Evaluating Robustness of Vision Transformers on Imbalanced Datasets

Kevin Li, Rahul Duggal, Duen Horng Chau

Georgia Institute of Technology
North Ave NW, Atlanta, GA 30332
{kevin.li rduggal polo}@gatech.edu

Abstract

Data in the real world is commonly imbalanced across classes. Training neural networks on imbalanced datasets often leads to poor performance on rare classes. Existing work in this area has primarily focused on Convolution Neural Networks (CNN), which are increasingly being replaced by Self-Attention-based Vision Transformers (ViT). Fundamentally, ViTs differ from CNNs in that they offer the flexibility in learning the appropriate inductive bias conducive to improving performance. This work is among the first to evaluate the performance of ViTs under class imbalance. We find that accuracy degradation in the presence of class imbalance is much more prominent in ViTs compared to CNNs. This degradation can be partially mitigated through loss reweighting—a popular strategy that increases the loss contributed by rare classes. We investigate the impact of loss reweighting on different components of a ViT, namely, the patch embedding, self-attention backbone, and linear classifier. Our ongoing investigations reveal that loss reweighting impacts mostly the linear classifier and self-attention backbone while having a small and negligible effect on the embedding layer.

Introduction

Transformer-based architectures have recently been modified to tackle image recognition, starting with the original Vision Transformer (ViT) (Dosovitskiy et al. 2021). With ViT-based models, there is a fundamental difference in how they capture inductive bias. While convolutional neural networks (CNNs) utilize convolutional and pooling layers to identify local patterns, ViTs gain a global representation through a patch embedding layer and self-attention mechanisms. Despite the gaining popularity of ViTs, very little work has explored their ability to perform on long-tailed data distributions where a large majority of classes constitute a small portion of the dataset. Given the lack of literature evaluating ViT-based architectures, our ongoing work makes the following contributions:

1. **Benchmarking ViT models on imbalanced datasets with and without loss reweighting.** To answer how robust ViTs are on imbalanced datasets, we compare ViTs to CNNs on balanced and imbalanced variants of CIFAR-10, CIFAR-100, and ImageNet. We then evaluate how well class-level loss reweighting (Cui et al. 2019), developed to mitigate imbalance for CNNs, transfers over to ViTs. We find that ViTs perform worse than CNNs on imbalanced datasets (Table 1), and reweighting works best as fine-tuning for ViTs (Table 2).

2. **Impact of loss reweighting on ViT architectural components.** We conduct further experiments that compare the impact of loss reweighting on the three components of ViT (patch embedding, self-attention, and linear classifier) by freezing the components during training, thus isolating their effects. We conclude that the impact to each component during reweighting fine-tuning from most to least significant are as follows: linear classifier, self-attention, and patch embedding (Table 3).

	CIFAR10			CIFAR100			ImageNet	
Model	$\rho = 1\times$	$10\times$	$100\times$	$\rho = 1\times$	$10\times$	$100\times$	bal	imbal
CNN	0.930	0.867	0.712	0.712	0.567	0.389	0.760	0.442
ViT	0.903	0.721	0.467	0.704	0.463	0.254	0.775	0.311

Table 1: Comparing the Top-1 accuracy of a Convolutional Neural Network (ResNet-32 on CIFAR, ResNet-50 on ImageNet) to a Vision Transformer (DeiT-T on CIFAR, DeiT-S on ImageNet) under varying levels of class imbalance on three datasets. The performance of the Vision Transformer degrades rapidly with increasing levels of class imbalance.

Experiments

Benchmarked datasets and models. We compared ResNet-32 against DeiT-T on CIFAR-10 and CIFAR-100 (Krizhevsky 2009) and ResNet-50 against DeiT-S on ImageNet-1K (Russakovsky et al. 2015). We created imbalanced versions of CIFAR-10 and CIFAR-100 by randomly sub-sampling images from different classes to achieve an imbalance ratio (ρ), i.e., the ratio of the sample size of the majority class to the sample size of the minority class, of 10 and 100. An exponential decay distribution similar to (Cao et al. 2019) was achieved. ImageNet-LT (Liu et al. 2019), a long-tailed distribution of ImageNet-1K, was used as the imbalanced variant for ImageNet.

