# Do ImageNet Classifiers Generalize to ImageNet?

**Benjamin Recht**[*1] **Rebecca Roelofs**[1] **Ludwig Schmidt**[1] **Vaishaal Shankar**[1]

## Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% – 15% on CIFAR-10 and 11% – 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.

## 1. Introduction

The overarching goal of machine learning is to produce models that *generalize*. We usually quantify generalization by measuring the performance of a model on a held-out test set. What does good performance on the test set then imply? At the very least, one would hope that the model also performs well on a new test set assembled from the same data source by following the same data cleaning protocol.

In this paper, we realize this thought experiment by replicating the dataset creation process for two prominent benchmarks, CIFAR-10 and ImageNet (Deng et al., 2009; Krizhevsky, 2009). In contrast to the ideal outcome, we find that a wide range of classification models fail to reach their original accuracy scores. The accuracy drops range from 3% to 15% on CIFAR-10 and 11% to 14% on ImageNet. On ImageNet, the accuracy loss amounts to approximately five years of progress in a highly active period of machine learning research.

Conventional wisdom suggests that such drops arise because the models have been adapted to the specific images in the original test sets, e.g., via extensive hyperparameter tuning. However, our experiments show that the relative order of models is almost exactly preserved on our new test sets: the models with highest accuracy on the original test sets are still the models with highest accuracy on the new test sets. Moreover, there are no diminishing returns in accuracy. In fact, every percentage point of accuracy improvement on the original test set translates to a *larger* improvement on our new test sets. So although later models could have been adapted more to the test set, they see smaller drops in accuracy. These results provide evidence that exhaustive test set evaluations are an effective way to improve image classification models. Adaptivity is therefore an unlikely explanation for the accuracy drops.

Instead, we propose an alternative explanation based on the relative difficulty of the original and new test sets. We demonstrate that it is possible to recover the original ImageNet accuracies almost exactly if we only include the easiest images from our candidate pool. This suggests that the accuracy scores of even the best image classifiers are still highly sensitive to minutiae of the data cleaning process. This brittleness puts claims about human-level performance into context (He et al., 2015; Karpathy, 2011; Russakovsky et al., 2015). It also shows that current classifiers still do not generalize reliably even in the benign environment of a carefully controlled reproducibility experiment.

Figure 1 shows the main result of our experiment. Before we describe our methodology in Section 3, the next section provides relevant background. To enable future research, we release both our new test sets and the corresponding code.[1]

## 2. Potential Causes of Accuracy Drops

We adopt the standard classification setup and posit the existence of a "true" underlying data distribution $\mathcal{D}$ over labeled examples $(x, y)$. The overall goal in classification

[*]Authors ordered alphabetically. Ben did none of the work. [1]Department of Computer Science, University of California Berkeley, Berkeley, California, USA. Correspondence to: Benjamin Recht <brecht@berkeley.edu>.

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. 

[1]https://github.com/modestyachts/CIFAR-10.1 and https://github.com/modestyachts/ImageNetV2

