

# Do Better ImageNet Models Transfer Better?

Simon Kornblith\*, Jonathon Shlens, and Quoc V. Le  
 Google Brain  
 {skornblith, shlens, qvl}@google.com

## Abstract

*Transfer learning is a cornerstone of computer vision, yet little work has been done to evaluate the relationship between architecture and transfer. An implicit hypothesis in modern computer vision research is that models that perform better on ImageNet necessarily perform better on other vision tasks. However, this hypothesis has never been systematically tested. Here, we compare the performance of 16 classification networks on 12 image classification datasets. We find that, when networks are used as fixed feature extractors or fine-tuned, there is a strong correlation between ImageNet accuracy and transfer accuracy ( $r = 0.99$  and  $0.96$ , respectively). In the former setting, we find that this relationship is very sensitive to the way in which networks are trained on ImageNet; many common forms of regularization slightly improve ImageNet accuracy but yield penultimate layer features that are much worse for transfer learning. Additionally, we find that, on two small fine-grained image classification datasets, pretraining on ImageNet provides minimal benefits, indicating the learned features from ImageNet do not transfer well to fine-grained tasks. Together, our results show that ImageNet architectures generalize well across datasets, but ImageNet features are less general than previously suggested.*

## 1. Introduction

The last decade of computer vision research has pursued academic benchmarks as a measure of progress. No benchmark has been as hotly pursued as ImageNet [18, 67]. Network architectures measured against this dataset have fueled much progress in computer vision research across a broad array of problems, including transferring to new datasets [20, 65], object detection [37], image segmentation [31, 7] and perceptual metrics of images [40]. An implicit assumption behind this progress is that network architectures that perform better on ImageNet necessarily perform better on other vision tasks. Another assumption is that bet-

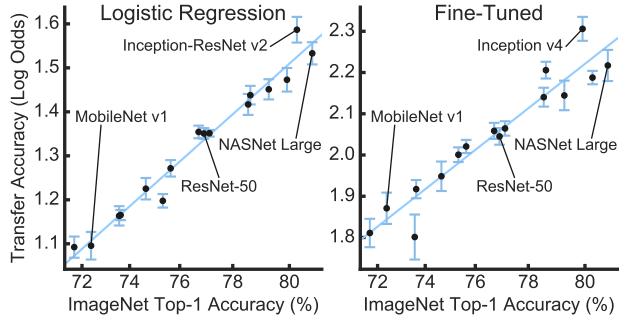


Figure 1. Transfer learning performance is highly correlated with ImageNet top-1 accuracy for fixed ImageNet features (left) and fine-tuning from ImageNet initialization (right). The 16 points in each plot represent transfer accuracy for 16 distinct CNN architectures, averaged across 12 datasets after logit transformation (see Section 3). Error bars measure variation in transfer accuracy across datasets. These plots are replicated in Figure 2 (right).

ter network architectures learn better features that can be transferred across vision-based tasks. Although previous studies have provided some evidence for these hypotheses (e.g. [6, 71, 37, 35, 31]), they have never been systematically explored across network architectures.

In the present work, we seek to test these hypotheses by investigating the transferability of both ImageNet features and ImageNet classification architectures. Specifically, we conduct a large-scale study of transfer learning across 16 modern convolutional neural networks for image classification on 12 image classification datasets in 3 different experimental settings: as fixed feature extractors [20, 65], fine-tuned from ImageNet initialization [1, 28, 6], and trained from random initialization. Our main contributions are as follows:

- Better ImageNet networks provide better penultimate layer features for transfer learning with linear classification ( $r = 0.99$ ), and better performance when the entire network is fine-tuned ( $r = 0.96$ ).
- Regularizers that improve ImageNet performance are highly detrimental to the performance of transfer learning based on penultimate layer features.
- Architectures transfer well across tasks even when

\*Work done as a member of the Google AI Residency program ([g.co/airesidency](http://g.co/airesidency)).

weights do not. On two small fine-grained classification datasets, fine-tuning does not provide a substantial benefit over training from random initialization, but better ImageNet architectures nonetheless obtain higher accuracy.

## 2. Related work

ImageNet follows in a succession of progressively larger and more realistic benchmark datasets for computer vision. Each successive dataset was designed to address perceived issues with the size and content of previous datasets. Torralba and Efros [80] showed that many early datasets were heavily biased, with classifiers trained to recognize or classify objects on those datasets possessing almost no ability to generalize to images from other datasets.

Early work using convolutional neural networks (CNNs) for transfer learning extracted fixed features from ImageNet-trained networks and used these features to train SVMs and logistic regression classifiers for new tasks [20, 65, 6]. These features could outperform hand-engineered features even for tasks very distinct from ImageNet classification [20, 65]. Following this work, several studies compared the performance of AlexNet-like CNNs of varying levels of computational complexity in a transfer learning setting with no fine-tuning. Chatfield et al. [6] found that, out of three networks, the two more computationally expensive networks performed better on PASCAL VOC. Similar work concluded that deeper networks produce higher accuracy across many transfer tasks, but wider networks produce lower accuracy [2]. More recent evaluation efforts have investigated transfer from modern CNNs to medical image datasets [58], and transfer of sentence embeddings to language tasks [12].

A substantial body of existing research indicates that, in image tasks, fine-tuning typically achieves higher accuracy than classification based on fixed features, especially for larger datasets or datasets with a larger domain mismatch from the training set [1, 6, 28, 86, 2, 49, 38, 9, 58]. In object detection, ImageNet-pretrained networks are used as backbone models for Faster R-CNN and R-FCN detection systems [66, 16]. Classifiers with higher ImageNet accuracy achieve higher overall object detection accuracy [37], although variability across network architectures is small compared to variability from other object detection architecture choices. A parallel story likewise appears in image segmentation models [7], although it has not been as systematically explored.

Several authors have investigated how properties of the original training dataset affect transfer accuracy. Work examining the performance of fixed image features drawn from networks trained on subsets of ImageNet have reached conflicting conclusions regarding the importance of the number of classes vs. number of images per class [38, 2]. Yosinski et al. [86] showed that the first layer of AlexNet can be frozen

when transferring between natural and manmade subsets of ImageNet without performance impairment, but freezing later layers produces a substantial drop in accuracy. Other work has investigated transfer from extremely large image datasets to ImageNet, demonstrating that transfer learning can be useful even when the target dataset is large [75, 54]. Finally, a recent work devised a strategy to transfer when labeled data from many different domains is available [88].

## 3. Statistical methods

Much of the analysis in this work requires comparing accuracies across datasets of differing difficulty. When fitting linear models to accuracy values across multiple datasets, we consider effects of model and dataset to be additive. In this context, using untransformed accuracy as a dependent variable is problematic: The meaning of a 1% additive increase in accuracy is different if it is relative to a base accuracy of 50% vs. 99%. Thus, we consider the log odds, i.e., the accuracy after the logit transformation  $\text{logit}(p) = \log(p/(1-p)) = \text{sigmoid}^{-1}(p)$ . The logit transformation is the most commonly used transformation for analysis of proportion data, and an additive change  $\Delta$  in logit-transformed accuracy has a simple interpretation as a multiplicative change  $\exp \Delta$  in the odds of correct classification:

$$\begin{aligned} \text{logit}\left(\frac{n_{\text{correct}}}{n_{\text{correct}} + n_{\text{incorrect}}}\right) + \Delta &= \log\left(\frac{n_{\text{correct}}}{n_{\text{incorrect}}}\right) + \Delta \\ &= \log\left(\frac{n_{\text{correct}}}{n_{\text{incorrect}}} \exp \Delta\right) \end{aligned}$$

We plot all accuracy numbers on logit-scaled axes.

We computed error bars for model accuracy averaged across datasets, using the procedure from Morey [57] to remove variance due to inherent differences in dataset difficulty. Given logit-transformed accuracies  $x_{md}$  of model  $m \in \mathcal{M}$  on dataset  $d \in \mathcal{D}$ , we compute adjusted accuracies  $\text{acc}(m, d) = x_{md} - \sum_{n \in \mathcal{M}} x_{nd}/|\mathcal{M}|$ . For each model, we take the mean and standard error of the adjusted accuracy across datasets, and multiply the latter by a correction factor  $\sqrt{|\mathcal{M}|}/(|\mathcal{M}| - 1)$ .

When examining the strength of the correlation between ImageNet accuracy and accuracy on transfer datasets, we report  $r$  for the correlation between the logit-transformed ImageNet accuracy and the logit-transformed transfer accuracy averaged across datasets. We report the rank correlation (Spearman's  $\rho$ ) in Appendix A.1.2.

We tested for significant differences between pairs of networks on the same dataset using a permutation test or equivalent binomial test of the null hypothesis that the predictions of the two networks are equally likely to be correct, described further in Appendix A.1.1. We tested for significant differences between networks in average performance across datasets using a t-test.

Dataset	Classes	Size (train/test)	Accuracy metric
Food-101 [5]	101	75,750/25,250	top-1
CIFAR-10 [43]	10	50,000/10,000	top-1
CIFAR-100 [43]	100	50,000/10,000	top-1
Birdsnap [4]	500	47,386/2,443	top-1
SUN397 [84]	397	19,850/19,850	top-1
Stanford Cars [41]	196	8,144/8,041	top-1
FGVC Aircraft [55]	100	6,667/3,333	mean per-class
PASCAL VOC 2007 Cls. [22]	20	5,011/4,952	11-point mAP
Describable Textures (DTD) [10]	47	3,760/1,880	top-1
Oxford-IIIT Pets [61]	37	3,680/3,369	mean per-class
Caltech-101 [24]	102	3,060/6,084	mean per-class
Oxford 102 Flowers [59]	102	2,040/6,149	mean per-class

Table 1. Datasets examined in transfer learning

## 4. Results

We examined 16 modern networks ranging in ImageNet (ILSVRC 2012 validation) top-1 accuracy from 71.6% to 80.8%. These networks encompassed widely used Inception architectures [77, 39, 78, 76]; ResNets [33, 30, 29]; DenseNets [36]; MobileNets [35, 68]; and NASNets [92]. For fair comparison, we retrained all models with scale parameters for batch normalization layers and without label smoothing, dropout, or auxiliary heads, rather than relying on pretrained models. Appendix A.3 provides training hyperparameters along with further details of each network, including the ImageNet top-1 accuracy, parameter count, dimension of the penultimate layer, input image size, and performance of retrained models. For all experiments, we rescaled images to the same image size as was used for ImageNet training.

We evaluated models on 12 image classification datasets ranging in training set size from 2,040 to 75,750 images (20 to 5,000 images per class; Table 1). These datasets covered a wide range of image classification tasks, including superordinate-level object classification (CIFAR-10 [43], CIFAR-100 [43], PASCAL VOC 2007 [22], Caltech-101 [24]); fine-grained object classification (Food-101 [5], Birdsnap [4], Stanford Cars [41], FGVC Aircraft [55], Oxford-IIIT Pets [61]); texture classification (DTD [10]); and scene classification (SUN397 [84]).

Figure 2 presents correlations between the top-1 accuracy on ImageNet vs. the performance of the same model architecture on new image tasks. We measure transfer learning performance in three settings: (1) training a logistic regression classifier on the *fixed* feature representation from the penultimate layer of the ImageNet-pretrained network, (2) fine-tuning the ImageNet-pretrained network, and (3) training the same CNN architecture from scratch on the new image task.

### 4.1. ImageNet accuracy predicts performance of logistic regression on fixed features, but regularization settings matter

We first examined the performance of different networks when used as *fixed* feature extractors by training an  $L_2$ -regularized logistic regression classifier on penultimate layer activations using L-BFGS [50] without data augmentation.<sup>1</sup> As shown in Figure 2 (top), ImageNet top-1 accuracy was highly correlated with accuracy on transfer tasks ( $r = 0.99$ ). Inception-ResNet v2 and NASNet Large, the top two models in terms of ImageNet accuracy, were statistically tied for first place.

Critically, results in Figure 2 were obtained with models that were all trained on ImageNet with the same training settings. In experiments conducted with publicly available checkpoints, we were surprised to find that ResNets and DenseNets consistently achieved higher accuracy than other models, and the correlation between ImageNet accuracy and transfer accuracy with fixed features was low and not statistically significant (Appendix B). Further investigation revealed that the poor correlation arose from differences in regularization used for these public checkpoints.

