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

Supplementary Material

Table of Contents

A. Experimental Details

A.1. Datasets

CIFAR-10 LT and CIFAR-100 LT. We use the imbalanced CIFAR-10 and CIFAR-100 datasets with an exponential decay in sample size across classes. This decay is guided by the Imbalance Ratio ($\rho = \frac{\max_i N_i}{\min_j N_j}$). For our experiments on CIFAR-10 LT and CIFAR-100 LT, we show the results on $\rho = 100$ and $\rho = 50$. CIFAR-10 LT comprises 12,406 training images across 10 classes (ρ = 100). Out of the 10 classes, the first 3 classes are considered *Head* classes with more than 1500 images per class, the following 4 classes are *Mid* (medium) classes with more than 250 images each class, and the last 3 classes account for the *Tail* classes, with each class containing less than 250 images each. Following a similar decay, the 100 classes of CIFAR-100 LT (10,847 training samples with ρ = 100) are also divided into three subcategories: the first 36 classes are considered as the *Head* classes, *Mid* contains the following 35 classes, and the remaining 29 classes are labeled as *Tail* classes. Both CIFAR-10 LT and CIFAR-100 LT datasets are evaluated on held-out sets of 10,000 images each, equally distributed across all classes.

ImageNet-LT. We use the standard LT dataset created out of ImageNet [42]. ImageNet-LT consists of 115,846 training images, with 1280 images in the class with the most images and 5 images in the class with the least images. Out of the 1,000 classes sorted in the descending order of sample frequency, we consider classes with more than 100 samples as *Head* classes, the classes with samples between 20 and 100 to be *Mid* classes and the classes with less than 20 samples as the *Tail* classes as done in Cui et al. [8].

iNaturalist-2018. iNaturalist-2018 [52] is a real-world imbalanced dataset with 437,513 training images. Out of the 8,142 classes sorted in the descending order of sample frequency, we consider classes with more than 100 samples as *Head* classes, the classes with samples between 20 and 100 to be *Mid* classes and the classes with less than 20 samples as the *Tail* classes, similar to ImageNet-LT.

A.2. Training Configuration

In this subsection, we detail the strategies adopted to train DeiT-LT Base (B) model on four benchmark datasets, namely CIFAR-10 LT, CIFAR-100 LT, ImageNet-LT, and iNaturalist-2018. We use the AdamW optimizer to train DeiT-LT from scratch across all the datasets. These runs use a cosine learning rate decay schedule with an initial learning rate of 5×10^{-4}. All the runs use a linear learning rate warm-up schedule for the initial five epochs. Furthermore, we deploy label smoothing with $\varepsilon = 0.1$ for all our experiments where the ground truth labels are used to train the `CLS` expert. Under label smoothing, the true label is assigned a $(1-\varepsilon)$ probability, and the remaining ε is distributed amongst the other labels. We use hard labels as distillation targets from the teacher network to train the `DIST` expert classifier via distillation from CNN teacher (Fig. 2). For training the teacher networks with SAM optimizer, we follow the setup mentioned in [38]

CIFAR-10 LT and CIFAR-100 LT : We train DeiT-LT for 1200 epochs on imbalanced versions of CIFAR datasets. DRW loss is added to the training of the `DIST` expert classifier after 1100 epochs. Mixup and Cutmix are used during the initial 1100 epochs of the training. As suggested in [48], we use Repeated Augmentation to improve the performance of the DeiT-LT training. The (32× 32) images of CIFAR datasets are resized to (224 × 224) before feeding into the transformer architecture. For CIFAR-10 LT and CIFAR-100 LT datasets, ResNet-32 is used as the teacher network. The

Table S.1. Summary of our training procedures used to train DeiT-LT Base (B) from scratch on CIFAR-10 LT, CIFAR-100 LT, ImageNet-LT and iNaturalist-2018.