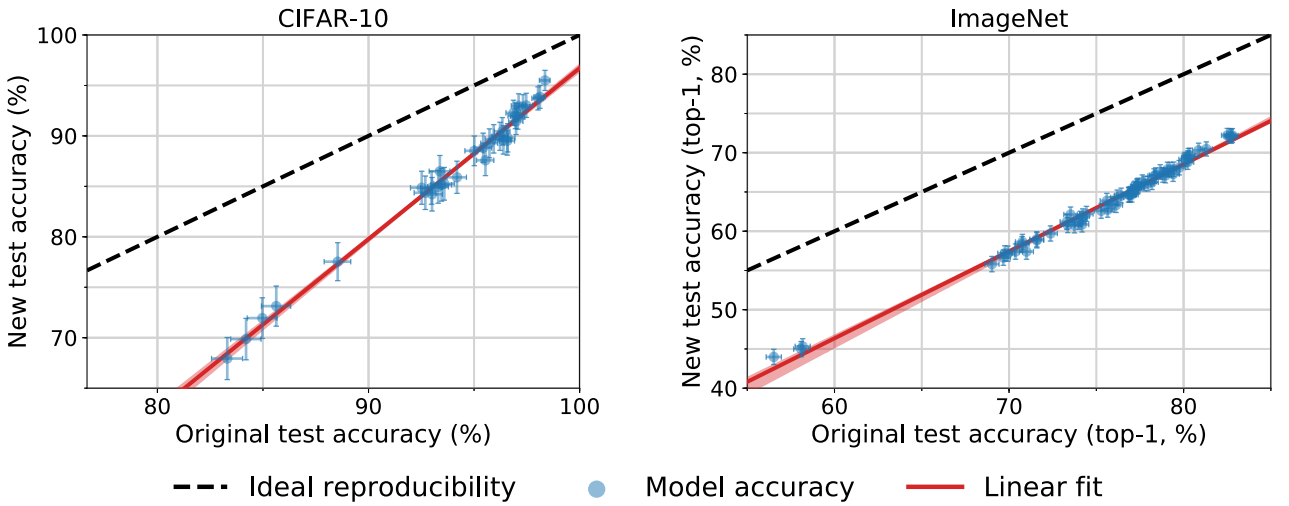


*Figure 1.* Model accuracy on the original test sets vs. our new test sets. Each data point corresponds to one model in our testbed (shown with 95% Clopper-Pearson confidence intervals). The plots reveal two main phenomena: (i) There is a significant drop in accuracy from the original to the new test sets. (ii) The model accuracies closely follow a linear function with slope *greater* than 1 (1.7 for CIFAR-10 and 1.1 for ImageNet). This means that every percentage point of progress on the original test set translates into more than one percentage point on the new test set. The two plots are drawn so that their aspect ratio is the same, i.e., the slopes of the lines are visually comparable. The red shaded region is a 95% confidence region for the linear fit from 100,000 bootstrap samples.

is to find a model $\hat{f}$ that minimizes the population loss

$$L_{\mathcal{D}}(\hat{f}) \;=\; \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\Big[\mathbb{I}[\hat{f}(x) \neq y]\Big] \;. \tag{1}$$

Since we usually do not know the distribution $\mathcal{D}$, we instead measure the performance of a trained classifier via a *test set* $S$ drawn from the distribution $\mathcal{D}$:

$$L_S(\hat{f}) \;=\; \frac{1}{|S|}\sum_{(x,y)\in S} \mathbb{I}[\hat{f}(x) \neq y] \;. \tag{2}$$

We then use this test error $L_S(\hat{f})$ as a proxy for the population loss $L_{\mathcal{D}}(\hat{f})$. If a model $\hat{f}$ achieves a low test error, we assume that it will perform similarly well on future examples from the distribution $\mathcal{D}$. This assumption underlies essentially all empirical evaluations in machine learning since it allows us to argue that the model $\hat{f}$ generalizes.

In our experiments, we test this assumption by collecting a new test set $S'$ from a data distribution $\mathcal{D}'$ that we carefully control to resemble the original distribution $\mathcal{D}$. Ideally, the original test accuracy $L_S(\hat{f})$ and new test accuracy $L_{S'}(\hat{f})$ would then match up to the random sampling error. In contrast to this idealized view, our results in Figure 1 show a large drop in accuracy from the original test set $S$ set to our new test set $S'$. To understand this accuracy drop in more detail, we decompose the difference between $L_S(\hat{f})$ and $L_{S'}(\hat{f})$ into three parts (dropping the dependence on $\hat{f}$ to simplify notation):

$$L_S - L_{S'} \;=\; \underbrace{(L_S - L_{\mathcal{D}})}_{\text{Adaptivity gap}} + \underbrace{(L_{\mathcal{D}} - L_{\mathcal{D}'})}_{\text{Distribution Gap}} + \underbrace{(L_{\mathcal{D}'} - L_{S'})}_{\text{Generalization gap}} .$$

We now discuss to what extent each of the three terms can lead to accuracy drops.