Standardizing tests across imbalance (Table 1). We started by training the ResNet-32 and ResNet-50 models to their comparable accuracies on the balanced datasets using hyperparameters and data augmentation regimes from prior

Model	Reweight	CIFAR10 10×	CIFAR10 100×	CIFAR100 10×	CIFAR100 100×	ImageNet imbal
CNN	RW	0.872	0.707	0.559	0.350	0.390
ViT	RW	0.720	0.538	0.404	0.202	0.270
CNN	DRW	0.877	0.757	0.578	0.415	0.476
ViT	DRW	0.759	0.556	0.494	0.302	0.356

Table 2: Top-1 accuracies for CNN and ViT models on varying levels of imbalance of three datasets with differing loss reweighting starts: immediately during training (RW) and after 80% of training (DRW). DRW outperforms RW for all.

work (He et al. 2016). We then fine-tuned the learning rate and clip gradient for the ViT counterparts while keeping the other hyperparameters and data augmentation regime constant with (Touvron et al. 2021) to reach similar Top-1 accuracies. After finalizing hyperparameters, we used the same values to train on the imbalanced dataset versions (Table 1).

Deploying class-level reweighting (Table 2). We apply class-level loss reweighting (Cui et al. 2019) during the training of the ViT and CNN models. Under reweighting, loss values are scaled based on the frequency of the true class label. We experimented with reweighting hyperparameter β and varied the initialization of class-level reweighting: starting immediately (RW) or deferring until after 80% of epochs had finished (DRW). All reweighting runs for a model type on a certain dataset variant used the same β that led to the highest Top-1 accuracy for DRW[1](Table 2).

Isolating ViT components during reweighting (Table 3). We evaluate the reweighting impact on ViTs' key components – patch embedding, self-attention, and linear classifier – by freezing these layers during reweighting. The ablation runs used the same models as the DRW runs, and as reweighting began, the weights and biases of the targeted components were frozen. This resulted in three designs: FR-PE, where only the patch embedding was frozen; FR-SA, where the patch embedding and self-attention were frozen; and FR-ALL, where patch embedding, self-attention, and linear classifier were frozen, i.e., the entire model (Table 3).

Discovery and Conclusion

Accuracy degradation of data imbalance more prominent in ViTs than CNNs. We discover that after training ViTs and CNNs to comparable accuracies on balanced datasets, the same ViTs perform worse on imbalanced variants, leading to differences of more than 10% in Top-1 accuracy (Table 1). However, this may be due to ViTs' inherent need for more data; imbalanced versions of datasets would provide fewer samples than their balanced counterparts. Future experiments will run re-sampling to mitigate the difference in training samples to determine if this is the reason.

ViT fine-tuning using class-level reweighting for imbalanced datasets. Our experiments show that DRW outperforms RW in all ViT runs, and RW may actually underperform compared to the baseline (Table 2). These findings

[1] $\beta = 0.9999$ for all CNNs and ViTs except for ViTs on CIFAR-10 $\rho = 100\times$ and CIFAR-100 $\rho = 100\times$ which $\beta = 0.999$.

Targeted Component	Reweight Type	CIFAR10 10×	CIFAR10 100×	CIFAR100 10×	CIFAR100 100×
N/A, baseline	DRW	0.759	0.556	0.494	0.302
Patch Embed.	FR-PE	0.756	0.555	0.493	0.301
Self-Attn.	FR-SA	0.738	0.534	0.479	0.294
Linear Class.	FR-ALL	0.699	0.447	0.446	0.247

Table 3: Investigating the impact of loss reweighting on the architectural components of ViT. Different components of ViTs are frozen during DRW: patch embedding (FR-PE); patch embedding and self-attention (FR-SA); and patch embedding, self-attention, and linear classifier (FR-ALL). A larger drop from the baseline indicates more impactfulness.

are similar to CNN-based deferred reweighting (Cao et al. 2019). DRW is also noted to lead to larger gains for ViTs than CNNs, although the reason is not apparent: perhaps it is due to ViTs having more room to recover. Further experiments show that starting reweighting too early leads to missed gains; however, future work will focus on if there is a "general" time/epoch to begin reweighting.

