowledge distillation, by utilizing embeddings as teachers. Typical knowledge distillation frameworks require running forward passes through a teacher model, which is often prohibitive in the case of billion or trillion parameter teachers. Our initial findings show that using only the embeddings of the teacher models to guide distillation, can outperform full-scale knowledge distillation using $9\times$ less memory and $8\times$ less training time.**

I. Introduction

Contrastive Language-Image Pre-training (CLIP) [1], involves pre-training an image encoder along with a text encoder to predict image-text pairings within a dataset. CLIP utilizes a constrastive objective function that computes scaled pair-wise cosine similarity between image embeddings (of the image encoder) and text embeddings (of the text encoder) to generate output logits. When training, the logits are compared to labels that match input images to their corresponding ground truth text embeddings, via a cross entropy loss. This alignment of image and text modalities enables CLIP to achieve robust zero-shot performances at image classification.

We explore the application of CLIP for *computationally efficient* knowledge distillation (KD) using only teacher embeddings. Knowledge distillation is the process of transferring (or *distilling*) the knowledge of a larger teacher model into a smaller, more compressed student model, by comparing the outputs of the teacher and student models using a *distillation loss*. Existing KD approaches require performing several forward passes through both the teacher and student models for comparing the corresponding outputs [2]. This can be prohibitive when the teacher models are billion parameters in size, and we wish to run KD on limited computational resources (e.g., mobile devices). Intuitively, aligning the teacher and student feature maps can be likened to the alignment of text and image embeddings with CLIP. Hence, we investigate: ***Can pre-computed embeddings obtained from the teacher model be used to train the student model in knowledge distillation?***

II. CLIP-Embed-KD

Our approach is shown in Figure 1. We begin by randomly sampling N samples of data for each class in the dataset. We obtain teacher embeddings of the [CLS] token (used in models like ViT [3]) for the sampled data and compute a cumulative representation of each class' embeddings by averaging the collected embeddings of the class along the embedding dimension. This gives a pre-computed, *"averaged"* embedding for each class, representative of the teacher embeddings for that class. These embeddings are projected into the embedding dimensions of the student model through a learnable projection layer. We normalize the resultant embeddings and compute the dot product of the normalized teacher and student embeddings. We pass the resultant dot product with ground truth labels into a cross entropy loss ($\mathcal{L}_{clip}$). The ground truth represents a one-hot encoding of the labels for each sample in the batch. The distillation loss is a weighted combination of cross entropy loss ($\mathcal{L}_{CE}$) of the student logits and ground truth labels (weighted by α_1), and $\mathcal{L}_{clip}$ (weighted by α_2). We use $\alpha_1 = 0.5$ and $\alpha_2 = (1 - \alpha_1) = 0.5$ in our experiments. Our detailed code is available at: https://github.com/lnairGT/CLIP-Distillation/.

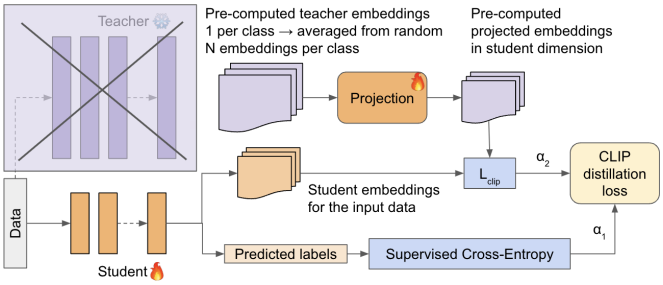

Fig. 1. **CLIP-Embed-KD** uses pre-computed teacher embeddings thus avoiding the need to run forward passes through the teacher for every sample.

III. Results

We evaluate our method using Vision Transformers (ViT) [3], on CIFAR100 image classification. Teacher models are HuggingFace ViT checkpoints (architectures in Table I). We compare the computational efficiency of CLIP-Embed-KD using teacher embeddings to baseline KD using the full teacher model (referred to as CLIP-Teacher-KD). CLIP-Teacher-KD computes embeddings for each input (no averaged embeddings used), for computing $\mathcal{L}_{clip}$. The teacher models use patch sizes 16, 32 and student models use patch size 4. We use batch size of 64. We compute averaged embeddings over $N = 100$. We train for 200 epochs, with a learning rate of 0.0001.

Table II compares the accuracy of CLIP-Teacher-KD vs. CLIP-Embed-KD for different teacher sizes and patch sizes, given the base student model. We note that for CLIP-Embed-

TABLE I
ARCHITECTURE SPECIFICATION OF THE MODELS.

Student	Layers	Embed dim	Heads	MLP
Base	6	256	8	1024
Large	10	512	8	2048
Teacher	**Layers**	**Embed dim**	**Heads**	**MLP**
Base-16/32	12	768	12	3072
Large-16/32	24	1024	16	4096

TABLE II
PERFORMANCE ON BASE STUDENT WITH IMAGE 32×32.