**Generalization Gap.** By construction, our new test set $S'$ is independent of the existing classifier $\hat{f}$. Hence the third term $L_{\mathcal{D}'} - L_{S'}$ is the standard *generalization gap* commonly studied in machine learning. It is determined solely by the random sampling error.

A first guess is that this inherent sampling error suffices to explain the accuracy drops in Figure 1 (e.g., the new test set $S'$ could have sampled certain "harder" modes of the distribution $\mathcal{D}$ more often). However, random fluctuations of this magnitude are unlikely for the size of our test sets. With 10,000 data points (as in our new ImageNet test set), a Clopper-Pearson 95% confidence interval for the test accuracy has size of at most $\pm 1\%$. Increasing the confidence level to 99.99% yields a confidence interval of size at most $\pm 2\%$. Moreover, these confidence intervals become smaller for higher accuracies, which is the relevant regime for the best-performing models. Hence random chance alone cannot explain the accuracy drops observed in our experiments.[2]

**Adaptivity Gap.** We call the term $L_S - L_{\mathcal{D}}$ the *adaptivity gap*. It measures how much adapting the model $\hat{f}$ to the test set $S$ causes the test error $L_S$ to underestimate the population loss $L_{\mathcal{D}}$. If we assumed that our model $\hat{f}$ is independent of the test set $S$, this terms would follow the

[2] We remark that the sampling process for the new test set $S'$ could indeed *systematically* sample harder modes more often than under the original data distribution $\mathcal{D}$. Such a systematic change in the sampling process would not be an effect of random chance but captured by the distribution gap described below.

same concentration laws as the generalization gap $L_{\mathcal{D}'} - L_{S'}$ above. But this assumption is undermined by the common practice of tuning model hyperparameters directly on the test set, which introduces dependencies between the model $\hat{f}$ and the test set $S$. In the extreme case, this can be seen as training directly on the test set. But milder forms of adaptivity may also artificially inflate accuracy scores by increasing the gap between $L_S$ and $L_{\mathcal{D}}$ beyond the purely random error.

**Distribution Gap.** We call the term $L_{\mathcal{D}} - L_{\mathcal{D}'}$ the *distribution gap*. It quantifies how much the change from the original distribution $\mathcal{D}$ to our new distribution $\mathcal{D}'$ affects the model $\hat{f}$. Note that this term is not influenced by random effects but quantifies the systematic difference between sampling the original and new test sets. While we went to great lengths to minimize such systematic differences, in practice it is hard to argue whether two high-dimensional distributions are exactly the same. We typically lack a precise definition of either distribution, and collecting a real dataset involves a plethora of design choices.

### 2.1. Distinguishing Between the Two Mechanisms

For a single model $\hat{f}$, it is unclear how to disentangle the adaptivity and distribution gaps. To gain a more nuanced understanding, we measure accuracies for *multiple* models $\hat{f}_1, \ldots, \hat{f}_k$. This provides additional insights because it allows us to determine how the two gaps have evolved over time.

For both CIFAR-10 and ImageNet, the classification models come from a long line of papers that incrementally improved accuracy scores over the past decade. A natural assumption is that later models have experienced more adaptive overfitting since they are the result of more successive hyperparameter tuning on the same test set. Their higher accuracy scores would then come from an increasing adaptivity gap and reflect progress only on the specific examples in the test set $S$ but not on the actual distribution $\mathcal{D}$. In an extreme case, the population accuracies $L_{\mathcal{D}}(\hat{f}_i)$ would plateau (or even decrease) while the test accuracies $L_S(\hat{f}_i)$ would continue to grow for successive models $\hat{f}_i$.

