Transfer learning with fewer ImageNet classes

Michal Kucer
ISR-3 Space Data Science and Systems
Los Alamos National Laboratory
Los Alamos, NM 87544
michal@lanl.gov

Diane Oyen
CCS-3 Information Sciences
Los Alamos National Laboratory
Los Alamos, NM 87544
doyen@lanl.gov

Abstract

Though much previous work tried to uncover the best practices for transfer learning, much is left unexplored. Our preliminary work explores the effect of removing a portion of the ImageNet classes with low per-class validation accuracy on the accuracy of the remaining classes. Furthermore, we explore if models trained with a reduced number of classes are suitable for transfer learning.

Introduction Models initialized with ImageNet trained weights serve as a basis for exploration of various problems, e.g. image retrieval [7] and video classification [4]. Recent methods based on weakly supervised training [16] and self-supervised training [25] achieve state of the art results on the ImageNet validation accuracies, however they use 3.5 billion and 300 million images respectively in addition to the the 1.3 million images in ImageNet. Though these are important results, it is hard to imagine one being able to achieve them without the vast computational resources of large companies. On the other hand, ImageNet with 1.3M images is a more accessible dataset requiring a machine with 2-4 GPUs to be able to run quality experiments. Therefore it is important to explore various methods of improving the representations learned on the ImageNet dataset in order for them to be used as a basis for other tasks. We are interested in understanding and improving learned representations for transfer learning. Much work in previous years has been put into understanding the best way to perform transfer learning [12, 21], however much remains to be understood. We explore the effect of removing 100 or 200 "hard" classes from the ImageNet dataset, and how this removal affects the validation accuracy on the remaining 800 classes and its effect on the learned representations for transfer-learning. The following are two questions we are interested in understanding:

Should removing "hard" classes improve the accuracy of ImageNet models? Looking at this question through the lens of curriculum learning [1], it can be argued that the answer is *yes*. We postulate that if one removes a portion of the lowest performing classes, then training the network will improve the accuracy on the remaining classes.

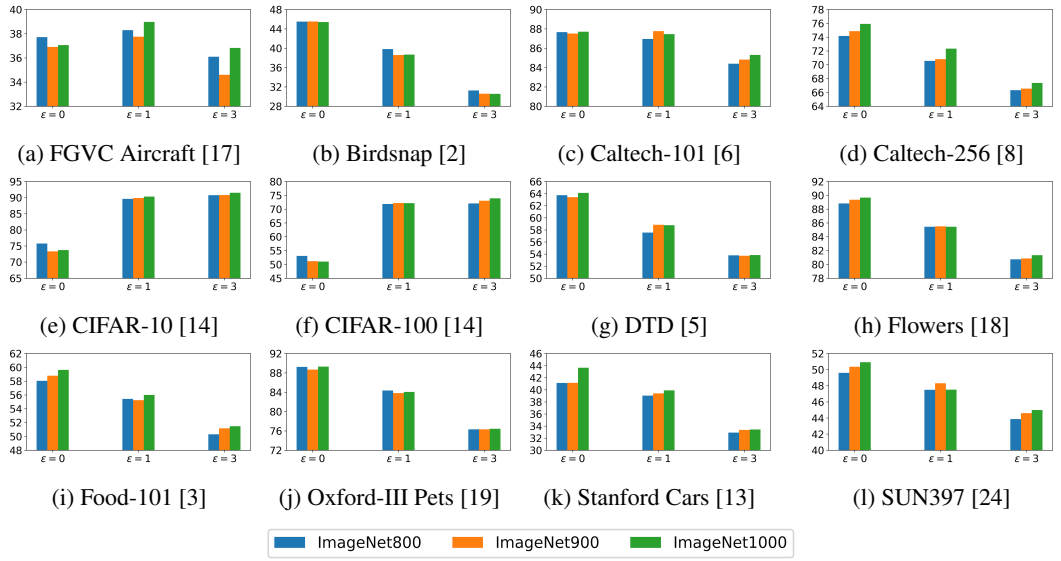
Would improving the accuracy on a subset of the images improve the transfer-learning performance on different datasets? If we look at the recent work by [12], the answer is seemingly yes. [12] shows that models with higher accuracy on the ImageNet dataset achieve better transfer-learning performance. However, by removing a certain number of classes from the training set, one is giving up the volume of data and diversity of images. Therefore, we set out to understand the transfer-learning performance of models which were trained with top 800, 900, or 1000 classes based on validation accuracy.