Figure 3 shows the transfer learning performance of Inception models with different training settings. We identify 4 choices made in the Inception training procedure and subsequently adopted by several other models that are detrimental to transfer accuracy: (1) The absence of scale parameter ( $\gamma$ ) for batch normalization layers; the use of (2) label smoothing [78] and (3) dropout [74]; and (4) the presence of an auxiliary classifier head [77]. These settings had a small (< 1%) impact on the overall ImageNet top-1 accuracy of each model (Figure 3, inset). However, in terms of average transfer accuracy, the difference between the default and

<sup>1</sup>We also repeated these experiments with support vector machine classifiers in place of logistic regression, and when using data augmentation for logistic regression; see Appendix G. Findings did not change.

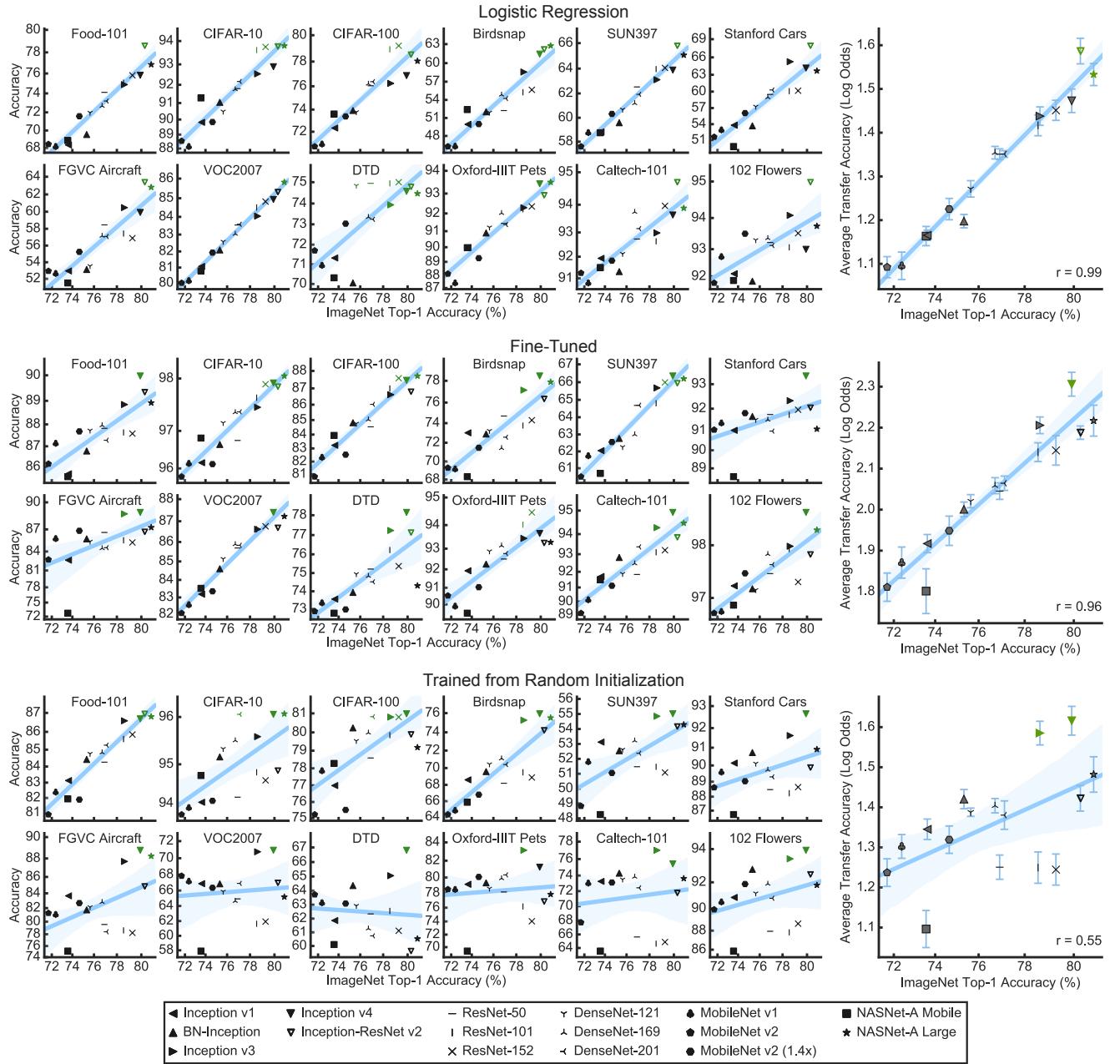


Figure 2. ImageNet accuracy is a strong predictor of transfer accuracy for logistic regression on penultimate layer features and fine-tuning. Each set of panels measures correlations between ImageNet accuracy and transfer accuracy across fixed ImageNet features (top), fine-tuned networks (middle) and networks trained from scratch (bottom). Left: Relationship between classification accuracy on transfer datasets (y-axis) and ImageNet top-1 accuracy (x-axis) in different training settings. Axes are logit-scaled (see text). The regression line and a 95% bootstrap confidence interval are plotted in blue. Right: Average log odds of correct classification across datasets. Error bars are standard error. Points corresponding to models not significantly different from the best model ( $p > 0.05$ ) are colored green.

optimal training settings was approximately equal to the difference between the worst and best ImageNet models trained with optimal settings. This difference was visible not only in transfer accuracy, but also in t-SNE embeddings of the features (Figure 4). Differences in transfer accuracy between settings were apparent earlier in training than differences

in ImageNet accuracy, and were consistent across datasets (Appendix C.1).

Label smoothing and dropout are regularizers in the traditional sense: They are intended to improve generalization accuracy at the expense of training accuracy. Although auxiliary classifier heads were initially proposed to alleviate

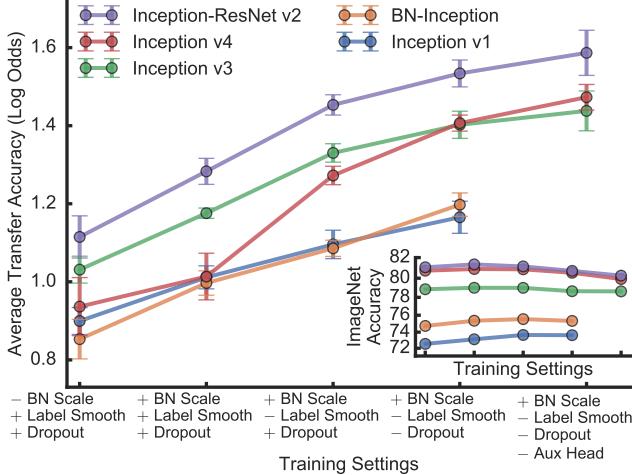


Figure 3. ImageNet training settings have a large effect upon performance of logistic regression classifiers trained on penultimate layer features. In the main plot, each point represents the logit-transformed transfer accuracy averaged across the 12 datasets, measured using logistic regression on penultimate layer features from a specific model trained with the training configuration labeled at the bottom. "+" indicates that a setting was enabled, whereas "-" indicates that a setting was disabled. The leftmost, most heavily regularized configuration is typically used for Inception models [78]; the rightmost is typically used for ResNets and DenseNets. The inset plot shows ImageNet top-1 accuracy for the same training configurations. See also Appendix C.1. Best viewed in color.

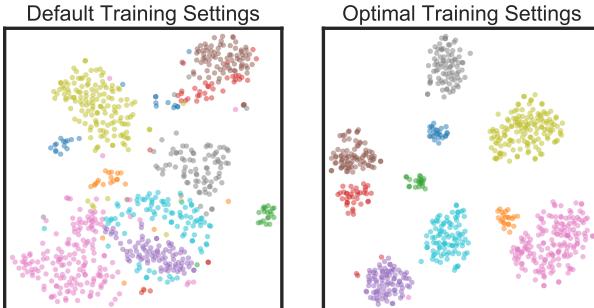


Figure 4. The default Inception training settings produce a suboptimal feature space. Low dimensional embeddings of Oxford 102 Flowers using t-SNE [53] on features from the penultimate layer of Inception v4, for 10 classes from the test set. Best viewed in color.

issues related to vanishing gradients [46, 77], Szegedy et al. [78] instead suggest that they also act as regularizers. The improvement in transfer performance when incorporating batch normalization scale parameters may relate to changes in effective learning rates [81, 90].

## 4.2. ImageNet accuracy predicts fine-tuning performance

We also examined performance when fine-tuning ImageNet networks (Figure 2, middle). We initialized each network from the ImageNet weights and fine-tuned for 20,000

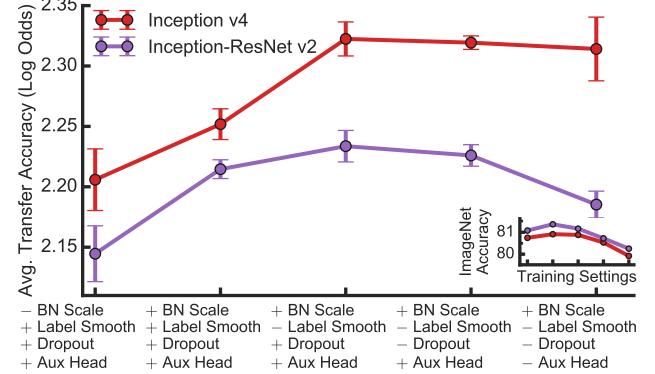


Figure 5. ImageNet training settings have only a minor impact on fine-tuning performance. Each point represents transfer accuracy for a model pretrained and fine-tuned with the same training configuration, labeled at the bottom. Axes follow Figure 3. See Appendix C.2 for performance of models pretrained with regularization but fine-tuned without regularization.

steps with Nesterov momentum and a cosine decay learning rate schedule at a batch size of 256. We performed grid search to select the optimal learning rate and weight decay based on a validation set (for details, see Appendix A.5). Again, we found that ImageNet top-1 accuracy was highly correlated with transfer accuracy ( $r = 0.96$ ).

Compared with the logistic regression setting, regularization and training settings had smaller effects upon the performance of fine-tuned models. Figure 5 shows average transfer accuracy for Inception v4 and Inception-ResNet v2 models with different regularization settings. As in the logistic regression setting, introducing a batch normalization scale parameter and disabling label smoothing improved performance. In contrast to the logistic regression setting, dropout and the auxiliary head sometimes improved performance, but only if used during fine-tuning. We discuss these results further in Appendix C.2.

Overall, fine-tuning yielded better performance than classifiers trained on fixed ImageNet features, but the gain differed by dataset. Fine-tuning improved performance over logistic regression in 179 out of 192 dataset and model combinations (Figure 6; see also Appendix E). When averaged across the tested architectures, fine-tuning yielded significantly better results on all datasets except Caltech-101 (all  $p < 0.01$ , Wilcoxon signed rank test; Figure 6). The improvement was generally larger for larger datasets. However, fine-tuning provided substantial gains on the smallest dataset, 102 Flowers, with 102 classes and 2,040 training examples.

## 4.3. ImageNet accuracy predicts performance of networks trained from random initialization

One confound of the previous results is that it is not clear whether ImageNet accuracy for transfer learning is due to the weights derived from the ImageNet training or the archi-

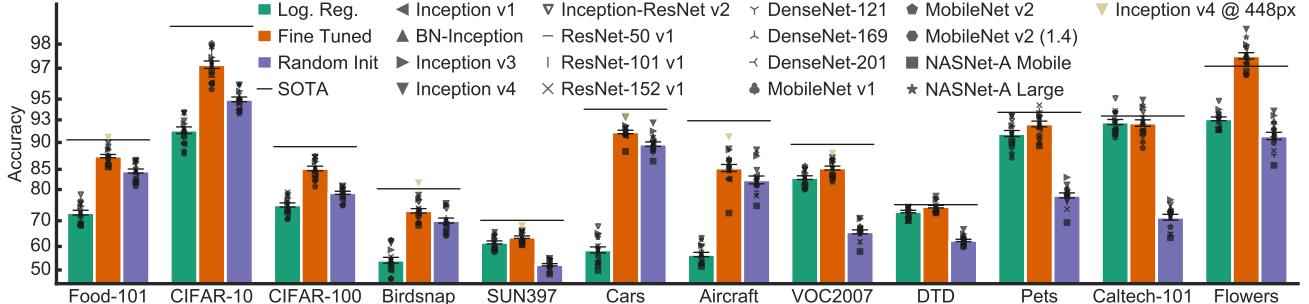


Figure 6. Performance comparison of logistic regression, fine-tuning, and training from random initialization. Bars reflect accuracy across models (excluding VGG) for logistic regression, fine-tuning, and training from random initialization. Error bars are standard error. Points represent individual models. Lines represent previous state-of-the-art. Best viewed in color.

ecture itself. To remove the confound, we next examined architectures trained from random initialization, using a similar training setup as for fine-tuning (see Appendix A.6). In this setting, the correlation between ImageNet top-1 accuracy and accuracy on the new tasks was more variable than in the transfer learning settings, but there was a tendency toward higher performance for models that achieved higher accuracy on ImageNet ( $r = 0.55$ ; Figure 2, bottom).

