DeiT-LT: Distillation Strikes Back for Vision Transformer Training on Long-Tailed Datasets

Harsh Rangwani[*1] Pradipto Mondal[*1,2] Mayank Mishra[*1]
Ashish Ramayee Asokan[1] R. Venkatesh Babu[1]

[1]Indian Institute of Science, Bangalore [2] Indian Institute of Technology, Kharagpur

CVPR 2024

Background

1. Unlike CNN, ViT's simple architecture has no informative inductive bias (e.g., locality, etc.). ViT requires a large amount of data for pre-training

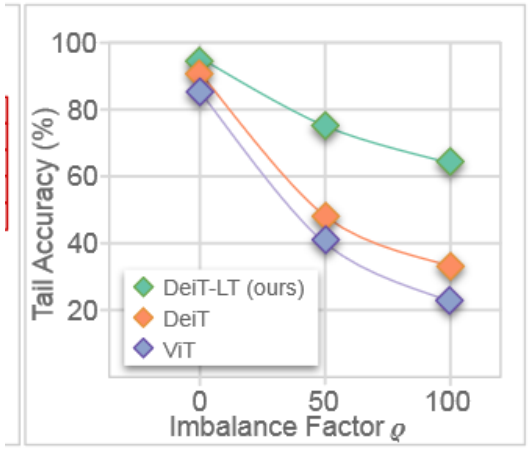
2. Re-weighting/re-sampling: focus on the tail classes, often lead to some performance degradation in the head. To mitigate this, multiple expert networks specialize in different portions of the data distribution. However, all these efforts have been restricted to CNNs.

Motivation

The data-efficient transformers (DeiT) aimed to reduce this requirement for pre-training by distilling information from a pre-trained CNN.

However, all these improvements have been primarily based on increasing performance on the balanced ImageNet dataset.

insufficient for long-tailed datasets.



(c) Tail acc. for DeiT-LT (ours) vs baselines

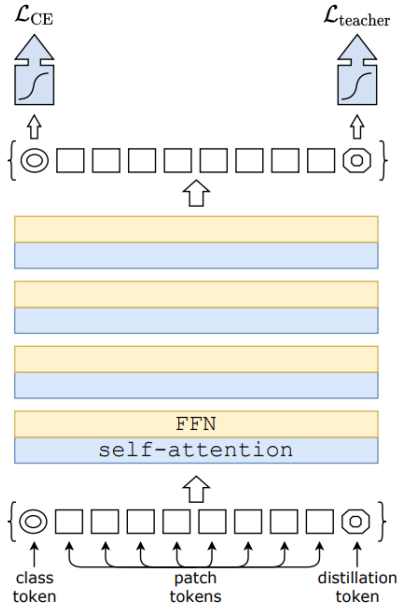
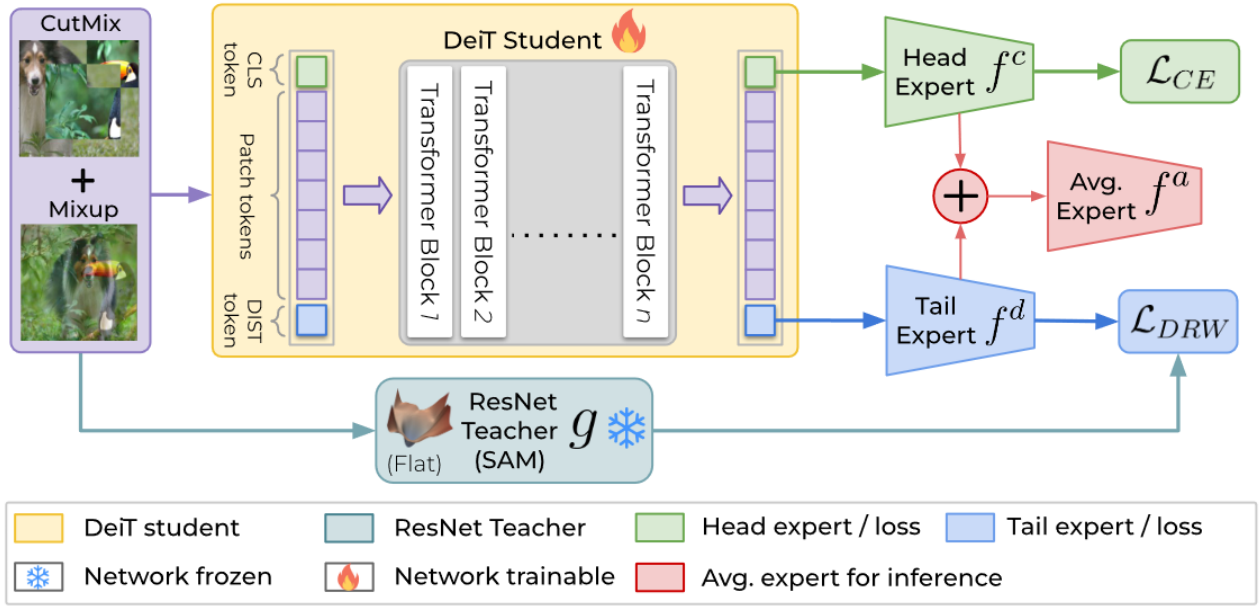


