Do Adversarially Robust ImageNet Models Transfer Better?

Hadi Salman[*]
hadi.salman@microsoft.com
Microsoft Research

Andrew Ilyas[*]
ailyas@mit.edu
MIT

Logan Engstrom
engstrom@mit.edu
MIT

Ashish Kapoor
akapoor@microsoft.com
Microsoft Research

Aleksander Mądry
madry@mit.edu
MIT

Abstract

Transfer learning is a widely-used paradigm in which models pre-trained on standard datasets can efficiently adapt to downstream tasks. Typically, better pre-trained models yield better transfer results, suggesting that initial accuracy is a key aspect of transfer learning performance. In this work, we identify another such aspect: we find that adversarially robust models, while less accurate, often perform better than their standard-trained counterparts when used for transfer learning. Specifically, we focus on adversarially robust ImageNet classifiers, and show that they yield improved accuracy on a standard suite of downstream classification tasks. Further analysis uncovers more differences between robust and standard models in the context of transfer learning. Our results are consistent with (and in fact, add to) recent hypotheses stating that robustness leads to improved feature representations. Our code and models are available at https://github.com/Microsoft/robust-models-transfer.

1 Introduction

Deep neural networks currently define state-of-the-art performance across many computer vision tasks. When large quantities of labeled data and computing resources are available, models perform well when trained from scratch. However, in many practical settings there is insufficient data or compute for this approach to be viable. In these cases, *transfer learning* [Don+14; Sha+14] has emerged as a simple and efficient way to obtain performant models. Broadly, transfer learning refers to any machine learning algorithm that leverages information from one ("source") task to better solve another ("target") task. A prototypical transfer learning pipeline in computer vision (and the focus of our work) starts with a model trained on the ImageNet-1K dataset [Den+09; Rus+15], and then refines this model for the target task.

Though the exact underpinnings of transfer learning are not fully understood, recent work has identified factors that make pre-trained ImageNet models amenable to transfer learning. For example, [HAE16; Kol+19] investigate the effect of the source dataset; Kornblith, Shlens, and Le [KSL19] find that pre-trained models with higher ImageNet accuracy also tend to transfer better; Azizpour et al. [Azi+15] observe that increasing depth improves transfer more than increasing width.

Our contributions. In this work, we identify another factor that affects transfer learning performance: adversarial robustness [Big+13; Sze+14]. We find that despite being less accurate on ImageNet, adversarially robust neural networks match or improve on the transfer performance of their standard counterparts. We first establish this trend in the "fixed-feature" setting, in which one trains

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Table 1: Transfer learning performance of robust and standard ImageNet models on 12 downstream classification tasks. For each type of model, we compute maximum accuracy (averaged over three random trials) over training parameters, architecture, and (for robust models) robustness level ε.

		Dataset											
Mode	**Model**	Aircraft	Birdsnap	CIFAR-10	CIFAR-100	Caltech-101	Caltech-256	Cars	DTD	Flowers	Food	Pets	SUN397
Fixed-feature	**Robust**	**44.14**	**50.72**	**95.53**	**81.08**	**92.76**	**85.08**	**50.67**	**70.37**	**91.84**	**69.26**	**92.05**	**58.75**
	Standard	38.69	48.35	81.31	60.14	90.12	82.78	44.63	**70.09**	**91.90**	65.79	**91.83**	55.92
Full-network	**Robust**	**86.24**	**76.55**	**98.68**	**89.04**	**95.62**	**87.62**	**91.48**	**76.93**	**97.21**	**89.12**	**94.53**	**64.89**
	Standard	**86.57**	75.71	97.63	85.99	94.75	86.55	**91.52**	75.80	**97.04**	88.64	**94.20**	63.72

a linear classifier on top of features extracted from a pre-trained network. Then, we show that this trend carries forward to the more complex "full-network" transfer setting, in which the pre-trained model is entirely fine-tuned on the relevant downstream task. We carry out our study on a suite of image classification tasks (summarized in Table 1), object detection, and instance segmentation.

Our results are consistent with (and in fact, add to) recent hypotheses suggesting that adversarial robustness leads to improved feature representations [Eng+19a; AL20]. Still, future work is needed to confirm or refute such hypotheses, and more broadly, to understand what properties of pre-trained models are important for transfer learning.

2 Motivation: Fixed-Feature Transfer Learning

In one of the most basic variants of transfer learning, one uses the source model as a feature extractor for the target dataset, then trains a simple (often linear) model on the resulting features. In our setting, this corresponds to first passing each image in the target dataset through a pre-trained ImageNet classifier, and then using the outputs from the penultimate layer as the image's feature representation. Prior work has demonstrated that applying this "fixed-feature" transfer learning approach yields accurate classifiers for a variety of vision tasks and often out-performs task-specific handcrafted features [Sha+14]. However, we still do not completely understand the factors driving transfer learning performance.

How can we improve transfer learning? Both conventional wisdom and evidence from prior work [Cha+14; SZ15; KSL19; Hua+17] suggests that accuracy on the source dataset is a strong indicator of performance on downstream tasks. In particular, Kornblith, Shlens, and Le [KSL19] find that pre-trained ImageNet models with higher accuracy yield better fixed-feature transfer learning results.

Still, it is unclear if improving ImageNet accuracy is the only way to improve performance. After all, the behaviour of fixed-feature transfer is governed by models' learned representations, which are not fully described by source-dataset accuracy. These representations are, in turn, controlled by the *priors* that we put on them during training. For example, the use of architectural components [UVL17], alternative loss functions [Mur+18], and data augmentation [VM01] have all been found to put distinct priors on the features extracted by classifiers.