Experimental details and results All of our experiments are implemented using PyTorch Deep Learning Framework [20], and extend the codebase from [21] (MIT License). Our initial experiments utilize ResNet-18 [9] as the architecture for all of our experiments. Training of the ImageNet models is done on a single node, each running 2 18-core Xeon CPUs and 4 NVIDIA P100s GPU. Training takes approximately 32 hours for the non-robust model, and 80 hours for a robust model (utilizing a 3-step PGD attack [15] to generate adversarial examples during training). Each ImageNet model is

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Table 1: Table showing the validation accuracies (Top-1 / Top-5 accuracy) of the various models on the ImageNet dataset on the subset of 800 classes with the best validation accuracy. The left column of the table shows the number of classes that were used to train the ImageNet model. Columns 3 and 4 show the validation accuracies of l_2-robust ImageNet models - left pair shows the accuracy on the clean ImageNet validation set, and right pair shows the accuracy on the adversarial validation set with ϵ equal to the robustness of the particular model.

Number of classes	$\epsilon = 0.$	$\epsilon = 1.$	$\epsilon = 3.$
800	78.27 / 92.88	71.44 / 88.81 | 56.34 / 81.34	60.70 / 80.86 | 37.50 / 64.87
900	77.42 / 92.56	69.97 / 87.87 | 55.25 / 80.09	60.35 / 81.29 | 36.47 / 63.02
1000	76.23 / 91.80	68.65 / 87.01 | 54.05 / 79.05	59.30 / 80.37 | 35.61 / 61.78



(a) FGVC Aircraft [17] (b) Birdsnap [2] (c) Caltech-101 [6] (d) Caltech-256 [8]

(e) CIFAR-10 [14] (f) CIFAR-100 [14] (g) DTD [5] (h) Flowers [18]

(i) Food-101 [3] (j) Oxford-III Pets [19] (k) Stanford Cars [13] (l) SUN397 [24]

Figure 1: Graphs comparing the fixed-feature transfer learning accuracy across 12 datasets at different levels of robustness and number of ImageNet classes used for training of the baseline models.

trained for 100 epochs, with initial learning rate of 0.1 and decreased by 10 every 30 epochs. The classes removed and defined as "hard" are those with the lowest validation accuracies on the ImageNet validation set with 1000 classes. We denote *ImageNet800* and *ImageNet900* the models trained with the 800 and 900 classes with the highest validation accuracies respectively. *ImageNet1000* is a model trained will all classes. [21] support the hypothesis that adversarial robustness improves features representations, and therefore we train additional models which are l_2 robust with $\epsilon = 1$ and $\epsilon = 3$. For transfer learning, we use the same 12 datasets used by [21].

ImageNet experiments We first explore the difference in validation accuracy of ImageNet800, ImageNet900, and ImageNet1000 models on the validation set of the 800 common classes. The results can be seen in the Table 1. Table 1 shows the validation accuracy for varying number of classes and degrees of robustness for the validation set with 800 classes (with best validation accuracy). As we can see, in both robust and non-robust models the validation accuracy decreases as more classes are added to the training set. This is likely due to the additional classes being confused with some of the original classes.

Transfer learning experiments In our transfer learning experiments, we follow the setup used in [21], and test both "fixed-feature" and full-network transfer. Figure 1 shows the transfer learning results for the "fixed-features" setting, in which the weights of the network are frozen, and only the top layer is fine-tuned. Interestingly, only in half of the non-robust case is the full model with 1000 classes more accurate on the target dataset and ImageNet800 or ImageNet900. And in some cases, robustifying the various models removed the advantage of haveing more classes, e.g. the Flowers and SUN-397 datasets. Figure 2 shows the full-network transfer learning results, where all of the weights were fine-tuned. As we can see from the Figure 2, transfer-learning benefited from being