Examining these results further, we found that a single correlation averages over a large amount of variability. For the 7 datasets with  $<10,000$  examples, the correlation was low and did not reach statistical significance ( $r = 0.29$ ; see also Appendix D). However, for the larger datasets, the correlation between ImageNet top-1 accuracy and transfer learning performance was markedly stronger ( $r = 0.86$ ). Inception v3 and v4 were among the top-performing models across all dataset sizes.

#### 4.4. Benefits of better models are comparable to specialized methods for transfer learning

Given the strong correlation between ImageNet accuracy and transfer accuracy, we next sought to compare simple approaches to transfer learning with better ImageNet models with baselines from the literature. We achieve state-of-the-art performance on half of the 12 datasets if we evaluate using the same image sizes as the baseline methods (Figure 6; see full results in Appendix F). Our results suggest that the ImageNet performance of the pretrained model is a critical factor in transfer performance.

Several papers have proposed methods to make better use of CNN features and thus improve the efficacy of transfer learning [49, 11, 48, 27, 85, 73, 15, 47, 63]. On the datasets we examine, we outperform all such methods simply by fine-tuning state-of-the-art CNNs (Appendix F). Moreover, in some cases a better CNN can make up for dataset deficiencies: By fine-tuning ImageNet-pretrained Inception v4, we outperform the best reported single-model results for networks pretrained on the Places dataset [34, 91], which more closely matches the domain of SUN397.

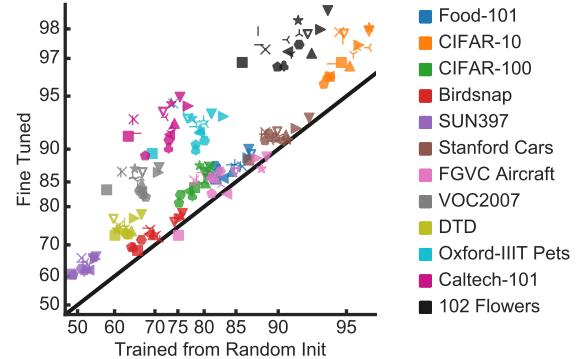


Figure 7. For some datasets and networks, the gap between fine-tuning and training from random initialization is small. Each point represents a dataset/model combination. Axes are logit-scaled. See Figure 6 for network legend and Appendix E for scatter plots of other settings. Best viewed in color.

It is likely that improvements obtained with better models, specialized transfer learning methods, and pretraining datasets with greater domain match are complementary. Combining these approaches could lead to even better performance. Nonetheless, it is surprising that simply using a better model can yield gains comparable to specialized techniques.

#### 4.5. ImageNet pretraining does not necessarily improve accuracy on fine-grained tasks

Fine-tuning was more accurate than training from random initialization for 189 out of 192 dataset/model combinations, but on Stanford Cars and FGVC Aircraft, the improvement was unexpectedly small (Figures 6 and 7). In both settings, Inception v4 was the best model we tested on these datasets. When trained at the default image size of  $299 \times 299$ , it achieved 92.7% on Stanford Cars when trained from scratch vs. 93.3% when fine-tuned, and 88.8% on FGVC Aircraft when trained from scratch vs. 89.0% when fine-tuned.

ImageNet pretraining thus appears to have only marginal accuracy benefits for fine-grained classification tasks where

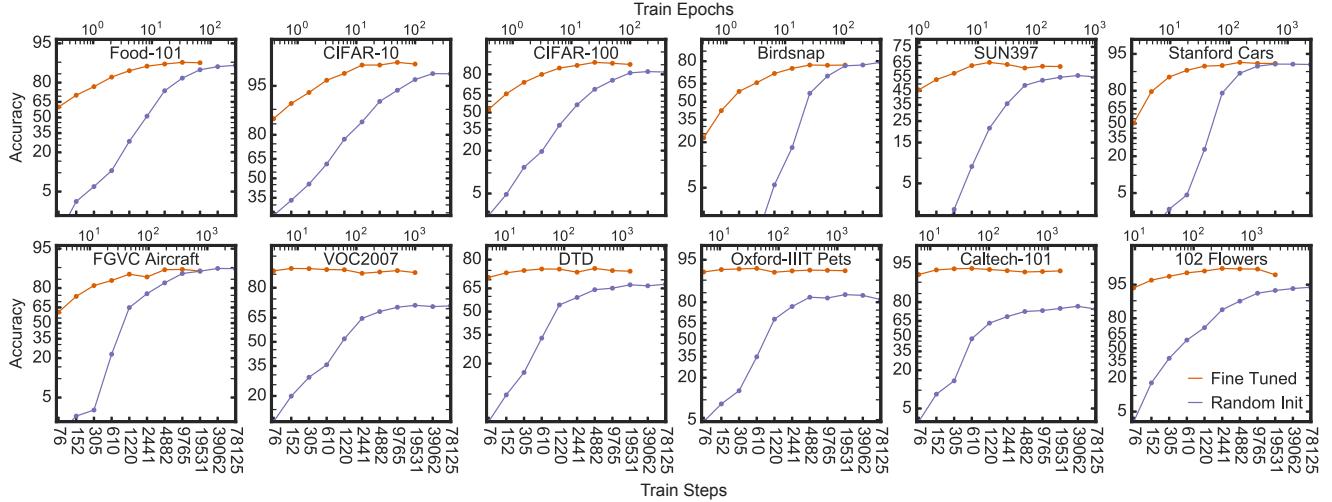


Figure 8. Networks pretrained on ImageNet converge faster, even when final accuracy is the same as training from random initialization. Each point represents an independent Inception v4 model trained with optimized hyperparameters. For fine-tuning, we initialize with the public TensorFlow Inception v4 checkpoint. Axes are logit-scaled.

labels are not well-represented in ImageNet. At 100+ classes and <10,000 examples, Stanford Cars and FGVC Aircraft are much smaller than most datasets used to train CNNs [26]. In fact, the ImageNet training set contains more car images than Stanford Cars (12,455 vs. 8,144). However, ImageNet contains only 10 high-level car classes (e.g., sports car), whereas Stanford Cars contains 196 car classes by make, model, and year. Four other datasets (Oxford 102 Flowers, Oxford-IIIT Pets, Birdsnap, and Food-101) require similarly fine-grained classification, but the classes contained in the latter three datasets are much better-represented in ImageNet. Most of the cat and dog breeds present in Oxford-IIIT Pets correspond directly to ImageNet classes, and ImageNet contains 59 classes of birds and around 45 classes of fruits, vegetables, and prepared dishes.

#### 4.6. ImageNet pretraining accelerates convergence

Given that fine-tuning and training from random initialization achieved similar performance on Stanford Cars and FGVC Aircraft, we next asked whether fine-tuning still posed an advantage in terms of training time. In Figure 8, we examine performance of Inception v4 when fine-tuning or training from random initialization for different numbers of steps. Even when fine-tuning and training from scratch achieved similar final accuracy, we could fine-tune the model to this level of accuracy in an order of magnitude fewer steps. To quantify this acceleration, we computed the number of epochs and steps required to reach 90% of the maximum odds of correct classification achieved at any number of steps, and computed the geometric mean across datasets. Fine-tuning reached this threshold level of accuracy in an average of 26 epochs/1151 steps (inter-quartile ranges 267-4882 steps, 12-58 epochs), whereas training from scratch

required 444 epochs/19531 steps (inter-quartile ranges 9765-39062 steps, 208-873 epochs) corresponding to a 17-fold speedup on average.

#### 4.7. Accuracy benefits of ImageNet pretraining fade quickly with dataset size

Although all datasets benefit substantially from ImageNet pretraining when few examples are available for transfer, for many datasets, these benefits fade quickly when more examples are available. In Figure 9, we show the behavior of logistic regression, fine-tuning, and training from random initialization in the regime of limited data, i.e., for dataset subsets consisting of different numbers of examples per class. When data is sparse (47-800 total examples), logistic regression is a strong baseline, achieving accuracy comparable to or better than fine-tuning. At larger dataset sizes, fine-tuning achieves higher performance than logistic regression, and, for fine-grained classification datasets, the performance of training from random initialization begins to approach results of pre-trained models. On FGVC Aircraft, training from random initialization achieved parity with fine-tuning at only 1600 total examples (16 examples per class).

### 5. Discussion

Has the computer vision community overfit to ImageNet as a dataset? In a broad sense, our results suggest the answer is no: We find that there is a strong correlation between ImageNet top-1 accuracy and transfer accuracy, suggesting that better ImageNet architectures are capable of learning better, transferable representations. But we also find that a number of widely-used regularizers that improve ImageNet performance do not produce better representations. These

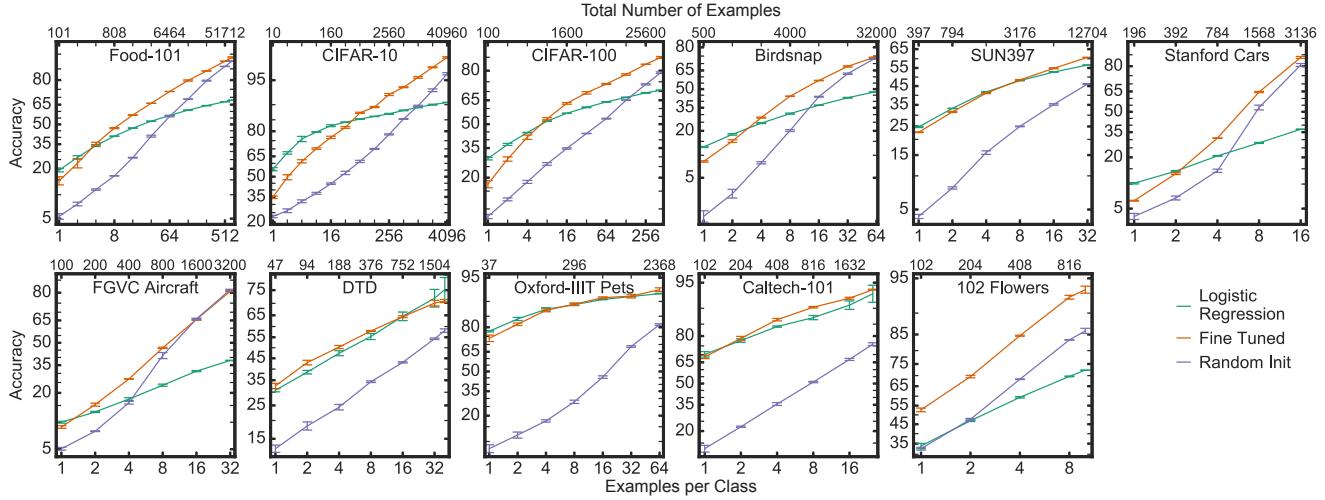


Figure 9. Pretraining on ImageNet improves performance on fine-grained tasks with small amounts of data, but the gap narrows quickly as dataset size increases. Performance of transfer learning with the public Inception v4 model at different dataset sizes. Error bars reflect standard error over 3 subsets. Note that the maximum dataset size shown is not the full dataset. Best viewed in color.

regularizers are harmful to the penultimate layer feature space, and have mixed effects when networks are fine-tuned.

More generally, our results reveal clear limits to transferring features, even among natural image datasets. ImageNet pretraining accelerates convergence and improves performance on many datasets, but its value diminishes with greater training time, more training data, and greater divergence from ImageNet labels. For some fine-grained classification datasets, a few thousand labeled examples, or a few dozen per class, are all that are needed to make training from scratch perform competitively with fine-tuning. Surprisingly, however, the value of architecture persists.

