Transferability of ImageNet Robustness to Downstream Tasks

Yutaro Yamada
Yale University
yutaro.yamada@yale.edu

Mayu Otani
CyberAgent, Inc.
otani_mayu@cyberagent.co.jp

Abstract

As clean ImageNet accuracy nears its ceiling, the research community is increasingly more concerned about robust accuracy under distributional shifts. While a variety of methods have been proposed to robustify neural networks, these techniques often target models trained on ImageNet classification. At the same time, it is a common practice to use ImageNet pretrained backbones for downstream tasks such as object detection, semantic segmentation, and image classification from different domains. This raises a question: Can these robust image classifiers transfer robustness to downstream tasks? For object detection and semantic segmentation, we find that a vanilla Swin Transformer, a variant of Vision Transformer tailored for dense prediction tasks, transfers robustness better than Convolutional Neural Networks that are trained to be robust to the corrupted version of ImageNet. For CIFAR10 classification, we find that models that are robustified for ImageNet do not retain robustness when fully fine-tuned. These findings suggest that current robustification techniques tend to emphasize ImageNet evaluations. Moreover, network architecture is a strong source of robustness when we consider transfer learning.

1. Introduction

ImageNet A newly proposed vision architecture, including recent Vision Transformer [1], is first tested against ImageNet to demonstrate a good performance before it gains popularity within the community. While accuracy on ImageNet has been considered as a surrogate for measuring progress in machine vision systems, the research community is now aware of the lack of robustness of vision models towards small input perturbations. [10] first reported that imperceptible adversarial perturbation can easily fool image classifiers. Recent studies show that even simpler, more natural noises such as blur, contrast change, and snow can significantly degrade the performance of models [4]. A typical strategy to increase robustness is data augmentation, where a vision model is trained with additional data, which are artificially corrupted during training. Examples include ANT [9], AugMix [5], and DeepAug [3]. However, these techniques often focus on improving robust accuracy for ImageNet classification. In fact, there are now a variety of ImageNet-scale robustness benchmarks, and the community is striving to improve accuracy on these benchmarks [3].

Due to the scale of ImageNet, it is a common practice to use ImageNet pretrained weights for downstream tasks such as object detection and image segmentation. This practice of using pretrained ImageNet weights for transfer learning raises a fundamental question from a robustness perspective: When we use pretrained weights that are made to be robust to ImageNet benchmarks, do these models necessarily show robustness to downstream tasks as well?

Scope. While there are various kinds of distributional shifts and robustness that the vision community studies, we focus on common corruption robustness in this paper, because we are interested in robustness transfer from ImageNet classification to downstream tasks such as object detection and segmentation. See Section 3.1 for more details about why we specifically choose common corruptions as a topic of our study.

2. Background

Ensuring robustness in downstream tasks such as object detection and semantic segmentation is equally, if not more, important than achieving robustness in image classification. Especially for safety-critical applications such as self-driving cars, vision systems that are vulnerable to image perturbations can lead to dire consequences. In such real-world applications, classification is only the first step of the pipeline, and ensuring robustness through the entire system of object detection and segmentation needs further care.

When we consider how to ensure robustness for downstream tasks, there are two viable approaches. One is to transfer robustness effectively from a pretrained, robustified classifier backbone to each downstream task, which is our focus of this paper. The other approach is to apply an ex-

Method	Noise	Blur	Digital	Weather
Regular	36.09	44.00	17.36	17.59
ANT	21.90	39.25	14.43	16.22
DeepAug+	**16.39**	**29.25**	**12.64**	11.27
Swin-T	18.01	38.18	14.66	**10.12**

Table 1. Accuracy drops across models and noise types are presented for fixed-feature transfer learning from ImageNet to COCO Object Detection. Regular represents a regular ImageNet-pretrained ResNet50, while DeepAug+ and ANT are ResNet50s that are robustified during ImageNet pretraining. Swin-T is a Swin Transformer (Tiny), where the model size is similar to ResNet50.

