Abstract

Self-supervised learning has been widely applied to train high-quality vision transformers (ViT). Unleashing their excellent performance on memory and compute constraint devices is therefore an important research topic. However, how to distill knowledge from one self-supervised ViT to another has not yet been explored. Moreover, existing self-supervised knowledge distillation (SSKD) methods focus on ConvNet architectures and are suboptimal for ViT knowledge distillation. In this paper, we study knowledge distillation of self-supervised vision transformers (ViT-SSKD). We show that directly distilling information from the crucial attention mechanism from teacher to student can significantly narrow the performance gap between both. In experiments on ImageNet-Subset and ImageNet-1K, we show that our method AttnDistill outperforms existing self-supervised knowledge distillation (SSKD) methods and achieves state-of-the-art $k$-NN accuracy compared with self-supervised learning (SSL) methods learning from scratch (with the ViT-S model). We are also the first to apply the tiny ViT-T model for self-supervised learning. Moreover, AttnDistill is independent of self-supervised learning algorithms, and it can be adapted to ViT-based SSL methods to improve performance in future research.

1 Introduction

Vision transformers [18] have been widely applied in computer vision tasks, including image classification [55, 62, 70], object recognition [4, 13, 19, 52, 76] and semantic segmentation [12, 50, 64, 73]. ViTs contain a self-attention mechanism [59] that allows for information exchange between distant patches and consequently leads to a more holistic understanding of image content. Another important aspect of transformers is that they are often pretrained in a self-supervised manner, followed by a finetuning stage to adapt to the downstream task [18, 26]. ViTs suffer from high memory requirements and substandard optimizability [11, 14, 25, 41, 60, 68], making them unsuitable for applications on memory or computation constraint devices. Consequently, methods that can reduce the memory footprint while maintaining the performance of ViTs are in demand.

One transfer learning technique is knowledge distillation [32]. Initial works focused on knowledge transfer for networks trained in a supervised manner [7, 57, 66, 69]. Recently, the
theory was extended to distill knowledge of self-supervised feature representations generated by large networks [21, 42, 43]. Since these networks do not output a conditional probability over a label set, but rather a feature representation, alternative distillation techniques needed to be developed [1, 21]. With the advent of transformers, supervised knowledge distillation for transformers has recently been investigated [33, 38, 54]. However, methods that can transfer self-supervised ViTs to smaller variants have not yet been explored.

Therefore, we explore knowledge distillation of self-supervised ViTs. We find that existing theory designed to transfer knowledge of ConvNets trained in a self-supervised manner results in a significant performance gap between teacher and student. To address this problem, we explore attention distillation that focuses on transferring the information present in the self-attention mechanism. Rather than just communicating the teacher’s conclusion which is the focus of most traditional knowledge distillation methods, attention distillation provides more guidance to the student network by identifying the important regions for understanding the image content. The potential of attention distillation has been explored for ConvNet [24, 71], however, since for these networks attention is not explicitly computed, additional computation and attention definition are needed. Since the attention mechanism is an integral and crucial part of transformers and no additional computation is required, we argue that attention distillation is a natural extension of the existing distillation theory for transformer networks.

In this paper, we focus on self-supervised knowledge distillation of self-supervised vision transformers (ViT-SSKD). First, we propose to use a projector alignment (PA) module to align the class tokens from teacher and student models. Second, we propose attention guidance (AG) with the Kullback–Leibler divergence to guide the student to obtain similar attention maps as the teacher model to further enhance the distillation. With these two modules, we can obtain state-of-the-art performance compared with self-supervised algorithms. Furthermore, we are the first to successfully train a small ViT-T model based on self-supervised learning (SSL) with knowledge distillation. More importantly, there might be more complex and outperforming SSL pretrained models in the future. In that case our method can be applied directly to obtain a smaller model while keeping competitive performance. Our main contributions are:

- We are the first to study the important ViT-SSKD problem allowing to transfer knowledge to small transformers in a self-supervised fashion.
- We propose an attention distillation loss for improved guidance of the student during knowledge distillation. Our method, AttnDistill, significantly reduces the gap between teacher and student models.
- We are the first to train a self-supervised ViT-T model. It obtains a performance almost (-0.3%) at par with the supervised ViT-T model.

