Class Updates

- Paper assignments posted online.
- Student paper presentations will begin on Monday, January 29th (a week from today).
- All paper presentations (except for paper battles) shortened to ~30min to give more time for a discussion.
- Project team members list due on January 31st, 11:59pm.

Image Classification

- The goal is to identify the category of a given image.

Input:

Classification

Output:

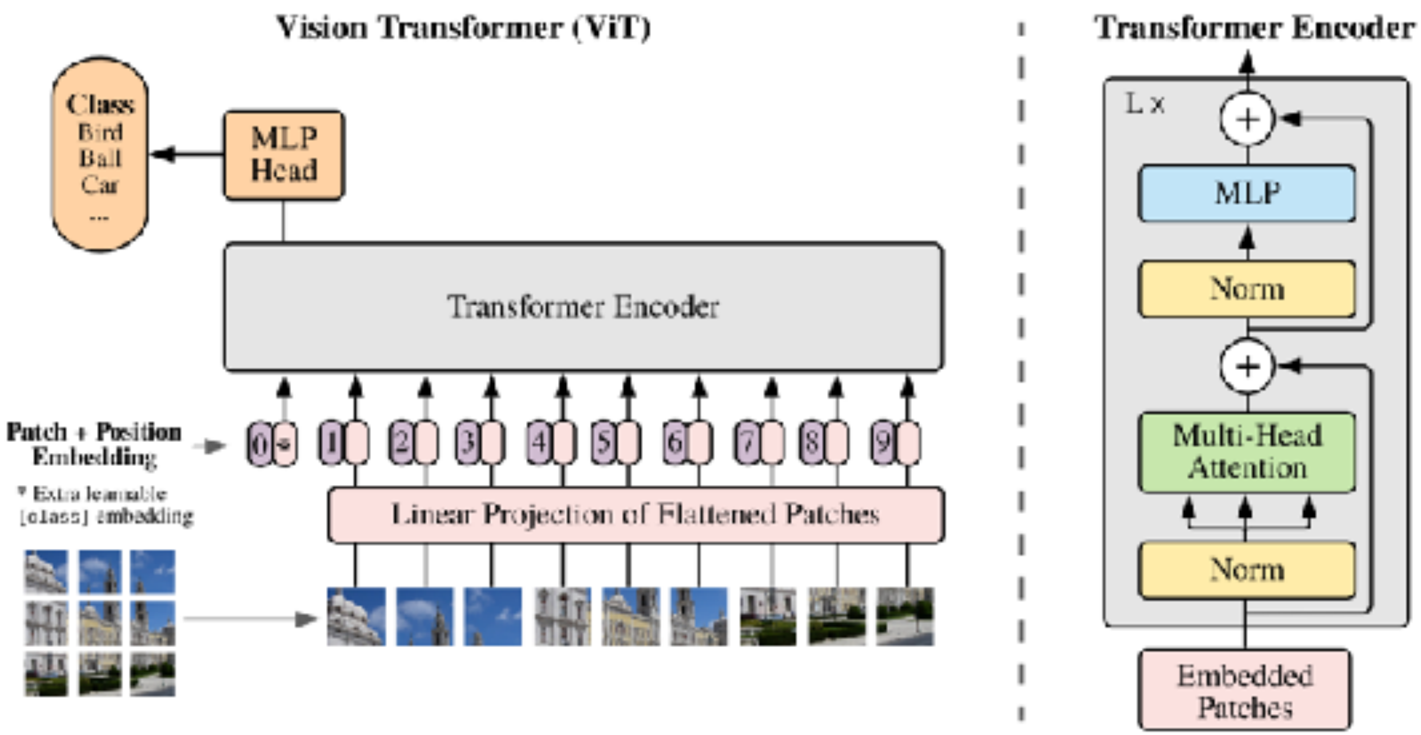
Training data-efficient image transformers & distillation through attention

ICML 2021

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

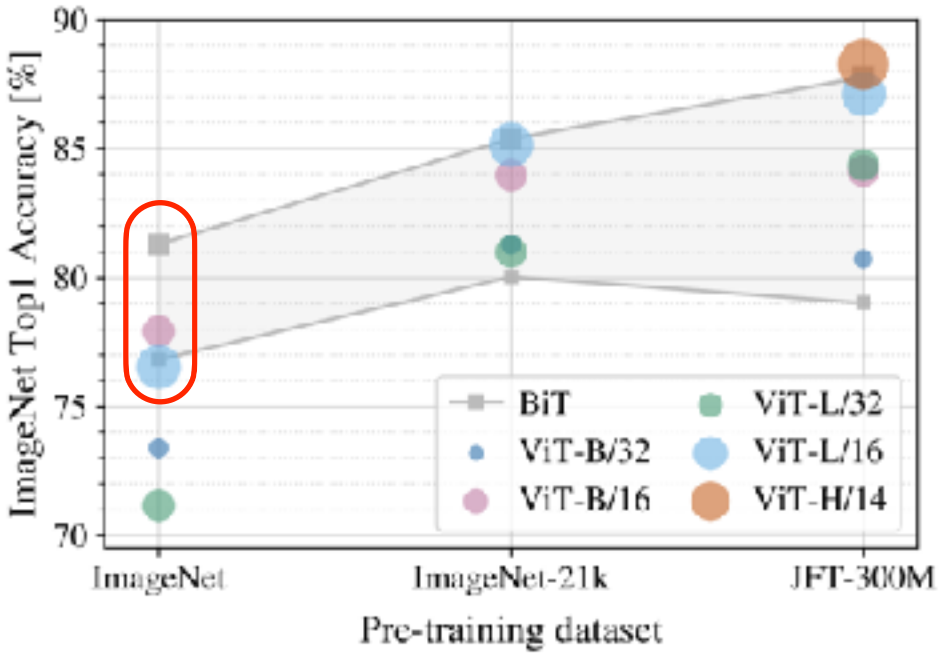
Vision Transformer (ViT)

- The authors split an image into fixed-size patches, linearly embed each of them, and add position embeddings.
- The resulting sequence of vectors is then fed into a standard Transformer encoder.



