Evaluating Robustness of Vision Transformers on Imbalanced Datasets (Student Abstract)

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 with the original Vision Transformer (ViT) (Kolesnikov 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 popularity of ViTs, little work has explored their performance 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 (He et al. 2016) against DeiT-T (Touvron et al. 2021) on CIFAR-10 and CIFAR-100 (Krizhevsky 2009) and ResNet-50 against DeiT-S on ImageNet-1K (Russakovsky et al. 2015). We created imbalanced CIFAR-10 and CIFAR-100 by sub-sampling images from classes to achieve an imbalance ratio ρ (ratio of majority to minority class sample size) of 10 and 100. An exponential decay distribution similar to (Cao et al. 2019) was achieved. ImageNet-LT (Liu et al. 2019) was used as the long-tailed variant of ImageNet.
Standardizing tests across imbalance (Table 1). We started by training ResNet-32 and ResNet-50 models to comparable accuracies on the balanced datasets using training regimes from prior work (He et al. 2016). We then fine-

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.

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 fine-tuning, we used the final values to train on the imbalanced dataset (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 tested with reweighting hyperparameter β (adjusts the scaling factor) and varied the initialization of 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 best 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 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.

[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.

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. In *CVPR*.

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

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

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. *Technical Report TR-2009*.

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

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. In *IJCV*.

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*.