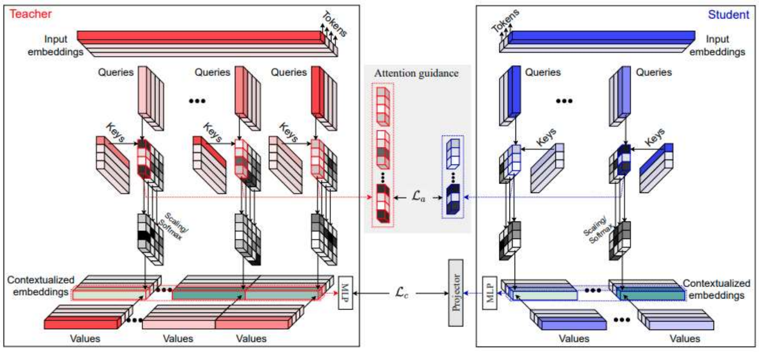
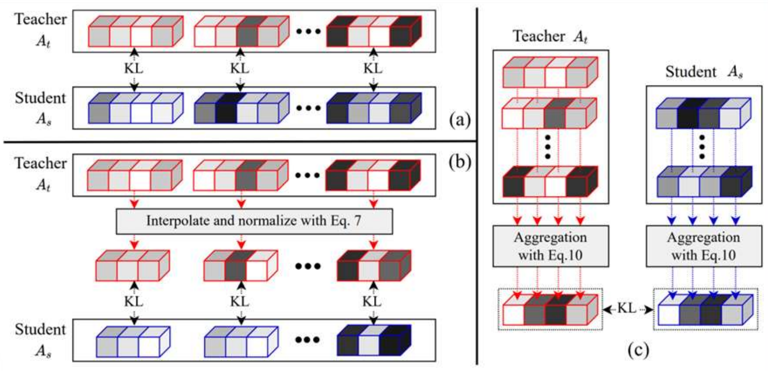
Attention Distillation: self-supervised vision transformer students need more guidance

Kai Wang, Fei Yang, Joost van de Weijer





- *AttnDistill* for ViT-SSKD on the *last block* of the ViT.
- It is composed of the projector alignment loss and attention guidance loss.
- 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.

(a) The teacher and student models have the same number of heads

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

(b) The teacher and student models have the same number of heads but a different number of patches

$$(a_j^h)'_t = \mathbf{NR}_{1-(a_0^h)_t}(\mathbf{IP}((a_j^h)_t))$$
$$\mathcal{L}_a = \sum_{h\in[1,H]} \mathcal{KL}((A^h)'_t || A_s^h)$$

(c) The teacher and student models have the same number of patches N but a different number of heads

$$a_j = \frac{1}{T}\cdot\sum_{h\in[1,H]} log(a_j^h) = \frac{1}{T}\cdot log(\prod_{h\in[1,H]} a_j^h)$$
$$A = \mathtt{Softmax}([a_0, a_1, ..., a_N])$$
$$\mathcal{L}_a = \mathcal{KL}(A_t || A_s)$$

Experiment configurations and training strategy

	PE	model	layers	dim	heads	patch size	#tokens	#params
Networks configurations for experiments on ImageNet-1K								
Teacher	learnable	Mugs (ViT-S/16)	12	384	6	16	197	22M
		DINO (ViT-S/8)	12	384	6	8	785	22M
		Mugs (ViT-B/16)	12	768	12	16	197	85M
Student	learnable	AttnDistill (ViT-T/16)	12	192	3	16	197	5.7M
		AttnDistill (ViT-S/16)	12	384	6	16	197	22M
Networks configurations for experiments on ImageNet-Subset								
Teacher	sin-cos	MAE (ViT-S/16)	12	384	6	16	197	22M
Student	sin-cos	AttnDistill (ViT-T)	12	192	6	16	197	5.7M
		AttnDistill (ViT-T)	12	192	3	16	197	5.7M
		AttnDistill (ViT-T)	12	192	3	32	65	5.7M
		AttnDistill (ViT-T)	8	192	3	16	197	3.8M
		AttnDistill (ViT-T)	8	192	3	32	65	3.8M

config	value
optimizer	AdamW [8]
base learning rate	1.5e-4
weight decay	0.05
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$ [2]
batch size	4096
learning rate schedule	cosine decay [7]
warmup epochs	40
training epochs	500 (ViT-T/16 ImageNet-1K) 800 (ViT-S/16 ImageNet-1K) 3200 (ViT-T/16 ImageNet-Subset)
augmentation	RandomResizedCrop