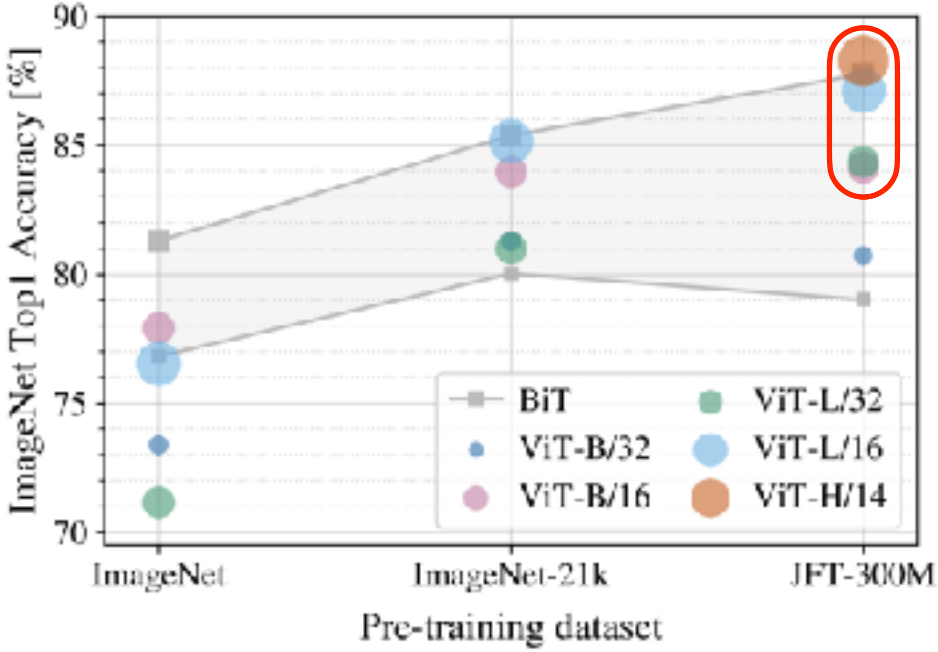
Pre-training Data Requirements

- The ViT models are pre-trained on datasets of increasing size: ImageNet, ImageNet-21k, and JFT300M.
- ImageNet accuracy is reported after finetuning on ImageNet.



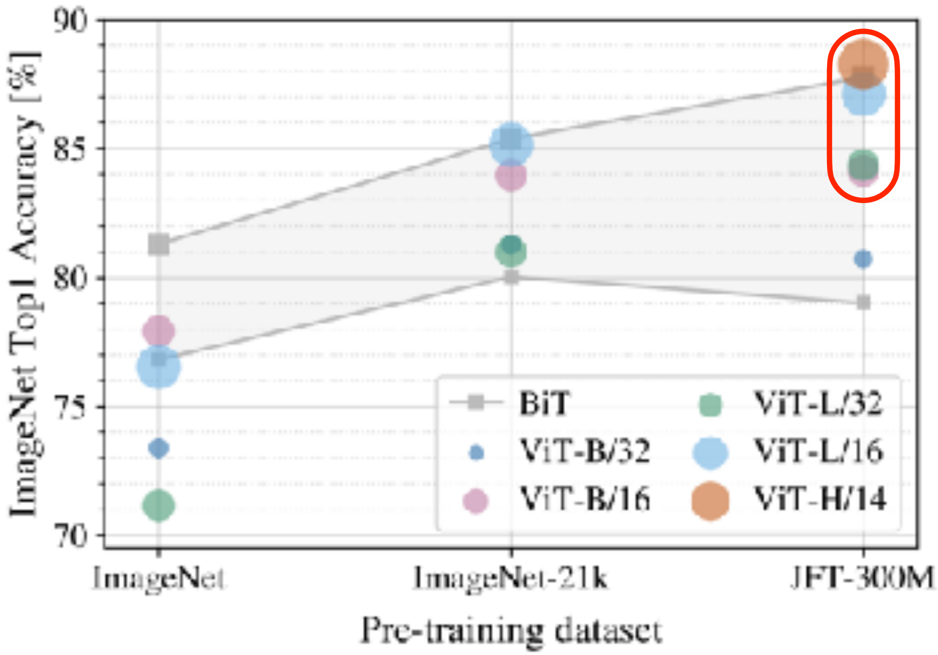
Pre-training Data Requirements

- The ViT models are pre-trained on datasets of increasing size: ImageNet, ImageNet-21k, and JFT300M.
- ImageNet accuracy is reported after finetuning on ImageNet.



Pre-training Data Requirements

- The ViT models are pre-trained on datasets of increasing size: ImageNet, ImageNet-21k, and JFT300M.
- ImageNet accuracy is reported after finetuning on ImageNet.



Can we effectively train ViTs on medium-sized datasets?