Procedure	CIFAR-10 LT	CIFAR-100 LT	ImageNet-LT	iNaturalist-2018
Epochs	1200	1200	1400	1000
Optimizer	AdamW	AdamW	AdamW	AdamW
Effective Batch Size	1024	1024	2048	2048
LR	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}
LR schedule	cosine	cosine	cosine	cosine
Warmup Epochs	5	5	5	5
DRW starting epoch	1100	1100	1200	900
Mixup (α)	0.8	0.8	0.8	0.8
Cutmix (α)	1.0	1.0	1.0	1.0
Mixup and Cutmix during DRW	×	×	✓	✓
Horizontal Flip	✓	✓	✓	✓
Color Jitter	✓	✓	✓	✓
Random Erase	✓	✓	×	×
Label smoothing	0.1	0.1	0.1	0.1
Solarization	×	×	✓	✓
Random Grayscale	×	×	✓	✓
Repeated Aug	✓	✓	×	×
Auto Aug	✓	✓	×	×

teacher is trained from scratch on these imbalanced datasets using LDAM+DRW+SAM [38] and contrastive PaCo+SAM (training PaCo [8] with SAM [13] optimizer) frameworks. The input images to the teacher are of size (32× 32), with the same augmentation used as input images to the teacher network during DeiT-LT training.

ImageNet-LT and iNaturalist-2018. DeiT-LT is trained from scratch for 1400 epochs on ImageNet-LT and for 1000 epochs on iNaturalist-2018. DRW loss for distillation head (`DIST` expert classifier) is initialized from epochs 1200 and 900 for ImageNet-LT and iNaturalist-2018, respectively. Mixup and Cutmix are used throughout the training, including the DRW training phase. More details regarding the training configuration can be found in Table S.1.

For the ImageNet-LT and iNaturalist-2018 datasets, the ResNet-50 teacher is trained from scratch on the respective datasets using the LDAM+DRW+SAM [38] and contrastive PaCo+SAM (training PaCo [8] with SAM [13] optimizer) methods. The input image size is (224 × 224) for both the student and teacher network.

A.3. Additional Baselines

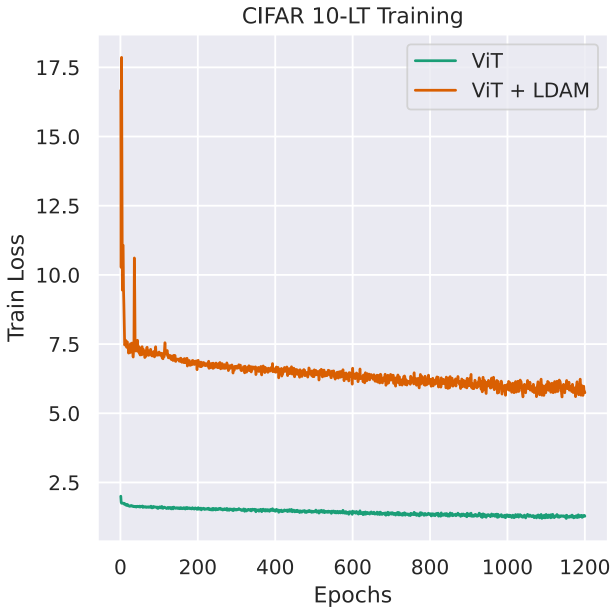
We want to highlight that we attempted training baselines, like LDAM for vanilla ViT. However, we find that the LDAM baseline (52.75%) performs inferiorly to the vanilla ViT baselines (62.62%). We find that the loss for the LDAM baseline gets plateaued very early, and the model does not fit to the training dataset (Fig. S.1). To make the comparison fair with DeiT baselines, we used similar augmentation and other hyperparameters for the ViT Baselines. We think this can be one reason for the non-convergence of the ViT-LDAM baseline. We find that similar abysmal performance for LDAM baseline is also reported by the recent work [59], which also resonates with our finding. We think that investigation into this behavior is a good direction for future work.

Additionally, for a fair comparison, we do not compare against baselines that use pre-training for long-tailed recognition tasks. RAC [46] uses a ViT-B encoder for their retrieval module with weights obtained from pre-training on ImageNet-21K. The authors do not report on small-scale datasets, as they acknowledge the unfair advantage of using the information present in the pretrained encoder. Similarly, for small-scale datasets, LiVT [58] method pretrains the encoder via Masked Generative Pretraining on ImageNet-1k. On the contrary, our DeiT-LT method enables training *ViT from scratch* for both small-scale and large-scale datasets.

