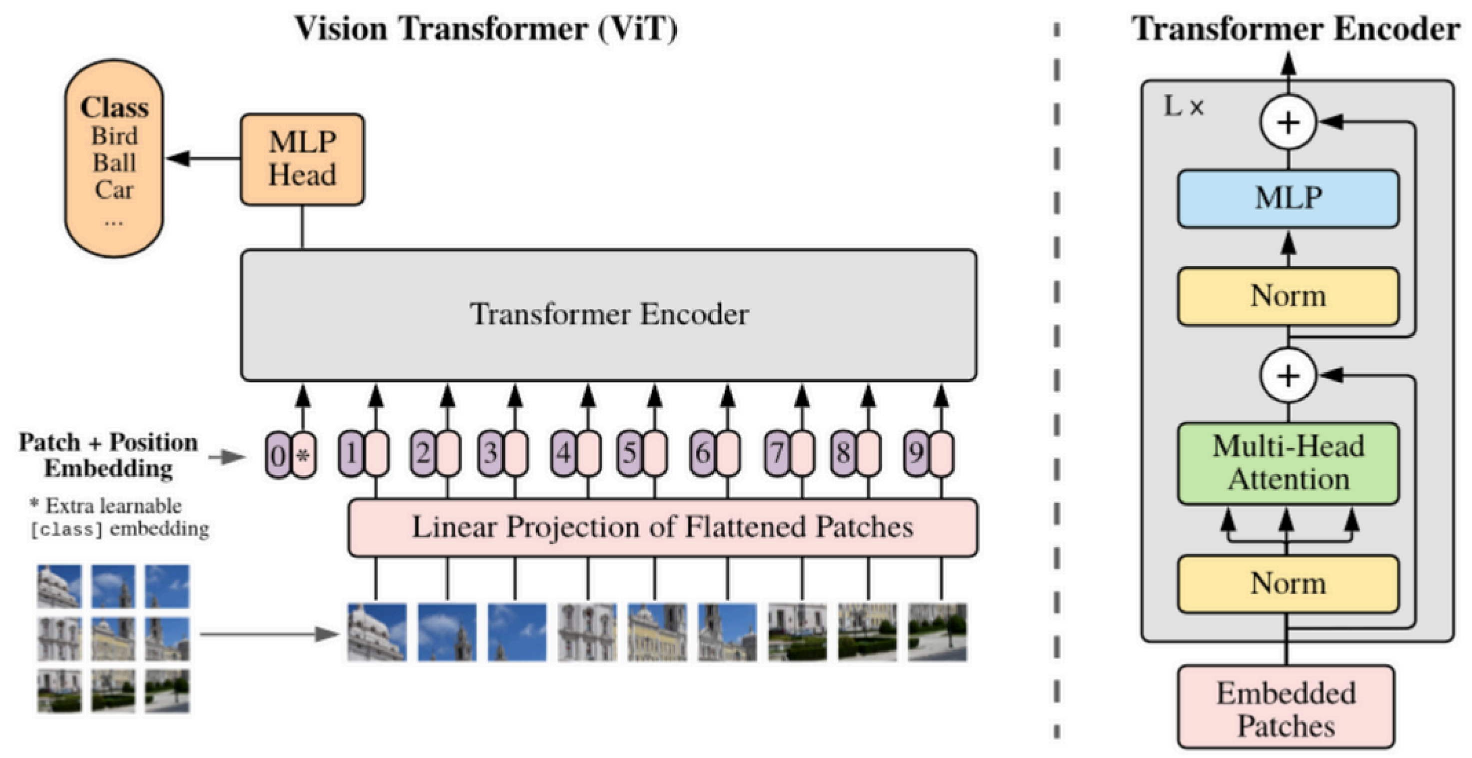
DeiT: Training ViTs Using Distillation through Attention