Goals

1. To design a data-efficient transformer trained only on ImageNet (eliminating JFT or ImageNet-21K pretraining).

Goals

1. To design a data-efficient transformer trained only on ImageNet (eliminating JFT or ImageNet-21K pretraining).

2. Instead of using hundreds of GPUs/TPUs, use a single GPU node/machine.

Goals

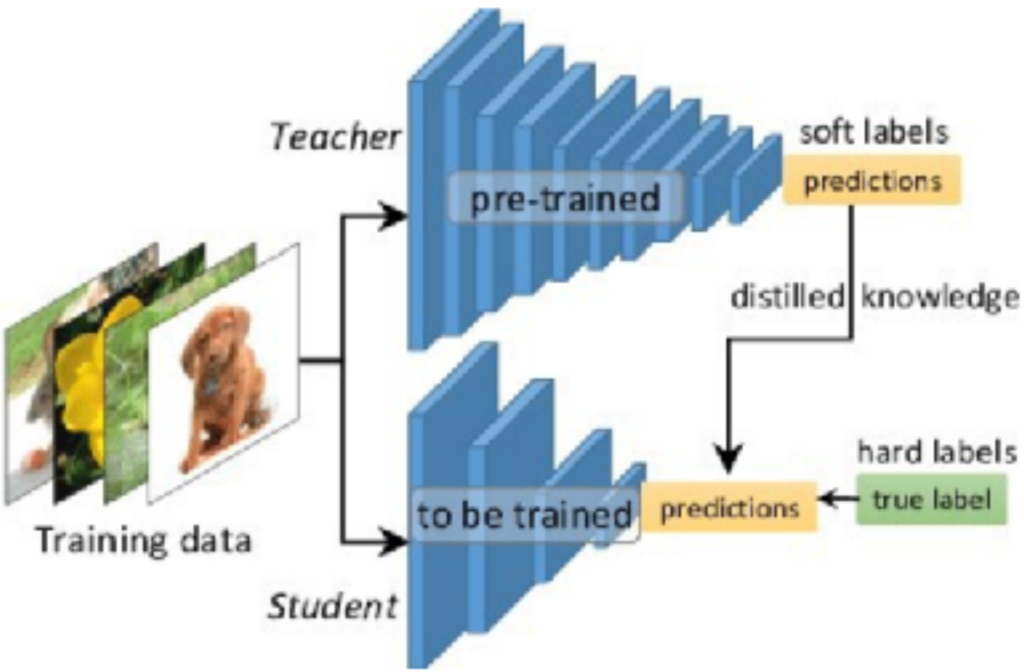
1. To design a data-efficient transformer trained only on ImageNet (eliminating JFT or ImageNet-21K pretraining).

2. Instead of using hundreds of GPUs/TPUs, use a single GPU node/machine.

3. Achieve competitive image classification performance on par or even better than the standard ViT.

Knowledge Distillation

- Knowledge distillation refers to the idea of using a pretrained network to supervise another smaller/less powerful network.

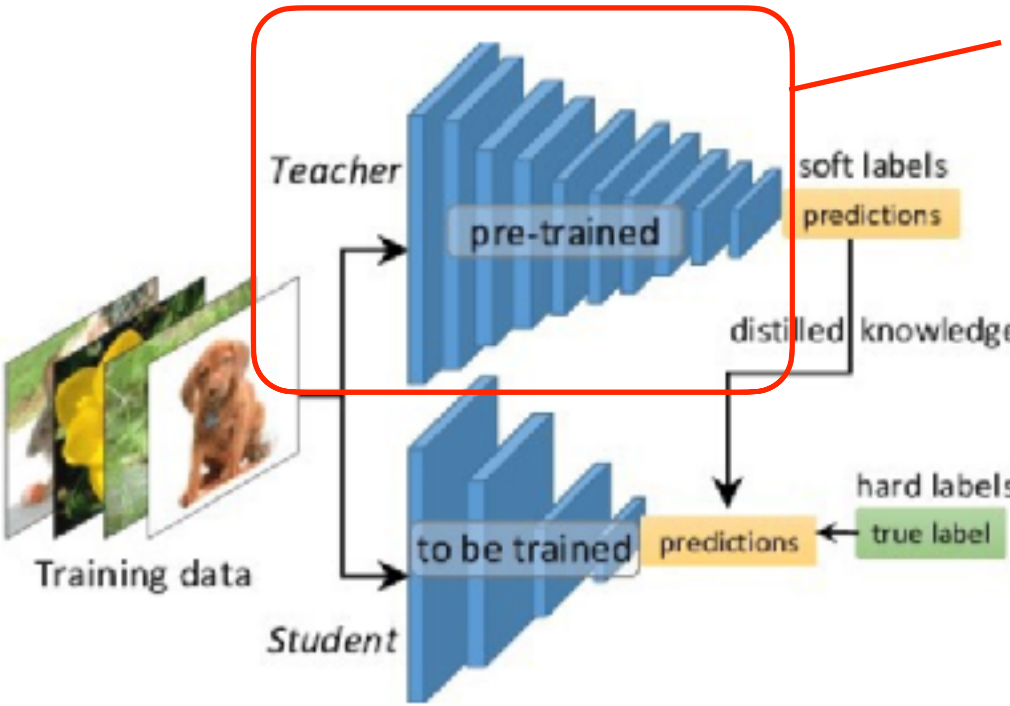


Hinton et al., "Distilling the Knowledge in a Neural Network", NIPS 2014 Deep Learning Workshop

Knowledge Distillation

- Knowledge distillation refers to the idea of using a pretrained network to supervise another smaller/less powerful network.

Train the teacher model in a standard supervised manner using ground truth labels.