The last decade of computer vision research has demonstrated the superiority of image features learned from data over generic, hand-crafted features. Before the rise of convolutional neural networks, most approaches to image understanding relied on hand-engineered feature descriptors [52, 17, 3]. Krizhevsky et al. [44] showed that, given the training data provided by ImageNet [18], features learned by convolutional neural networks could substantially outperform these hand-engineered features. Soon after, it became clear that intermediate representations learned from ImageNet also provided substantial gains over hand-engineered features when transferred to other tasks [20, 65].

Is the general enterprise of learning widely-useful features doomed to suffer the same fate as feature engineering? Given differences between datasets [80], it is not entirely surprising that features learned on one dataset benefit from some amount of adaptation when applied to another. However, given the history of attempts to build general natural-image feature descriptors, it is surprising that common transfer learning approaches cannot always profitably adapt features learned from a large natural-image to a much smaller natural-

image dataset.

ImageNet weights provide a starting point for features on a new classification task, but perhaps what is needed is a way to learn adaptable features. This problem is closely related to few-shot learning [45, 82, 64, 72, 25, 72, 56], but these methods are typically evaluated with training and test classes from the same distribution. Common few-shot learning methods do not seem to outperform classifiers trained on fixed features when domain shift is present [8], but it may be possible to obtain better results with specialized methods [21] or by combining few-shot learning methods with fine-tuning [69]. It thus remains to be seen whether methods can be developed or repurposed to adapt visual representations learned from ImageNet to provide larger benefits across natural image tasks.

## Acknowledgements

We thank George Dahl, Boyang Deng, Sara Hooker, Pieter-jan Kindermans, Rafael Müller, Jiquan Ngiam, Ruoming Pang, Daiyi Peng, Kevin Swersky, Vishy Tirumalashetty, Vijay Vasudevan, and Emily Xue for comments on the experiments and manuscript, and Aliza Elkin and members of the Google Brain team for support and ideas.

## References

- [1] Pulkit Agrawal, Ross B. Girshick, and Jitendra Malik. Analyzing the performance of multilayer neural networks for object recognition. In *European Conference on Computer Vision (ECCV)*, 2014.
- [2] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson. Factors of transferability for a generic convnet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1790–1802, Sept 2016.

- [3] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.
- [4] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2019–2026. IEEE, 2014.
- [5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 — mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, pages 446–461. Springer, 2014.
- [6] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.
- [7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.
- [8] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019.
- [9] Brian Chu, Vashisht Madhavan, Oscar Beijbom, Judy Hoffman, and Trevor Darrell. Best practices for fine-tuning visual classifiers to new domains. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 435–442, Cham, 2016. Springer International Publishing.
- [10] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613. IEEE, 2014.
- [11] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3828–3836. IEEE, 2015.
- [12] Alexis Conneau and Douwe Kiela. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*, 2018.
- [13] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.
- [14] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [15] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [16] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.
- [17] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE, 2005.
- [18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.
- [19] Birk Diedenhofen and Jochen Musch. cocor: A comprehensive solution for the statistical comparison of correlations. *PLOS ONE*, 10(4):1–12, 04 2015.
- [20] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, pages 647–655, 2014.
- [21] Nanqing Dong and Eric P Xing. Domain adaption in one-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 573–588. Springer, 2018.
- [22] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.
- [23] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9(Aug):1871–1874, 2008.
- [24] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on Generative-Model Based Vision*, 2004.
- [25] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135, 2017.
- [26] Blair Hanley Frank. Google Brain chief: Deep learning takes at least 100,000 examples. In *VentureBeat*. <https://venturebeat.com/2017/10/23/google-brain-chief-says-100000-examples-is-enough-data-for-deep-learning/>, 2017.
- [27] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 317–326, 2016.
- [28] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.
- [29] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large mini-batch SGD: training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.
- [30] Sam Gross and Michael Wilber. Training and investigating residual nets. In *The Torch Blog*. <http://torch.ch/blog/2016/02/04/resnets.html>, 2016.

- [31] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE, 2017.
- [32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision (ECCV)*, pages 346–361. Springer, 2014.
- [33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [34] Luis Herranz, Shuqiang Jiang, and Xiangyang Li. Scene recognition with CNNs: objects, scales and dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 571–579, 2016.
- [35] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [36] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.
- [37] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [38] Mi-Young Huh, Pulkit Agrawal, and Alexei A. Efros. What makes ImageNet good for transfer learning? *CoRR*, abs/1608.08614, 2016.
- [39] Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.
- [40] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016.
- [41] Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. In *Second Workshop on Fine-Grained Visual Categorization*, 2013.
- [42] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 301–320, Cham, 2016. Springer International Publishing.
- [43] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- [44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.
- [45] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.
- [46] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570, 2015.
- [47] Zhichao Li, Yi Yang, Xiao Liu, Feng Zhou, Shilei Wen, and Wei Xu. Dynamic computational time for visual attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1199–1209, 2017.
- [48] Tsung-Yu Lin and Subhransu Maji. Visualizing and understanding deep texture representations. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2791–2799, 2016.
- [49] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1449–1457, 2015.
- [50] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.
- [51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.
- [52] David G Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157. Ieee, 1999.
- [53] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.
- [54] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *European Conference on Computer Vision (ECCV)*, pages 181–196, 2018.
- [55] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.
- [56] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *International Conference on Learning Representations*, 2018.
- [57] Richard D. Morey. Confidence intervals from normalized data: A correction to cousins (2005). *Tutorials in Quantitative Methods for Psychology*, 4(2):61–64, 2008.
- [58] 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*, pages 2262–2271, 2018.
- [59] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on*, pages 722–729. IEEE, 2008.
- [60] Augustus Odena, Avital Oliver, Colin Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of semi-supervised learning algorithms. In *ICLR Workshops*, 2018.

- [61] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3498–3505. IEEE, 2012.
- [62] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.
- [63] Yuxin Peng, Xiangteng He, and Junjie Zhao. Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing*, 27(3):1487–1500, 2018.
- [64] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Machine Learning*, 2016.
- [65] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 512–519. IEEE, 2014.
- [66] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [67] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec 2015.
- [68] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.
- [69] Tyler Scott, Karl Ridgeway, and Michael C Mozer. Adapted deep embeddings: A synthesis of methods for k-shot inductive transfer learning. In *Advances in Neural Information Processing Systems*, pages 76–85, 2018.
- [70] N. Clayton Silver, James B. Hittner, and Kim May. Testing dependent correlations with nonoverlapping variables: A monte carlo simulation. *Journal of Experimental Education*, 73(1):53–69, 2004.
- [71] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.
- [72] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4080–4090, 2017.
- [73] Yang Song, Fan Zhang, Qing Li, Heng Huang, Lauren J O’Donnell, and Weidong Cai. Locally-transferred fisher vectors for texture classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4912–4920, 2017.
- [74] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [75] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 843–852. IEEE, 2017.
- [76] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017.
- [77] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [78] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.
- [79] The TensorFlow Authors. `inception_preprocessing.py`.
- [80] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1521–1528. IEEE, 2011.
- [81] Twan van Laarhoven. L2 regularization versus batch and weight normalization. *CorR*, abs/1706.05350, 2017.
- [82] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016.
- [83] Xiu-Shen Wei, Chen-Wei Xie, Jianxin Wu, and Chunhua Shen. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognition*, 76:704 – 714, 2018.
- [84] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492. IEEE, 2010.
- [85] Hantao Yao, Shiliang Zhang, Yongdong Zhang, Jintao Li, and Qi Tian. Coarse-to-fine description for fine-grained visual categorization. *IEEE Transactions on Image Processing*, 25(10):4858–4872, 2016.
- [86] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.
- [87] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2403–2412, 2018.
- [88] Amir R Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3712–3722, 2018.
- [89] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833. Springer, 2014.

- [90] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. In *International Conference on Learning Representations*, 2019.
- [91] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [92] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8697–8710, 2018.

# Supplementary Material for ‘‘Do Better ImageNet Models Transfer Better?’’

Simon Kornblith, Jonathon Shlens, and Quoc V. Le  
Google Brain  
`{skornblith, shlens, qvl}@google.com`

## A. Supplementary experimental procedures

### A.1. Statistical methods

#### A.1.1 Comparison of two models on the same dataset

To test for superiority of one model over another on a given dataset, we constructed permutations where, for each example, we randomly exchanged the predictions of the two networks. For each permutation, we computed the difference in accuracy between the two networks. (For VOC2007, we considered the accuracy of predictions across labels.) We computed a p-value as the proportion of permutations where the difference is at least as extreme as the observed difference in accuracy. For top-1 accuracy, this procedure is equivalent to a binomial test sometimes called the “exact McNemar test,” and a p-value can be computed exactly. For mean per-class accuracy, we approximated a p-value based on 10,000 permutations. These tests assess whether one trained model performs better than another on data drawn from the test set distribution. However, they are tests between trained models, rather than tests between architectures, since we do not measure variability arising from training networks from different random initializations or from different orderings of the training data.

#### A.1.2 Measures of correlation

Setting	$r^2$	$r$	$\rho$	p-value
Logistic regression	0.97	0.99	0.99	$< 10^{-11}$
Fine-tuned	0.91	0.96	0.97	$< 10^{-8}$
Trained from scratch	0.30	0.55	0.59	0.03
Logistic regression (public checkpoints)	0.14	0.37	0.48	0.16

Table A.1. Correlations between ImageNet accuracy and average transfer accuracy (Pearson  $r$  and  $r^2$  and Spearman’s  $\rho$ ), as well as p-values for the null hypothesis that  $r = 0$ .

Table A.1 shows the Pearson correlation (as  $r^2$  and  $r$ ) as well as the Spearman rank correlation ( $\rho$ ) in each of the three transfer settings we examine. We believe that Pearson correlation is the more appropriate measure, given that it is less dependent on the specific CNNs chosen for the study and the effects are approximately linear, but our results are similar in either case.

### A.2. Datasets

All datasets had a median image size on the shortest side of at least 331 pixels (the highest ImageNet-native input image size out of all networks tested), except Caltech-101, for which the median size is 225 on the shortest side and 300 on the longer side, and CIFAR-10 and CIFAR-100, which consist of  $32 \times 32$  pixel images.

For datasets with a provided validation set (FGVC Aircraft, VOC2007, DTD, and 102 Flowers), we used this validation set to select hyperparameters. For other datasets, we constructed a validation set by subsetting the original training set. For the DTD and SUN397 datasets, which provide multiple train/test splits, we used only the first provided split. For the Caltech-101 dataset, which specifies no train/test split, we trained on 30 images per class and tested on the remainder, as in previous works [20, 71, 89, 6]. With the exception of dataset subset results (Figure 9), all results indicate the performance of models retrained on the combined training and validation set.

### A.3. Networks and ImageNet training procedure

Table A.2 lists the parameter count, penultimate layer feature dimension, and input image size for each network examined. Unless otherwise stated, our results were obtained with networks we trained, rather than publicly available checkpoints. We

Model	Parameters <sup>a</sup>	Features	Image Size	Paper	ImageNet Top-1 Accuracy	
					Public Checkpoint <sup>b</sup>	Retrained
Inception v1 <sup>c</sup> [77]	5.6M	1024	224	73.2	69.8	73.6
BN-Inception <sup>d</sup> [39]	10.2M	1024	224	74.8	74.0	75.3
Inception v3 [78]	21.8M	2048	299	78.8	78.0	78.6
Inception v4 [76]	41.1M	1536	299	80.0	80.2	79.9
Inception-ResNet v2 [76]	54.3M	1536	299	80.1	80.4	80.3
ResNet-50 v1 <sup>e</sup> [33, 30, 29]	23.5M	2048	224	76.4	75.2	76.9
ResNet-101 v1 [33, 30, 29]	42.5M	2048	224	77.9	76.4	78.6
ResNet-152 v1 [33, 30, 29]	58.1M	2048	224	N/A	76.8	79.3
DenseNet-121 [36]	7.0M	1024	224	75.0	74.8	75.6
DenseNet-169 [36]	12.5M	1664	224	76.2	76.2	76.7
DenseNet-201 [36]	18.1M	1920	224	77.4	77.3	77.1
MobileNet v1 [35]	3.2M	1024	224	70.6	70.7	72.4
MobileNet v2 [68]	2.2M	1280	224	72.0	71.8	71.6
MobileNet v2 (1.4) [68]	4.3M	1792	224	74.7	75.0	74.7
NASNet-A Mobile [92]	4.2M	1056	224	74.0	74.0	73.6
NASNet-A Large [92]	84.7M	4032	331	82.7	82.7	80.8

<sup>a</sup>Excludes logits layer.