A.4. Augmentations for OOD distillation

While both DeiT and our DeiT-LT pass images with strong augmentations to the teacher network for distilling into the student network, the set of augmentations used to train the teacher network itself differs between the two approaches. DeiT first trains a large teacher CNN (RegNetY-16GF) using the same set of strong augmentations as that used for the student network. However, we find that distilling from a small teacher CNN (such as ResNet32) trained with weak aug-

Figure S.1. Comparison of training loss for vanilla ViT and ViT+LDAM training on CIFAR-10 LT



mentations gives better performance (see Sec. 3.1) for more details). Table S.2 compares the augmentations used to train the teacher for DeiT (RegNetY-16GF) and for our method DeiT-LT. Our experiments use ResNet32 as the teacher network for CIFAR-10 LT and CIFAR-100 LT, and ResNet50 for the Imagenet-LT and iNaturalist-2018 datasets. For the PaCo teacher, we utilize the mildly strong augmentations used by the PaCo [8] method itself. We would like to convey, that the PaCo training does not utilize the Mixup and CutMix augmentation in particular while training, which helps us to create OOD samples for this using Mixup and CutMix itself. Distilling via out-of-distribution (OOD) images enables the student to learn the inductive biases of the teacher effectively. This is particularly helpful in improving the performance on the tail classes that have significantly fewer training images.

B. Detailed Results

Performance of individual experts: Our approach focuses on training diverse experts, where the CLS expert classifier is able to perform well on *Head* (majority) classes, while the DIST expert classifier is able to perform well on the *Tail* (minority) classes. By averaging the output of the individual classifiers, we are able to exploit the benefit of both.

In this portion, we discuss the individual performance of the CLS and DIST expert classifiers of our proposed DeiT-LT method on CIFAR-10 LT, CIFAR-100 LT, ImageNet-LT, and iNaturalist-2018. As can be seen in Table S.3 and Table S.4, the CLS and DIST classifiers give a contrasting performance on the head and tail classes, supporting our claim of expert classifiers. For CIFAR-10 LT ($\rho = 100$), the CLS expert classifier is able to report an accuracy of 96.5% on images of the head classes, whereas the DIST expert classifier settles with 72.8% on the same set of classes. On the other hand, the DIST expert classifier reports 93.0% accuracy on the tail classes, which is almost 33% more than that of the CLS expert classifier. Like CIFAR-10 LT, the CLS expert classifier performs better on the head classes of CIFAR-100 LT ($\rho = 100$) than the DIST, whereas the DIST expert classifier reports much higher accuracy on the tail classes. The CLS classifier achieves an accuracy of 73.7% on the head classes, and the DIST expert classifier secures 43.1% accuracy on the tail classes. We notice that by averaging the output of the classifiers, we are able to report good performance in both the majority and the minority classes. CIFAR-10 LT reaches an overall accuracy of 87.3%, with 93.8% on head classes and 85.7% on tail classes. Similarly, with 72.8% on the head and 31.0% on the tail, DeiT-LT is able to secure an overall 54.8% on CIFAR-100 LT. The results demonstrate that there is a parallel trend in the performance of experts for both CIFAR-10 LT and CIFAR-100 LT when ρ is set to 50.

Table S.2. Comparing augmentation used to train RegNetY-16GF (teacher for DeiT training) and ResNet32 (teacher for DeiT-LT training) for CIFAR-10 LT.

Procedure	RegNetY-16GF (Strong)	ResNet32 (Weak)
Image Size	224×224	32×32
Random Crop	✓	✓
Horizontal Flip	✓	✓
Mixup (α)	0.8	×
Cutmix (α)	1.0	×
Color Jitter	0.3	×
Random Erase	✓	×
Auto Aug	✓	×
Repeated Aug	✓	×

A similar trend is seen for large-scale ImageNet-LT and iNaturalist-2018 in Table S.4. For ImageNet-LT, the CLS expert classifier reports 68.3% accuracy on the head classes, which is approximately 11% more reported by the DIST expert classifier. At the same time, we observe that the DIST expert classifier is able to get an accuracy of 46.6% on the tail, which is significantly higher than the 13.5% of the CLS expert classifier. For iNaturalist-2018 as well, the CLS expert classifier achieves a high accuracy of 73.8% on the head classes, and the DIST expert classifier reaches 77.0% on the tail classes. After averaging the outputs of the two classifiers, DeiT-LT reports an overall accuracy of 59.1% for ImageNet-LT and 75.1% for iNaturalist-2018, which would not have been possible by training a standard Vision

Table S.3. Accuracy of expert classifiers on Head, Mid, and Tail classes for CIFAR-10(100) LT.