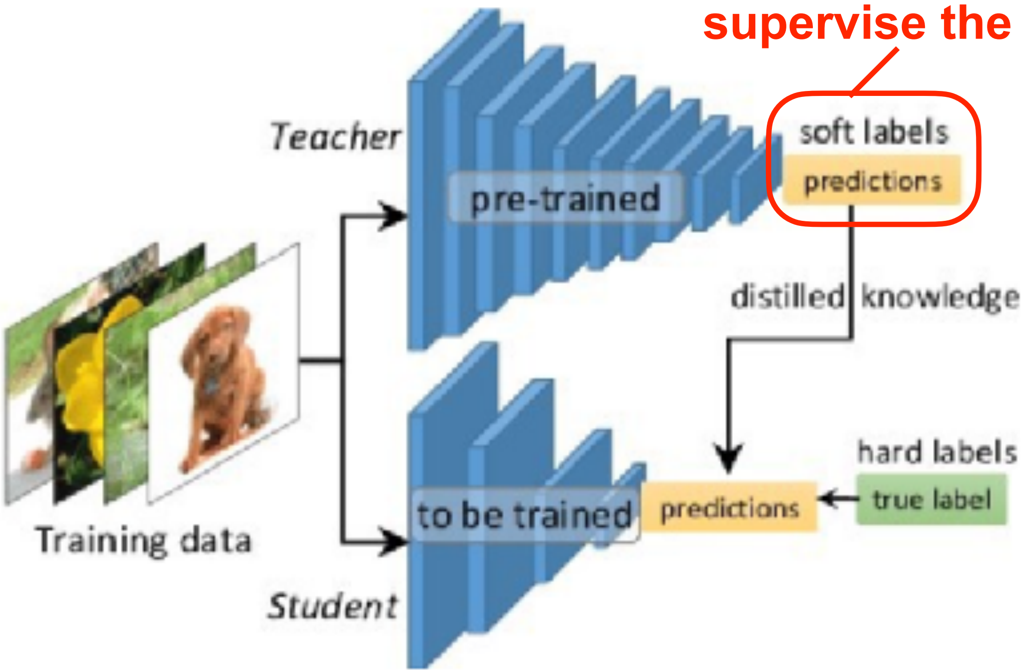


Hinton et al., "Distilling the Knowledge in a Neural Network", NIPS 2014 Deep Learning Workshop

Knowledge Distillation

- Knowledge distillation refers to the idea of using a pretrained network to supervise another smaller/less powerful network.

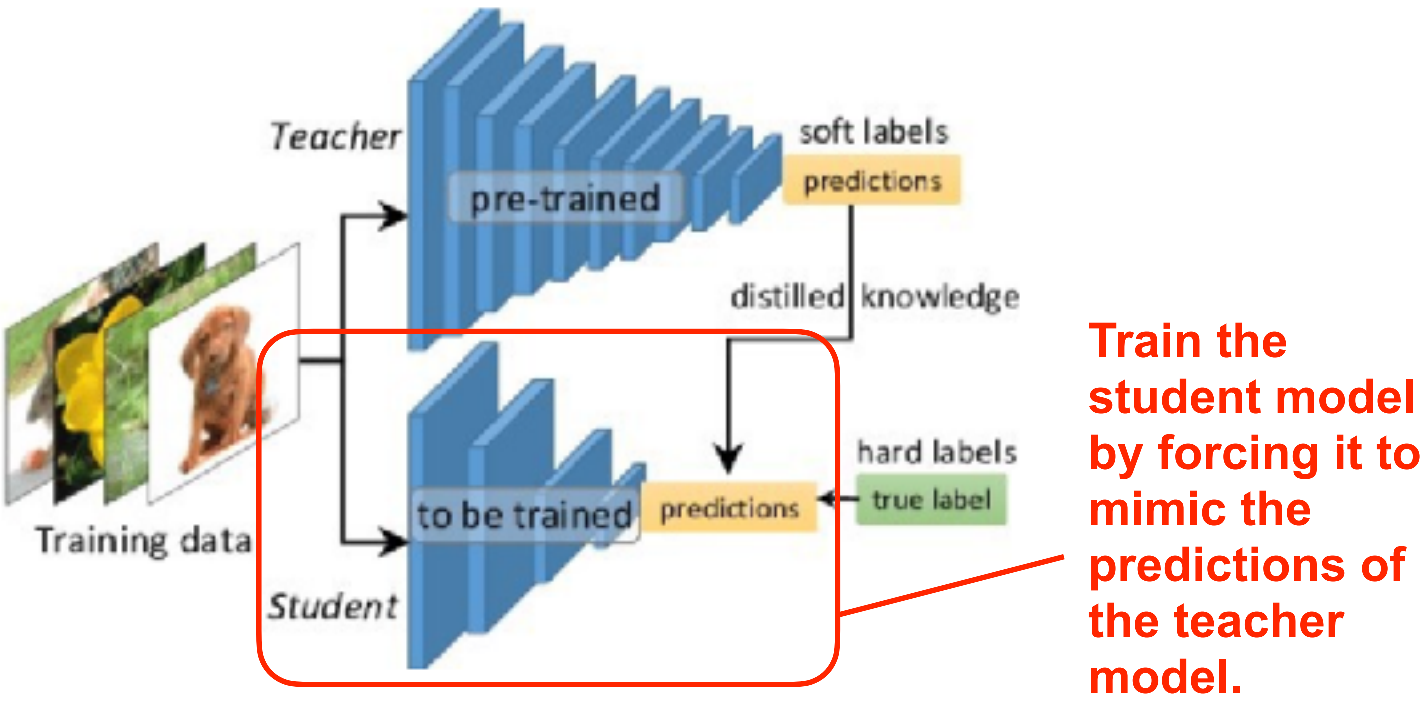
The outputs of the teacher model can then be used to supervise the student model.