<sup>b</sup>Performance of checkpoint from TF-Slim repository (<https://github.com/tensorflow/models/tree/master/research/slim>), or, for DenseNets, from Keras applications (<https://keras.io/applications/>).

<sup>c</sup>We used Inception model code from the TF-Slim repository, which uses batch normalization layers for Inception v1. Additionally, the models in this repository contain minor modifications compared to the models described in the original papers. We cite the performance number for BN-GoogLeNet from Szegedy et al. [78].

<sup>d</sup>This model is called "Inception v2" in TF-Slim model repository, but matches the model described in Ioffe and Szegedy [39], rather than the model that Szegedy et al. [78] call "Inception v2."

<sup>e</sup>The ResNets we train incorporate two common modifications to the original ResNet v1 model: Stride-2 downsampling on the  $3 \times 3$  convolution instead of the first  $1 \times 1$  convolution in the block [30, 29] and initialization of the batch normalization  $\gamma$  to 0 in the last batch normalization layer of each block [29]. We report the numbers from Goyal et al. [29] as the original accuracy. No public TensorFlow checkpoints are available for these models, so, for public checkpoint results, we use the TF-Slim ResNet v1 checkpoints, which were converted from the original He et al. [33] model.

Table A.2. ImageNet classification networks

trained all networks with a batch size of 4096 using Nesterov momentum of 0.9 and weight decay of  $8 \times 10^{-5}$ , taking an exponential moving average of the weights with a decay factor of 0.9999. We performed linear warmup to a learning rate of 1.6 over the first 10 epochs, and then continuously decayed the learning rate by a factor of 0.975 per epoch. We used the preprocessing and data augmentation from [79]. To determine how long to train each network, we trained a separate model for up to 300 epochs with approximately 50,000 ImageNet training images held out as a validation set, and then trained a model on the full ImageNet training set for the number of steps that yielded the highest performance. Except in experiments explicitly studying the effects of these choices, for all networks, we used scale parameters for batch normalization layers, and did not use label smoothing, dropout, or an auxiliary head. For NASNet-A Large, we additionally disabled drop path regularization.

When training on ImageNet, we did not optimize hyperparameters for each network individually because we were able to achieve ImageNet top-1 performance comparable to publicly available checkpoints without doing so. (When fine-tuning and training from random initialization, we found that hyperparameters were more important and performed extensive tuning; see below.) For all networks except NASNet-A Large, our retrained models achieved accuracy no more than 0.5% lower than the original reported results and public checkpoint, and sometimes substantially higher (Table A.2). Given that we disabled the regularizers used in the original model, we expected a larger performance drop. Our experiments indicate that these regularizers further improve accuracy, but are evidently not necessary to achieve performance close to the published results.

For NASNet-A Large, there was a substantial gap between the performance of the published model and our retrained model (82.7% vs. 80.8%). As a sanity check, we enabled label smoothing, dropout, the auxiliary head, and drop path, and retrained NASNet-A Large with the same hyperparameters described above. This regularized model achieved 82.5% accuracy, suggesting that most of the loss in accuracy in our setup is due to disabling regularization. For other models, we could further improve ImageNet top-1 accuracy over published results by applying regularizers: A retrained Inception-ResNet v2 model with label smoothing, dropout, and the auxiliary head enabled achieved 81.4% top-1 accuracy, 1.1% better than the unregularized model and 1.3% better than the published result [76]. However, because these regularizers clearly hurt results in the logistic regression setting, and because our goal was to compare all models and settings fairly, we report results for models trained and fine-tuned without regularization unless otherwise specified.

#### A.4. Logistic regression

For each dataset, we extracted features from the penultimate layer of the network. We trained a multinomial logistic regression classifier using L-BFGS, with an L2 regularization parameter applied to the sum of the per-example losses, selected from a range of 45 logarithmically spaced values from  $10^{-6}$  to  $10^5$  on the validation set. Since the optimization problem is convex, we used the solution at the previous point along the regularization path as a warm start for the next point, which greatly accelerated the search. For these experiments, we did not perform data augmentation or scale aggregation, and we used the entire image, rather than cropping the central 87.5% as is common for testing on ImageNet.

#### A.5. Fine-tuning

For fine-tuning experiments in Figure 2, we initialized networks with ImageNet-pretrained weights and trained for 20,000 steps at a batch size of 256 using Nesterov momentum with a momentum parameter of 0.9. We selected the optimal learning rate and weight decay on the validation set by grid search. Our early experiments indicated that the optimal weight decay at a given learning rate varied inversely with the learning rate, as has been recently reported [51]. Thus, our grid consisted of 7 logarithmically spaced learning rates between 0.0001 and 0.1 and 7 logarithmically spaced weight decay to learning rate ratios between  $10^{-6}$  and  $10^{-3}$ , as well as no weight decay. We found it useful to decrease the batch normalization momentum parameter from its ImageNet value to  $\max(1 - 10/s, 0.9)$  where  $s$  is the number of steps per epoch. We found that the maximum performance on the validation set at any step during training was very similar to the maximum performance at the last step, presumably because we searched over learning rate, so we did not perform early stopping. On the validation set, we evaluated on both uncropped images and images cropped to the central 87.5% and picked the approach that gave higher accuracy for evaluation on the test set. Cropped images typically yielded better performance, except on CIFAR-10 and CIFAR-100, where results differed by model.

When examining the effect of dataset size (Section 4.7), we fine-tuned for at least 1000 steps or 100 epochs (following guidance from our analysis of training time in Section 4.6) at a batch size of 64, with the learning rate range scaled down by a factor of 4. Otherwise, we used the same settings as above. Because we chose hyperparameters based on a large validation set, the results may not reflect what can be accomplished in practice when training on datasets of this size [60]. In Sections 4.7 and 4.6, we fine-tuned models from the publicly available Inception v4 checkpoint rather than using the model trained as above.

#### A.6. Training from random initialization

We used a similar training protocol for training from random initialization as for fine-tuning, i.e., we trained for 20,000 steps at a batch size of 256 using Nesterov momentum with a momentum parameter of 0.9. Training from random initialization generally achieved optimal performance at higher learning rates and with greater weight decay, so we adjusted the learning rate range to span from 0.001 to 1.0 and the weight decay to learning rate ratio range to span from  $10^{-5}$  to  $10^{-2}$ .

When examining the effect of dataset size (Section 4.7), we trained from random initialization for at least 78,125 steps or 200 epochs at a batch size of 16, with the learning rate range scaled down by a factor of 16. We chose these parameters because investigation of effects of training time (Section 4.6) indicated that training from random initialization always benefited from increased training time, whereas fine-tuning did not. Additionally, pilot experiments indicated that training from random initialization, but not fine-tuning, benefited from a reduced batch size with very small datasets.

## B. Logistic regression performance of public checkpoints

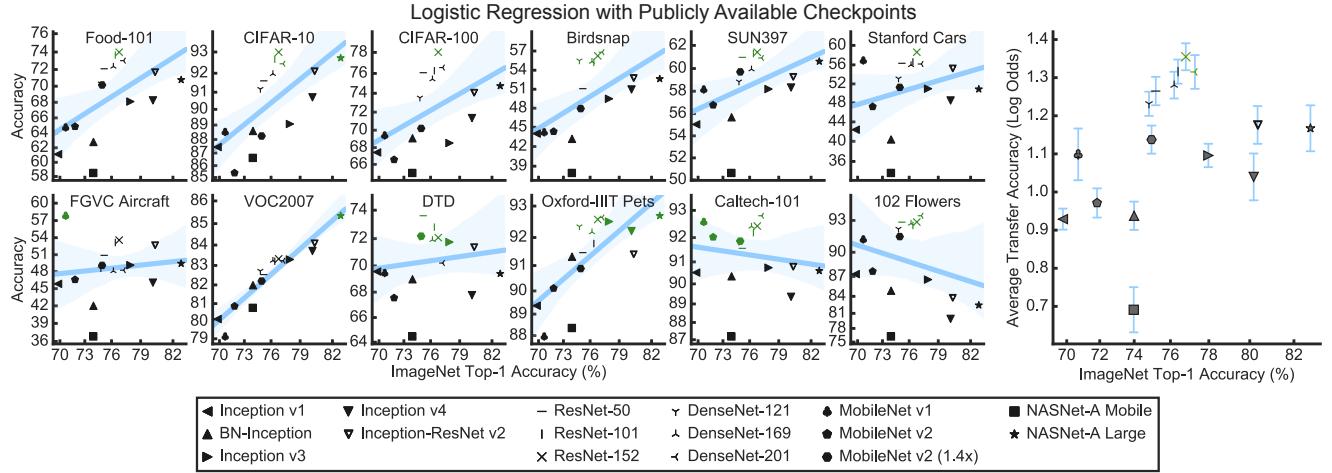


Figure B.1. Accuracy of logistic regression classifiers on fixed features from publicly available checkpoints, rather than retrained models. See also Figure 2.

We present results of logistic regression with features extracted from publicly available checkpoints in Figure B.1. With these checkpoints, ResNets and DenseNets were consistently among the top performing models. The correlation between ImageNet top-1 accuracy and accuracy across transfer tasks was weak and did not reach statistical significance ( $r = 0.37$ ,  $p = 0.16$ ). By contrast, the correlation with between ImageNet top-1 accuracy and accuracy across transfer tasks with retrained models ( $r = 0.99$ ) was much higher ( $p < 10^{-4}$ ,  $z = 5.2$ , test of equality of nonoverlapping correlations based on dependent groups [70, 19]).

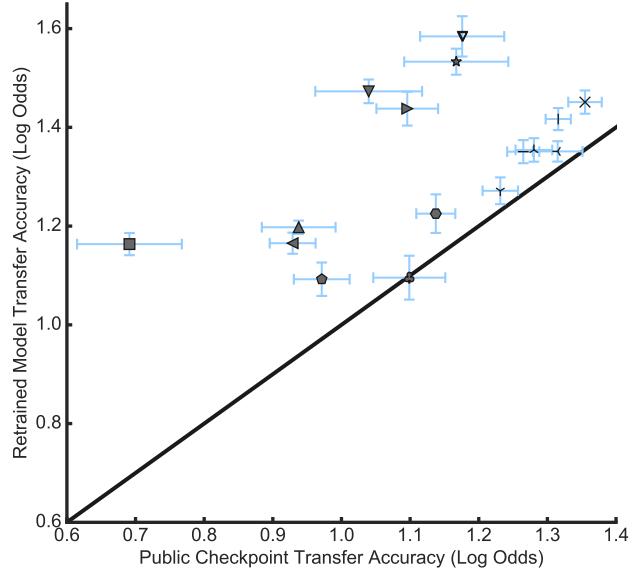


Figure B.2. Our retrained models consistently outperform public checkpoints for logistic regression. See Figure B.1 for legend.

Retrained models achieved higher transfer accuracy than publicly available checkpoints. For 11 of the 12 datasets investigated (all but Oxford Pets), features from the best retrained model achieved higher accuracy than features from the best publicly available checkpoint. Retrained models achieved higher accuracy for 84% of dataset/model pairs (162/192), and transfer accuracy averaged across datasets was higher for retrained models for all networks except MobileNet v1 (Figure B.2). The best retrained model, Inception-ResNet v2, achieved an average log odds of 1.58, whereas the best public checkpoint, ResNet v1 152, achieved an average log odds of 1.35 ( $t(11) = 5.6$ ,  $p = 0.0002$ ).