Imbalance	Expert	CIFAR-10 LT				CIFAR-100 LT			
		Overall	Head	Mid	Tail	Overall	Head	Mid	Tail
100	Average	$87.3_{\pm 0.10}$	$93.8_{\pm 0.33}$	$83.7_{\pm 0.26}$	$85.7_{\pm 0.33}$	$54.8_{\pm 0.42}$	$72.8_{\pm 0.16}$	$55.9_{\pm 0.51}$	$31.0_{\pm 0.73}$
	`CLS`	$78.6_{\pm 0.15}$	$96.5_{\pm 0.06}$	$79.4_{\pm 0.39}$	$59.7_{\pm 0.20}$	$43.3_{\pm 0.39}$	$73.7_{\pm 0.19}$	$41.7_{\pm 0.73}$	$7.5_{\pm 0.26}$
	`DIST`	$79.9_{\pm 0.31}$	$72.8_{\pm 0.92}$	$75.4_{\pm 0.18}$	$93.0_{\pm 0.15}$	$42.5_{\pm 0.48}$	$39.3_{\pm 1.64}$	$45.1_{\pm 0.47}$	$43.1_{\pm 0.33}$
50	Average	$89.9_{\pm 0.17}$	$94.5_{\pm 0.18}$	$87.2_{\pm 0.26}$	$88.8_{\pm 0.34}$	$60.6_{\pm 0.03}$	$74.6_{\pm 0.10}$	$60.5_{\pm 0.10}$	$43.1_{\pm 0.06}$
	`CLS`	$84.1_{\pm 0.33}$	$96.5_{\pm 0.12}$	$83.3_{\pm 0.66}$	$72.8_{\pm 0.55}$	$49.6_{\pm 0.21}$	$76.0_{\pm 0.31}$	$50.5_{\pm 0.46}$	$15.9_{\pm 0.41}$
	`DIST`	$83.2_{\pm 0.23}$	$74.6_{\pm 0.51}$	$81.8_{\pm 0.21}$	$93.6_{\pm 0.08}$	$48.0_{\pm 0.20}$	$44.0_{\pm 0.25}$	$48.4_{\pm 0.36}$	$52.6_{\pm 0.07}$

Table S.4. Accuracy of experts on Head, Mid and Tail classes for ImageNet-LT and iNaturalist-2018.

Expert	ImageNet-LT				iNaturalist-2018			
	Overall	Head	Mid	Tail	Overall	Head	Mid	Tail
Average	59.1	66.7	58.3	40.0	75.1	70.3	75.2	76.2
`CLS` expert classifier	47.5	68.3	40.0	13.5	65.6	73.8	65.8	63.1
`DIST` expert classifier	56.4	57.2	58.6	46.6	72.9	56.1	73.2	77.0

Transformer (ViT) with a single classifier.

C. Comparison with CLIP based methods

Recently, some approaches such as VL-LTR [46] and PEL [43] have adopted a pre-trained CLIP backbone to address long-tailed recognition challenges. As indicated originally, and also reinforced by [57], CLIP is trained on large-scale balanced dataset (400 M Image-Text pair). As there is a lot of *overlapping concepts between balanced CLIP data and long-tailed datasets (ImageNet-LT and iNat-18)*, the performance of the CLIP fine-tuned methods *does not indicate meaningful progress on long-tail learning tasks*, as CLIP has already seen tail concepts in abundance. Due to this unfairness in training datasets used, we refrain from comparing the CLIP fine-tuned models (i.e., VL-LTR, PEL etc.) with DeiT-LT models trained from scratch.

D. Visualization of Attention

To demonstrate the effect of distillation in DeiT-LT, we visualize the attention of baseline methods on ImageNet-LT without distillation (ViT and DeiT-III) and compare it with DeiT-LT. As DeiT-LT contains both the `DIST` token and the `CLS` token, for visualization we average the attention across both. We use the Attention Rollout [45] method for visualization. Fig. S.2 shows the result of attention for different methods. It can be clearly observed that DeiT-LT is able to localize attention at the correct position of objects, across almost all cases. We find that DeiT-III attention maps are better in comparison to ViT, but it also often gets confused (eg. Bell Pepper, Sea Snake etc.) compared to DeiT-LT.

E. Statistical Significance of Experiments

In this section, we present the results of our experiments on CIFAR-10 LT and CIFAR-100 LT (ρ = 100, 50)(as in Table S.3), with three different random seeds. In Table S.3, we report the average performance of the expert classifiers along with the standard error for each. The low error demonstrates that the DeiT-LT training procedure is stable and quite robust across random seeds.