Hinton et al., "Distilling the Knowledge in a Neural Network", NIPS 2014 Deep Learning Workshop

Knowledge Distillation

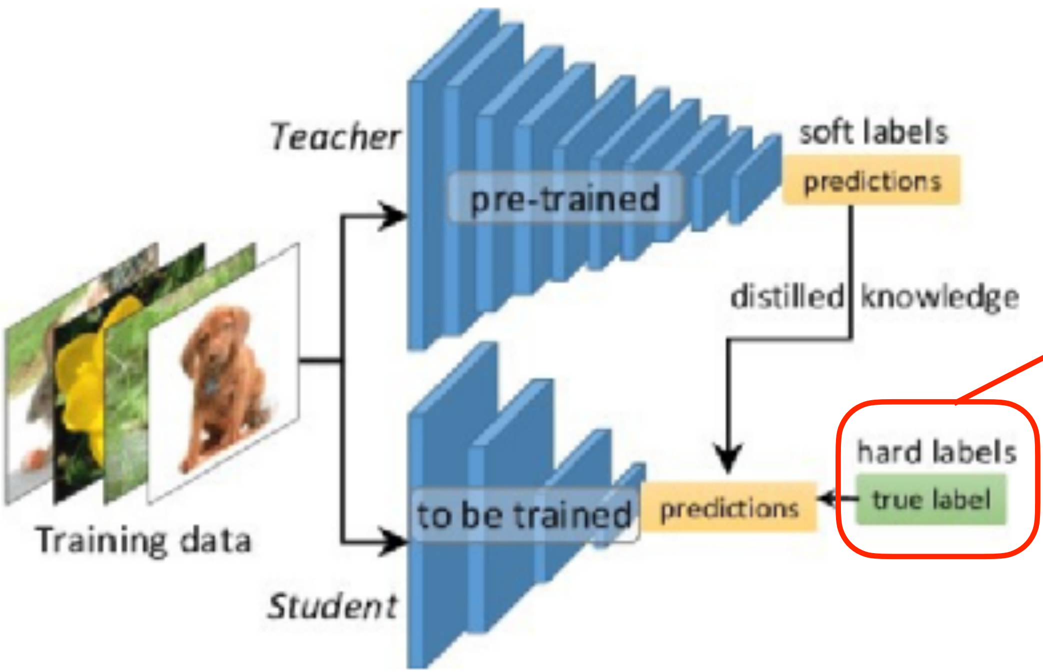
- Knowledge distillation refers to the idea of using a pretrained network to supervise another smaller/less powerful network.



Hinton et al., "Distilling the Knowledge in a Neural Network", NIPS 2014 Deep Learning Workshop

Knowledge Distillation

- Knowledge distillation refers to the idea of using a pretrained network to supervise another smaller/less powerful network.

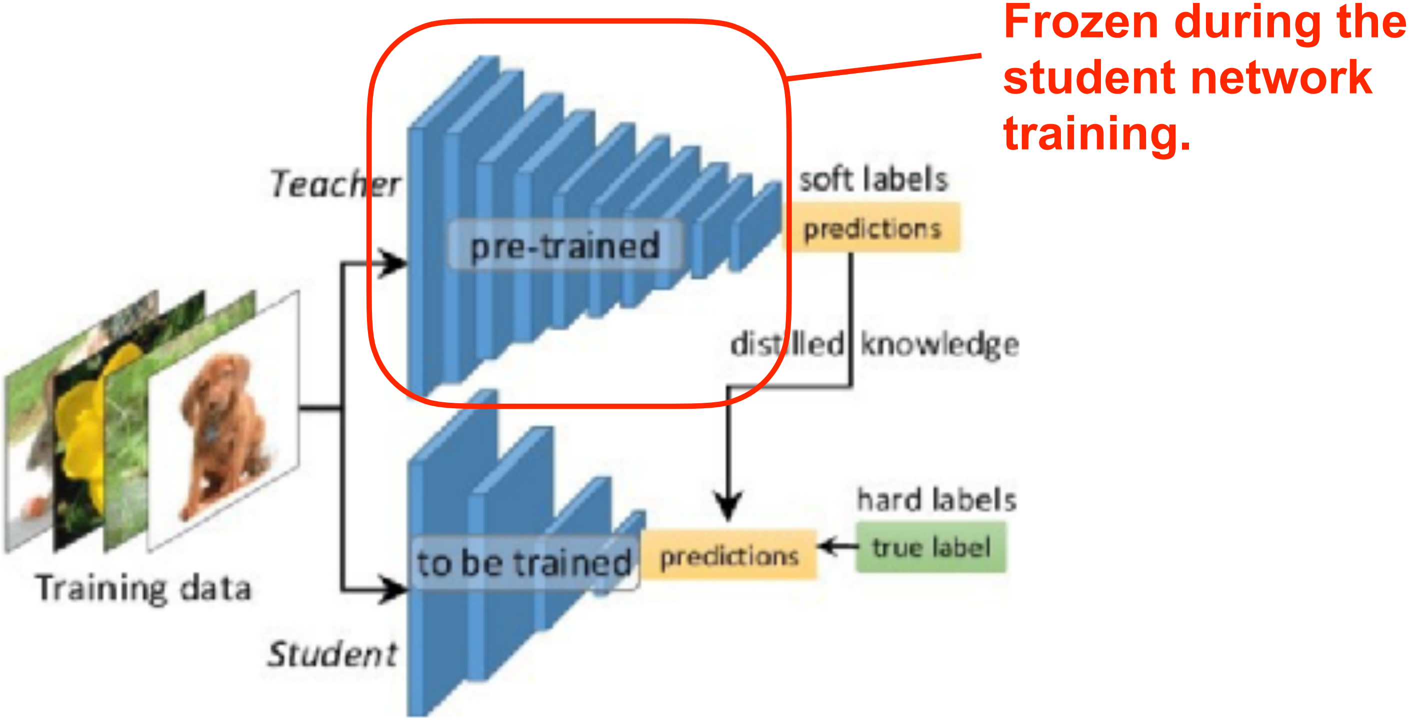


Optionally use ground truth labels as an additional source of supervision.