Figure 2: Our distillation procedure: we simply include a new *distillation token*. It interacts with the class and patch tokens through the self-attention layers. This distillation token is employed in a similar fashion as the class token, except that on output of the network its objective is to reproduce the (hard) label predicted by the teacher, instead of true label. Both the class and distillation tokens input to the transformers are learned by back-propagation.

Method

In this work, we aim to investigate and improve the training of Vision Transformers from scratch without the need for large-scale pre-training on diverse long-tailed datasets, varying in image size and resolution.

Recent works show improved performance for ViTs on long-tailed recognition tasks, but they often need expensive pre-training on large-scale extra datasets and do not generalize well to other domains like medical, synthetic etc.



a) the effective distillation via OOD images.
b) training Tail Expert classifier using DRW loss.
c) learning of low-rank generalizable features from flat teachers via distillation.

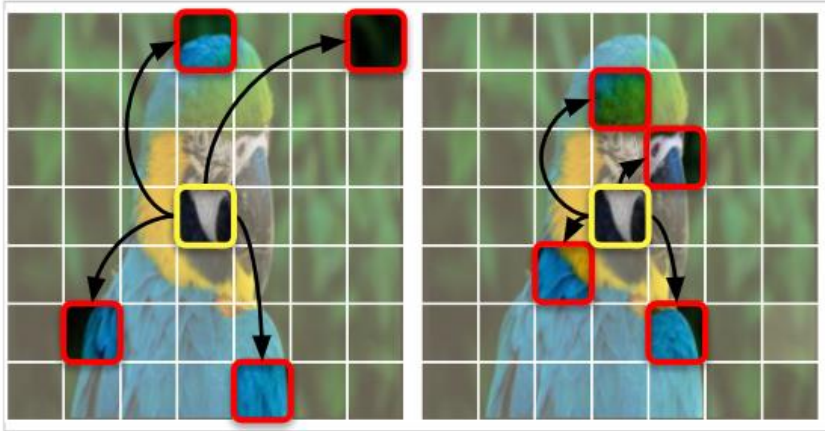
Method/Distillation via Out of Distribution Images

Table 1. **Effect of augmentations:** Comparison of teacher (*Tch*) and student (*Stu*) accuracy (%) and training time (in hours) on CIFAR-10 LT ($\rho = 100$) using various augmentation strategies with mixup (✓) and without mixup (✗). Despite low teacher training accuracy on the out-of-distribution images, the student (Stu.) performs better on the validation set.