2 Related work

Self-supervised learning. SSL [8, 9, 16, 23, 29, 30, 58, 63] automatically derives a supervisory signal for the training of high-quality feature representations, preventing the need of large labeled datasets. The common paradigm here is to pretrain on ImageNet [49] and then evaluated on downstream tasks, on which it has reached excellent performance, closing the gap with supervised methods. Recent popular SSL methods can be divided into two streams. *Contrastive learning* [6, 8, 27, 29] is the most popular stream. Another stream of
representation learning, named masked image encoding [3, 17, 30, 74], learns representations from corrupted images. In this paper, we study knowledge distillation for SSL based on both technical streams. From the aspect of backbone architectures, the previous methods are all based on ConvNet [5, 8, 26]. Recently, with the appearance of ViT, these are also applied for SSL (DINO [6], MoCo [10], etc). Compared with ConvNet, the attention-based ViTs suffer less from an image-specific inductive bias and have a larger potential when training on large scale datasets. In this paper we focus on the original ViT design, but the method could also be generalized to Swin Transformer [39] based SSL methods [37, 65].

Knowledge distillation for self-supervised models. Most knowledge distillation [32] techniques are proposed under the supervised learning scenario [2, 7, 44, 46, 48, 53, 57, 69, 71, 72]. Under the SSL settings, CC [43] exploit pseudo labels from clustering teacher embeddings as distillation signals. Then SEED [21] and CompRes [1] maintain memory banks to store a huge number of samples to calculate instance-level similarity score distributions for aligning the teacher and student models. SimReg [42] has similar projector architecture as our method, where they use the projector to align the teacher and student features. However, in some cases when the student ViT architectures become quite different from the teacher model, only projector regression is not sufficient to transfer knowledge from the teacher to the student model. Reg [67] is specified for metric learning, which could also be applied to self-supervised representation distillation. Recently, KDEP [31] propose the power temperature scaling to distill representation from a supervised teacher model.

Except for these examples in computer vision, there are several distillation attempts in NLP [20, 34, 51, 61]. However, these methods are limited to the case that teacher-student models share similar architectures. Also, Pelosin et al. [45] apply attention distillation between similar transformer architectures for continual learning. Our proposal is a more generalizable framework and allows for attention distillation between different ViT architectures.

3 Methodology

3.1 Preliminaries

Vision Transformers Architecture. Here we consider the ViT proposed in [18] but the theory is general and can be extended to other transformer architectures. The ViT consists of a patch embedding part, where the transformer encoder is a stack of $L$ multi-head self-attention blocks (MH-SAB). In each MH-SAB, there are two parts: a multi-head self-attention module (MSA) and a fully connected feedforward module (MLP). Each self-attention module has $H$ heads. We will further use $d$ as the output dimension for each head and $N$ as the number of patches. Also considering the class token, we can denote the output of the $l$th MH-SAB as $z_l \in \mathbb{R}^{(N+1)\times(dH)}, l \in [1, L]$. $z^0$ is the encoded patch embedding of image $x$ (i.e., from $z^0_1$ to $z^0_N$) and the initial class token (i.e., $z^0_0$). For the $h$th head in the self-attention module, learnable parameters $W_Q^{l,h}, W_K^{l,h}, W_V^{l,h}$ implemented as FC layers, map one slice of the input tokens $z^l_{-1,h}$ into the queries, keys and values ($Q^{l,h}, K^{l,h}, V^{l,h} \in \mathbb{R}^{(N+1)\times(d)}$). We obtain the attention map with Eq. 1, where $A^{l,h} \in \mathbb{R}^{(N+1)\times(N+1)}$, and the output of this head is obtained by Eq. 2. Combining the multiple head outputs, we obtain the final output of this multi-head self-attention layer (see Eq. 3). Finally, with the MLP and layer normalization (LN), the output tokens produced by the $l$th MH-SAB are given by Eq. 4. We have shown the process of the last MH-SAB from both the teacher and student ViT separately in Fig. 1 (the design is
Figure 1: AttnDistill for ViT-SSKD on the last block of the ViT. It is composed of the projector alignment loss $L_c$ and attention guidance loss $L_a$. The class tokens are taken from the last layer of the teacher and student ViT. We only consider the attention vectors that are formed by the interaction of the class token query with all keys for distillation.