Hinton et al., "Distilling the Knowledge in a Neural Network", NIPS 2014 Deep Learning Workshop

Knowledge Distillation

- Knowledge distillation refers to the idea of using a pretrained network to supervise another smaller/less powerful network.



Hinton et al., "Distilling the Knowledge in a Neural Network", NIPS 2014 Deep Learning Workshop

Distillation Objective

- Soft distillation minimizes the Kullback-Leibler divergence between the softmax of the teacher and the softmax of the student model.

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

Distillation Objective

- Soft distillation minimizes the Kullback-Leibler divergence between the softmax of the teacher and the softmax of the student model.

$$\mathcal{L}_{\text{global}} = \underbrace{(1-\lambda)\mathcal{L}_{\text{CE}}(\psi(Z_{\text{s}}), y)}_{\text{Standard supervised cross-entropy loss}} + \lambda\tau^2 \text{KL}(\psi(Z_{\text{s}}/\tau), \psi(Z_{\text{t}}/\tau))$$

Distillation Objective

- Soft distillation minimizes the Kullback-Leibler divergence between the softmax of the teacher and the softmax of the student model.

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

Ground truth labels

Distillation Objective

- Soft distillation minimizes the Kullback-Leibler divergence between the softmax of the teacher and the softmax of the student model.

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

The softmax function

The logits of the student model

Distillation Objective

- Soft distillation minimizes the Kullback-Leibler divergence between the softmax of the teacher and the softmax of the student model.

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

Distillation Objective

- Soft distillation minimizes the Kullback-Leibler divergence between the softmax of the teacher and the softmax of the student model.

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

The softmax function

The logits of the teacher model

Distillation Objective

- Soft distillation minimizes the Kullback-Leibler divergence between the softmax of the teacher and the softmax of the student model.

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

The softmax function

The logits of the student model

Distillation Objective

- Soft distillation minimizes the Kullback-Leibler divergence between the softmax of the teacher and the softmax of the student model.

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

the temperature for the distillation

Distillation Objective

- Soft distillation minimizes the Kullback-Leibler divergence between the softmax of the teacher and the softmax of the student model.

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

The coefficient balancing the KL divergence loss and the cross-entropy loss

Hard Distillation Objective

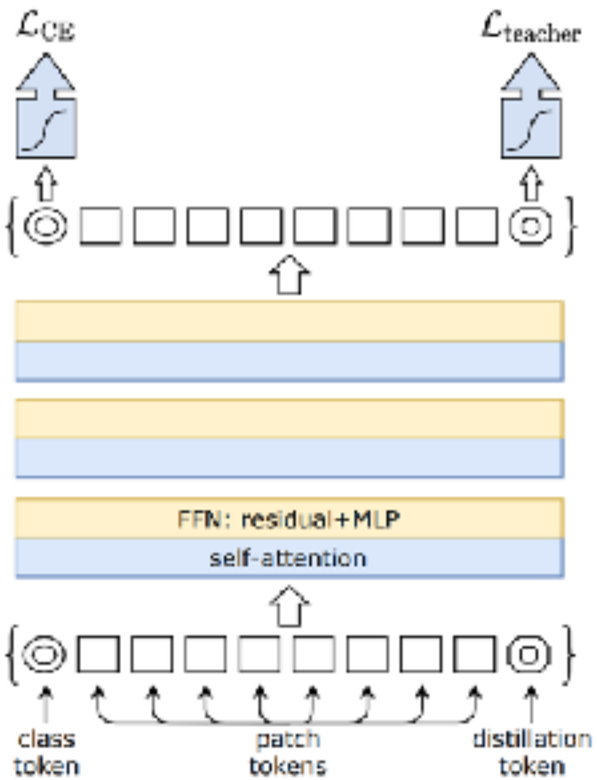
- Hard distillation uses the hard decision of the teacher as a supervisory signal to the student model.

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

$$y_{\text{t}} = \text{argmax}_c Z_{\text{t}}(c)$$

DeiT Architecture

- The authors include a new distillation token, which interacts with the class and patch tokens through self-attention layers.



DeiT Architecture Variants

- Summary of the Data-efficient image Transformer (DeiT) architectures considered in this paper.

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

Distillation Experiments

- ImageNet-1k top-1 accuracy of the student as a function of the teacher model used for distillation.

Teacher		Student: DeiT-B	
Models	acc.	pretrain	↑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	83.0	83.9
RegNetY-16GF	82.9	83.0	84.0

Radosavovic et al., ”Designing Network Design Spaces”, CVPR 2020

Distillation Experiments

- ImageNet-1k top-1 accuracy of the student as a function of the teacher model used for distillation.

Teacher Models	acc.	Student: DeiT-B pretrain	↑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	83.0	83.9
RegNetY-16GF	82.9	83.0	84.0

Image transformers learn more from a CNN than from another transformer with comparable performance

Distillation Experiments

- Distillation experiments on ImageNet-1k with DeiT, 300 epochs of pre-training.

DeiT: method ↓	supervision label	supervision teacher	ImageNet top-1 (%) Ti 224	S 224	B 224	B↑384
no distillation	✓	✗	72.2	79.8	81.8	83.1
usual distillation	✗	soft	72.2	79.8	81.8	83.2
hard distillation	✗	hard	74.3	80.9	83.0	84.0
class embedding	✓	hard	73.9	80.9	83.0	84.2
distil. embedding	✓	hard	74.6	81.1	83.1	84.4
DeiT⚗: class+distil.	✓	hard	74.5	81.2	83.4	84.5

Distillation Experiments

- Distillation experiments on ImageNet-1k with DeiT, 300 epochs of pre-training.