F. Details on Local Connectivity Analysis

We compute the mean attention distance for samples of tail classes (i.e. 7,8,9 class for CIFAR-10) using the method proposed by Raghu et al. [36]. For each head present in self-attention blocks, we calculate the distance of the patches it attends to. More specifically, we weigh the distance in the pixel space with the attention value and then average it. This is averaged for all the images present in the tail classes. We utilize the code provided here as our reference [1]. We show in Fig. 4b that for early blocks (1 and 2) of ViT, the proposed DeiT-LT method contains local features. As we go from ViT to distilled DeiT to proposed DeiT-LT, we find that features

[1] https://github.com/sayakpaul/probing-vits

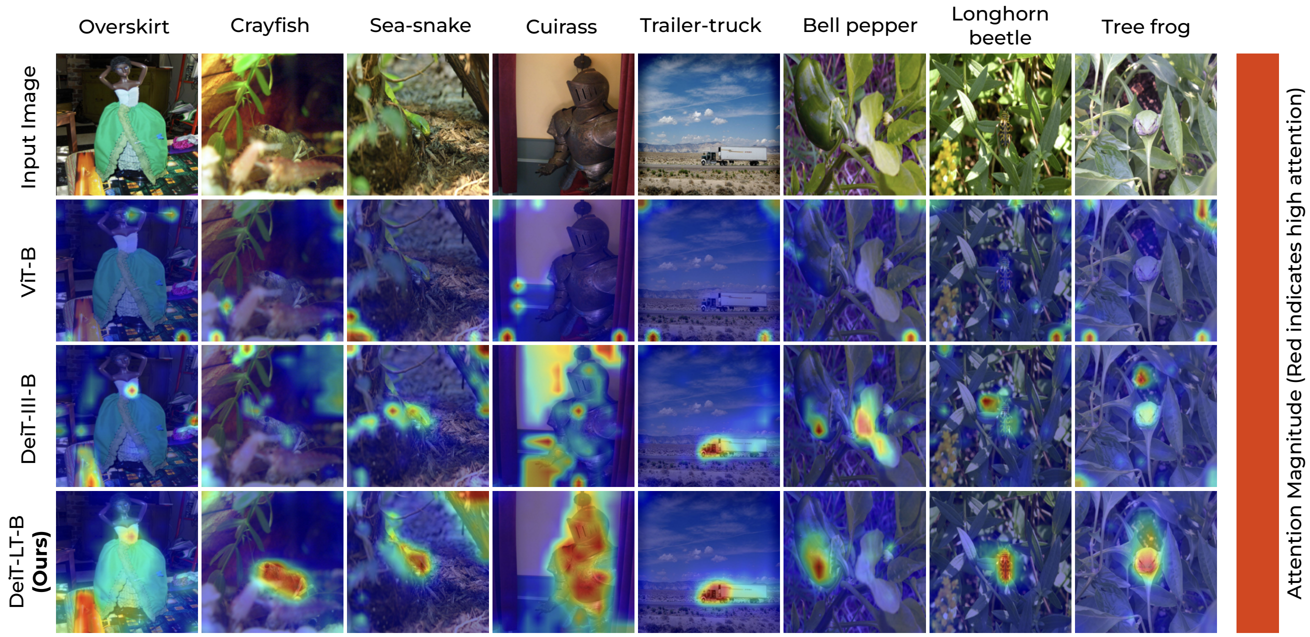


Figure S.2. Visual comparison of the attention maps from ViT-B, DeiT-III [51] and DeiT-LT *(ours)* on the ImageNet-LT dataset, computed using the method of *Attention Rollout* [1].

become more local, which explains the generalizability of DeiT-LT for tail classes. To further confirm our observations, we also provide local connectivity plots for the tail classes of the CIFAR-100 dataset (Fig. S.3). We observe that DeiT-LT produces highly local features. Further, we find that the DeiT baseline (Table 2), which is inferior to ViT for CIFAR-100, shows the presence of global features. Hence, the local connectivity correlates well with generalization on tail classes. The correlation of locality of features to generalization has also been observed by [36], who find that using the ImageNet-21k dataset for pre-training leads to more local and generalizable features in comparison to networks pre-trained on ImageNet-1k data.