Tch Model	Stu Augs.	Tch Augs.	Tch Acc.	Stu Acc.	Train Time
RegNetY 16GF	Strong (✓)	Strong (✓)	79.1	70.2	33.3
ResNet-32	Strong (✗)	Weak (✗)	97.2	54.2	17.8
	Strong (✗)	Strong (✗)	71.9	69.6	17.8
	Strong (✓)	Strong (✓)	56.6	79.4	19.0

This works because the ViT student learns to mimic the incorrect predictions of the CNN teacher on the out-of-distribution images, which in turn enables the student to learn the inductive biases (locality) of the teacher.

$$f^d(X) \approx g(X), X \sim A(x)$$



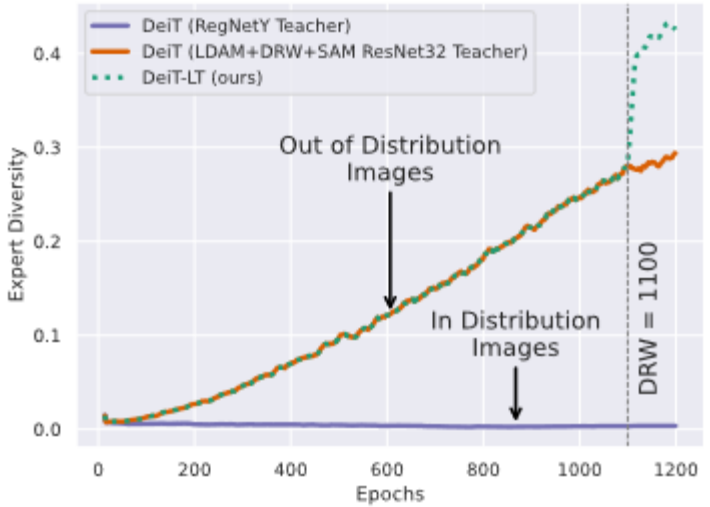
(a) Comparison b/w attention maps of DeiT (left) and DeiT-LT (ours, right)

Method/Distillation via Out of Distribution Images

Due to distillation via out-of-distribution images, the teacher predictions y_t often diffe from the ground truth y. Hence, the classification token (CLS) and distillation token (DIST) representations diverge while training.

Our observation debunks the myth that it is required for the CLS token predictions to be similar to DIST for effective distillation in transformer, as observed by Touvron et al. [48].

$$\mathcal{L} = \frac{1}{2}\mathcal{L}_{CE}(f^c(x), y) + \frac{1}{2}\mathcal{L}_{DRW}(f^d(x), y_t),$$
$$\text{where } \mathcal{L}_{DRW} = -w_{y_t}\ log(f^d(x)_{y_t})$$



(a) Diversity for CLS and DIST experts

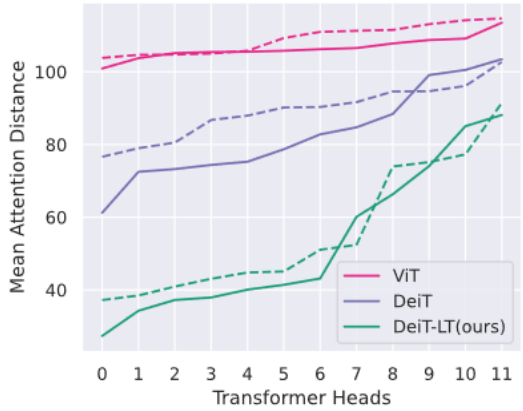
Method/Distillation via Out of Distribution Images

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

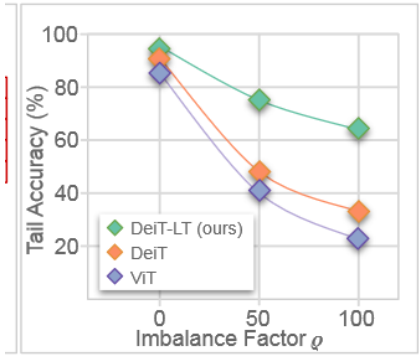
To gain insights into the generality and effectiveness of OOD Distillation, we take a closer look at the tail features produced by DeiT-LT.