The adversarial robustness prior. In this work, we turn our attention to another prior: *adversarial robustness*. Adversarial robustness refers to a model's invariance to small (often imperceptible) perturbations of its inputs. Robustness is typically induced at training time by replacing the standard empirical risk minimization objective with a robust optimization objective [Mad+18]:

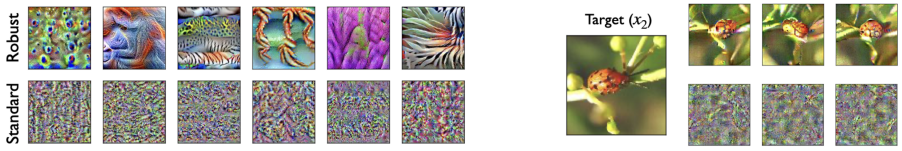
$$\min_\theta \mathbb{E}_{(x,y)\sim D}\left[\mathcal{L}(x,y;\theta)\right] \implies \min_\theta \mathbb{E}_{(x,y)\sim D}\left[\max_{\|\delta\|_2\leq\varepsilon} \mathcal{L}(x+\delta,y;\theta)\right], \tag{1}$$

where ε is a hyperparameter governing how invariant the resulting "adversarially robust model" (more briefly, "robust model") should be. In short, this objective asks the model to minimize risk on the training datapoints while also being locally stable in the (radius-ε) neighbourhood around each of these points. (A more detailed primer on adversarial robustness is given in Appendix E.)

Adversarial robustness was originally studied in the context of machine learning security [Big+13; BR18; CW17; Ath+18] as a method for improving models' resilience to adversarial examples [GSS15; Mad+18]. However, a recent line of work has studied adversarially robust models in their own right, casting (1) as a prior on learned feature representations [Eng+19a; Ily+19; Jac+19; ZZ19].

Should adversarial robustness help fixed-feature transfer? It is, a priori, unclear what to expect from an "adversarial robustness prior" in terms of transfer learning. On one hand, robustness to adversarial examples may seem somewhat tangential to transfer performance. In fact, adversarially robust models are known to be significantly less accurate than their standard counterparts [Tsi+19; Su+18; Rag+19; Nak19], suggesting that using adversarially robust feature representations should hurt transfer performance.

On the other hand, recent work has found that the feature representations of robust models carry several advantages over those of standard models. For example, adversarially robust representations typically have better-behaved gradients [Tsi+19; San+19; ZZ19; KCL19] and thus facilitate regularization-free feature visualization [Eng+19a] (cf. Figure 1a). Robust representations are also approximately invertible [Eng+19a], meaning that unlike for standard models [MV15; DB16], an image can be approximately reconstructed directly from its robust representation (cf. Figure 1b). More broadly, Engstrom et al. [Eng+19a] hypothesize that by forcing networks to be invariant to signals that humans are also invariant to, the robust training objective leads to feature representations that are more similar to what humans use. This suggests, in turn, that adversarial robustness might be a desirable prior from the point of view of transfer learning.



(a) Perceptually aligned gradients

(b) Representation invertibility

Figure 1: Adversarially robust (top) and standard (bottom) representations: robust representations allow (a) feature visualization without regularization; (b) approximate image inversion by minimizing distance in representation space. Figures reproduced from Engstrom et al. [Eng+19a].

Experiments. To resolve these two conflicting hypotheses, we use a test bed of 12 standard transfer learning datasets (all the datasets considered in [KSL19] as well as Caltech-256 [GHP07]) to evaluate fixed-feature transfer on standard and adversarially robust ImageNet models. We considere four ResNet-based architectures (ResNet-{18,50}, WideResNet-50-x{2,4}), and train models with varying robustness levels ε for each architecture (for the full experimental setup, see Appendix A).

In Figure 2, we compare the downstream transfer accuracy of a standard model to that of the best robust model with the same architecture (grid searching over ε). The results indicate that robust networks consistently extract better features for transfer learning than standard networks—this effect is most pronounced on Aircraft, CIFAR-10, CIFAR-100, Food, SUN397, and Caltech-101. Due to computational constraints, we could not train WideResNet-50-4x models at the same number of robustness levels ε, so a coarser grid was used. It is thus likely that a finer grid search over ε would further improve results (we discuss the role of ε in more detail in Section 4.3).

3 Adversarial Robustness and Full-Network Fine Tuning

A more expensive but often better-performing transfer learning method uses the pre-trained model as a weight initialization rather than as a feature extractor. In this "full-network" transfer learning setting, we update all of the weights of the pre-trained model (via gradient descent) to minimize loss on the target task. Kornblith, Shlens, and Le [KSL19] find that for standard models, performance on full-network transfer learning is highly correlated with performance on fixed-feature transfer learning. Therefore, we might hope that the findings of the last section (i.e., that adversarially robust models transfer better) also carry over to this setting. To resolve this conjecture, we consider

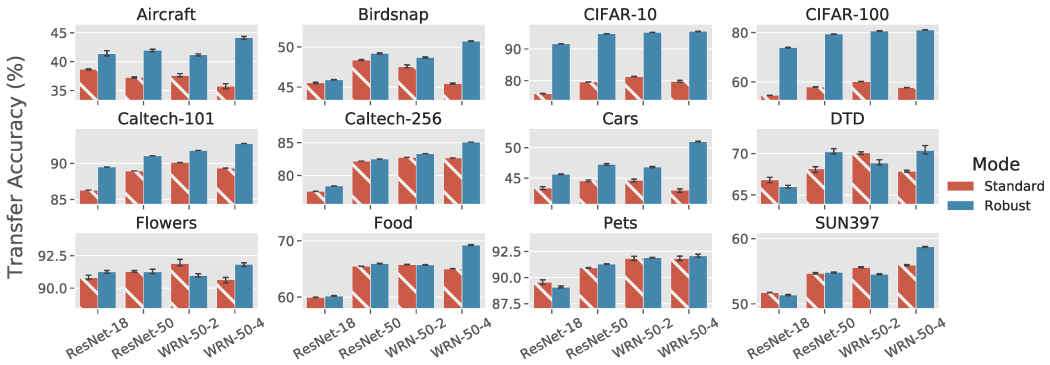


Figure 2: **Fixed-feature** transfer learning results using standard and robust models for the 12 downstream image classification tasks considered. Following [KSL19], we record re-weighted accuracy for the unbalanced datasets, and raw accuracy for the others (cf. Appendix A). Error bars denote the maximum and minimum error attained over three random trials. A similar plot with ten random trials is in Appendix F.

three applications of full-network transfer learning: downstream image classification (i.e., the tasks considered in Section 2), object detection, and instance segmentation.