Method	Noise	Blur	Digital	Weather
Regular	48.98	29.42	14.01	25.68
ANT	17.78	23.41	**10.67**	25.62
DeepAug+	20.07	**19.47**	10.70	19.12
Swin-T	**13.57**	23.50	13.48	**14.28**

Table 2. Accuracy drops across models and noise types are presented for fixed-feature transfer learning from ImageNet to ADE10K Semantic Segmentation.

isting robust data augmentation technique during transfer learning. While applying robustification techniques during finetuning for downstream tasks is an option, these robust methods often degrade performance on clean data , or require more training data to perform on par with models that are simply trained on clean data [8]. This is especially concerning, since data scarcity is common in downstream tasks , which is precisely why transfer learning is needed in the first place. Therefore, rather than entirely resorting to data augmentation during fine-tuning, it is critical to better understand robustness transfer to achieve both robustness and good clean accuracy in downstream tasks.

2.1. Transfer Learning for Dense Prediction Tasks

While image classification only requires a single feature map typically extracted from the last layer, object detection and semantic segmentation benefits a lot from multiresolution feature maps. These feature maps provide richer information that helps object detection at different scale and pixel-level semantic prediction. Most object detection and semantic segmentation systems uses a CNN as their backbone and exploit hierarchical feature maps that are extracted from different blocks of the model.

Motivated by the success of Transformer architecture in NLP, Vision Transformer (ViT) [1] was proposed. While the original ViT excels at image classification, it is not amenable to dense prediction tasks such as object detection and semantic segmentation. This is because the original ViT processes tokens at fixed scale, producing single low-resolution feature maps. Recently, a variant of ViT called Swin Transformer was proposed to address this limitation [7]. Swin Transformer uses a hierarchical architecture to build multiresolution feature maps, while achieving linear-time complexity with respect to the image size. Because of this, Swin Transformer achieves the state-of-the-art performance in both object detection and semantic segmentation. In this work, we use Swin Transformer for our ViT architecture.

3. Fixed-Feature Transfer Learning

When we consider transfer learning from image classifiers to object detection or segmentation, we can freeze the backbone, while only training the head of the detection or segmentation system. We refer to this approach as fixed-feature transfer learning. On the other hand, we can use pretrained image classifiers as initialization to train object detection or segmentation models, which we call full-network transfer learning.

Fixed-feature transfer learning from ImageNet to object detection and semantic segmentation is not a common practice because full-network transfer learning generally performs better. However, for *robustness transfer*, fixed-feature transfer learning is an important setup to consider because it allows us to directly leverage robustified ImageNet backbones and measure how much robustness the model carries over to downstream tasks after fine-tuning only the head of the entire model. Full-network transfer learning, on the other hand, potentially erases the robustness property of backbones during fine-tuning, which can confound our analysis of robustness transfer.

From earlier work on robustness of image classifiers, a vanilla Swin Transformer is known to perform better than CNNs on ImageNet-C. At the same time, we can robustify these CNNs by data augmentation, so that they perform well on ImageNet-C. Should we expect that robustified CNNs transfer their robustness automatically when we only finetune the head while fixing the backbone? Moreover, which source of robustness (architecture vs. data augmentation) is better suited in terms of robustness transfer?

To resolve this question, we prepare two CNNs that are robustified during ImageNet-1k pretraining using ANT [9] and DeepAug+AugMix [3] respectively, and a Swin Transformer, also pretrained on ImageNet-1k but without applying any robustification technique. To control for the model size, we use ResNet50 and Swin-T, where the parameter counts are 25M and 28M, respectively. For object detection, we use Mask-RCNN [2] and for semantic segmentation, we use UperNet [11] as the head.