inspired from [22].) which includes the key components in our approach.

$$A_{l,h}^{t,h} = \text{Softmax}(Q_{l,h}^{t,h} \cdot (K_{l,h}^{t,h})^T / \sqrt{d})$$

(1)

$$y_{l,h}^{t,h} = A_{l,h}^{t,h} \cdot V_{l,h}^{t,h}$$

(2)

$$y_{l} = \text{concat}(y_{l,1}, y_{l,2}, ..., y_{l,H})$$

(3)

$$z_{l} = \text{MLP}\left(\text{LN}\left(y_{l} + z_{l-1}\right)\right) + z_{l-1}.$$  

(4)

3.2 AttnDistill: Attention distillation

An outline of AttnDistill is given in Fig. 1. It can be divided into two parts: projector alignment (PA) and attention guidance (AG). For our distillation, we do not need to generate multiple views, since we focus on distilling the knowledge of the teacher. So other than contrastive-based SSL methods [6, 74, 75], in this phase we do not rely on multi-crop augmentations. Thus, we have less computation costs and the effective epoch is equal to the training epoch.

**Projector alignment (PA).** Suppose we have a teacher-student pair, each with a ViT architecture named $V_t$ and $V_s$. In self-supervised knowledge distillation, our aim is to distill knowledge from the teacher to the student model in a SSL way while maintaining its transferability. In most cases, we expect a smaller student model compared to the teacher. The parameter size is highly dependent on the feature dimension in the ViT. Thus, $V_t$ and $V_s$ typically have different feature dimensions. Therefore, we introduce a linear mapping projector $P$ to map the student to the teacher feature space for alignment. And since in ViT, the class token embedding $E_c = z_0^t$ is the most representative embedding for a classification decision, in AttnDistill, we only map the class token from the last layer for aligning the teacher and student model with a MSE loss to communicate with the final output from the teacher:

$$L_c = ||E_t^c - P(E_s^c)||_2$$

(5)
where ‘s/t’ subscripts represent ‘student/teacher’. In our ablation study, we also explored the influence of aligning the image patch embeddings $z^L_i, i \in [1, N]$ with the linear mapping $P$.  

**Attention guidance (AG).** However, aligning the class tokens can only tell the student model about "what" is in the image. More guidance from the teacher explaining "why" it reached this conclusion could be helpful. This extra guidance can be extracted from the multi-head self-attention (MSA) mechanism in ViTs. The MSA module pays attention to the decisive and informative parts in the image. Actually, we can observe that distillation with other methods will lead to *attention drift* where the student attention differs from the teacher (see the attention maps in Fig. 3.).

For a ViT, each $A^{l,h}$ given by layer $l$ head $h$ is an attention map from all tokens to all tokens. And since $A_{0,j}^{L,h}, j \in [0, N]$ (‘0’ represents the first row of the attention map $A^{L,h}$) contains the attention probabilities for the class token, it represents the importance of each token for the classification prediction of the image. By denoting $A = A_{0}^{L}, A^h = A_{0}^{L,h}, a^h_j = A_{0,j}^{L,h}$, we propose to apply Kullback–Leibler divergence ($\mathcal{KL}$) to make the student model pay attention to the same regions as the teacher model by aligning $A_s$ and $A_t$. We also study the performance with attentions from all layers in our ablation study section. To address knowledge transfer between ViTs with different designs, several cases need to be addressed. Here, we categorize them into four cases that might occur and discuss the solution below. The illustration of these variations are provided in Fig. 2.