3.1 Downstream image classification

We first recreate the setup of Section 2: we perform full-network transfer learning to adapt the robust and non-robust pre-trained ImageNet models to the same set of 12 downstream classification tasks. The hyperparameters for training were found via grid search (cf. Appendix A). Our findings are shown in Figure 3—just as in fixed-feature transfer learning, robust models match or improve on standard models in terms of transfer learning performance.

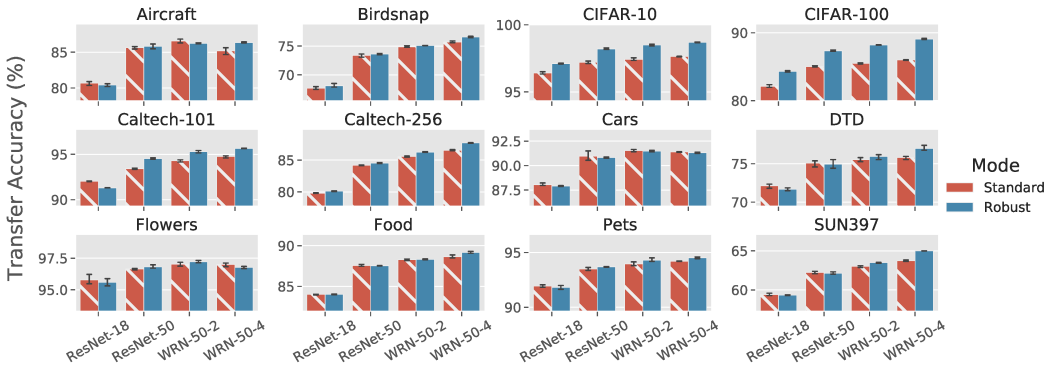
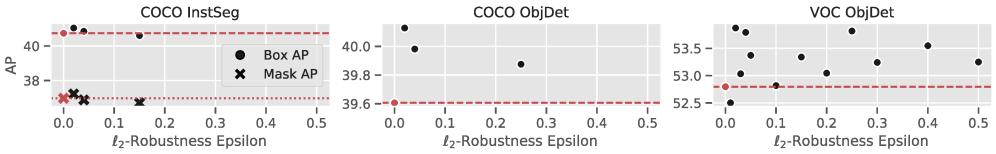


Figure 3: **Full-network** transfer learning results using standard and robust models for the 12 downstream image classification tasks considered. Following [KSL19], we record re-weighted accuracy for the unbalanced datasets, and raw accuracy for the others (cf. Appendix A). Error bars denote the maximum and minimum error attained over three random trials. A similar plot with ten random trials is in Appendix F.

3.2 Object detection and instance segmentation

It is standard practice in data-scarce object detection or instance segmentation tasks to initialize earlier model layers with weights from ImageNet-trained classification networks. We study the benefits of using robustly trained networks to initialize object detection and instance segmentation models, and find that adversarially robust networks consistently outperform standard networks.



Task	Box AP		Mask AP	
	Standard	Robust	Standard	Robust
VOC Object Detection	52.80	53.87	—	—
COCO Object Detection	39.80 ± 0.08	40.07 ± 0.10	—	—
COCO Instance Segmentation	40.67 ± 0.06	40.91 ± 0.15	36.92 ± 0.08	37.08 ± 0.10

Figure 4: AP of instance segmentation and object detection models with backbones initialized with ε-robust models before training. Robust backbones generally lead to better AP, and the best robust backbone always outperforms the standardly trained backbone for every task. COCO results averaged over four runs due to computational constraints; $\pm$ represents standard deviation.

Experimental setup. We evaluate with benchmarks in both object detection (PASCAL Visual Object Classes (VOC) [Eve+10] and Microsoft COCO [Lin+14]) and instance segmentation (Microsoft COCO). We train systems using default models and hyperparameter configurations from the Detectron2 [Wu+19] framework (i.e., we do not perform any additional hyperparameter search). Appendix C describes further experimental details and more results.

We first study object detection. We train Faster R-CNN FPN [Lin+17] models with varying ResNet-50 backbone initializations. For VOC, we initialize with one standard network, and twelve adversarially robust networks with different values of ε. For COCO, we only train with three adversarially robust models (due to computational constraints). For instance segmentation, we train Mask R-CNN FPN models [He+17] while varying ResNet-50 backbone initialization. We train three models using adversarially robust initializations, and one model from a standardly trained ResNet-50. Figure 4 summarizes our findings: the best robust backbone initializations outperform standard models.

4 Analysis and Discussion

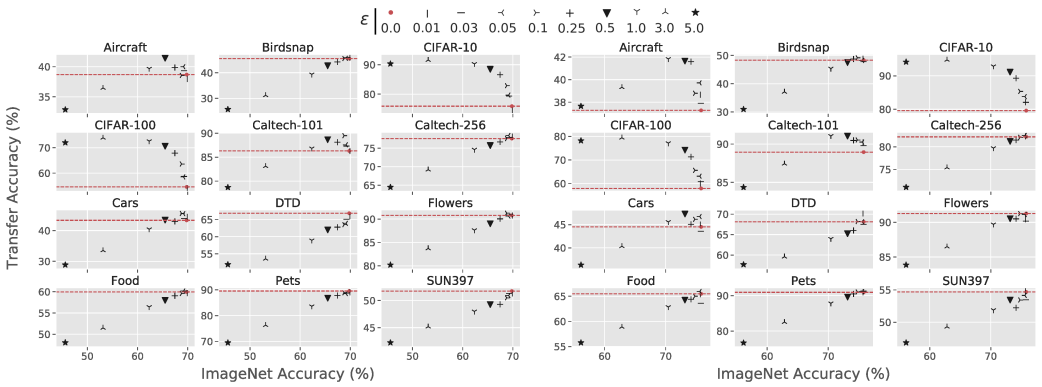
Our results from the previous section indicate that robust models match or improve on the transfer learning performance of standard ones. In this section, we take a closer look at the similarities and differences in transfer learning between robust networks and standard networks.