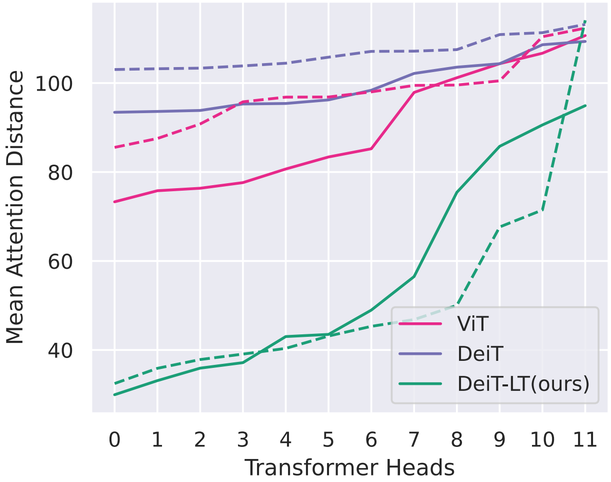


Figure S.3. Mean attention distance for early blocks (1,2) for CIFAR-100 LT tail training images.

G. Distilling low-rank features

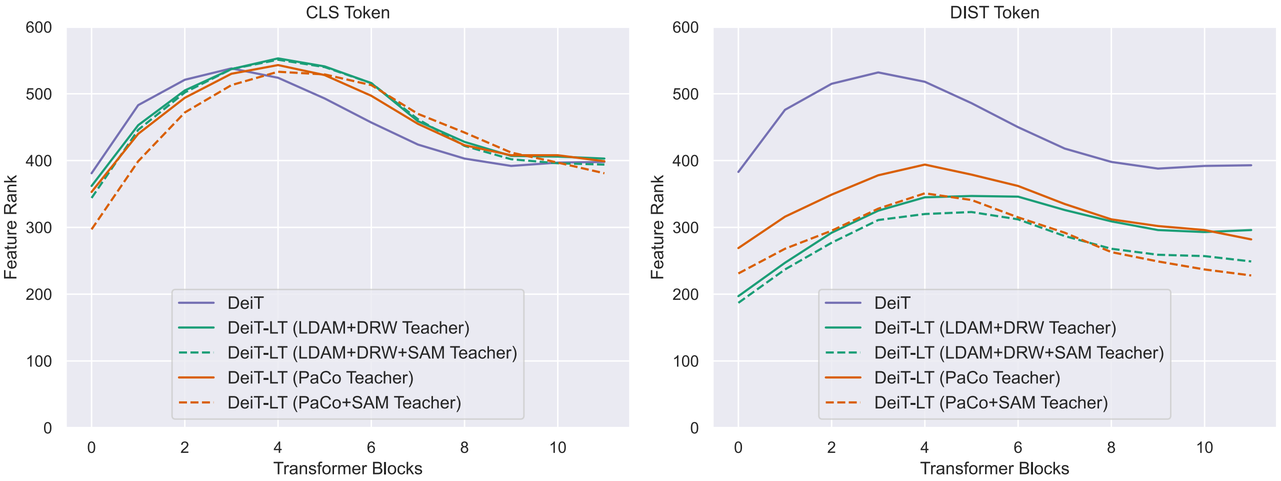
In our proposed method, as the DIST token serves as the expert on tail classes, it is important to ensure that it learns generalizable features for minority classes that are less prone to overfitting. As stated in [3], training a network with SAM optimizer leads to low-rank features. In this subsection, we investigate the feature rank of the DIST token that is distilled via a SAM-based teacher.

Calculating Feature Rank. Consider two sets of images $\mathcal{X}_{all}, \mathcal{X}_{min} \subset \mathcal{X}$, where $\mathcal{X}_{all}, \mathcal{X}_{min}$ refer to the set of images from all the classes and minority (tail) classes, respectively, with $\mathcal{X}$ being the set of all images. We construct feature matrices $F^{all}_{n_h,d}$ and $F^{min}_{n_t,d}$, where n_h and n_t are the number of images in $\mathcal{X}_{all}$ and $\mathcal{X}_{min}$ respectively, and d is the dimension of the feature representation from DIST token.

Upon centering the columns of $F^{all}_{n_h,d}$, we decompose the feature matrix as $U, S, V^T = \text{SVD}(F^{all}_{n_h,d})$, and project $F^{min}_{n_t,d}$ using the right singular vectors V as

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

where V_k contains the top k singular vectors (principal componenets). We calculate our rank as the least k that



(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).