(a) The teacher and student models have the same number of heads $H = H_t = H_s$ and patches $N = N_t = N_s$ (see Fig.2 (a)). This is the simplest case, where we align according to:

$$
\mathcal{L}_a = \sum_{h \in [1, H]} \mathcal{KL}(A^t_h || A^h_s)
$$

(b) The teacher and student models have the same number of heads $H$ but a different number of patches $N_t$ and $N_s$ (see Fig.2 (b)). In this case, we propose to interpolate (IP, by default we apply bicubic function $[35]$ as a smoother interpolation) the teacher model attention map $(a^h_j)_i, j \in [1, N_t]$ into $(a^h'_j)_i, j \in [1, N_s]$ ($N_t = w \times h, N_s = w' \times h'$), and then
normalize (NR) them into $1 - (a^h_0)_t$ by scaling up to have the attentions sum to 1 as in Eq. 7. Then the attention guidance loss is given by Eq. 8.

$$\left\langle a^j_h \right\rangle_t = NR_{1-(a^h_0)_t}(\text{IP}((a^h_j)_t))$$

$$L_a = \sum_{h \in [1,H]} KL((A^h_t)||A^h_s)$$

(c) The teacher and student models have the same number of patches $N$ but a different number of heads $H$ (see Fig.2 (c)). Here we merge the attentions from all heads for distillation. We have considered several aggregation functions, including mean, maximum and soft-maximum. We found that using the log summation to aggregate the attention probabilities for both teacher and student models (see Eq. 9 and Eq. 10) leads to slightly superior results compared to max-based fusion. Then the attention guidance loss $L_a$ is as Eq. 11.

$$a_j = \frac{1}{T} \cdot \sum_{h \in [1,H]} \log(a^j_h) = \frac{1}{T} \cdot \log(\prod_{h \in [1,H]} a^j_h)$$

$$A = \text{Softmax}([a_0, a_1, ..., a_N])$$

$$L_a = KL(A_t||A_s)$$

This aggregation could effectively highlight the maximum probabilities from the attention maps of the $H$ heads, as can be seen from our ablation study.

(d) The teacher and student models have a different number of heads $H$ and patches $N$. This case is a combination of the above two, thus we apply interpolation and aggregation sequentially and then apply distillation.

Finally, the self-supervised loss to update the student model $V_s$ is:

$$\mathcal{L} = \mathcal{L}_c + \lambda \cdot L_a$$

4 Experiments

4.1 Pre-Training setup

Datasets. In our experiments, ImageNet-Subset [49] is used for ablation study and to compare with other self-supervised knowledge distillation methods. This dataset contains 100 classes and $\approx 130k$ images in high resolution (resized to $224 \times 224$) [47]. For comparison with SSL methods, we employ the ImageNet-1K dataset [49].

Architecture. For the Teacher-Student pairs, we focus on knowledge distillation from a larger ViT teacher model to a smaller ViT student model. Due to the high computation demands of ViT, we select Teacher-Student pairs as below:

- On ImageNet-1K we select the following three pairs (Teacher $\rightarrow$ Student): (a) Mugs(ViT-S/16) $\rightarrow$ ViT-T/16; (b) Mugs(ViT-B/16) $\rightarrow$ ViT-S/16; (c) DINO(ViT-S/8) $\rightarrow$ ViT-S/16;
- On ImageNet-Subset, we fix the teacher model as MAE(ViT-S/16) with 12-Layer, 6-Head, 16-Patch and vary the design of the student model in the ablation study.
Table 1: Comparison with state-of-the-art SSL methods with $k$-NN and linear probing (LP.) on ImageNet-1K. "Effect Epo." is the effective pretraining epochs computed by multiplying number of views processed by the models following iBOT [74].

Moreover, since MAE/MoCo-v3 and DINO/iBOT/Mugs are following different position embedding strategies (fixed vs. learnable), our teacher-student pairing setups can show the effectiveness of AttnDistill for these two kind of position encodings.