4.1 ImageNet accuracy and transfer performance

In Section 2, we discussed a potential tension between the desirable properties of robust network representations (which we conjectured would improve transfer performance) and the decreased accuracy of the corresponding models (which, as prior work has established, should hurt transfer performance). We hypothesize that robustness and accuracy have counteracting yet separate effects: that is, higher accuracy improves transfer learning for a fixed level of robustness, and higher robustness improves transfer learning for a fixed level of accuracy.

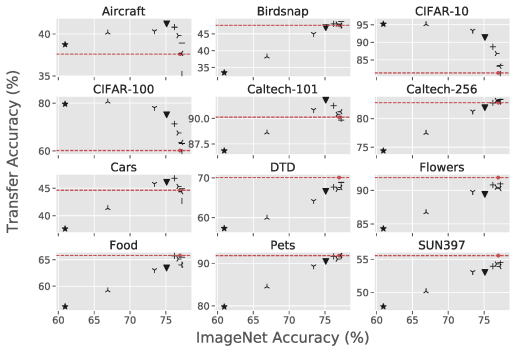
To test this hypothesis, we first study the relationship between ImageNet accuracy and transfer accuracy for each of the robust models that we trained. Under our hypothesis, we should expect to see a deviation from the direct linear accuracy-transfer relation observed by [KSL19], due to the confounding factor of varying robustness. The results (cf. Figure 5; similar results for full-network transfer in Appendix F) support this. Indeed, we find that the previously observed linear relationship between accuracy and transfer performance is often violated once robustness aspect comes into play.

In even more direct support of our hypothesis (i.e., that robustness and ImageNet accuracy have opposing yet separate effects on transfer), we find that when the robustness level is held fixed, the accuracy-transfer correlation observed by prior works for standard models actually holds for robust models too. Specifically, we train highly robust ($\varepsilon = 3$)—and thus less accurate—models with six different architectures, and compared ImageNet accuracy against transfer learning performance.

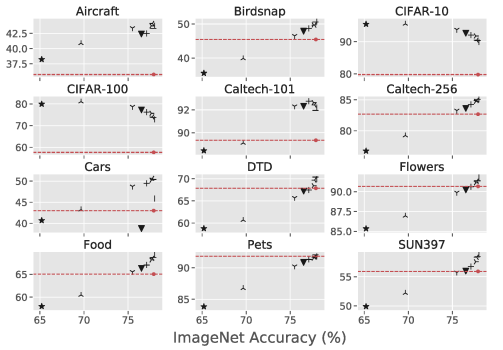


(a) ResNet-18

(b) ResNet-50



(c) WRN-50-2



(d) WRN-50-4

Figure 5: **Fixed-feature** transfer accuracies of standard and robust ImageNet models to various image classification datasets. The linear relationship between ImageNet and transfer accuracies does not hold.

Table 2: Source (ImageNet) and target (CIFAR-10) accuracies, fixing robustness (ε) but varying architecture. When robustness is controlled for, ImageNet accuracy is highly predictive of transfer performance. Similar trends for other datasets are shown in Appendix F.

		Architecture (see details in Appendix A.1)						
Robustness	**Dataset**	A	B	C	D	E	F	R^2
Std ($\varepsilon = 0$)	ImageNet	77.37	77.32	73.66	65.26	64.25	60.97	—
	CIFAR-10	97.84	97.47	96.08	95.86	95.82	95.55	0.79
Adv ($\varepsilon = 3$)	ImageNet	66.12	65.92	56.78	50.05	42.87	41.03	—
	CIFAR-10	98.67	98.22	97.27	96.91	96.23	95.99	0.97

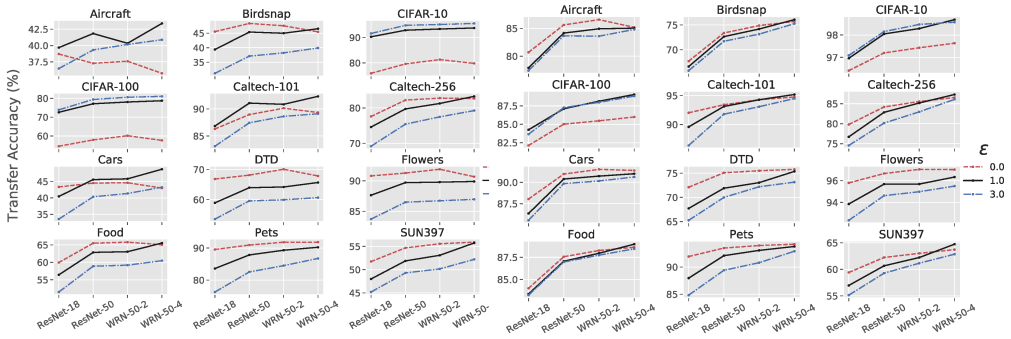
Table 2 shows that for these models improving ImageNet accuracy improves transfer performance at around the same rate as (and with higher R^2 correlation than) standard models.

These observations suggest that transfer learning performance can be further improved by applying known techniques that increase the accuracy of robust models (e.g. [BGH19; Car+19]). More broadly, our findings also indicate that accuracy is not a sufficient measure of feature quality or versatility. Understanding why robust networks transfer particularly well remains an open problem, likely relating to prior work that analyses the features these networks use [Eng+19a; Sha+19; AL20].