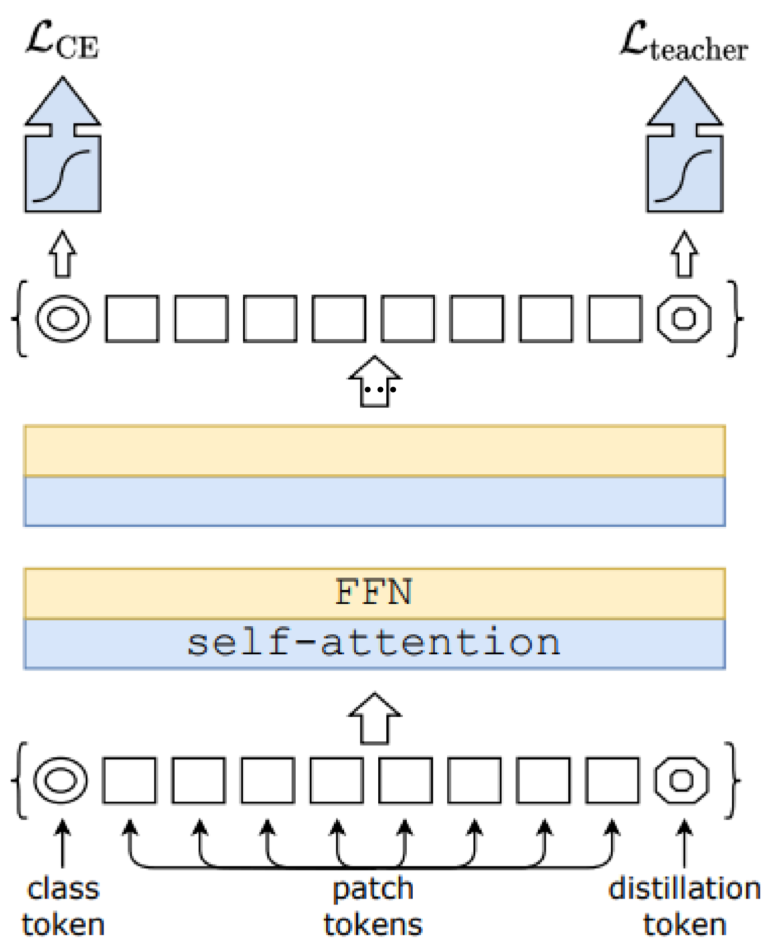
Yunoo Kim, Liesel Wong, Angela Xiao
Cornell University

Problem



ViTs require an intensive amount of data to be competitive with CNN (300+ million images in original paper for 88.55% accuracy)

Solution



A different model called DeiT with distillation and distillation attention achieved comparable or better results using 1.2 million images

Our Goal

Build a model using DeiT architecture, and reproduce the results and benchmarks described in the paper with more limited resources

References

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2020. Training data-efficient image transformers & distillation through attention. arXiv [cs.CV]. Retrieved from http://arxiv.org/abs/2012.12877

https://github.com/huyvnphan/PyTorch_CIFAR10

Dataset

We used the CIFAR-10 to train our model, which consisted of 10 classes and 60000 images

Models

method ↓	Supervision label	Supervision teacher	ImageNet top-1 (%) Ti 224	S 224	B 224	B↑384
DeiT– no distillation	✓	✗	72.2	79.8	81.8	83.1
DeiT– usual distillation	✗	soft	72.2	79.8	81.8	83.2
DeiT– hard distillation	✗	hard	74.3	80.9	83.0	84.0
DeiT⚗: distil. embedding	✓	hard	74.6	81.1	83.1	84.4
DeiT⚗: class+distillation	✓	hard	74.5	81.2	83.4	84.5

- The models with the alembic sign uses distillation token architecture, and the others use a ViT infrastructure.