DeiT: method ↓	supervision		ImageNet top-1 (%)			
	label	teacher	Ti 224	S 224	B 224	B↑384
no distillation	✓	✗	72.2	79.8	81.8	83.1
usual distillation	✗	soft	72.2	79.8	81.8	83.2
hard distillation	✗	hard	74.3	80.9	83.0	84.0
class embedding	✓	hard	73.9	80.9	83.0	84.2
distil. embedding	✓	hard	74.6	81.1	83.1	84.4
DeiT⚗: class+distil.	✓	hard	74.5	81.2	83.4	84.5

Distillation Experiments

- Distillation experiments on ImageNet-1k with DeiT, 300 epochs of pre-training.

	supervision		ImageNet top-1 (%)			
DeiT: method ↓	label	teacher	Ti 224	S 224	B 224	B↑384
no distillation	✓	✗	72.2	79.8	81.8	83.1
usual distillation	✗	soft	72.2	79.8	81.8	83.2
hard distillation	✗	hard	74.3	80.9	83.0	84.0
class embedding	✓	hard	73.9	80.9	83.0	84.2
distil. embedding	✓	hard	74.6	81.1	83.1	84.4
DeiT⚗: class+distil.	✓	hard	74.5	81.2	83.4	84.5

Distillation Experiments

- Distillation experiments on ImageNet-1k with DeiT, 300 epochs of pre-training.

DeiT: method ↓	supervision label	supervision teacher	ImageNet top-1 (%) Ti 224	S 224	B 224	B↑384
no distillation	✓	✗	72.2	79.8	81.8	83.1
usual distillation	✗	soft	72.2	79.8	81.8	83.2
hard distillation	✗	hard	74.3	80.9	83.0	84.0
class embedding	✓	hard	73.9	80.9	83.0	84.2
distil. embedding	✓	hard	74.6	81.1	83.1	84.4
DeiT⚗: class+distil.	✓	hard	74.5	81.2	83.4	84.5

Hard distillation outperforms standard supervision and standard soft distillation.

Distillation Experiments

- Distillation experiments on ImageNet-1k with DeiT, 300 epochs of pre-training.

DeiT: method ↓	supervision label	supervision teacher	ImageNet top-1 (%) Ti 224	S 224	B 224	B↑384
no distillation	✓	✗	72.2	79.8	81.8	83.1
usual distillation	✗	soft	72.2	79.8	81.8	83.2
hard distillation	✗	hard	74.3	80.9	83.0	84.0
class embedding	✓	hard	73.9	80.9	83.0	84.2
distil. embedding	✓	hard	74.6	81.1	83.1	84.4
DeiT⚗: class+distil.	✓	hard	74.5	81.2	83.4	84.5

The newly added distillation token leads to improved accuracy.

Data Augmentations

- Mixup
- CutMix
- Random Erasing

Mixup

- Mixup mixes two samples by interpolating both the image and labels.

```
# y1, y2 should be one-hot vectors
for (x1, y1), (x2, y2) in zip(loader1, loader2):
    lam = numpy.random.beta(alpha, alpha)
    x = Variable(lam * x1 + (1. - lam) * x2)
    y = Variable(lam * y1 + (1. - lam) * y2)
    optimizer.zero_grad()
    loss(net(x), y).backward()
    optimizer.step()
```

Input Image:

Label:
Dog 0.5
Cat 0.5

Zhang et al., "MIXUP: BEYOND EMPIRICAL RISK MINIMIZATION", ICLR 2018

CutMix

- Patches are cut and pasted among training images.
- The ground truth labels are also mixed proportionally to the area of the patches.

$$\tilde{x} = \mathbf{M} \odot x_A + (\mathbf{1} - \mathbf{M}) \odot x_B$$
$$\tilde{y} = \lambda y_A + (1 - \lambda) y_B,$$

where $\mathbf{M} \in \{0,1\}^{W \times H}$ denotes a binary mask indicating where to drop out and fill in from two images

Input Image:

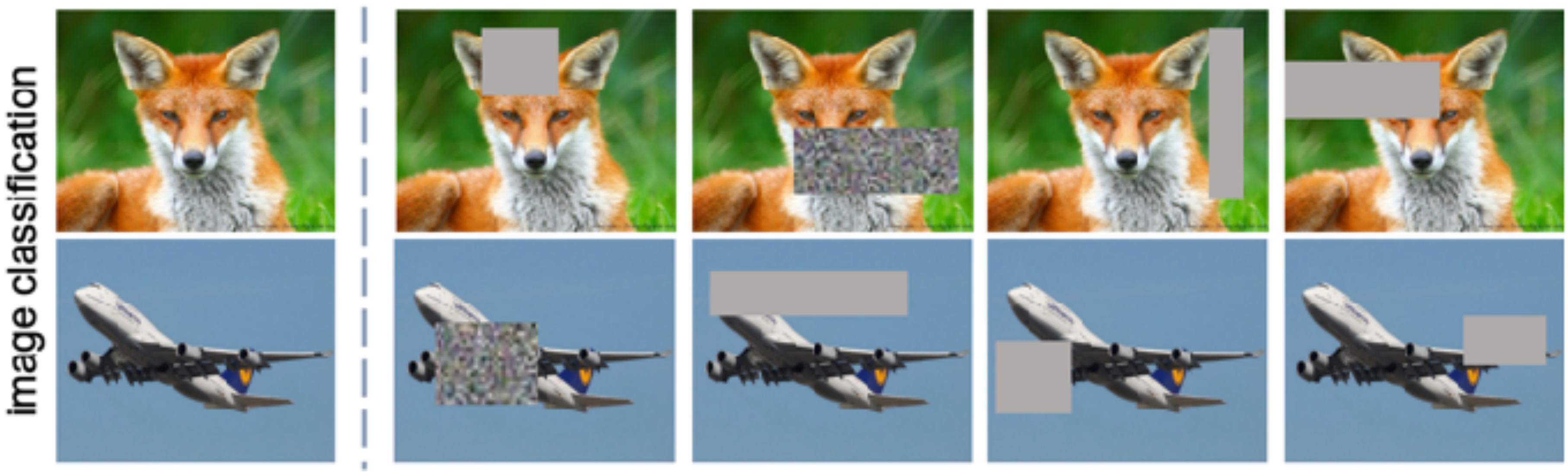
Label:

Dog 0.6
Cat 0.4

Yun et al., "CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features", ICCV 2019

Random Erasing

- Randomly choosing a rectangle region in the image and erase its pixels.
- Images with various levels of occlusion are generated.



Zhong et al., ”Random Erasing Data Augmentation”, AAAI 2020

Ablation Experiments

- Ablation study on data augmentations and regularization schemes evaluated on ImageNet.

Rand-Augment	AutoAug	Mixup	CutMix	Erasing	Stoch. Depth	Repeated Aug.	Dropout	Exp. Moving Avg.	pre-trained 224	fine-tuned 384
✓	✗	✓	✓	✓	✓	✓	✗	✗	81.8±[illegible]	83.1±[illegible]
✓	✗	✓	✓	✓	✓	✓	✗	✗	74.5	77.3
✓	✗	✓	✓	✓	✓	✓	✗	✗	81.8	83.1
✗	✗	✓	✓	✓	✓	✓	✗	✗	79.6	80.4
✗	✓	✓	✓	✓	✓	✓	✗	✗	81.2	81.9
✓	✗	✗	✓	✓	✓	✓	✗	✗	78.7	79.8
✓	✗	✓	✗	✓	✓	✓	✗	✗	80.0	80.6
✓	✗	✗	✗	✓	✓	✓	✗	✗	75.8	76.7
✓	✗	✓	✓	✗	✓	✓	✗	✗	4.3*	0.1
✓	✗	✓	✓	✓	✗	✓	✗	✗	3.4*	0.1
✓	✗	✓	✓	✓	✓	✗	✗	✗	76.5	77.4
✓	✗	✓	✓	✓	✓	✓	✓	✗	81.3	83.1
✓	✗	✓	✓	✓	✓	✓	✗	✓	81.9	83.1

Ablation Experiments

- Ablation study on data augmentations and regularization schemes evaluated on ImageNet.

Rand-Augment	AutoAug	Mixup	CutMix	Erasing	Stoch. Depth	Repeated Aug.	Dropout	Exp. Moving Avg.	pre-trained 224	fine-tuned 384
✓	✗	✓	✓	✓	✓	✓	✗	✗	81.8±[illegible]	83.1±[illegible]
✓	✗	✓	✓	✓	✓	✓	✗	✗	74.5	77.3
✓	✗	✓	✓	✓	✓	✓	✗	✗	81.8	83.1
✗	✗	✓	✓	✓	✓	✓	✗	✗	79.6	80.4
✗	✓	✓	✓	✓	✓	✓	✗	✗	81.2	81.9
✓	✗	✗	✓	✓	✓	✓	✗	✗	78.7	79.8
✓	✗	✓	✗	✓	✓	✓	✗	✗	80.0	80.6
✓	✗	✗	✗	✓	✓	✓	✗	✗	75.8	76.7
✓	✗	✓	✓	✗	✓	✓	✗	✗	4.3*	0.1
✓	✗	✓	✓	✓	✗	✓	✗	✗	3.4*	0.1
✓	✗	✓	✓	✓	✓	✗	✗	✗	76.5	77.4
✓	✗	✓	✓	✓	✓	✓	✓	✗	81.3	83.1
✓	✗	✓	✓	✓	✓	✓	✗	✓	81.9	83.1

Most data augmentations lead to significant boost in performance.

Ablation Experiments

- Ablation study on data augmentations and regularization schemes evaluated on ImageNet.

Rand-Augment	AutoAug	Mixup	CutMix	Erasing	Stoch. Depth	Repeated Aug.	Dropout	Exp. Moving Avg.	pre-trained 224	fine-tuned 384
✓	✗	✓	✓	✓	✓	✓	✗	✗	81.8±[illegible]	83.1±[illegible]
✓	✗	✓	✓	✓	✓	✓	✗	✗	74.5	77.3
✓	✗	✓	✓	✓	✓	✓	✗	✗	81.8	83.1
✗	✗	✓	✓	✓	✓	✓	✗	✗	79.6	80.4
✗	✓	✓	✓	✓	✓	✓	✗	✗	81.2	81.9
✓	✗	✗	✓	✓	✓	✓	✗	✗	78.7	79.8
✓	✗	✓	✗	✓	✓	✓	✗	✗	80.0	80.6
✓	✗	✗	✗	✓	✓	✓	✗	✗	75.8	76.7
✓	✗	✓	✓	✗	✓	✓	✗	✗	4.3*	0.1
✓	✗	✓	✓	✓	✗	✓	✗	✗	3.4*	0.1
✓	✗	✓	✓	✓	✓	✗	✗	✗	76.5	77.4
✓	✗	✓	✓	✓	✓	✓	✓	✗	81.3	83.1
✓	✗	✓	✓	✓	✓	✓	✗	✓	81.9	83.1

Several regularization schemes boost performance as well.