EXPERIMENTS — MAIN RESULTS

Teacher model	Method	Student Arch.	Par.(M)	Train Epo.	Effect Epo.	k-NN	LP.
✗	Supervised	ViT-T/16	5.7	-	-	72.2	72.2
SwAV (RN-50)	CRD	RN-18	11	240	240	44.7	58.2
SwAV (RN-50)	CC	RN-18	11	100	100	51.0	60.8
SwAV (RN-50)	Reg	RN-18	11	100	100	47.6	60.6
SwAV (RN-50)	CompRess-2q	RN-18	11	130	130	53.7	62.4
SwAV (RN-50)	CompRess-1q	RN-18	11	130	130	56.0	65.6
SwAV (RN-50)	SimReg	RN-18	11	130	130	59.3	65.8
SwAV (RN-50×2)	SEED	RN-18	11	200	200	55.3	63.0
SwAV (RN-50×2)	SEED	EffNet-B1	7.8	200	200	60.3	68.0
SwAV (RN-50×2)	SEED	EffNet-B0	5.3	200	200	57.4	67.6
SwAV (RN-50×2)	SEED	MbNet-v3	5.5	200	200	55.9	68.2
Mugs (ViT-S/16)	AttnDistill	ViT-T/16	5.7	500	500	**71.4**	**71.9**
✗	Supervised	ViT-S/16	22	-	-	79.8	79.8
✗	SimCLR	ViT-S/16	22	300	600	-	69
✗	BYOL	ViT-S/16	22	300	600	-	71
✗	MoCo v3	ViT-S/16	22	600	1200	-	73.4
✗	SwAV	ViT-S/16	22	800	2400	66.3	73.5
✗	DINO	ViT-S/16	22	800	3200	74.5	77.0
✗	iBOT	ViT-S/16	22	800	3200	75.2	77.9
✗	MUGS	ViT-S/16	22	800	3200	75.6	**78.9**
SwAV (RN-50×2)	SEED	RN-34	21	200	200	58.2	65.7
SwAV (RN-50×2)	SEED	RN-50	24	200	200	59.0	74.3
SimCLR (RN-50×4)	CompRess-1q	RN-50	24	130	130	63.3	71.9
SimCLR (RN-50×4)	CompRess-2q	RN-50	24	130	130	63.0	71.0
SimCLR (RN-50×4)	CC	RN-50	24	100	100	55.6	68.9
SimCLR (RN-50×4)	SimReg	RN-50	24	130	130	60.3	74.2
Mugs (ViT-B/16)	AttnDistill	ViT-S/16	22	800	800	76.8	78.6
DINO (ViT-S/8)	AttnDistill	ViT-S/16	22	800	800	**77.4**	78.8
Teacher Models statistics							
Mugs (ViT-S/16)	-	-	22	800	3200	75.6	78.9
DINO (ViT-S/8)	-	-	22	800	3200	78.3	79.7
Mugs (ViT-B/16)	-	-	85	400	1600	78.0	80.6
SwAV (RN-50)	-	-	24	800	2400	64.8	75.6
SwAV (RN-50×2)	-	-	94	800	2400	-	77.3
SimCLR (RN-50×4)	-	-	375	1000	2000	64.5	75.6

Conclusions

- We explored the ViT-based self-supervised knowledge distillation problem.
- 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.
- Our distilled ViT-S gets state-of-the-art in k-NN accuracy and is second in linear probing.
- AttnDistill is 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.