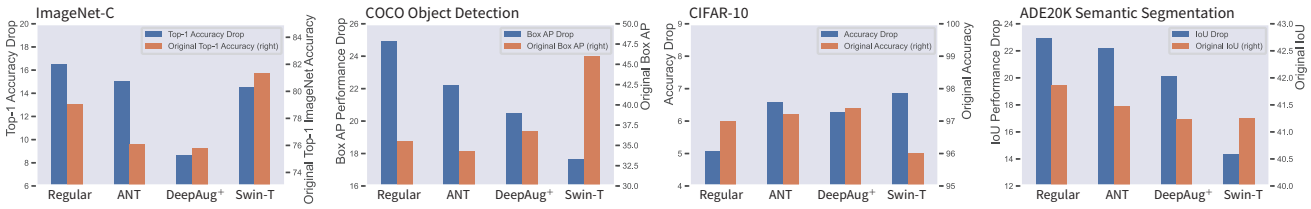

Figure 1. Robust models vs. mean performance drop under 15 corruption types in full-network fine-tuning. The lower the performance drop, the more robust models are to these image corruptions. Regular is a vanilla ResNet50. DeepAug+ and ANT refer to ResNet50 models robustified via DeepAug+AugMix and ANT, which are all data augmentation techniques to increase robustness against common corruptions [3, 9]. Swin-T is a vanilla Tiny Swin Transformer, where the parameter counts are similar to ResNet50. If robustness on ImageNet is transferable to other downstream tasks, we would see a similar pattern of ImageNet-C in object detection and semantic segmentation as well. However, we see that Swin-T performs much better than DeepAug+, the most robust model against ImageNet-C. This shows that the Swin Transformer as architecture is a stronger source of robustness transfer than robustification techniques that are used (e.g. DeepAug, AugMix, or ANT). Moreover, for CIFAR-10, Regular appears to be the most robust model, highlighting the difficulty of transferring ImageNet robustness effectively.

3.1. Robustness Transfer Benchmark

To measure how well a model transfers robustness from ImageNet classification to downstream tasks, we have to prepare the same set of distributional shifts that can be applied to both classification and downstream tasks. While there are a variety of ImageNet-related benchmarks to measure robustness against distributional shifts, most of these distributional shifts are not adoptable to our setting, because they are specifically designed for ImageNet classification. To measure the performance of robustness transfer to downstream tasks, we focus on 15 synthetic image corruption types, grouped into 4 categories: "noise," "blur," "weather," and "digital," introduced in ImageNet-C [4]. They measure corruption robustness of ImageNet classifiers by computing how much the original accuracy drops when these models are evaluated on corrupted images of the ImageNet test set. Since these image corruptions are algorithmically generated, they can be applied to images in both classification and downstream tasks such as object detection and segmentation. Therefore, these image corruptions allow us to compare the accuracy drop in classification with one in downstream tasks, which is useful to measure the degree of robustness transfer across different models.

Formally, we take ImageNet models, fine-tune them for downstream tasks. We calculate model performance on the clean test set in downstream tasks, and compute the performance drop after we apply image corruptions. We then compare the accuracy drop for classification and downstream tasks. We report the mean performance drop across the 15 image corruptions as our metric. The benchmark performance is computed in terms of mean performance under corruption: $mPC = \frac{1}{N_c}\sum_{c=1}^{N_c} P_c$, where N_c is 15, and P_c is the task-specific performance measure evaluated under corruption c on the test set. We then compute the relative performance under corruption: $rPC = \frac{mPC}{P_{clean}}$ where P_{clean} is the task-specific performance measure evaluated on the clean test set. We use $1 - rPC$ as our main metric to report and refer to this metric as Accuracy Drop or Performance Drop depending on the context. rPC allows us to compare the degree of robustness transfer from ImageNet to downstream tasks such as object detection and semantic segmentation.

Dataset. For object detection, we choose MS-COCO [6] and use the COCO 2017 validation set as our test split, following the convention. For semantic segmentation, we choose ADE20K [12] that consists of 20210 train, 2000 validation images, and 150 semantic classes. We use the following downstream-task specific performance measures:

Object Detection. We use the COCO-style mAP, which averages over IoUs between 50% and 95%.