Implementation details. We train our AttnDistill with the AdamW [40] optimizer. The learning rate is linearly ramped up during the first 40 epochs to the base learning rate $lr = (1.5e^{-4}) \times \frac{\text{batchsize}}{256}$. After the warming up epochs, we decay it with a cosine schedule till 800 epochs (except the distillation from Mugs(ViT-S/16) to ViT-T/16, where we train for 500 epochs because of performance saturation). By default, we set $T = 10.0$ and $\lambda = 0.1$, the projector $P$ is a 4-layer linear mapping. For the evaluations of the student model, we found it is optimal to perform with the features before the $P$. For the experiments on ImageNet-Subset, we fix the teacher model as a ViT-S/16 pre-trained with MAE [30] method with 3200 epochs. More details are in our supplementary materials.
4.2 Comparison with state-of-the-art

Comparison with SSKD methods on ViT. As shown in Table 2, we compare several methods to distill a MAE(ViT-S/16) teacher to a ViT-T student on ImageNet-Subset. Next to the common ViT-T architecture (with 12-layers, 6-heads, 16-patches), we also consider a harder variant (with 8-layers, 3-heads, 32-patches) as the student model. AttnDistill clearly outperforms them all in both cases. The margin is larger for the harder setup. This also indicates the importance of the attention distillation guidance. A comparison of the attention maps from the 8-layer 3-head 32-patch student case is shown in Fig. 3.

Comparison with SSL methods with linear probing, k-NN and finetuning accuracies. For linear probing and k-NN evaluation on ImageNet-1K, we follow the commonly used setting in DINO [6] and iBOT [74]. The comparison is shown in Table 1. We draw the following conclusions:

• Based on ViT-T/16 distilled from Mugs(ViT-S/16), our method AttnDistill gets state-of-the-art k-NN and Linear probing performance compared with previous knowledge distillation methods based on ConvNet. AttnDistill (ViT-T/16) is with only 5.7M parameters but outperforms the previous methods by a large margin and gets quite close to the supervised ViT-T/16 learning from scratch. In this case, we only train the student model for 500 epochs since we have observed marginal improvement on linear probing after that.

• Based on ViT-S/16 distilled from DINO(ViT-S/8), our method AttnDistill gets state-of-the-art in k-NN and the second in linear probing. This distillation decreases the ViT computational demand since there are 75% less patches in ViT-S/16 than ViT-S/8. Then, based on ViT-S/16 distilled from Mugs(ViT-B/16), AttnDistill gets the second in k-NN and the third in linear probing evaluations. In this case, the model size is decreased by 75%.

Moreover, we could further observe the advantage of ViT in k-NN evaluations, which means the extracted features from self-supervised pretrained ViT are more beneficial without learning an extra classifier as in linear probing evaluations. The accuracy curves during training are shown in Fig. 4.

For the finetuning comparison on ImageNet-1K shown in Table 3, we compared with existing methods working on ViT-T and ViT-S. AttnDistill (ViT/T) and AttnDistill (ViT/S) are distilled from Mugs (ViT-B) and Mugs (ViT-S) respectively. In both cases, AttnDistill works better than supervised learning methods and just marginally worse than the state-of-the-art with ViT-S. In conclusion, whereas AttnDistill is state-of-the-art for k-NN evaluation (Table 1), this is not the case when evaluating by means of finetuning.

Comparison with SSL methods on downstream tasks. For semi-supervised learning, results with SSL methods based on ViT-S/16 are shown in Table 4. In this setting, first, models are trained self-supervised on all ImageNet-1K data. Next labels for a small fraction of data (1% or 10%) are used to perform fine-tuning, linear probing or k-NN classification. Under all three settings with only 1% of the data, we can observe a considerable advantage of AttnDistill with a 3.9%/2.6%/5.8% improvement compared with Mugs(ViT-S/16). With 10% data, the improvement is less but still notable as 0.4%/2.7%/3.1%.