ViT linear classifier affected by reweighting most, followed by self-attention. Based on Table 3, several insights can be drawn. Patch embedding is changed minimally. There is little difference between DRW and FR-PE. Linear classifier, followed by self-attention, is impacted most during loss reweighting, as the stagnation of linear classifier leads to the most significant drop in Top-1 accuracy out of all three components. Our future work will incorporate these conclusions to develop a linear classifier-targeted reweighting strategy.

References

Cao, K.; Wei, C.; Gaidon, A.; Aréchiga, N.; and Ma, T. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *NeurIPS*.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. J. 2019. Class-Balanced Loss Based on Effective Number of Samples. *CVPR, 2019*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. *CVPR, 2016*.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-Scale Long-Tailed Recognition in an Open World. *CVPR, 2019*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV, 2015*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and J'egou, H. 2021. Training data-efficient image transformers & distillation through attention. In *ICML, 2021*.

Appendix

We provide supplemental details for our experiments and results of experiments that were not detailed in Table 1, 2, and 3.

Experimental Settings

We ran all our experiments on an NVIDIA RTX A6000, with 4 GPUs with ∼50GB of memory each. Our framework utilized PyTorch version "1.9.0+cu111" on Python 3.8.5. For our ViTs, we used a learning rate of 0.0015 and a clip gradient norm of 5 while keeping the other hyperparameters for training and image augmentation regime constant with (Touvron et al. 2021). Hyperparameters determined for each model during balanced dataset training are reused for the imbalanced variant trainings, i.e., the same hyperparameters are applied on both balanced and imbalanced versions. CIFAR models were trained with a batch size of 128 on a single GPU. For ImageNet, the CNN used a batch size of 256, and ViT a batch size of 512. ImageNet models were also trained on a single GPU. Code and data are available upon request.

Choosing β hyperparameter

We ran all RW and DRW experiments for both ViT and CNN on $\beta = \{0.99, 0.999, 0.9999\}$. For each combination of model type and dataset variant, e.g. DeiT-T on CIFAR-10 $\rho = 10$, ResNet-32 on CIFAR-100 $\rho = 100$, or DeiT-S on ImageNet-LT, the results from different experiments, e.g., RW, DRW, reported in the main text used the same β values. The β value that led to the best Top-1 accuracy for the DRW experiment was used for RW, DRW, FR-PE, FR-SA, and FR-ALL (the FR cases will be discussed later). This is due to the DRW performing better than RW; hence, we standardize experiments using the β values that will be used most often (largest imbalanced dataset accuracy recoveries).

β hyperparameter results for CIFAR-10

	$\beta = 0.99$		$\beta = 0.999$		$\beta = 0.9999$	
Model	$\rho = 10\times$	$100\times$	$\rho = 10\times$	$100\times$	$\rho = 10\times$	$100\times$
ResNet-32 RW	0.861	0.716	0.869	0.719	0.872	0.707
ResNet-32 DRW	0.863	0.700	0.868	0.731	**0.877**	**0.757**
DeiT-T RW	0.726	0.477	0.751	0.538	0.720	0.454
DeiT-T DRW	0.721	0.473	0.747	**0.556**	**0.759**	0.545

Table 4: Top-1 accuracies for ViTs and CNNs on imbalanced CIFAR-10 with varying β values. The best Top-1 accuracy for each model and dataset combination is highlighted.

In the main paper, for all runs reported of ResNet-32 on CIFAR-10 $\rho = 10\times$ and $\rho = 100\times$, we use $\beta = 0.9999$. For all runs of DeiT-T on CIFAR-10 $\rho = 10\times$, we use $\beta = 0.9999$, and for all runs of DeiT-T on CIFAR-10 $\rho = 100\times$, we use $\beta = 0.999$ (Table 4).