4.2 Robust models improve with width

Our experiments also reveal a contrast between robust and standard models in how their transfer performance scales with model width. Azizpour et al. [Azi+15], find that although increasing network

depth improves transfer performance, increasing width hurts it. Our results corroborate this trend for standard networks, but indicate that it does *not* hold for robust networks, at least in the regime of widths tested. Indeed, Figure 6 plots results for the three widths of ResNet-50 studied here (x1, x2, and x4), along with a ResNet-18 for reference: as width increases, transfer performance plateaus and decreases for standard models, but continues to steadily grow for robust models. This suggests that scaling network width may further increase the transfer performance gain of robust networks over the standard ones. (This increase comes, however, at a higher computational cost.)



(a) **Fixed-feature** transfer

(b) **Full-network** transfer

Figure 6: Varying width and model robustness while transfer learning from ImageNet to various datasets. Generally, as width increases, transfer learning accuracies of standard models generally plateau or level off while those of robust models steadily increase. More values of ε are in Appendix F.

4.3 Optimal robustness levels for downstream tasks

We observe that although the best robust models often outperform the best standard models, the optimal choice of robustness parameter ε varies widely between datasets. For example, when transferring to CIFAR-10 and CIFAR-100, the optimal ε values were 3.0 and 1.0, respectively. In contrast, smaller values of ε (smaller by an order of magnitude) tend to work better for the rest of the datasets.

One possible explanation for this variability in the optimal choice of ε might relate to dataset granularity. We hypothesize that on datasets where leveraging finer-grained features are necessary (i.e., where there is less norm-separation between classes in the input space), the most effective values of ε will be much smaller than for a dataset where leveraging more coarse-grained features suffices. To illustrate this, consider a binary classification task consisting of image-label pairs (x, y), where the correct class for an image $y \in \{0, 1\}$ is determined by a single pixel, i.e., $x_{0,0} = \delta \cdot y$, and $x_{i,j} = 0$, otherwise. We would expect transferring a standard model onto this dataset to yield perfect accuracy regardless of δ, since the dataset is perfectly separable. On the other hand, a robust model is trained to be invariant to perturbations of norm ε—thus, if $\delta < \varepsilon$, the dataset will not appear separable to the standard model and so we expect transfer to be less successful. So, the smaller the δ (i.e., the larger the "fine grained-ness" of the dataset), the smaller the ε must be for successful transfer.

Unifying dataset scale. We now present evidence in support of our above hypothesis. Although we lack a quantitative notion of granularity (in reality, features are not simply singular pixels), we consider image resolution as a crude proxy. Since we scale target datasets to match ImageNet dimensions, each pixel in a low-resolution dataset (e.g., CIFAR-10) image translates into several pixels in transfer, thus inflating datasets' separability. Drawing from this observation, we attempt to calibrate the granularities of the 12 image classification datasets used in this work, by first downscaling all the images to the size of CIFAR-10 (32×32), and then upscaling them to ImageNet size once more. We then repeat the fixed-feature regression experiments from prior sections, plotting the results in Figure 7 (similar results for full-network transfer are presented in Appendix F). After controlling for original dataset dimension, the datasets' epsilon vs. transfer accuracy curves all behave almost identically to CIFAR-10 and CIFAR-100 ones. Note that while this experimental data supports our hypothesis, we do not take the evidence as an ultimate one and further exploration is needed to reach definitive conclusions.

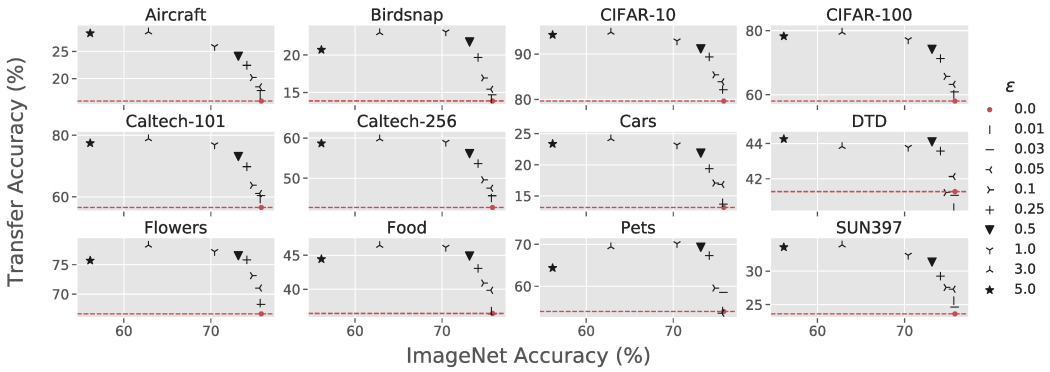
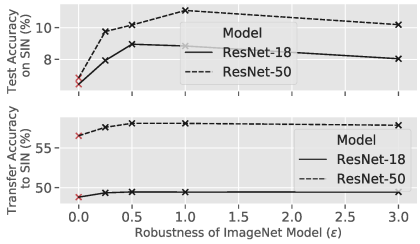
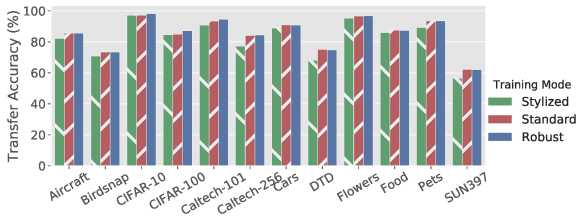


Figure 7: **Fixed-feature** transfer accuracies of various datasets that are down-scaled to 32×32 before being up-scaled again to ImageNet scale and used for transfer learning. The accuracy curves are closely aligned, unlike those of Figure 5, which illustrates the same experiment without downscaling.

4.4 Comparing adversarial robustness to texture robustness



(a) Stylized ImageNet Transfer



(b) Transfer to Standard Datasets using a ResNet-50

Figure 8: We compare standard, stylized and robust ImageNet models on standard transfer tasks (and to stylized ImageNet).