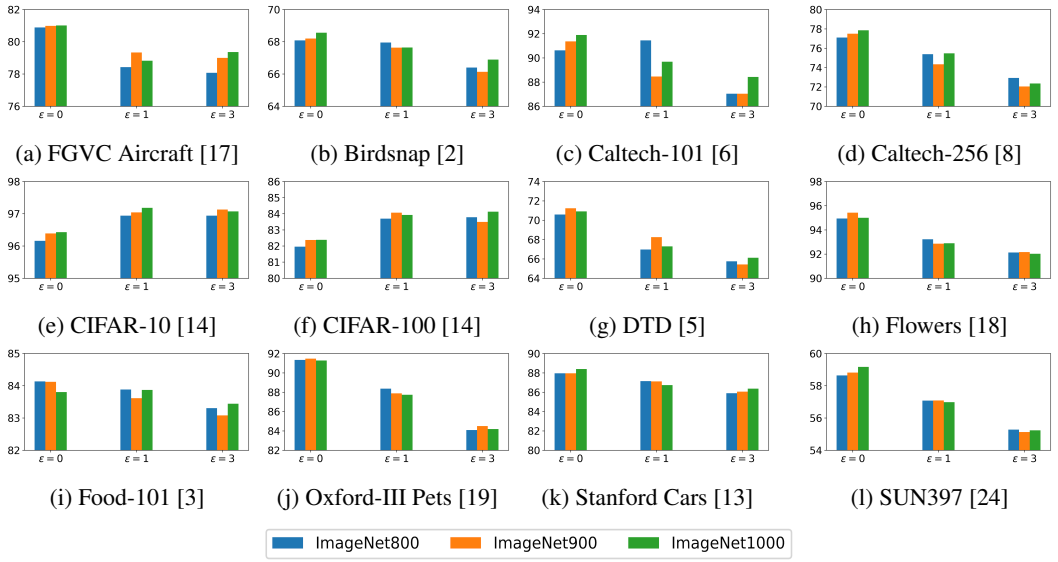


Figure 2: Graphs comparing the full-network transfer learning accuracy across 12 datasets at different levels of robustness and number of ImageNet classes used for training of the baseline models.

trained on larger number of classes in 7 out of the 12 cases. In case of DTD, Flowers, Food-101, and Oxford-III Pets datasets, we saw no or slight improvement in transfer learning performance when removing the classes with the worst validation accuracies. This however changes, when we look at the robust case - in 9 out of 12 datasets models initlaized with $l_2(\epsilon = 1)$ robust models with 800 or 900 classes performed better.

Limitations of our work. This is preliminary work in exploring methods for improving the learned representations and their suitability for transfer-learning. Due to computational and time constraints, our experiments are limited to a single architecture, a small number of robustness values, and a single run of transfer-learning experiments.

Conclusion Though previous work identifies some factors that effect transfer learning from ImageNet pre-trained models, much work still remains. In our preliminary experiments we show that even though we remove a significant portion of classes from the ImageNet dataset (up to 20 %), one can still achieve comparable or better performance in terms of transfer learning. In the following paragraphs, we highlight potential directions for future research.

Leveraging self-supervised learning for learning better representations. [21] provide evidence that adversarially robust models achieve better transfer learning in both a "fixed-features" setting, and full-network transfer learning. Examining Table 8. in the supplement of [22], we can see that in case of full-network transfer learning, for majority of the datasets and tested architectures, the robustness setting that provides the best transfer-learning is either $\epsilon = 0$ or $\epsilon < 0.1$. Only in case of the CIFAR-10 and CIFAR-100 datasets, the best performing robust models had $\epsilon \geq 1.0$. Recent work by [10] suggests that supplementing a supervised loss with a self-supervised objective improves the robustness and out-of-distribution detection. Therefore, it'd be interesting to explore to what extent do various self-supervised methods robustify ImageNet models (by way of observing the adversarial validation accuracy at different values of ϵ), and in turn what effect they have on transfer-learning performance of various datasets.

Using incremental learning to learn to improve accuracies. Table 1 shows that removing a portion of the worst performing classes improves the validation accuracy. However, we see a drop in this validation accuracy when we fine-tune these models with data containing more classes. Therefore, in order to preserve the higher accuracy of the original models, we propose to adopt various incremental learning approaches, e.g. work by [23], or leveraging knowledge distillation [11] as an auxiliary loss while learning with a larger number of classes.

Acknowledgements

Research presented in this article was supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project number 20210043DR.

References

[1] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48. Association for Computing Machinery, 2009.

[2] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2019–2026, 2014. doi: 10.1109/CVPR.2014.259.

[3] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101 – mining discriminative components with random forests. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 446–461, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10599-4.

[4] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[5] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. doi: 10.1109/CVPR.2014.461.

[6] L. Fei-Fei, R. Fergus, and P. 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*, pages 178–178, 2004. doi: 10.1109/CVPR.2004.383.

[7] A. Gordo, J. Almazan, J. Revaud, and D. Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 124(2):237–254, 2017.

[8] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[10] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song. Using self-supervised learning can improve model robustness and uncertainty. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[11] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[12] S. Kornblith, J. Shlens, and Q. V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2661–2671, 2019.

[13] J. Krause, J. Deng, M. Stark, and L. Fei-Fei. Collecting a large-scale dataset of fine-grained cars. 2013.

[14] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[16] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018.

[17] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[18] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*, pages 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

[19] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092.

[20] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

[21] H. Salman, A. Ilyas, L. Engstrom, A. Kapoor, and A. Madry. Do adversarially robust imagenet models transfer better? In *Advances in Neural Information Processing Systems*, volume 33, pages 3533–3545, 2020.

[22] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. []cnn] features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.

[23] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, and Y. Fu. Large scale incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 374–382, 2019.

[24] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[25] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.