β hyperparameter results for CIFAR-100

In the main text, for all runs reported of ResNet-32 on CIFAR-100 $\rho = 10\times$ and $\rho = 100\times$, we use $\beta = 0.9999$.

	$\beta = 0.99$		$\beta = 0.999$		$\beta = 0.9999$	
Model	$\rho = 10\times$	$100\times$	$\rho = 10\times$	$100\times$	$\rho = 10\times$	$100\times$
ResNet-32 RW	0.564	0.371	0.557	0.336	0.559	0.350
ResNet-32 DRW	0.572	0.409	0.578	0.414	**0.578**	**0.415**
DeiT-T RW	0.467	0.242	0.424	0.202	0.404	0.198
DeiT-T DRW	0.480	0.294	0.493	**0.302**	**0.494**	0.299

Table 5: Top-1 accuracies for ViTs and CNNs on imbalanced CIFAR-100 with varying β values. The best Top-1 accuracy for each model and dataset combination is highlighted.

For all runs of DeiT-T on CIFAR-100 $\rho = 10\times$, we use $\beta = 0.9999$, and for all runs of DeiT-T on CIFAR-100 $\rho = 100\times$, we use $\beta = 0.999$ (Table 5).

β hyperparameter results for ImageNet-LT

Model	$\beta = 0.99$	$\beta = 0.999$	$\beta = 0.9999$
ResNet-50 RW	0.431	0.395	0.390
ResNet-50 DRW	0.466	0.475	**0.476**
DeiT-S RW	0.298	0.271	0.270
DeiT-S DRW	0.340	0.355	**0.356**

Table 6: Top-1 accuracies for ViTs and CNNs on ImageNet-LT with varying β values. The best Top-1 accuracy for each model is highlighted.

In the main paper, for all runs reported of ResNet-50 and DeiT-T on ImageNet-LT, we use $\beta = 0.9999$ (Table 6).

Fine-tuning ViTs with reweighting

After observing that ViT trained with DRW performed better than those with RW (Table 2), we wanted to confirm that loss reweighting should be applied later rather than earlier for ViTs, e.g., used for fine-tuning models. To do so, we started loss reweighting earlier than our 80% training epoch point, as described in the main paper. For the following sections, **new notation will be used**. From here on out, we will reference deferred loss reweighting as *DRW-X*, where *X* is the training epoch where loss reweighting begins. Because the standard training epochs for ViTs is 300 epochs, DRW reported in the main paper can also be denoted by DRW-240 (80% of total training epochs). We also ran experiments with DRW-160 for ViT models, i.e., starting reweighting at epoch 160, and used all three β values $\{0.99, 0.999, 0.9999\}$ from before.

Fine-tuning DRW ViT on CIFAR-10
Fine-tuning DRW ViT on CIFAR-100

	$\beta = 0.99$		$\beta = 0.999$		$\beta = 0.9999$	
Model	$\rho = 10\times$	$100\times$	$\rho = 10\times$	$100\times$	$\rho = 10\times$	$100\times$
DeiT-T DRW-160	0.721	0.474	0.747	0.581	0.751	0.525
DeiT-T DRW-240	0.721	0.473	0.747	0.556	0.759	0.545

Table 7: Top-1 accuracies for ViT with DRW-160 and DRW-240 on imbalanced CIFAR-10 for varying β values.

	$\beta = 0.99$		$\beta = 0.999$		$\beta = 0.9999$	
Model	$\rho = 10\times$	$100\times$	$\rho = 10\times$	$100\times$	$\rho = 10\times$	$100\times$
DeiT-T DRW-160	0.480	0.282	0.483	0.271	0.481	0.269
DeiT-T DRW-240	0.480	0.294	0.493	0.302	0.494	0.299

Table 8: Top-1 accuracies for ViT with DRW-160 and DRW-240 on imbalanced CIFAR-100 for varying β values.

Fine-tuning DRW ViT on ImageNet-LT

Model	$\beta = 0.99$	$\beta = 0.999$	$\beta = 0.9999$
DeiT-S DRW-160	0.334	0.374	0.349
DeiT-S DRW-240	0.340	0.355	0.356

Table 9: Top-1 accuracies for for ViT with DRW-160 and DRW-240 on ImageNet-LT for varying β values.

