Training data-efficient image transformers & distillation through attention

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles

Proceedings of the 38th International Conference on Machine Learning, 2021.

경영과학연구실 김윤석

Introduction

- ViT demonstrated the success of Transformer in the field of computer vision.
- ViT has a disadvantage of requiring pre-training on large-scale datasets to achieve good performance.

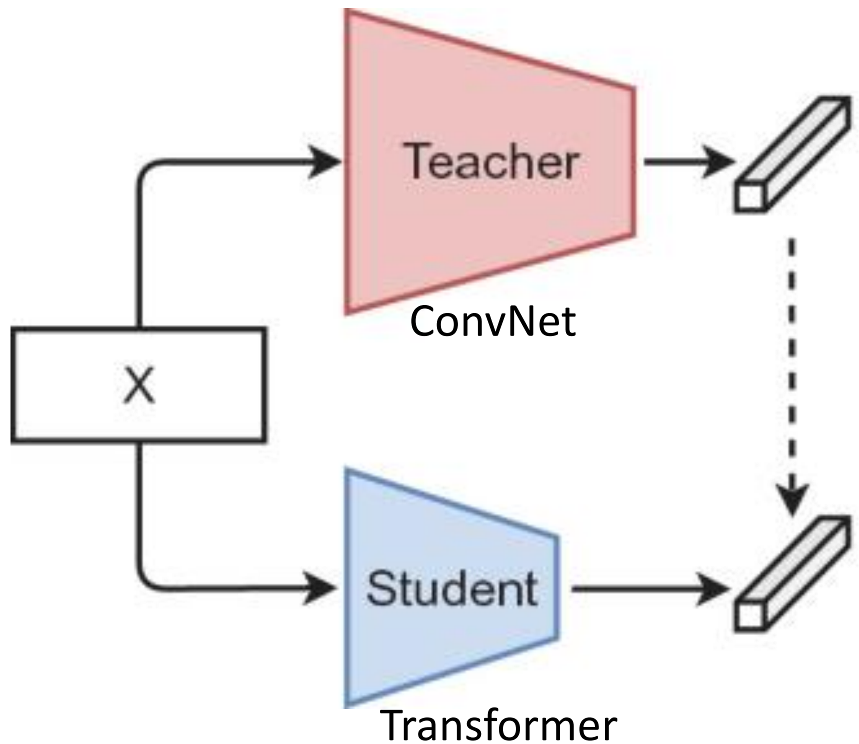
		ViT-B/16	ViT-B/32	ViT-L/16	ViT-L/32	ViT-H/14
ImageNet	CIFAR-10	98.13	97.77	97.86	97.94	-
	CIFAR-100	87.13	86.31	86.35	87.07	-
	ImageNet	77.91	73.38	76.53	71.16	-
	ImageNet ReaL	83.57	79.56	82.19	77.83	-
	Oxford Flowers-102	89.49	85.43	89.66	86.36	-
	Oxford-IIIT-Pets	93.81	92.04	93.64	91.35	-
ImageNet-21k	CIFAR-10	98.95	98.79	99.16	99.13	99.27
	CIFAR-100	91.67	91.97	93.44	93.04	93.82
	ImageNet	83.97	81.28	85.15	80.99	85.13
	ImageNet ReaL	88.35	86.63	88.40	85.65	88.70
	Oxford Flowers-102	99.38	99.11	99.61	99.19	99.51
	Oxford-IIIT-Pets	94.43	93.02	94.73	93.09	94.82
JFT-300M	CIFAR-10	99.00	98.61	99.38	99.19	99.50
	CIFAR-100	91.87	90.49	94.04	92.52	94.55
	ImageNet	84.15	80.73	87.12	84.37	88.04
	ImageNet ReaL	88.85	86.27	89.99	88.28	90.33
	Oxford Flowers-102	99.56	99.27	99.56	99.45	99.68
	Oxford-IIIT-Pets	95.80	93.40	97.11	95.83	97.56

Problem statement

- Development of Vision Transformer that can be used without pre-training on large-scale data
- A study on applying knowledge distillation technique to Vision Transformer