Without the OOD distillation, we find that the vanilla DeiT-III and ViT baselines overfit only on the spurious global features and do not generalize well for tail classes.

the early self attention:
block 1 (solid)
block 2 (dashed



(b) Locality of Attention Heads



(c) Tail acc. for DeiT-LT (ours) vs baselines

Method/Low-Rank Features via SAM teachers

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

$$\mathcal{X}_{all}, \mathcal{X}_{min} \subset \mathcal{X}, \text{where } \mathcal{X}_{all}, \mathcal{X}_{min}$$

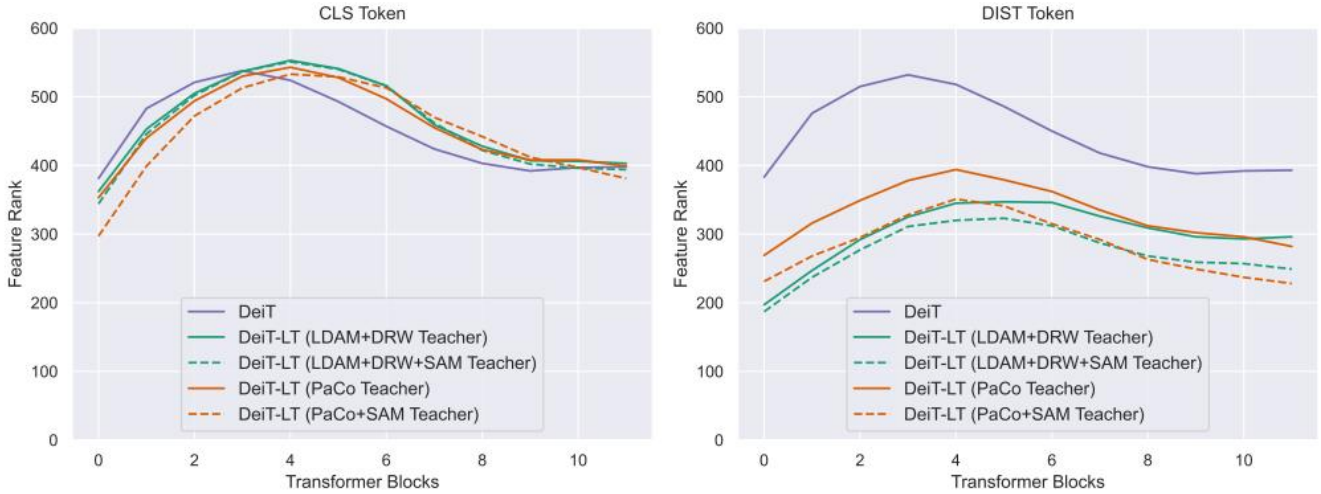
$$F^{all}_{n_h,d} \quad F^{min}_{n_t,d}$$

$$U, S, V^T = \text{SVD}(F^{all}_{n_h,d}),$$

$$F^{min}_{proj}(k) = F^{min}_{n_t,d} * V_k$$

$$\frac{||F^{min}_{n_t,d} - F^{min}_{recon}(k)||^2}{||F^{min}_{n_t,d}||^2} \leq 0.01$$

$$F^{min}_{recon}(k) = F^{min}_{proj}(k) * {V_k}^T.$$



(a) Rank of ViT from Distillation of CNN teachers using `CLS` token

(b) Rank of ViT from Distillation of CNN teachers using `DIST` token

Figure S.4. We compare the rank calculated using features from the a) `CLS` token and b) `DIST` token when trained on CIFAR-10 LT. Our DeiT-LT captures both fine-grained features (from high-rank `CLS` token) and generalizable features (from low-rank `DIST` token).

By learning semantic similar features, our training of DIST token ensures good representation learning for minority classes by leveraging the discriminative features learned from majority classes.