We now investigate the effects of adversarial robustness on transfer learning performance in comparison to other invariances commonly imposed on deep neural networks. Specifically, we consider texture-invariant [Gei+19] models, i.e., models trained on the texture-randomizing Stylized ImageNet (SIN) [Gei+19] dataset. Figure 8b shows that transfer learning from adversarially robust models outperforms transfer learning from texture-invariant models on all considered datasets.

Finally, we use the SIN dataset to further re-inforce the benefits conferred by adversarial robustness. Figure 8a top shows that robust models outperform standard imagenet models when evaluated (top) or fine-tuned (bottom) on Stylized-ImageNet.

5 Related Work

A number of works study transfer learning with CNNs [Don+14; Cha+14; Sha+14; Azi+15]. Indeed, transfer learning has been studied in varied domains including medical imaging [MGM18], language modeling [CK18], and various object detection and segmentation related tasks [Ren+15; Dai+16; Hua+17; Che+17]. In terms of methods, others [AGM14; Cha+14; Gir+14; Yos+14; Azi+15; LRM15; HAE16; Chu+16] show that fine-tuning typically outperforms frozen feature-based methods. As discussed throughout this paper, several prior works [Azi+15; HAE16; KSL19; Zam+18; Kol+19; Sun+17; Mah+18; Yos+14] have investigated factors improving or otherwise affecting transfer learning performance. Recently proposed methods have achieved state-of-the-art performance on downstream tasks by scaling up transfer learning techniques [Hua+18; Kol+19].

On the adversarial robustness front, many works—both empirical (e.g., [Mad+18; Miy+18; BGH19; Zha+19]) and certified (e.g., [Lec+19; Wen+18; WK18; RSL18; CRK19; Sal+19; Yan+20])—significantly increase model resilience to adversarial examples [Big+13; Sze+14]. A growing body

of research has studied the *features* learned by these robust networks and suggested that they improve upon those learned by standard networks (cf. [Ily+19; Eng+19a; San+19; AL20; KSJ19; KCL19] and references). On the other hand, prior studies have also identified theoretical and empirical tradeoffs between standard accuracy and adversarial robustness [Tsi+19; BPR18; Su+18; Rag+19]. At the intersection of robustness and transfer learning, Shafahi et al. [Sha+19] investigate transfer learning for increasing downstream-task adversarial robustness (rather than downstream accuracy, as in this work). Aggarwal et al. [Agg+20] find that adversarially trained models perform better at downstream zero-shot learning tasks and weakly-supervised object localization. Finally, concurrent to our work, [Utr+20] also study the transfer performance of adversarially robust networks. Our studies reach similar conclusions and are otherwise complementary: here we study a larger set of downstream datasets and tasks and analyze the effects of model accuracy, model width, and data resolution; Utrera et al. [Utr+20] study the effects of training duration, dataset size, and also introduce an influence function-based analysis [KL17] to study the representations of robust networks. For a detailed discussion of prior work, see Appendix D.

6 Conclusion

In this work, we propose using adversarially robust models for transfer learning. We compare transfer learning performance of robust and standard models on a suite of 12 classification tasks, object detection, and instance segmentation. We find that adversarial robust neural networks consistently match or improve upon the performance of their standard counterparts, despite having lower ImageNet accuracy. We also take a closer look at the behavior of adversarially robust networks, and study the interplay between ImageNet accuracy, model width, robustness, and transfer performance.

Acknowledgements

Work supported in part by the NSF awards CCF-1553428, CNS-1815221, the Open Philanthropy Project AI Fellowship, and the Microsoft Corporation. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0015.

7 Statement of Broader Impact

Our work attempts to improve upon standard techniques within computer vision, and as such comes with all of the positive and negative broader impacts of the larger field. More specifically, however, transfer learning allows researchers and practitioners to efficiently train models on their custom datasets starting from models pretrained on large-scale labeled datasets. In this way, transfer learning helps those who are compute-limited or otherwise resource-constrained competititive, and thus makes ML more accessible. We believe that our paper discovers new aspects of pretrained models that make them effective at transfer learning, therefore pushing our understanding of transfer learning and helping us to improve its performance.

References

[Agg+20] Gunjan Aggarwal et al. "On the Benefits of Models with Perceptually-Aligned Gradients". In: *Towards Trustworthy ML Workshop (ICLR)*. 2020.

[AGM14] Pulkit Agrawal, Ross Girshick, and Jitendra Malik. "Analyzing the performance of multilayer neural networks for object recognition". In: *European conference on computer vision*. 2014.

[AL20] Zeyuan Allen-Zhu and Yuanzhi Li. "Feature Purification: How Adversarial Training Performs Robust Deep Learning". In: 2020. arXiv: 2005.10190 [cs.LG].

[Ath+18] Anish Athalye et al. "Synthesizing Robust Adversarial Examples". In: *International Conference on Machine Learning (ICML)*. 2018.

[Azi+15] Hossein Azizpour et al. "Factors of transferability for a generic convnet representation". In: *IEEE transactions on pattern analysis and machine intelligence* (2015).

[Ber+14] Thomas Berg et al. "Birdsnap: Large-scale fine-grained visual categorization of birds". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.

[BGH19] Yogesh Balaji, Tom Goldstein, and Judy Hoffman. "Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets". In: *Arxiv preprint arXiv:1910.08051*. 2019.

[BGV14] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. "Food-101–mining discriminative components with random forests". In: *European conference on computer vision*. 2014.

[Big+13] Battista Biggio et al. "Evasion attacks against machine learning at test time". In: *Joint European conference on machine learning and knowledge discovery in databases (ECML-KDD)*. 2013.

[BPR18] Sébastien Bubeck, Eric Price, and Ilya Razenshteyn. "Adversarial examples from computational constraints". In: *arXiv preprint arXiv:1805.10204*. 2018.

[BR18] Battista Biggio and Fabio Roli. "Wild patterns: Ten years after the rise of adversarial machine learning". In: 2018.

[Car+19] Yair Carmon et al. "Unlabeled data improves adversarial robustness". In: *Neural Information Processing Systems (NeurIPS)*. 2019.