From these results, we observed that for the β hyperparameters that led to the highest DRW-240 Top-1 accuracy, DRW-160 outperformed RW. We also show that DRW-240 mostly outperforms or is equivalent to DRW-160 under differing reweighting hyperparameters (β) and differing levels of imbalance on various datasets. The difference is even more apparent when comparing RW with DRW-240. With both deferred reweighting variants (DRW-160 and DRW-240) outperforming reweighting (RW) (Table 7, 8, 9), we conclude that reweighting should be used as a fine-tuning method. However, it is important to note that more experiments should be run to determine which DRW epoch may lead to the best Top-1 accuracy. A general best starting point will probably fall past the 50% training mark since Top-1 accuracy generally increased from DRW-160 to DRW-240 and decreased from DRW-240 to the baseline, i.e., DRW-300/no reweighting at all. In terms of hyperparameter β for loss reweighting, it seems like using $\beta = 0.9999$ is a generally safe option as the Top-1 accuracies for $\beta = 0.9999$ are usually the best option or are close to the best accuracies.

Freezing ViTs components during reweighting

To evaluate the effect loss reweighting has on the key components of ViTs – patch embedding layer, self-attention backbone, and linear classifier layer – we gradually freeze the model throughout runs during loss reweighting to isolate certain components. These runs used the same models and hyperparameters as the DRW-240 runs. To freeze the target layer(s), the weights and biases of the targeted components were made unchanging by turning the back-propagation gradients of the components to zero. To evaluate the effect of each individual component, we decided to freeze the model from the bottom up, i.e., starting from patch embedding (furthest from prediction) to the linear classifier (prediction). Using this method, we froze the patch embedding first (FR-PE), then up to the self-attention backbone, i.e., patch embedding and self-attention blocks, (FR-SA), and then the linear classifier (FR-ALL), which froze the entire model given it is the last layer. The importance and impact on each layer would be evaluated by the change in accuracy from when the component was not frozen to when the component was frozen. We ran FR-PE, FR-SA, and FR-ALL on both DRW-160 and DRW-240 to confirm that observations would be consistent across different reweighting regimes.

		CIFAR10		CIFAR100	
Model	Reweight	$10\times$	$100\times$	$10\times$	$100\times$
ViT	DRW-160	0.751	0.581	0.481	0.271
ViT	FR-PE-160	0.746	0.584	0.479	0.272
ViT	FR-SA-160	0.663	0.466	0.425	0.252
ViT	FR-ALL-160	0.570	0.364	0.363	0.190
ViT	DRW-240	0.759	0.556	0.481	0.271
ViT	FR-PE-240	0.756	0.555	0.493	0.301
ViT	FR-SA-240	0.738	0.534	0.479	0.294
ViT	FR-ALL-240	0.699	0.447	0.446	0.247

Table 10: Impact of loss reweighting on key architectural components of ViTs.

Regular DRW-160 and DRW-240 are also included in Table 10 as baselines, since no components were frozen. As we gradually start freezing more components, we would expect the Top-1 accuracy to decrease as larger parts of the model are no longer updating. And we do observe this for the self-attention backbone and the linear classifier; Top-1 accuracies decrease when those respective components are frozen. But there is little difference between DRW and FR-PE, leading to the conclusion that the patch embedding is changed minimally and is not critical to accuracy gains for DRW-240. This can also be observed for DRW-160. In addition, some instances of FR-PE lead to higher accuracies than the DRW baselines, which may indicate overfitting of the patch embeddings caused by class-based reweighting. We also note that, in general, the freezing of the linear classifier leads to larger decreases in Top-1 (FR-SA to FR-ALL) than the freezing of the self-attention layer (FR-PE to FR-SA). With these insights, targeting the linear classifier layer during reweighting because it is the most impacted by loss reweighting may prove fruitful. Preventing overfitting of the patch embedding layer by potentially downscaling the gradients from back-propagation may also improve Top-1 accuracy of reweighting.