Experiment

Table 2. Results on CIFAR-10 LT and CIFAR-100 LT datasets with ρ=50 and ρ=100. We report the *overall* accuracy for available methods. (The teacher used to train the respective student (DeiT-LT) model can be identified by matching superscripts)

Method	CIFAR-10 LT		CIFAR-100 LT	
	$\rho = 100$	$\rho = 50$	$\rho = 100$	$\rho = 50$
ResNet32 Backbone				
CB Focal loss [9]	74.6	79.3	38.3	46.2
LDAM+DRW [5]	77.0	79.3	42.0	45.1
LDAM+DAP [19]	80.0	82.2	44.1	49.2
BBN [67]	79.8	82.2	39.4	47.0
CAM [64]	80.0	83.6	47.8	51.7
Log. Adj. [32]	77.7	-	43.9	-
RIDE [56]	-	-	49.1	-
MiSLAS [65]	82.1	85.7	47.0	52.3
Hybrid-SC [55]	81.4	85.4	46.7	51.9
SSD [27]	-	-	46.0	50.5
ACE [4]	81.4	84.9	49.6	51.9
GCL [26]	82.7	85.5	48.7	53.6
VS [23]	78.6	-	41.7	
VS+SAM [38]	82.4	-	46.6	-
[1]L-D-SAM [38]	81.9	84.8	45.4	49.4
[2]PaCo+SAM[8, 38]	86.8	88.6	52.8	56.6
ViT-B Backbone				
ViT [12]	62.6	70.1	35.0	39.0
ViT (cRT) [20]	68.9	74.5	38.9	42.2
DeiT [48]	70.2	77.5	31.3	39.1
DeiT-III [51]	59.1	68.2	38.1	44.1
[1]DeiT-LT(ours)	84.8	87.5	52.0	54.1
[2]DeiT-LT(ours)	**87.5**	**89.8**	**55.6**	**60.5**

Table 3. Results on ImageNet-LT. (The teacher used to train respective student (DeiT-LT) can be identified by matching superscripts)

Method	ImageNet-LT			
	Overall	Head	Mid	Tail
ResNet50 Backbone				
CB Focal loss [9]	33.2	39.6	32.7	16.8
LDAM [5]	49.8	60.4	46.9	30.7
c-RT [20]	49.6	61.8	46.2	27.3
τ-Norm [21]	49.4	59.1	46.9	30.7
Log. Adj. [32]	50.1	61.1	47.5	27.6
RIDE(3 exps) [56]	54.9	66.2	51.7	34.9
MiSLAS [65]	52.7	62.9	50.7	34.3
Disalign [63]	52.9	61.3	52.2	31.4
TSC [28]	52.4	63.5	49.7	30.4
GCL [26]	54.5	63.0	52.7	37.1
SAFA [17]	53.1	63.8	49.9	33.4
BCL [41]	57.1	67.9	54.2	36.6
ImbSAM [68]	55.3	63.2	53.7	38.3
CBD_{ENS} [18]	55.6	68.5	52.7	29.2
[1]L-D-SAM [38]	53.1	62.0	52.1	32.8
[2]PaCo+SAM [8, 38]	57.5	62.1	58.8	39.3
ViT-B Backbone				
ViT [12]	37.5	56.9	30.4	10.3
DeiT-III [51]	48.4	**70.4**	40.9	12.8
[1]DeiT-LT(ours)	55.6	65.2	54.0	37.1
[2]DeiT-LT(ours)	**59.1**	66.6	**58.3**	**40.0**

Table 4. Results on iNaturalist-2018. (The teacher used to train student (DeiT-LT) can be identified by matching superscripts)