[Cha+14] Ken Chatfield et al. "Return of the devil in the details: Delving deep into convolutional nets". In: *arXiv preprint arXiv:1405.3531* (2014).

[Che+17] Liang-Chieh Chen et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs". In: *IEEE transactions on pattern analysis and machine intelligence* (2017).

[Chu+16] Brian Chu et al. "Best practices for fine-tuning visual classifiers to new domains". In: *European conference on computer vision*. 2016.

[Cim+14] Mircea Cimpoi et al. "Describing textures in the wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.

[CK18] Alexis Conneau and Douwe Kiela. "Senteval: An evaluation toolkit for universal sentence representations". In: *arXiv preprint arXiv:1803.05449* (2018).

[CRK19] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. "Certified adversarial robustness via randomized smoothing". In: *arXiv preprint arXiv:1902.02918*. 2019.

[CW17] Nicholas Carlini and David Wagner. "Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods". In: *Workshop on Artificial Intelligence and Security (AISec)*. 2017.

[Dai+16] Jifeng Dai et al. "R-fcn: Object detection via region-based fully convolutional networks". In: *Advances in neural information processing systems*. 2016.

[Dan67] John M. Danskin. *The Theory of Max-Min and its Application to Weapons Allocation Problems*. 1967.

[DB16] Alexey Dosovitskiy and Thomas Brox. "Inverting visual representations with convolutional networks". In: *Computer Vision and Pattern Recognition (CVPR)*. 2016.

[Den+09] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *computer vision and pattern recognition (CVPR)*. 2009.

[Don+14] Jeff Donahue et al. "Decaf: A deep convolutional activation feature for generic visual recognition". In: *International conference on machine learning (ICML)*. 2014.

[Eng+19a] Logan Engstrom et al. "Learning Perceptually-Aligned Representations via Adversarial Robustness". In: *ArXiv preprint arXiv:1906.00945*. 2019.

[Eng+19b] Logan Engstrom et al. *Robustness (Python Library)*. 2019. URL: https://github.com/MadryLab/robustness.

[Eve+10] M. Everingham et al. "The Pascal Visual Object Classes (VOC) Challenge". In: *International Journal of Computer Vision*. 2010.

[Evt+18] Ivan Evtimov et al. "Robust Physical-World Attacks on Machine Learning Models". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[FFP04] Li Fei-Fei, Rob Fergus, and Pietro Perona. "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories". In: *2004 conference on computer vision and pattern recognition workshop*. IEEE. 2004, pp. 178–178.

[Gao+19] Ruiqi Gao et al. "Convergence of Adversarial Training in Overparametrized Networks". In: *arXiv preprint arXiv:1906.07916* (2019).

[Gei+19] Robert Geirhos et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." In: *International Conference on Learning Representations*. 2019.

[GHP07] Gregory Griffin, Alex Holub, and Pietro Perona. "Caltech-256 object category dataset". In: (2007).

[Gir+14] Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *computer vision and pattern recognition (CVPR)*. 2014, pp. 580–587.

[GSS15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *International Conference on Learning Representations (ICLR)*. 2015.

[HAE16] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. "What makes ImageNet good for transfer learning?" In: *arXiv preprint arXiv:1608.08614* (2016).

[He+17] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[Hua+17] Jonathan Huang et al. "Speed/accuracy trade-offs for modern convolutional object detectors". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[Hua+18] Yanping Huang et al. "GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism". In: *ArXiv preprint arXiv:1811.06965*. 2018.

[Ily+18] Andrew Ilyas et al. "Black-box Adversarial Attacks with Limited Queries and Information". In: *International Conference on Machine Learning (ICML)*. 2018.

[Ily+19] Andrew Ilyas et al. "Adversarial Examples Are Not Bugs, They Are Features". In: *Neural Information Processing Systems (NeurIPS)*. 2019.

[Jac+19] Jorn-Henrik Jacobsen et al. "Excessive Invariance Causes Adversarial Vulnerability". In: *International Contemporary on Learning Representations*. 2019.

[KCL19] Simran Kaur, Jeremy Cohen, and Zachary C. Lipton. "Are Perceptually-Aligned Gradients a General Property of Robust Classifiers?" In: *Arxiv preprint arXiv:1910.08640*. 2019.

[KL17] Pang Wei Koh and Percy Liang. "Understanding Black-box Predictions via Influence Functions". In: *ICML*. 2017.

[Kol+19] Alexander Kolesnikov et al. "Big Transfer (BiT): General Visual Representation Learning". In: *arXiv preprint arXiv:1912.11370* (2019).

[Kra+13] Jonathan Krause et al. "Collecting a large-scale dataset of fine-grained cars". In: (2013).

[Kri09] Alex Krizhevsky. "Learning Multiple Layers of Features from Tiny Images". In: *Technical report*. 2009.

[KSJ19] Beomsu Kim, Junghoon Seo, and Taegyun Jeon. "Bridging Adversarial Robustness and Gradient Interpretability". In: *International Conference on Learning Representations Workshop on Safe Machine Learning (ICLR SafeML)*. 2019.

[KSL19] Simon Kornblith, Jonathon Shlens, and Quoc V Le. "Do better imagenet models transfer better?" In: *computer vision and pattern recognition (CVPR)*. 2019.

[Lec+19] Mathias Lecuyer et al. "Certified robustness to adversarial examples with differential privacy". In: *Symposium on Security and Privacy (SP)*. 2019.

[Lin+14] Tsung-Yi Lin et al. "Microsoft coco: Common objects in context". In: *European conference on computer vision (ECCV)*. 2014.

[Lin+17] Tsung-Yi Lin et al. "Feature pyramid networks for object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[LRM15] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. "Bilinear cnn models for fine-grained visual recognition". In: *Proceedings of the IEEE international conference on computer vision*. 2015.