However, this idealized scenario is in stark contrast to our results in Figure 1. Later models do not see diminishing returns but an *increased* advantage over earlier models. Hence we view our results as evidence that the accuracy drops mainly stem from a large distribution gap. After presenting our results in more detail in the next section, we will further discuss this point in Section 5.

## 3. Summary of Our Experiments

We now give an overview of the main steps in our reproducibility experiment. Appendices C and D describe our methodology in more detail. We begin with the first decision, which was to choose informative datasets.

### 3.1. Choice of Datasets

We focus on image classification since it has become the most prominent task in machine learning and underlies a broad range of applications. The cumulative progress on ImageNet is often cited as one of the main breakthroughs in computer vision and machine learning (Malik, 2017). State-of-the-art models now surpass human-level accuracy by some measure (He et al., 2015; Russakovsky et al., 2015). This makes it particularly important to check if common image classification models can reliably generalize to new data from the same source.

We decided on CIFAR-10 and ImageNet, two of the most widely-used image classification benchmarks (Hamner, 2017). Both datasets have been the focus of intense research for almost ten years now. Due to the competitive nature of these benchmarks, they are an excellent example for testing whether adaptivity has led to overfitting. In addition to their popularity, their carefully documented dataset creation process makes them well suited for a reproducibility experiment (Deng et al., 2009; Krizhevsky, 2009; Russakovsky et al., 2015).

Each of the two datasets has specific features that make it especially interesting for our replication study. CIFAR-10 is small enough so that many researchers developed and tested new models for this dataset. In contrast, ImageNet requires significantly more computational resources, and experimenting with new architectures has long been out of reach for many research groups. As a result, CIFAR-10 has likely experienced more hyperparameter tuning, which may also have led to more adaptive overfitting.

On the other hand, the limited size of CIFAR-10 could also make the models more susceptible to small changes in the distribution. Since the CIFAR-10 models are only exposed to a constrained visual environment, they may be unable to learn a robust representation. In contrast, ImageNet captures a much broader variety of images: it contains about $24\times$ more training images than CIFAR-10 and roughly $100\times$ more pixels per image. So conventional wisdom (such as the claims of human-level performance) would suggest that ImageNet models also generalize more reliably .

As we will see, neither of these conjectures is supported by our data: CIFAR-10 models do not suffer from more adaptive overfitting, and ImageNet models do not appear to be significantly more robust.

### 3.2. Dataset Creation Methodology

One way to test generalization would be to evaluate existing models on new i.i.d. data from the original test distribution. For example, this would be possible if the original dataset authors had collected a larger initial dataset and randomly

split it into two test sets, keeping one of the test sets hidden for several years. Unfortunately, we are not aware of such a setup for CIFAR-10 or ImageNet.

In this paper, we instead mimic the original distribution as closely as possible by repeating the dataset curation process that selected the original test set[3] from a larger data source. While this introduces the difficulty of disentangling the adaptivity gap from the distribution gap, it also enables us to check whether independent replication affects current accuracy scores. In spite of our efforts, we found that it is astonishingly hard to replicate the test set distributions of CIFAR-10 and ImageNet. At a high level, creating a new test set consists of two parts:

**Gathering Data.** To obtain images for a new test set, a simple approach would be to use a different dataset, e.g., Open Images (Krasin et al., 2017). However, each dataset comes with specific biases (Torralba and Efros, 2011). For instance, CIFAR-10 and ImageNet were assembled in the late 2000s, and some classes such as `car` or `cell_phone` have changed significantly over the past decade. We avoided such biases by drawing new images from the same source as CIFAR-10 and ImageNet. For CIFAR-10, this was the larger Tiny Image dataset (Torralba et al., 2008). For ImageNet, we followed the original process of utilizing the Flickr image hosting service and only considered images uploaded in a similar time frame as for ImageNet. In addition to the data source and the class distribution, both datasets also have rich structure *within* each class. For instance, each class in CIFAR-10 consists of images from multiple specific keywords in Tiny Images. Similarly, each class in ImageNet was assembled from the results of multiple queries to the Flickr API. We relied on the documentation of the two datasets to closely match the sub-class distribution as well.