Attributes

- Used pre-trained CNN as the 'teacher' for our model
- Used Cross Entopy Loss to match class token with output
- Used distillation loss to match teacher token
- Ran model for only 20 epochs, used DeiT-Ti which is smaller than DeiT

Soft vs Hard Distillation Loss

Soft Distillation - KL divergence, teacher probs

$$\mathcal{L}_{\text{global}} = (1-\lambda)\mathcal{L}_{\text{CE}}(\psi(Z_s), y) + \lambda\tau^2 \text{KL}(\psi(Z_s/\tau), \psi(Z_t/\tau)).$$

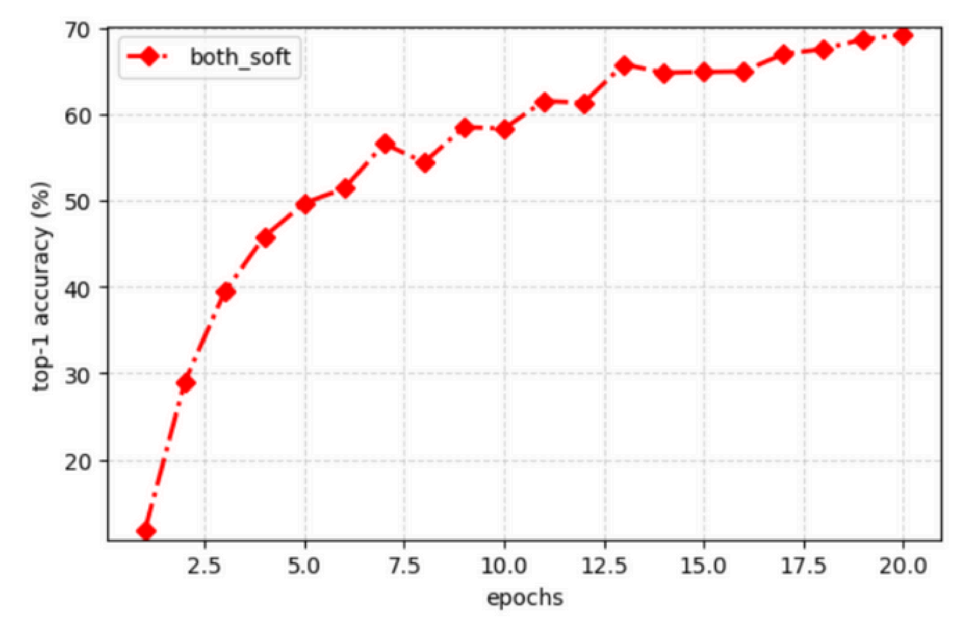
Hard Distillation - cross entropy, teacher decision

$$\mathcal{L}_{\text{global}}^{\text{hardDistill}} = \frac{1}{2}\mathcal{L}_{\text{CE}}(\psi(Z_s), y) + \frac{1}{2}\mathcal{L}_{\text{CE}}(\psi(Z_s), y_t).$$

Results

Paper-proposed architecture achieved 69.19% test acc for epoch 20 with soft distillation loss; better than ViT with no distillation

Method	Supervision Label	Supervision Teacher	CIFAR-10 Accuracy
DeiT-Ti - no distillation	✓	✗	67.69%
DeiT-Ti - usual distillation	✗	soft	68.85%
DeiT-Ti - hard distillation	✗	hard	40.19%
DeiT-Ti ⚗ - distil embedding	✓	hard	8.07%
DeiT-Ti ⚗ - class embedding	✓	hard	61.69%
DeiT-Ti ⚗ - class+distillation	✓	soft	69.19%
DeiT-Ti ⚗ - class+distillation	✓	hard	46.11%
ResNet-18 (Pretrained)	n/a	n/a	88.46%



For DeiT-Ti with class and distillation tokens, test acc ↑ throughout 20 epochs, does not plateau yet

Conclusion

- Distillation token architecture (class + distillation) overall performs better compared to the ViT.
- Teacher domain mismatch causes low accuracy for certain distillation methods
- Did not achieve comparable performance to a CNN due to low epochs.

Future Work

- Implementing data augmentation methods
- Training for more epochs
- Implementing fine tuning on different sized images using upscaling algorithm