## C. Extended analysis of effect of training/regularization settings

### C.1. Performance of penultimate layer features

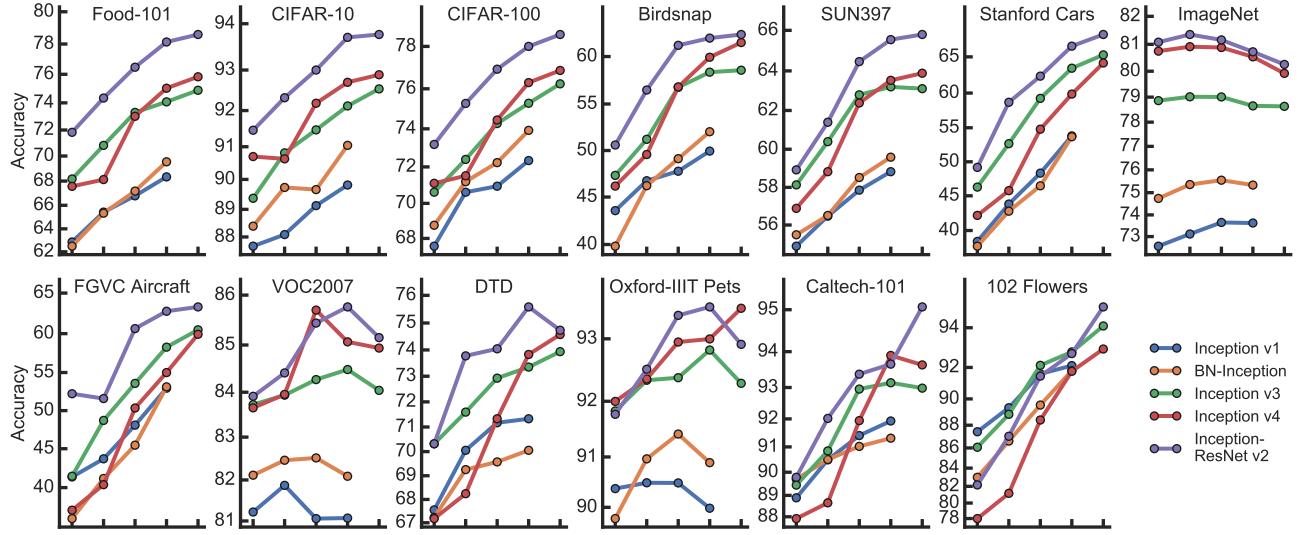


Figure C.1. When performing logistic regression on penultimate layer features, all datasets and models benefit from removal of regularization. Each subplot represents transfer performance on one of the datasets investigated. The top right plot shows ImageNet top-1 accuracy of the models. Points along the x-axis represent different training settings (presence/absence of batch normalization scale parameter, label smoothing, dropout, and presence/absence of auxiliary head), following the same convention as in Figure 3. The leftmost setting is the Inception default, and uses no batch normalization scale parameter, but includes label smoothing, dropout, and an auxiliary head. From left to right, we enable the batch normalization scale parameter; disable label smoothing; disable dropout; and disable the auxiliary head.

Figure C.1 shows performance of penultimate layer features in each of the training settings in Figure 3, broken down by dataset. Across nearly all datasets and models, the least-regularized models achieved the highest performance, even though these models were not the best in terms of ImageNet top-1 accuracy. We also experimented with removing weight decay, but found that this yielded substantially lower performance on both ImageNet and all transfer datasets except for FGVC Aircraft.

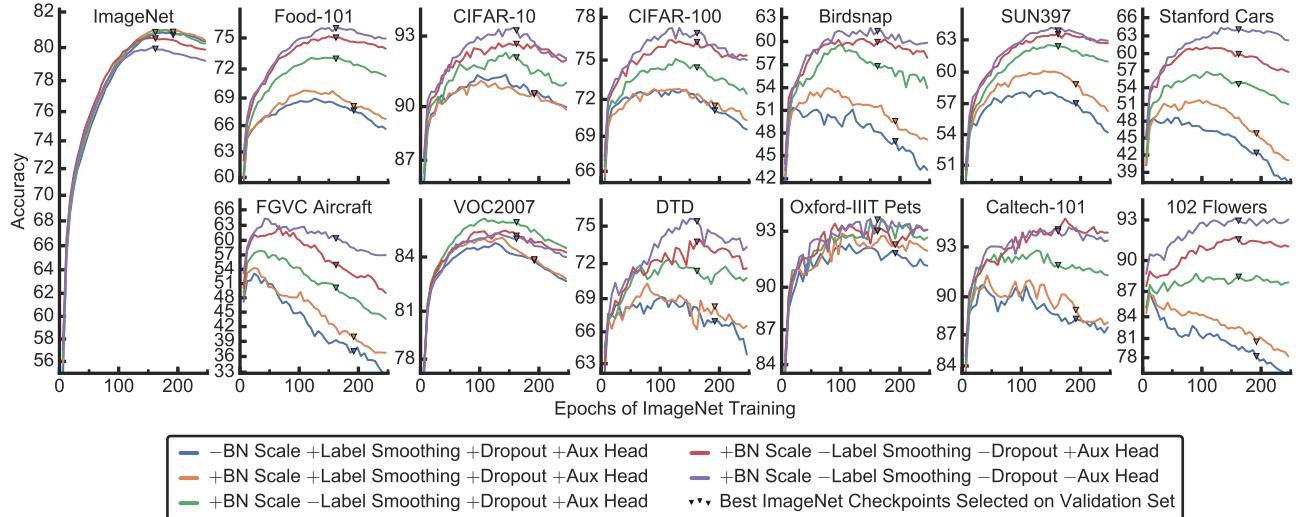


Figure C.2. Regularization affects transfer learning with fixed features earlier in training than ImageNet top-1 accuracy. Transfer learning performance of Inception v4 checkpoints evaluated every 12 epochs. Triangles represent the checkpoint that optimized performance on a validation set split from the training set, in a separate training run.

To investigate whether the effect of regularization upon the performance of fixed features was mediated by training time,

rather than the regularization itself, we performed logistic regression on penultimate layer features from Inception v4 at different epochs over training (Figure C.2). For models with more regularizers, checkpoints from earlier in training typically performed better than the checkpoint that achieved the best accuracy on ImageNet. However, on most datasets, the best checkpoint without regularization outperformed all checkpoints with regularization. For most datasets, the best transfer accuracy was achieved at around the same number of training epochs as the best ImageNet top-1 accuracy, but on FGVC Aircraft, we found that a checkpoint early in training yielded much higher accuracy.

## C.2. Fine-tuning performance

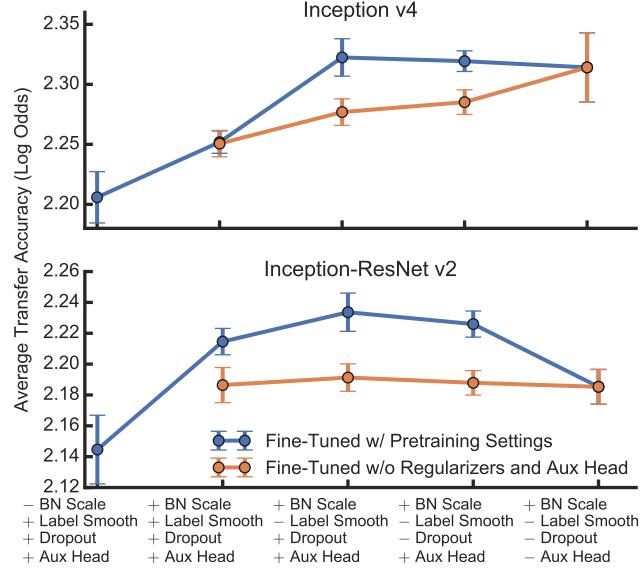


Figure C.3. Using regularizers at ImageNet pretraining time does not benefit fine-tuning performance unless the same regularizers are used to fine-tune. Blue points represent models pretrained and fine-tuned with the same training configuration, as in Figure 5. Orange points represent models pretrained with different configurations but fine-tuned without regularization or an auxiliary head (the rightmost configuration in the plot). Accuracy is averaged across 3 fine-tuning runs each for 2 rounds of hyperparameter tuning.

In this section, we present an expanded analysis of the effect of regularization upon fine-tuning analysis. Figure C.3 shows average fine-tuning both performance across datasets when fine-tuning with the same settings as used for pretraining (blue, same data as in Figure 5), and when pretraining with regularization but fine-tuning without any regularization (orange). For all regularization settings, benefits are only clearly observed when the regularization is used for both pretraining and fine-tuning. Figure C.4 shows results broken down by dataset for pretraining and fine-tuning with the same settings.

Overall, the effect of regularization upon fine-tuning performance was much smaller than the effect upon the performance of logistic regression on penultimate layer features. As in the logistic regression setting, enabling batch normalization scale parameters and disabling label smoothing improved performance. Effects of dropout and the auxiliary head were not entirely consistent across models and datasets (Figure C.4). Inception-ResNet v2 clearly performed better when the auxiliary head was present. For Inception v4, the auxiliary head improved performance on some datasets (Food-101, FGVC Aircraft, VOC2007, and Oxford Pets) but worsened performance on others (CIFAR-100, DTD, Oxford 102 Flowers). However, because improvements were only observed when the auxiliary head was used both for pretraining and fine-tuning, it is unclear whether the auxiliary head leads to better weights or representations. It may instead improve fine-tuning performance by acting as a regularizer at fine-tuning time.

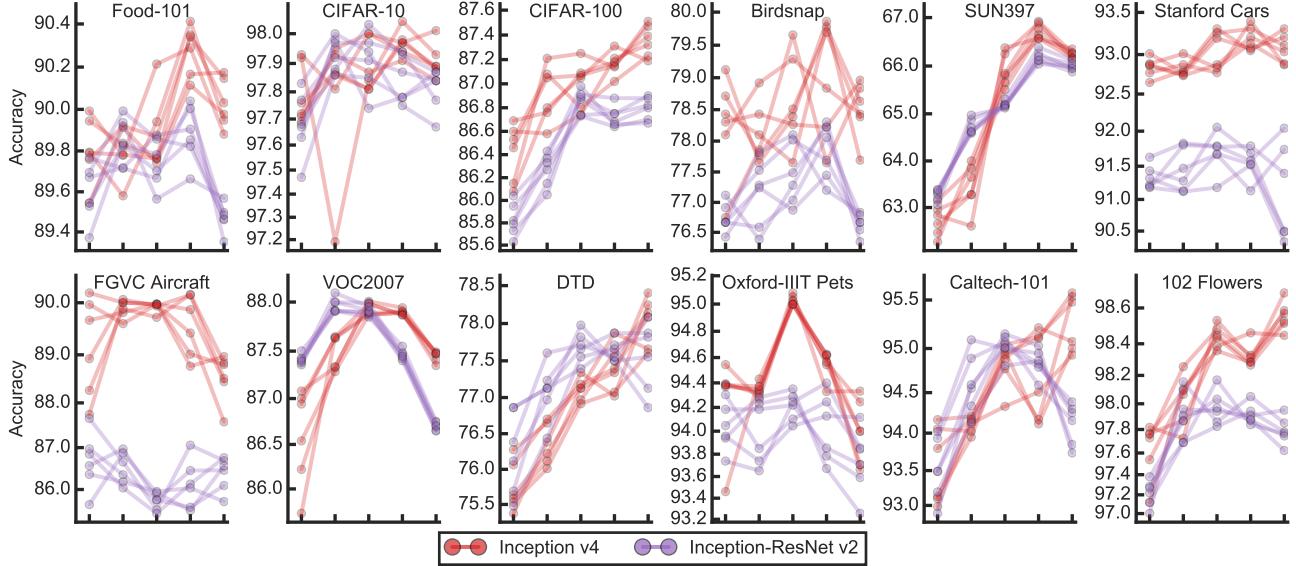


Figure C.4. For fine-tuning, different training settings are best on different datasets. Points along the x-axis represent different training settings (presence/absence of batch normalization scale parameter, label smoothing, dropout, and presence/absence of auxiliary head), following the same convention as in Figure C.3. The leftmost setting is the Inception default, and uses no batch normalization scale parameter, but includes label smoothing, dropout, and an auxiliary head. From left to right, we enable the batch normalization scale parameter; disable label smoothing; disable dropout; and disable the auxiliary head. Each line shows performance for a different fine-tuning run.