Semantic Segmentation. We use the mean IoU, which indicates the intersection-over-union between the predicted and ground truth pixels, averaged over all the classes.

Table 1 and 2 summarize the results for the fixed feature transfer learning experiment. While ANT and DeepAug+ transfer robustness well across both downstream tasks, we also notice that for some noise types, Swin-T outperforms the robust CNNs (e.g. Noise, Weather in Table 2 and Weather in Table 1.) This suggests that, to our surprise, a vanilla Swin Transformer has a potential to transfer robustness better than robust CNNs. In the next section, we investigate to what extent these phenomena can be observed in the full-network transfer learning setting.

4. Full-Network Transfer Learning

A more common practice to perform transfer learning is to use ImageNet pretrained weights as initialization and fine-tune the entire network for downstream tasks. Even though it takes more computational resources than the fixed-feature case, full-network transfer learning generally performs better.

However, when we take robustness into consideration, full-network transfer learning can be detrimental, because gradient updates during fine-tuning can erase robustified features acquired during ImageNet pretraining. This possibility is especially concerning for robustification techniques that rely on data augmentation during pretraining such as DeepAug, AugMix, and ANT. Thus, one may argue that robustness arising from these data augmentation techniques might be less effective when we fine-tune the entire network for downstream tasks. On the other hand, robustness arising from the architecture itself can be more resilient to full-network fine-tuning, because the robustness property is not directly encoded into weights, but rather stems from the topology of architecture. Thus, we do not need to worry about erasing robustness that arises from architecture during transfer learning. As we see that a vanilla Swin Transformer outperforms robustified CNNs for some noise types in the Section 3, architecture indeed plays some role in transferring robustness. Therefore, we hypothesize that in the setting of full-network transfer learning, Transformer architectures might be more effective than CNNs that are robustified via data augmentation.

To resolve this hypothesis, we repeat the same set of experiments as in the Section 3, but now train all weights for object detection, semantic segmentation, and image classification. For downstream image classification tasks, we choose CIFAR10. The results are shown in Figure 1. As a reference, we also plot the original ImageNet accuracy as well as the Top-1 Accuracy Drop on ImageNet-C for all ImageNet models we use. We can confirm that the two robust CNNs (DeepAug+ and ANT) indeed demonstrate higher robustness than Regular. It is noteworthy that a vanilla Swin-T shows slightly higher robustness than ANT (represented as a lower accuracy drop in the blue bar). More surprisingly, Swin-T performs best in object detection and semantic segmentation. This shows that DeepAug+ and ANT are less successful to transfer their ImageNet-C robustness to downstream tasks than Swin-T, verifying our hypothesis. Moreover, when we test robust transfer from ImageNet-C to CIFAR10, we find that these robust models fail to outperform Regular. This shows that robustness from ImageNet for downstream image classification seems to be harder to transfer than object detection and semantic segmentation.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 2

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[3] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3

[4] Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 3

[5] Dan Hendrycks, Norman Mu, E. D. Cubuk, Barret Zoph, J. Gilmer, and Balaji Lakshminarayanan. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *International Conference on Learning Representations (ICLR)*, 2020. 1

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, 2014. 3

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *International Conference on Computer Vision (ICCV)*, 2021. 2

[8] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[9] Evgenia Rusak, Lukas Schott, Roland S. Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A Simple Way to Make Neural Networks Robust Against Diverse Image Corruptions. In *European Conference on Computer Vision (ECCV)*, volume 12348, pages 53–69, 2020. 1, 2, 3

[10] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 1

[11] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified Perceptual Parsing for Scene Understanding. In *European Conference on Computer Vision (ECCV)*, volume 11209, pages 432–448, 2018. 2

[12] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing through ADE20K Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 3