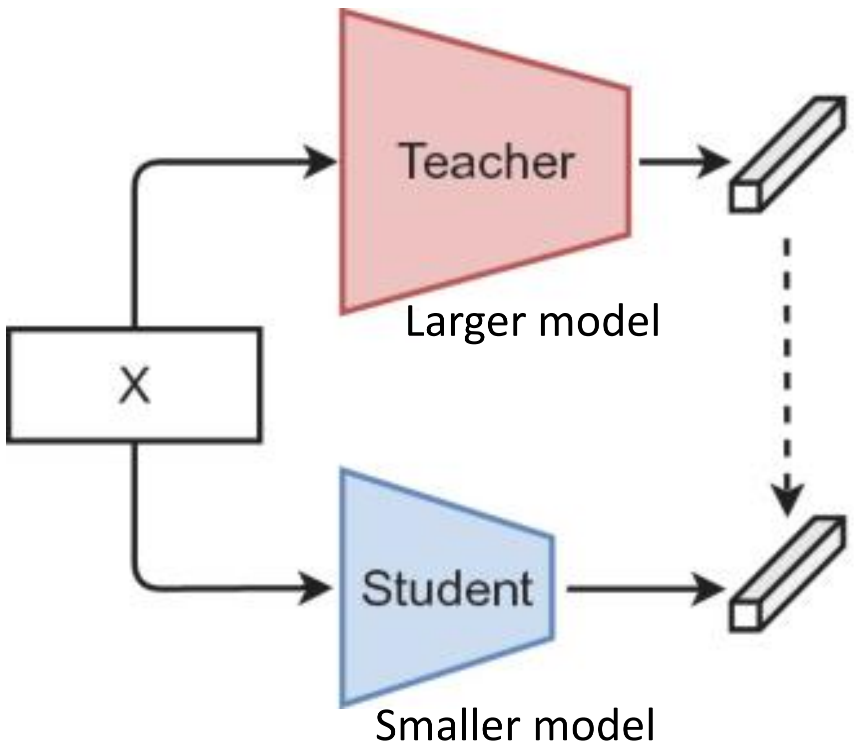
Key idea

- Application of knowledge distillation technique using ConvNet as a teacher network



Knowledge Distillation

- Knowledge Distillation is a technique of training a smaller model using a well-trained larger model



Soft/Hard Label

- Soft label is a result of smoothing the model's prediction results.
- Hard label is the ground truth.

0.1	0.4	0.8	0.1

Model's prediction

0.24	0.25	0.27	0.24

Soft label

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

$$q_i = \frac{exp(z_i)}{\sum_j exp(z_j)}$$

0.1	0.4	0.8	0.1

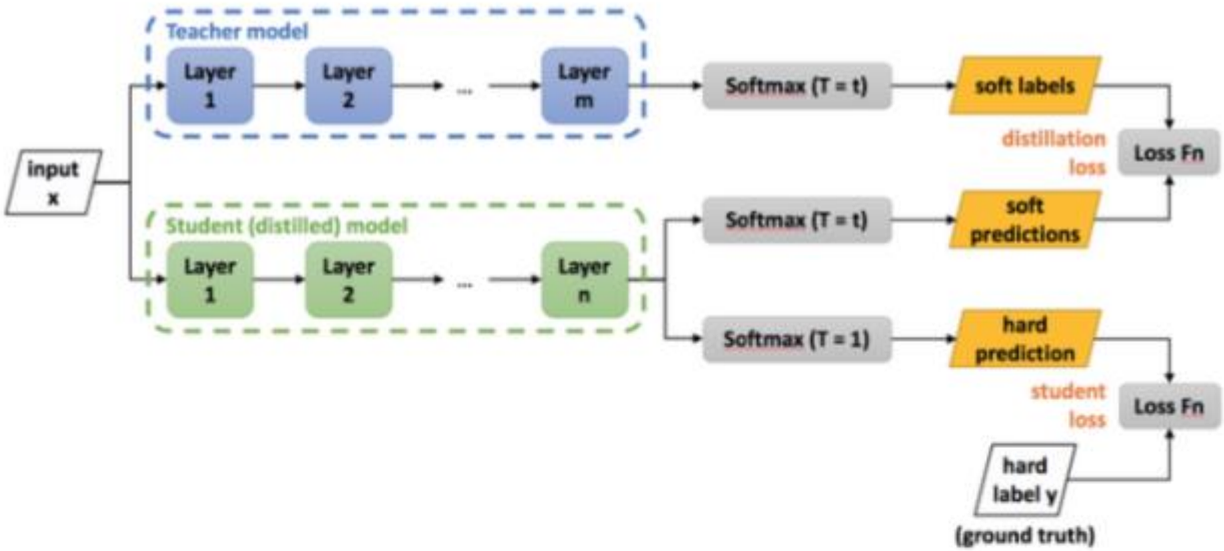
Model's prediction

0.19	0.24	0.38	0.19

label

Distillation Loss

- Both soft-labels from the Teacher network and hard labels, which are ground truth values, are used to compute the cross-entropy loss.



$$L = \sum_{(x,y)\in\mathbb{D}} L_{KD}(S(x,\theta_S,\tau),T(x,\theta_T,\tau)) + \lambda L_{CE}(\hat{y}_S,y)$$

Distillation using Convolutional Neural Networks

- ViT requires pre-training on large-scale data for good performance due to its limited generalization ability
- DeiT utilizes ConvNet as a teacher network to train the generalization ability of ConvNet.

Proposed method

- The authors cover two axes of distillation.
  - hard distillation versus soft distillation
  - classical distillation versus the distillation token

Soft distilation

- Soft distillation involves calculating cross-entropy loss with ground truth and KL-divergence loss with teacher model's temperature-scaled softmax function values.