## D. Relationship between dataset size and predictive power of ImageNet accuracy

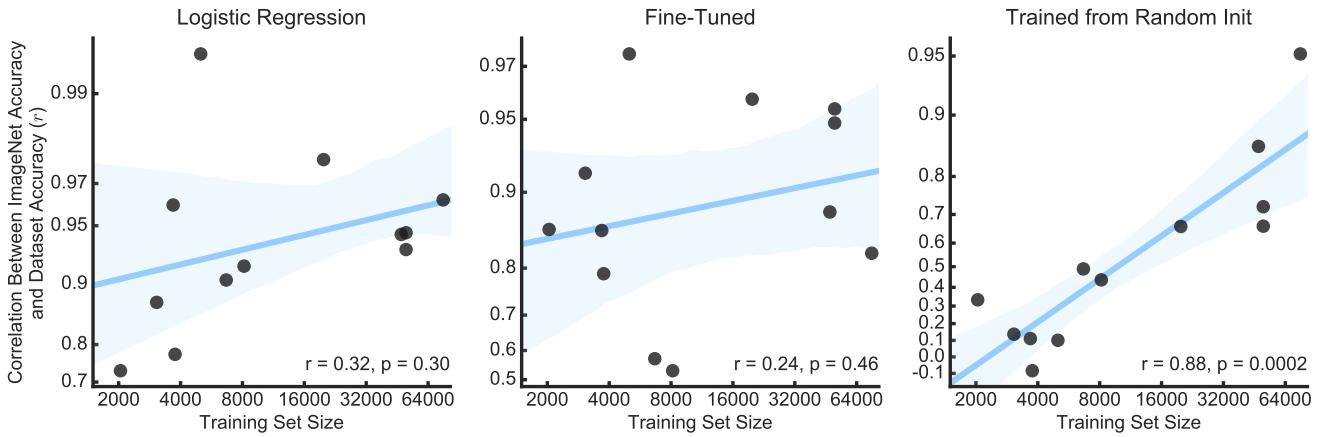


Figure D.1. When training from random initialization, the correlation between ImageNet top-1 accuracy and transfer accuracy is higher for larger datasets. Each point represents one of the 12 datasets investigated. The y-axis represents the Pearson correlation between ImageNet top-1 accuracy and accuracy for that dataset, based on the 16 ImageNet networks investigated, and is scaled according to the Fisher z-transformation. The x-axis is log-scaled.  $r$  and  $p$  values in bottom left reflect the correlation between the log-transformed training set size and Fisher z-transformed correlations.

As datasets get larger, ImageNet accuracy becomes a better predictor of the performance of models trained from scratch. Figure D.1 shows the relationship between the dataset size and the correlation between ImageNet accuracy and accuracy on other datasets, for each of the 12 datasets investigated. We found that there was a significant relationship when networks were trained from random initialization ( $p = 0.0002$ ), but there were no significant relationships in the transfer learning settings.

One possible explanation for this behavior is that ImageNet performance measures both inductive bias and capacity. When training from scratch on smaller datasets, inductive bias may be more important than capacity.

## E. Additional comparisons of logistic regression, fine-tuning, and training from random initialization

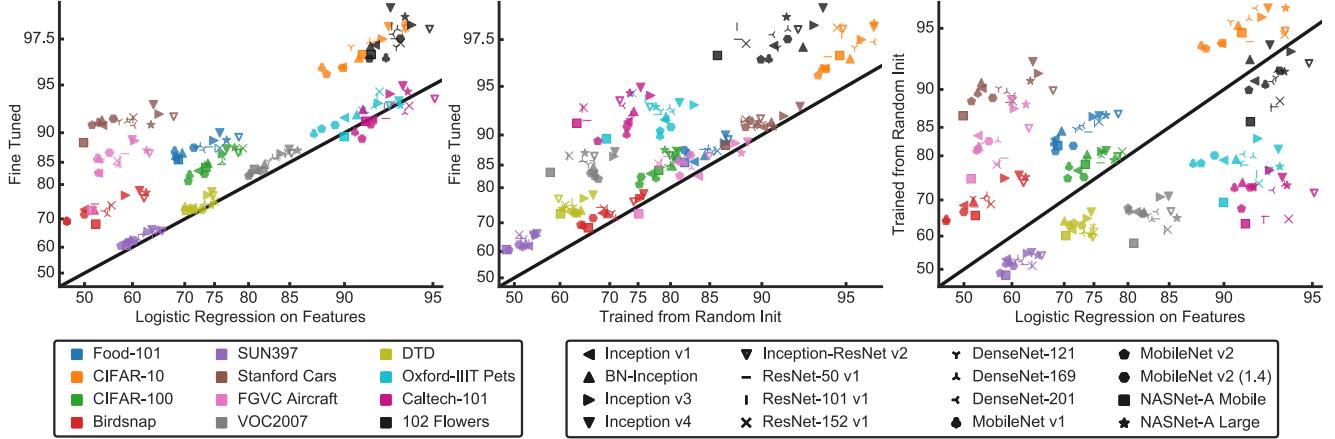


Figure E.1. Scatter plots comparing trained models in each pair of settings investigated. Axes are logit-scaled.

Figure E.1 presents additional scatter plots comparing performance in the three settings we investigated. Fine-tuning usually achieved higher accuracy than logistic regression on top of fixed ImageNet features or training from randomly initialized models, but for some datasets, the gap was small. The performance of logistic regression on fixed ImageNet features vs. networks trained from random initialization was heavily dependent on the dataset.

## F. Comparison versus state-of-the-art

Dataset	Acc.	Previously reported	Acc.	Current work
		Method		Best network
Food-101	90.4	Domain-specific transfer @ 448 [14]	90.0 ( <b>90.8<sup>a</sup></b> )	Inception v4, fine-tuned
CIFAR-10	98.5	AutoAugment [13]	98.0 <sup>b</sup>	NASNet-A Large, fine-tuned
CIFAR-100	89.3	AutoAugment [13]	87.5 <sup>b</sup>	NASNet-A Large, fine-tuned
Birdsnap	80.2 <sup>c</sup>	Mask-CNN @ 448 [83]	78.4 ( <b>81.8<sup>a</sup></b> )	Inception v4, fine-tuned
SUN397	<b>70.2</b>	Places/ImageNet-pretrained multi-scale VGG ensemble [34]	66.4 (68.3 <sup>a</sup> )	Inception v4, fine-tuned
Stanford Cars	<b>94.8</b>	AutoAugment @ 448 [13]	93.3 (93.4 <sup>a</sup> )	Inception v4, fine-tuned
FGVC Aircraft	<b>92.9<sup>c</sup></b>	Deep layer aggregation @ 448 [87]	89.0 (90.9 <sup>a</sup> )	Inception v4, fine-tuned
VOC 2007 Cls.	<b>89.7</b>	VGG multi-scale ensemble [71]	87.4 (88.2 <sup>a</sup> )	Inception v4, fine-tuned
DTD	75.5	FC+FV-CNN+D-SIFT [11]	<b>78.1</b>	Inception v4, fine-tuned
Oxford-IIIT Pets	93.8	Object-part attention [63]	<b>94.5</b>	ResNet-152 v1, fine-tuned
Caltech-101	93.4	Spatial pyramid pooling [32]	<b>95.1</b>	Inception-ResNet v2, log. regression
Oxford 102 Flowers	97.7	Domain-specific transfer [14]	<b>98.5</b>	Inception v4, fine-tuned

<sup>a</sup>For datasets where the best published result evaluated at  $448 \times 448$  or at multiple scales, we provide results at  $448 \times 448$  in parentheses.

<sup>b</sup>Accuracy excludes images duplicated between the ImageNet training set and CIFAR test sets; see Appendix H. A previous version of this paper achieved accuracies of 98.4% on CIFAR-10 and 88.2% on CIFAR-100 by fine-tuning the public NASNet checkpoint with the auxiliary head, dropout, and drop path. The difference in this version is due to the change in settings; the previous results remain valid.

<sup>c</sup>Krause et al. [42] achieve 85.4% on Birdsnap and 95.9% on Aircraft using bird and aircraft images collected from Google image search.

Table F.1. Performance of best models.

Table F.1 shows the best previously reported results of which we are aware on each of the datasets investigated. We achieve state-of-the-art performance on either 4 datasets at networks’ native image sizes, or 6 if we retrain networks at  $448 \times 448$ , as some previous transfer learning works have done. For CIFAR-10, CIFAR-100, and Stanford Cars, the best result was trained from scratch; for all other datasets, the baselines use some form of ImageNet pretraining.

## G. Comparison of alternative classifiers

In addition to the logistic regression without data augmentation setting described in the main text, we investigated transfer learning performance using support vector machines without data augmentation, and using logistic regression with data augmentation. Results are shown in Figure G.1.

### G.1. SVM

Although a logistic regression classifier trained on the penultimate layer activations has a natural interpretation as retraining the last layer of the neural network, many previous studies have reported results with support vector machines [20, 65, 6, 71]. Thus, we examine performance in this setting as well (Figure G.1A-D). Following previous work [71, 6], we  $L_2$ -normalized the input to the model along the feature dimension. We used the SVM implementation from scikit-learn [23, 62], selecting the value of the hyperparameter  $C$  from 26 logarithmically spaced values between 0.001 and 100. SVM and logistic regression results were highly correlated ( $r = 0.998$ ). For most (146/192) dataset/model pairs, logistic regression outperformed SVM, but differences were small (average log odds 1.32 vs. 1.28,  $p < 10^{-19}$ , t-test).

### G.2. Logistic regression with data augmentation

Finally, we trained a logistic regression classifier with data augmentation, in the same setting we use for fine-tuning. We trained for 40,000 steps with Nesterov momentum and a batch size of 256. Because the optimization problem is convex, we did not optimize over learning rate, but instead fixed the initial learning rate at 0.1 and used a cosine decay schedule. We optimized over L2 regularization parameters for the final layer, applied to the mean of the per-example losses, selected from a range of 10 logarithmically spaced values between  $10^{-10}$  and 0.1. Results are shown in Figure G.1E-H. No findings changed. Transfer accuracy with data augmentation was highly correlated with ImageNet accuracy (Figure G.1F) and with results without data augmentation (Figure G.1G;  $r = 0.99$  for both correlations). Fine-tuning remained clearly superior to logistic regression with data augmentation, achieving better results for 188/192 dataset/model pairs (Figure G.1H).

Logistic regression with data augmentation performed better for 100/192 dataset/model pairs. Data augmentation gave a slight improvement in average log odds (1.35 vs. 1.32), but the best performing model without data augmentation was better than the best performing model with data augmentation on half of the 12 datasets.

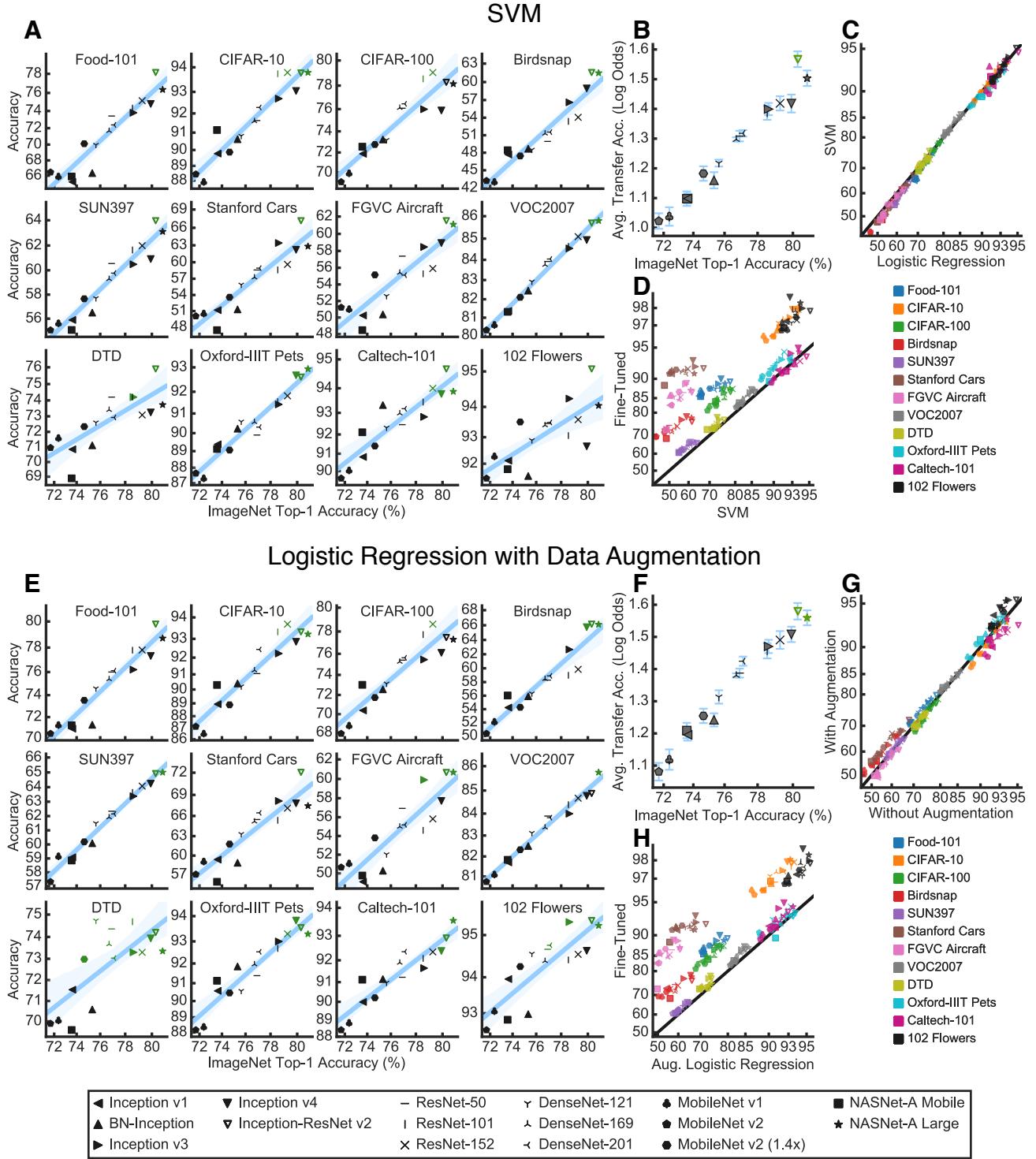


Figure G.1. Analysis of SVM and logistic regression with data augmentation, performed on fixed features. **A** and **E**: Scatter plots of ImageNet top-1 accuracy versus transfer accuracy on each of the 12 datasets examined. See also Figure 2. **B** and **F**: ImageNet top-1 accuracy versus average transfer accuracy for each network investigated. **C** and **G**: Performance of logistic regression without data augmentation versus SVM (**C**) or logistic regression with data augmentation (**G**). **D** and **H**: Performance of SVM (**D**) or logistic regression with data augmentation (**H**) versus fine-tuning. See also Figure E.1.