Teacher	CLIP-Embed-KD	CLIP-Teacher-KD
Base-16	49.32	51.68
Large-16	49.67	51.92
Base-32	49.34 ($17\times \downarrow$ mem)	52.61
Large-32	**50.33** ($59\times \downarrow$ mem)	51.95

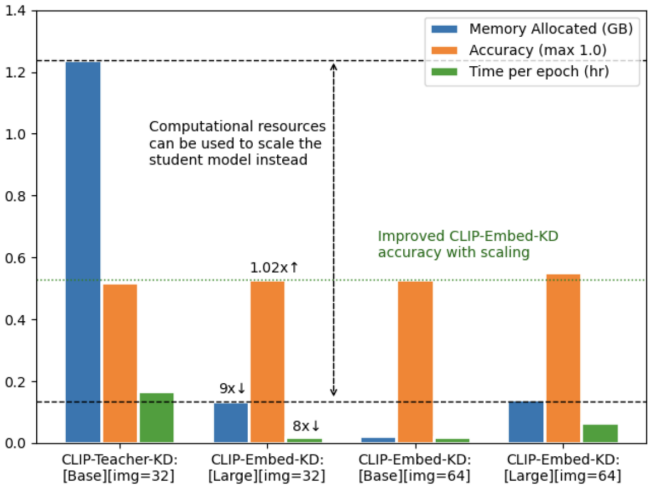

Fig. 2. Training resource utilization (with Large-32 teacher) of CLIP-Embed-KD, CLIP-Teacher-KD [student-size][image-size]: CLIP-Embed-KD scales well (to larger models and larger images) to outperform CLIP-Teacher-KD for much less memory.

KD, the final accuracy of the student seems to have a small improvement when using the larger teacher models over the base ones. This pattern is not explicit with CLIP-Teacher-KD where the base teacher model has a slightly improved student accuracy compared to the large teacher at patch size 32. Since CLIP-Embed-KD uses pre-computed teacher embeddings to train the student, larger teacher sizes potentially contribute to improved quality of the average embeddings.

CLIP-Embed-KD is computationally more resource efficient

TABLE III
SCALING CLIP-EMBED-KD TO LARGER MODELS AND IMAGES.

Student	Teacher	Image sz	CLIP-Embed-KD
Base		64	53.28
Large	Large-16	32	52.92
Large		64	**54.36**
Base		64	53.40
Large	Large-32	32	52.77
Large		64	**54.92**

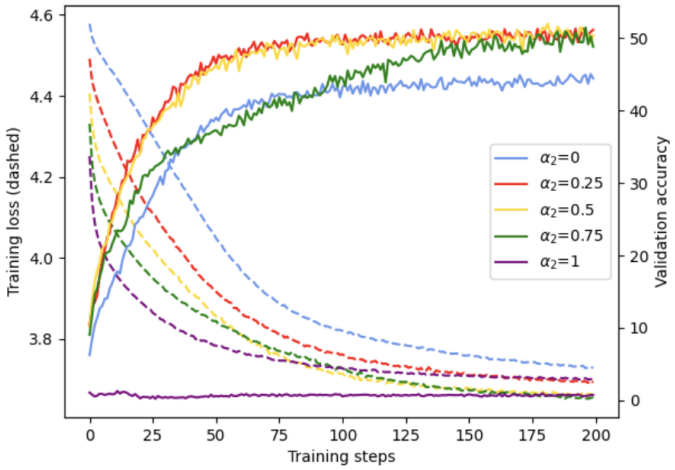

Fig. 3. Accuracy vs. α_2 ($\alpha_1 = 1 - \alpha_2$): $\alpha_2 = 0$ is the typical supervised learning that uses cross-entropy loss with ground truth labels (no $\mathcal{L}_{clip}$).

than CLIP-Teacher-KD, and can scale better to outperform CLIP-Teacher-KD. In Table II, CLIP-Embed-KD achieves roughly about $\approx$2% lesser accuracy than CLIP-Teacher-KD, since the averaged embeddings result in some loss of information per sample compared to CLIP-Teacher-KD. However, CLIP-Teacher-KD uses more memory as teacher size grows, whereas, CLIP-Embed-KD uses fixed memory for embeddings alone, leading to improved scaling behavior ($17\times$, $59\times$ less memory used). In Figure 2 and Table III, we see that CLIP-Embed-KD can achieve higher accuracy than CLIP-Teacher-KD, with larger student models as well as larger image sizes *while staying at a significantly lower computational budget*. In essence, eliminating the need to store the teacher model and run repeated forward passes through it, allows the freed resources to be better utilized for training larger student models instead. Even with slightly larger students, the memory used is much lesser than the teacher model, and CLIP-Embed-KD outperforms CLIP-Teacher-KD in accuracy.

To emphasize importance of the CLIP distillation loss $\mathcal{L}_{clip}$, we measure validation accuracy for different α_2 values, shown in Figure 3 (for large student; large-32 teacher; and image size 32). CLIP-Embed-KD using non-zero weighting of $\mathcal{L}_{CE}$ and $\mathcal{L}_{clip}$ ($\alpha_2 = 0.5, 0.25, 0.75$) outperforms regular supervised training ($\alpha_2 = 0$). This highlights that CLIP distillation indeed improves the student accuracy over typical supervised learning (that uses only $\mathcal{L}_{CE}$). Using only $\mathcal{L}_{clip}$ (i.e., $\alpha_2 = 1$) leads to overfitting, due to imbalanced reliance on embeddings alone.

REFERENCES

[1] 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*. PMLR, 2021, pp. 8748–8763.

[2] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.