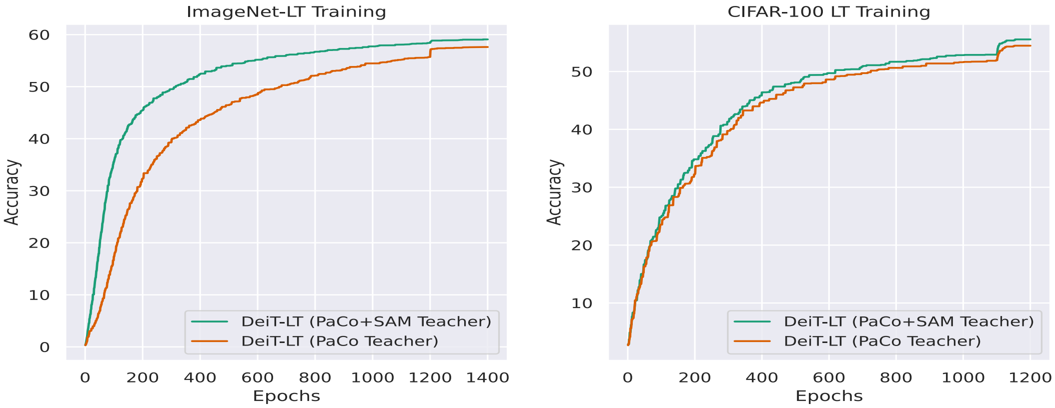


Figure S.5. Validation Accuracy Plots for the ImageNet-LT *(left)* and CIFAR-100 LT *(right)*. It can be observed that DeiT-LT trained with SAM teachers converges faster than vanilla teachers.

satisfies

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

where $F^{min}_{recon}(k)$ is an approximate reconstructed feature matrix given by $F^{min}_{recon}(k) = F^{min}_{proj}(k) * {V_k}^T$.

As shown in Fig. S.4b, we find that the `DIST` token trained with a SAM-based teacher reports a lower rank. As we are able to use the same principal components to represent both the majority and minority classes' feature representation, it signifies that the `DIST` token learns generalizable characteristics relevant across different categories of images in an imbalanced dataset. 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.

On the other hand, we observe that `CLS` token learns high-rank feature representations (Fig. S.4a), signifying that it captures intricately detailed information. Our DeiT-LT, thus, captures a wide range of information by using the predictions made using both fine-grained details from `CLS` token and generalizable features from `DIST` token.

G.1. Convergence Analysis with SAM Teachers

We find that models distilled from the teachers trained using SAM [13] converge faster than the usual CNN teachers. We provide the analysis for the Deit-LT(PaCo+SAM) and DeiT-LT(PaCo) on the ImageNet-LT and CIFAR-100 datasets in Fig. S.5. We observe that models with SAM, coverage much faster, particularly for the ImageNet-LT dataset, demonstrating the increased convergence speed for the distillation. This can be attributed to the fact that low-rank models are simpler in structure and are much easier to distill to the transformer.

H. Computation Requirement

For training our proposed DeiT-LT method on CIFAR-10 LT and CIFAR-100 LT, we use two NVIDIA RTX 3090 GPU cards with 24 GiB memory each, with both datasets requiring about 15 hours to train to train the ViT student. We train the DeiT-LT student network on four NVIDIA RTX A5000 GPU cards for the large-scale ImageNet-LT dataset and on four NVIDIA A100 GPU cards for the iNaturalist-2018 dataset, in 61 and 63 hours, respectively.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 6, 8, 5

[2] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021. 2

[3] Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. *arXiv preprint arXiv:2305.16292*, 2023. 5

[4] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *ICCV*, 2021. 6

[5] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019. 1, 2, 4, 5, 6, 7

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[7] Jun Chen, Aniket Agarwal, Sherif Abdelkarim, Deyao Zhu, and Mohamed Elhoseiny. Reltransformer: A transformer-based long-tail visual relationship recognition. In *CVPR*, 2022. 2, 7

[8] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *ICCV*, 2021. 2, 6, 7, 1, 3

[9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 1, 2, 4, 6, 7

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[11] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE TPAMI*, 2015. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 3, 6, 7, 8

[13] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 2, 3, 5, 6, 7

[14] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[16] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *CVPR*, 2021. 2

[17] Yan Hong, Jianfu Zhang, Zhongyi Sun, and Ke Yan. Safa:sample-adaptive feature augmentation for long-tailed image classification. In *ECCV*, 2022. 7

[18] Ahmet Iscen, André Araujo, Boqing Gong, and Cordelia Schmid. Class-balanced distillation for long-tailed visual recognition. 2021. 7

[19] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *CVPR*, 2020. 6

[20] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 2, 6, 7

[21] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020. 7

[22] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *CVPR*, 2020. 2

[23] Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. In *Advances in Neural Information Processing Systems*, pages 18970–18983. Curran Associates, Inc., 2021. 1, 2, 6

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[25] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested collaborative learning for long-tailed visual recognition. In *CVPR*, pages 6949–6958, 2022. 1

[26] Mengke Li, Yiu-ming Cheung, and Yang Lu. Long tail visual recognition via gaussian clouded logit adjustment. In *CVPR*, 2022. 6, 7

[27] Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *ICCV*, 2021. 6

[28] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6918–6928, 2022. 7

[29] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019. 5

[30] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *CVPR*, 2022. 2, 3, 6

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[32] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar.

Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020. 1, 2, 6, 7

[33] Gaurav Kumar Nayak, Konda Reddy Mopuri, and Anirban Chakraborty. Effectiveness of arbitrary transfer sets for data-free knowledge distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1430–1438, 2021. 4

[34] Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, 2021. 2

[35] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 4

[36] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 5, 4

[37] Harsh Rangwani, Konda Reddy Mopuri, and R Venkatesh Babu. Class balancing gan with a classifier in the loop. In *Conference on Uncertainty in Artificial Intelligence (**UAI**)*, 2021. 2

[38] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, and Venkatesh Babu R. Escaping saddle points for effective generalization on class-imbalanced data. In *Advances in Neural Information Processing Systems*, pages 22791–22805. Curran Associates, Inc., 2022. 4, 5, 6, 7, 1, 2

[39] Harsh Rangwani, Naman Jaswani, Tejan Karmali, Varun Jampani, and R. Venkatesh Babu. Improving gans for long-tailed data through group spectral regularization. In *European Conference on Computer Vision (**ECCV**)*, 2022. 2

[40] Harsh Rangwani*, Lavish Bansal*, Kartik Sharma, Tejan Karmali, Varun Jampani, and R. Venkatesh Babu. Noisytwins: Class-consistent and diverse image generation through style-GANs. In *Conference on Computer Vision and Pattern Recognition (**CVPR**)*, 2023. 2

[41] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. *arXiv preprint arXiv:2007.10740*, 2020. 2, 7

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5, 1

[43] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Xin-Yan Han, Jie-Jing Shao, and Yu-Feng Li. Parameter-efficient long-tailed recognition. *arXiv preprint arXiv:2309.10019*, 2023. 4

[44] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 1

[45] Zichang Tan, Yang Yang, Jun Wan, Hanyuan Hang, Guodong Guo, and Stan Z Li. Attention-based pedestrian attribute analysis. *TIP*, 28(12):6126–6140, 2019. 4

[46] Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *European Conference on Computer Vision*, pages 73–91. Springer, 2022. 6, 2, 4

[47] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 1

[48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers amp; distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021. 1, 2, 3, 4, 5, 6, 8

[49] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021. 6

[50] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Herve Jegou. Three things everyone should know about vision transformers. *arXiv preprint arXiv:2203.09795*, 2022. 1

[51] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 516–533. Springer, 2022. 1, 5, 6, 7

[52] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 1, 5

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 3

[54] Angelina Wang and Olga Russakovsky. Overwriting pretrained bias with finetuning data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3957–3968, 2023. 2, 3

[55] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *CVPR*, 2021. 2, 6

[56] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2021. 1, 3, 6, 7

[57] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 4

[58] Zhengzhuo Xu, Ruikang Liu, Shuo Yang, Zenghao Chai, and Chun Yuan. Learning imbalanced data with vision transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6, 7, 2

[59] Zhengzhuo Xu, Shuo Yang, Xingjun Wang, and Chun Yuan. Rethink long-tailed recognition with vision transforms. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 6, 2

[60] Han-Jia Ye, Hong-You Chen, De-Chuan Zhan, and Wei-Lun Chao. Identifying and compensating for feature deviation in imbalanced deep learning. *arXiv preprint arXiv:2001.01385*, 2020. 2

[61] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 4

[62] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2, 4

[63] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, 2021. 7

[64] Yongshun Zhang, Xiu-Shen Wei, Boyan Zhou, and Jianxin Wu. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *AAAI*, 2021. 6

[65] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *CVPR*, 2021. 6, 7

[66] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 2017. 2

[67] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, 2020. 1, 6

[68] Yixuan Zhou, Yi Qu, Xing Xu, and Hengtao Shen. Imbsam: A closer look at sharpness-aware minimization in class-imbalanced recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11345–11355, 2023. 7