We also evaluate AttnDistill (ViT-T) for transfer learning. We compare with supervised ViT models, since there are no papers using ViT-T for SSL. Results are summarized in Table 5 for several small datasets. AttnDistill gets a 2.3% improvement compared with the previous best supervised learning method CCT-7/3x1 [28]. Next, we consider transfer learning of AttnDistill (ViT-S) in Table 6. Here we compare with previous SSL methods, we are only
Table 2: Compare with SSKD methods on ImageNet Subset with top-1/top-5 (LP.)

Table 3: Finetuning comparison on ImageNet1K.

Table 5: Transfer learning comparison on CIFAR10/CIFAR100. AttnDistill (ViT-T) is distilled from Mugs(ViT-S/16).

Table 4: Semi-supervised learning on ImageNet1K. AttnDistill (ViT-S) is distilled from the teacher model Mugs (ViT-B/16) for 800 epochs.

Table 6: Compared with SSL methods on four small datasets. AttnDistill (ViT-S) is distilled from the teacher Mugs (ViT-B/16).

marginally worse than the state-of-the-art and much better than the supervised distillation method DEiT [54].

4.3 Ablation study

To prove the generalizability of AttnDistill, we perform an ablation study on ImageNet-Subset with a fixed MAE(ViT-S/16) teacher and vary the architecture of the student model. Extended ablation studies are in the supplementary. In Fig. 5, our ablation study contains three parts:

(a) The architecture of ViT (in Fig. 5-(a)) : To verify the effectiveness of AttnDistill for various architectures of ViT, we modify the number of heads, patch sizes and the number of block layers. In all cases, AttnDistill significantly improves the PA baseline and closes the gap with the teacher performance. Especially, for the smaller student architectures and those with fewer tokens, attention distillation is shown to be crucial leading to improvements of over 5%.

(b) The various aggregation functions (in Fig. 5-(b)) : Here we fix the design of the student model and vary the strategy to compute the attention guidance loss. To verify the superiority of the used log summation in Eq. 9, we replace it with MEAN/MIN/MAX strategies to aggregate attention maps from different heads. However, they are all suboptimal.

(c) Alternative self-supervised losses (in Fig. 5-(c)) : A recent work for distillation of self-supervised representations of ConvNet is CompRess [1]. Here, we apply the knowledge distillation loss from CompRess [1] to our PA module, we can clearly observe that this
Figure 4: More results of AttnDistill with different training epochs on ImageNet-1K.

Figure 5: Ablation study for ViT architectures, aggregation functions and alternative losses.

KD loss is an obstacle in ViT distillation since when combined with our PA module it leads to a performance drop. Except distilling the attention maps from the last layer, we also experiment the distillation over attention maps from all layers. This is 0.6% lower than AttnDistill based on only the last layer. Finally, we also align the patch tokens with our PA module. This is 2.4% worse than without the patch token alignment, thus aligning patch tokens is not necessary.

5 Conclusion

In this paper, we explored the ViT-based self-supervised knowledge distillation problem. Observing that the previous SSKD methods focussed on ConvNet do not work well on ViT, we proposed AttnDistill to distill the knowledge from a pretrained teacher model to its student model. The experiments clearly show that AttnDistill outperforms other SSKD methods. Furthermore, our distilled ViT-S gets state-of-the-art in k-NN accuracy and is second in linear probing compared with SSL methods. Also, our method AttnDistill is especially advantageous in semi-supervised learning evaluation and competitive in transfer learning evaluation. To prove the effectiveness of AttnDistill, we also implement various ablation studies on ImageNet-Subset. For future work, we are interested to explore AttnDistill for knowledge distillation between ConvNets and ViT.

Limitations. A drawback of the attention mechanism is that it is tailored for transformer usage and requires additional computation when applied to ConvNets (namely the computation of the attention maps). A further limitation is that the theory only applies to a single teacher-student pair. In case of multiple teacher models, further thought has to be given on how the multiple attention maps can be meaningfully communicated with the student.
Acknowledgement

We acknowledge the support from Huawei Kirin Solution, and the Spanish Government funded project PID2019-104174GB-I00/AEI/10.13039/501100011033.

References


[30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022.


