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
\[ L_d = \sum_{h, |H|} K.L(A^h_t|A^h_s) \]

(b) The teacher and student models have the same number of heads but a different number of patches
\[ a_j^h = \frac{1}{|H|} \sum_{h, |H|} \log(a_j^h) = \frac{1}{|H|} \log(\prod_{h, |H|} a_j^h) \]
\[ A = \text{Softmax}(a_0, a_1, ..., a_N) \]
\[ L_d = \sum_{h, |H|} K.L(A^h_t|A^h_s) \]

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

ExpErimEnts — main rEsults

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.

The tables and figures illustrate the configurations and training strategies for various experiments on ImageNet-1K and ImageNet-Subset. The network configurations include layers, dimension, heads, patch size, number of tokens, and parameters. The experiments evaluate different configurations and training strategies for both the teacher and student models in ViT-SSKD.