[LSK19] Juncheng Li, Frank R. Schmidt, and J. Zico Kolter. "Adversarial camera stickers: A physical camera-based attack on deep learning systems". In: *Arxiv preprint arXiv:1904.00759*. 2019.

[Mad+18] Aleksander Madry et al. "Towards deep learning models resistant to adversarial attacks". In: *International Conference on Learning Representations (ICLR)*. 2018.

[Mah+18] Dhruv Mahajan et al. "Exploring the limits of weakly supervised pretraining". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

[Maj+13] Subhransu Maji et al. "Fine-grained visual classification of aircraft". In: *arXiv preprint arXiv:1306.5151* (2013).

[MGM18] Romain Mormont, Pierre Geurts, and Raphaël Marée. "Comparison of deep transfer learning strategies for digital pathology". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018.

[Miy+18] Takeru Miyato et al. "Virtual adversarial training: a regularization method for supervised and semi-supervised learning". In: 2018.

[Mur+18] Nikhil Muralidhar et al. "Incorporating prior domain knowledge into deep neural networks". In: *2018 IEEE International Conference on Big Data (Big Data)*. 2018.

[MV15] Aravindh Mahendran and Andrea Vedaldi. "Understanding deep image representations by inverting them". In: *computer vision and pattern recognition (CVPR)*. 2015.

[Nak19] Preetum Nakkiran. "Adversarial robustness may be at odds with simplicity". In: *arXiv preprint arXiv:1901.00532*. 2019.

[NZ08] Maria-Elena Nilsback and Andrew Zisserman. "Automated flower classification over a large number of classes". In: *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. 2008.

[Pap+17] Nicolas Papernot et al. "Practical black-box attacks against machine learning". In: *Asia Conference on Computer and Communications Security*. 2017.

[Par+12] Omkar M Parkhi et al. "Cats and dogs". In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 3498–3505.

[Rag+19] Aditi Raghunathan et al. "Adversarial Training Can Hurt Generalization". In: *arXiv preprint arXiv:1906.06032* (2019).

[Ren+15] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems*. 2015.

[RSL18] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. "Certified defenses against adversarial examples". In: *International Conference on Learning Representations (ICLR)*. 2018.

[Rus+15] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)*. 2015.

[RWK20] Leslie Rice, Eric Wong, and J. Zico Kolter. "Overfitting in adversarially robust deep learning". In: *Arxiv preprint arXiv:2002.11569*. 2020.

[Sal+19] Hadi Salman et al. "Provably robust deep learning via adversarially trained smoothed classifiers". In: *Advances in Neural Information Processing Systems*. 2019.

[San+19] Shibani Santurkar et al. "Image Synthesis with a Single (Robust) Classifier". In: *Neural Information Processing Systems (NeurIPS)*. 2019.

[Sha+14] Ali Sharif Razavian et al. “CNN features off-the-shelf: an astounding baseline for recognition”. In: *conference on computer vision and pattern recognition (CVPR) workshops*. 2014.

[Sha+19] Ali Shafahi et al. “Adversarially robust transfer learning”. In: *arXiv preprint arXiv:1905.08232* (2019).

[Su+18] Dong Su et al. “Is Robustness the Cost of Accuracy? A Comprehensive Study on the Robustness of 18 Deep Image Classification Models”. In: *European Conference on Computer Vision (ECCV)*. 2018.

[Sun+17] Chen Sun et al. “Revisiting unreasonable effectiveness of data in deep learning era”. In: *Proceedings of the IEEE international conference on computer vision*. 2017.

[SZ15] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *International Conference on Learning Representations (ICLR)*. 2015.

[Sze+14] Christian Szegedy et al. “Intriguing properties of neural networks”. In: *International Conference on Learning Representations (ICLR)*. 2014.

[Tsi+19] Dimitris Tsipras et al. “Robustness May Be at Odds with Accuracy”. In: *International Conference on Learning Representations (ICLR)*. 2019.

[Utr+20] Francisco Utrera et al. “Adversarially-Trained Deep Nets Transfer Better”. In: *ArXiv preprint arXiv:2007.05869*. 2020.

[UVL17] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. “Deep Image Prior”. In: *ArXiv preprint arXiv:1711.10925*. 2017.

[VM01] David A Van Dyk and Xiao-Li Meng. “The art of data augmentation”. In: *Journal of Computational and Graphical Statistics*. 2001.

[Wal45] Abraham Wald. “Statistical Decision Functions Which Minimize the Maximum Risk”. In: *Annals of Mathematics*. 1945.

[Wen+18] Tsui-Wei Weng et al. “Towards fast computation of certified robustness for ReLU networks”. In: *International Conference on Machine Learning (ICML)*. 2018.

[WK18] Eric Wong and J Zico Kolter. “Provable defenses against adversarial examples via the convex outer adversarial polytope”. In: *International Conference on Machine Learning (ICML)*. 2018.

[Wu+19] Yuxin Wu et al. *Detectron2*. https://github.com/facebookresearch/detectron2. 2019.

[Xia+10] Jianxiong Xiao et al. “Sun database: Large-scale scene recognition from abbey to zoo”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2010.

[Yan+20] Greg Yang et al. *Randomized Smoothing of All Shapes and Sizes*. 2020.

[Yos+14] Jason Yosinski et al. “How transferable are features in deep neural networks?” In: *Advances in neural information processing systems*. 2014.

[Zam+18] Amir R Zamir et al. “Taskonomy: Disentangling task transfer learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[Zha+19] Hongyang Zhang et al. “Theoretically Principled Trade-off between Robustness and Accuracy”. In: *International Conference on Machine Learning (CIML)*. 2019.

[Zha+20] Yi Zhang et al. “Over-parameterized Adversarial Training: An Analysis Overcoming the Curse of Dimensionality”. In: *Arxiv preprint arXiv:2002.06668*. 2020.

[ZZ19] Tianyuan Zhang and Zhanxing Zhu. “Interpreting Adversarially Trained Convolutional Neural Networks”. In: *International Conference on Machine Learning (ICML)*. 2019.