## H. Duplicate images

Dataset	Train Size	Test Size	Train Dups	Test Dups	Train Dup %	Test Dup %
Food-101	75,750	25,250	2	1	0.00%	0.00%
CIFAR-10	50,000	10,000	703	137	1.41%	1.37%
CIFAR-100	50,000	10,000	1,134	229	2.27%	2.29%
Birdsnap	47,386	2,443	431	23	0.91%	0.94%
SUN397	19,850	19,850	113	95	0.57%	0.48%
Stanford Cars	8,144	8,041	10	14	0.12%	0.17%
FGVC Aircraft	6,667	3,333	0	1	0.00%	0.03%
VOC2007	5,011	4,952	46	38	0.92%	0.77%
DTD	3,760	1,880	14	9	0.37%	0.48%
Oxford-IIIT Pets	3,680	3,669	227	58	6.17%	1.58%
Caltech-101	3,060	6,084	28	21	0.92%	0.35%
102 Flowers	2,040	6,149	1	0	0.05%	0.00%

Table H.1. Prevalence of images duplicated between the ImageNet training set and datasets investigated for transfer.

We used a CNN-based duplicate detector trained on synthesized image triplets to detect images that were present in both the ImageNet training set and the datasets we examine. Because the duplicate detector is optimized for speed, it is imperfect. We used a threshold that was conservative based on manual examination, i.e., it resulted in some false positives but very few false negatives. Thus, the results below represent a worst-case scenario for overlap in the datasets examined. Generally, there are relatively few duplicates. For most of these datasets, standard practice is to fine-tune an ImageNet pretrained network without special handling of duplicates, so the presence of duplicates does not affect the comparability of our results to previous work. However, for CIFAR-10 and CIFAR-100, we compare against networks trained from scratch and there are a substantial number of duplicates, so we exclude duplicates from the test set.

On CIFAR-10, we achieve an accuracy of 98.04% when fine-tuning NASNet Large (the best model) on the full test set. We also achieve an accuracy of 98.02% on the 9,863 example test set that is disjoint with the ImageNet training set. We achieve an accuracy of 99.27% on the 137 duplicates. On CIFAR-100, we achieve an accuracy of 87.7% on the full test set. We achieve an accuracy of 87.5% on the 9,771 example test set that is disjoint from the ImageNet training set, and an accuracy of 95.63% on the 229 duplicates.

## I. Numerical performance results

We present the numerical results for logistic regression, fine-tuning, and training from random initialization in Table I.1. Bold-faced numbers represent best models, or models insignificantly different from the best, in each training setting.

Logistic regression

Network	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech101	Flowers
Inception v1	68.3	89.8	72.3	49.9	58.8	53.8	53.0	81.1	71.3	90.0	91.93	92.1
BN-Inception	69.5	91.0	73.9	52.0	59.6	53.7	53.1	82.1	70.1	90.9	91.32	91.8
Inception v3	74.9	92.5	76.2	58.6	63.1	65.3	60.5	84.0	<b>73.9</b>	92.3	92.98	94.1
Inception v4	75.8	92.9	76.9	<b>61.4</b>	63.9	64.2	59.9	84.9	<b>74.6</b>	<b>93.4</b>	93.65	93.0
Inception-ResNet v2	<b>78.6</b>	<b>93.8</b>	<b>78.5</b>	<b>62.3</b>	<b>65.8</b>	<b>67.9</b>	<b>63.3</b>	85.2	<b>74.7</b>	<b>92.9</b>	<b>95.06</b>	<b>94.9</b>
ResNet-50 v1	74.1	91.8	76.0	52.2	62.5	59.5	58.5	83.5	<b>74.9</b>	91.5	92.74	93.2
ResNet-101 v1	75.1	<b>93.6</b>	<b>78.9</b>	55.3	64.0	60.1	57.4	84.5	<b>74.9</b>	92.2	92.65	93.1
ResNet-152 v1	75.8	<b>93.8</b>	<b>79.2</b>	55.7	64.1	60.2	56.9	84.8	<b>75.0</b>	92.4	93.96	93.5
DenseNet-121	72.0	90.5	73.8	51.9	60.7	57.3	53.5	82.6	<b>74.8</b>	91.2	92.13	93.3
DenseNet-169	72.7	91.8	76.2	54.9	61.2	59.0	57.2	83.0	<b>73.4</b>	92.0	93.75	93.4
DenseNet-201	73.2	92.2	76.4	54.2	61.9	60.3	57.1	83.6	73.2	91.4	93.15	93.1
MobileNet v1	68.2	88.2	70.9	46.3	58.8	52.9	52.6	80.2	71.0	87.4	90.77	92.7
MobileNet v2	68.4	88.6	70.6	46.3	57.6	51.6	52.9	80.0	71.7	88.1	91.26	91.7
MobileNet v2 (1.4)	71.6	89.8	73.4	50.0	60.3	56.1	55.2	81.9	73.0	89.3	91.83	93.5
NASNet-A Mobile	68.9	91.3	73.6	52.4	58.8	49.8	51.5	80.8	70.3	90.0	91.52	91.8
NASNet-A Large	76.9	<b>93.8</b>	78.0	<b>62.8</b>	65.1	63.7	<b>62.8</b>	<b>85.8</b>	<b>74.5</b>	<b>93.5</b>	93.89	93.8

Fine-tuned

Network	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech101	Flowers
Inception v1	85.6	96.17	83.2	73.0	62.0	91.0	82.7	83.2	73.6	91.9	91.7	97.26
BN-Inception	86.8	96.67	84.8	72.9	62.8	91.7	85.8	84.6	73.9	92.3	92.8	97.2
Inception v3	88.8	97.5	86.6	<b>77.2</b>	65.7	92.3	<b>88.8</b>	86.6	<b>77.2</b>	93.5	<b>94.3</b>	97.98
Inception v4	<b>90.0</b>	<b>97.93</b>	<b>87.5</b>	<b>78.4</b>	<b>66.4</b>	<b>93.3</b>	<b>89.0</b>	<b>87.4</b>	<b>78.1</b>	93.7	<b>94.9</b>	<b>98.45</b>
Inception-ResNet v2	89.4	<b>97.87</b>	86.8	76.3	<b>65.9</b>	92.0	86.7	86.7	<b>77.1</b>	93.3	<b>93.9</b>	97.85
ResNet-50 v1	87.8	96.77	84.5	74.7	64.7	91.7	86.6	85.7	75.2	92.5	91.8	97.51
ResNet-101 v1	87.6	97.68	87.0	73.8	64.8	91.7	85.6	86.6	76.2	<b>94.0</b>	93.1	97.94
ResNet-152 v1	87.6	<b>97.91</b>	<b>87.6</b>	74.3	<b>66.0</b>	92.0	85.3	86.8	75.4	<b>94.5</b>	93.2	97.35
DenseNet-121	87.7	97.18	84.8	73.2	62.3	91.5	85.4	85.1	74.9	92.9	91.9	97.18
DenseNet-169	88.0	97.4	85.0	71.4	63.0	91.5	84.5	85.9	74.8	93.1	92.5	97.86
DenseNet-201	87.3	97.41	86.0	72.6	64.7	91.0	84.6	85.8	74.5	92.8	93.4	97.68
MobileNet v1	87.1	96.15	82.3	69.2	61.7	91.4	85.8	82.6	73.4	89.9	90.1	96.67
MobileNet v2	86.2	95.74	80.8	69.3	60.5	91.0	82.8	82.1	72.9	90.5	89.1	96.63
MobileNet v2 (1.4)	87.7	96.13	82.5	71.5	62.6	91.8	86.8	83.4	73.0	91.0	91.1	97.52
NASNet-A Mobile	85.5	96.83	83.9	68.3	60.7	88.5	72.8	83.5	72.8	89.4	91.5	96.83
NASNet-A Large	88.9	<b>98.04</b>	<b>87.7</b>	<b>77.9</b>	<b>66.2</b>	91.1	87.2	87.2	74.3	93.3	<b>94.5</b>	<b>98.22</b>

Trained from random initialization

Network	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech101	Flowers
Inception v1	83.1	94.03	77.0	68.6	53.1	90.1	83.7	66.9	61.9	79.1	73.3	90.9
BN-Inception	84.4	95.17	80.2	69.6	52.5	90.7	81.7	66.9	64.4	79.3	74.3	92.8
Inception v3	86.6	95.61	<b>80.8</b>	<b>75.3</b>	<b>54.9</b>	91.6	87.7	70.8	65.1	<b>83.2</b>	<b>77.0</b>	<b>93.5</b>
Inception v4	<b>86.7</b>	<b>96.05</b>	<b>81.0</b>	<b>75.9</b>	<b>55.0</b>	<b>92.7</b>	<b>88.8</b>	<b>70.9</b>	<b>66.8</b>	81.2	<b>75.4</b>	<b>93.9</b>
Inception-ResNet v2	<b>87.0</b>	94.85	79.9	74.2	54.2	89.9	84.9	67.0	59.6	76.9	71.8	92.5
ResNet-50 v1	84.3	94.17	78.6	68.2	51.5	88.5	79.6	64.9	62.3	78.1	65.6	87.9
ResNet-101 v1	85.6	94.81	79.9	69.5	51.5	88.2	78.6	61.6	62.6	76.2	64.6	87.8
ResNet-152 v1	85.9	94.61	<b>80.8</b>	68.9	51.1	88.6	78.2	61.9	61.1	74.0	64.9	88.7
DenseNet-121	84.8	95.35	79.5	70.4	52.6	90.1	82.1	65.9	62.9	78.6	73.5	91.2
DenseNet-169	84.8	95.53	80.0	71.1	53.2	89.7	82.8	64.6	61.3	79.9	73.8	91.9
DenseNet-201	85.3	<b>96.05</b>	<b>80.8</b>	70.4	52.4	89.3	78.4	66.9	60.7	80.3	72.4	90.8
MobileNet v1	82.4	93.88	77.9	64.8	51.8	89.6	81.1	67.2	63.1	78.5	73.0	90.5
MobileNet v2	80.9	93.68	75.2	64.3	48.8	88.6	81.3	67.8	63.7	78.5	67.7	89.9
MobileNet v2 (1.4)	81.9	94.07	75.5	66.8	51.1	89.0	82.7	66.3	63.1	80.1	73.1	91.9
NASNet-A Mobile	81.9	94.73	78.3	65.9	48.3	86.7	75.1	57.9	60.1	69.4	63.5	85.8
NASNet-A Large	<b>86.8</b>	<b>96.06</b>	79.2	<b>75.5</b>	54.3	90.9	<b>88.2</b>	65.2	60.5	77.8	73.6	91.8

Table I.1. Model performance