Soft distillation loss function

$$\mathcal{L}_{\mathrm{global}} = (1-\lambda)\mathcal{L}_{\mathrm{CE}}(\psi(Z_{\mathrm{s}}), y) + \lambda\tau^2 \mathrm{KL}(\psi(Z_{\mathrm{s}}/\tau), \psi(Z_{\mathrm{t}}/\tau))$$

- λ: The coefficient balancing the Kullback–Leibler divergence loss (KL) and the cross-entropy (LCE) on ground truth labels y
- Z: The logits
- ψ: softmax function
- τ: temperature

Hard-label distilation

- Hard label distillation involves using the teacher model's predicted values as hard labels for training.

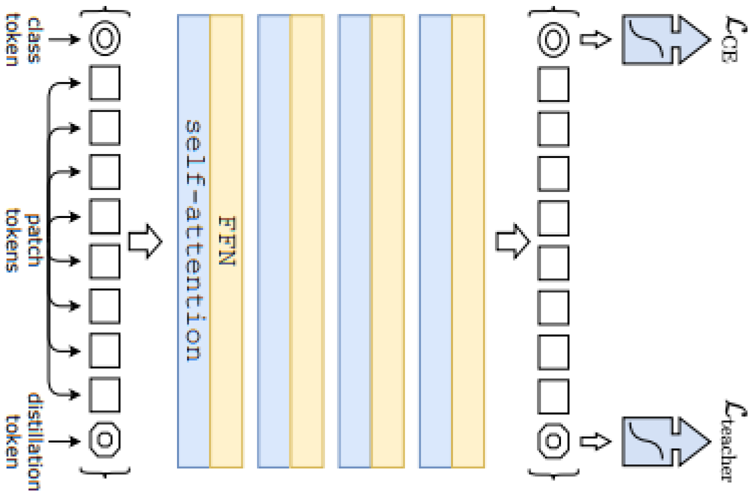
Hard-label distillation loss function

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

- Z: The logits
- ψ: softmax function
- $y_t = \mathrm{argmax}_c \mathrm{Z_t}(\mathrm{c})$: Hard-label of the teacher model's prediction

Distilation token

- Distillation token is a token added at the embedding layer before the transformer input.
- Distillation token is a token added at the embedding layer before the transformer input, which is compared to the output of the teacher model, similar to the CLS token in transformers.



Model setup

- The DeiT model is experimented at the same number of layers as ViT-B and determines the model size by adjusting the embedding vector and attention head numbers.

Variants of DeiT architecture

Model	ViT model	embedding dimension	#heads	#layers	#params	training resolution	throughput (im/sec)
DeiT-Ti	N/A	192	3	12	5M	224	2536
DeiT-S	N/A	384	6	12	22M	224	940
DeiT-B	ViT-B	768	12	12	86M	224	292

Experiments on Convnets teachers

- Convnet teachers perform better than transformer teacher models.
- Due to distillation, the inductive bias of convnets is transferred to DeiT.

Teacher Models	Teacher acc.	Student: DeiT-B ⚗ pretrain	Student: DeiT-B ⚗ ↑384
DeiT-B	81.8	81.9	83.1
RegNetY-4GF	80.0	82.7	83.6
RegNetY-8GF	81.7	82.7	83.8
RegNetY-12GF	82.4	83.1	84.1
RegNetY-16GF	82.9	83.1	84.2

Comparison of distillation methods

- Comparison between Soft distillation and Hard-label distillation
- Comparison between traditional distillation and the proposed distillation method

method ↓	Supervision label	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⚗: class embedding	✓	hard	73.9	80.9	83.0	84.2
DeiT⚗: distil. embedding	✓	hard	74.6	81.1	83.1	84.4
DeiT⚗: class+distillation	✓	hard	74.5	81.2	83.4	84.5

Agreement with the teacher

- Learning with the distillation token provides agreement between the DeiT and the teacher model

	groundtruth	no distillation		DeiT⚗ student (of the convnet)		
		convnet	DeiT	class	distillation	DeiT⚗
groundtruth	0.000	0.171	0.182	0.170	0.169	0.166
convnet (RegNetY)	0.171	0.000	0.133	0.112	0.100	0.102
DeiT	0.182	0.133	0.000	0.109	0.110	0.107
DeiT⚗– class only	0.170	0.112	0.109	0.000	0.050	0.033
DeiT⚗– distil. only	0.169	0.100	0.110	0.050	0.000	0.019
DeiT⚗– class+distil.	0.166	0.102	0.107	0.033	0.019	0.000

Conclusion

- Convnet-based distillation training for Vision Transformer is proposed in this paper.
- By utilizing distillation with Convnet, the inductive bias of Convnet is learned, showing good performance without the need for large-scale pre-training of data.
- They propose a Vision Transformer model that can be used with limited resources.