Method	iNaturalist-2018			
	Overall	Head	Mid	Tail
ResNet50 Backbone				
c-RT [20]	65.2	69.0	66.0	63.2
τ-Norm [21]	65.6	65.6	65.3	65.9
RIDE(3 exps) [56]	72.2	70.2	72.2	72.7
MiSLAS [65]	71.6	**73.2**	72.4	70.4
Disalign [63]	70.6	69.0	71.1	70.2
TSC [28]	69.7	72.6	70.6	67.8
GCL [26]	71.0	67.5	71.3	71.5
ImbSAM [68]	71.1	68.2	72.5	72.9
CBD_{ENS} [18]	73.6	75.9	74.7	71.5
[1]L-D-SAM [38]	70.1	64.1	70.5	71.2
[2]PaCo+SAM [38]	73.4	66.3	73.6	75.2
ViT-B Backbone				
ViT [12]	54.2	64.3	53.9	52.1
DeiT-III [51]	61.0	72.9	62.8	55.8
[1]DeiT-LT(ours)	72.9	69.0	73.3	73.3
[2]DeiT-LT(ours)	**75.1**	70.3	**75.2**	**76.2**

Experiment

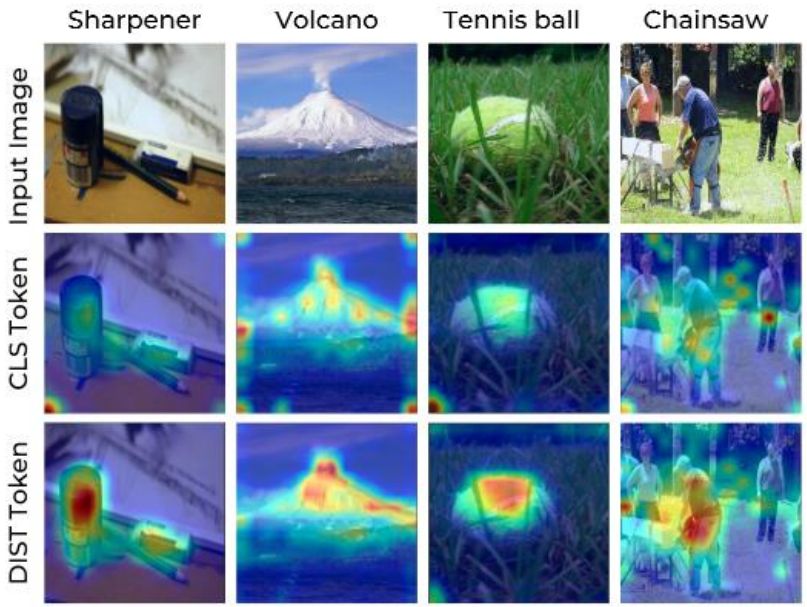


Figure 5. Visual comparison of the attention maps with respect to the CLS and DIST tokens for *tail* images from the ImageNet-LT dataset. The attention maps are computed by *Attention Rollout* [1].

Table 5. Table showing ablations for various components in DeiT-LT for CIFAR-10 LT and CIFAR-100 LT.

OOD Distill	DRW	SAM	C10 LT	C100 LT
✗	✗	✗	70.2	31.3
✓	✗	✗	84.5	48.9
✓	✓	✗	87.3	54.5
✓	✓	✓	87.5	55.6

Table 6. Analysis across transformer capacity for CIFAR-10 LT and CIFAR-100 LT for DeiT-LT student($\rho = 100$) with PaCo teacher.

Model	Overall	Head	Mid	Tail
CIFAR-10 LT ($\rho = 100$)				
DeiT-LT Tiny (Ti)	80.8	89.7	75.1	79.4
DeiT-LT Small (S)	85.5	92.7	81.5	83.7
DeiT-LT Base (B)	87.5	94.5	84.1	85.0
CIFAR-100 LT ($\rho = 100$)				
DeiT-LT Tiny (Ti)	49.3	66.3	50.0	27.3
DeiT-LT Small (S)	54.3	72.6	54.8	31.1
DeiT-LT Base (B)	55.6	73.1	56.9	32.1

Thank you