**Cleaning Data.** Many images in Tiny Images and the Flickr results are only weakly related to the query (or not at all). To obtain a high-quality dataset with correct labels, it is therefore necessary to manually select valid images from the candidate pool. While this step may seem trivial, our results in Section 4 will show that it has major impact on the model accuracies.

The authors of CIFAR-10 relied on paid student labelers to annotate their dataset. The researchers in the ImageNet project utilized Amazon Mechanical Turk (MTurk) to handle the large size of their dataset. We again replicated both annotation processes. Two graduate students authors of this paper impersonated the CIFAR-10 labelers, and we employed MTurk workers for our new ImageNet test set. For both datasets, we also followed the original labeling instructions, MTurk task format, etc.

After collecting a set of correctly labeled images, we sampled our final test sets from the filtered candidate pool. We decided on a test set size of 2,000 for CIFAR-10 and 10,000 for ImageNet. While these are smaller than the original test sets, the sample sizes are still large enough to obtain 95% confidence intervals of about $\pm 1\%$. Moreover, our aim was to avoid bias due to CIFAR-10 and ImageNet possibly leaving only "harder" images in the respective data sources. This effect is minimized by building test sets that are small compared to the original datasets (about 3% of the overall CIFAR-10 dataset and less than 1% of the overall ImageNet dataset).

### 3.3. Results on the New Test Sets

After assembling our new test sets, we evaluated a broad range of image classification models spanning a decade of machine learning research. The models include the seminal AlexNet (Krizhevsky et al., 2012), widely used convolutional networks (He et al., 2016a; Huang et al., 2017; Simonyan and Zisserman, 2014; Szegedy et al., 2016), and the state-of-the-art (Cubuk et al., 2018; Liu et al., 2018). For all deep architectures, we used code previously published online. We relied on pre-trained models whenever possible and otherwise ran the training commands from the respective repositories. In addition, we also evaluated the best-performing approaches preceding convolutional networks on each dataset. These are random features for CIFAR-10 (Coates et al., 2011; Rahimi and Recht, 2009) and Fisher vectors for ImageNet (Perronnin et al., 2010).[4] We wrote our own implementations for these models, which we also release publicly.[5]

Overall, the top-1 accuracies range from 83% to 98% on the original CIFAR-10 test set and 21% to 83% on the original ImageNet validation set. We refer the reader to Appendices D.4.3 and C.3.2 for a full list of models and source repositories.

Figure 1 in the introduction plots original vs. new accuracies, and Table 1 in this section summarizes the numbers of key models. The remaining accuracy scores can be found in Appendices C.3.3 and D.4.4. We now briefly describe the

[3] For ImageNet, we repeat the creation process of the *validation set* because most papers developed and tested models on the validation set. We discuss this point in more detail in Appendix D.1. In the context to this paper, we use the terms "validation set" and "test set" interchangeably for ImageNet.

[4] We remark that our implementation of Fisher vectors yields top-5 accuracy numbers that are 17% lower than the published numbers in ILSVRC 2012 (Russakovsky et al., 2015). Unfortunately, there is no publicly available reference implementation of Fisher vector models achieving this accuracy score. Hence our implementation should not be seen as an exact reproduction of the state-of-the-art Fisher vector model, but as a baseline inspired by this approach. The main goal of including Fisher vector models in our experiment is to investigate if they follow the same overall trends as convolutional neural networks.

[5] https://github.com/modestyachts/nondeep

**CIFAR-10**

| Orig. Rank | Model | Orig. Accuracy | New Accuracy | Gap | New Rank | Δ Rank |
|---|---|---|---|---|---|---|
| 1 | autoaug_pyramid_net_tf | 98.4 [98.1, 98.6] | 95.5 [94.5, 96.4] | 2.9 | 1 | 0 |
| 6 | shake_shake_64d_cutout | 97.1 [96.8, 97.4] | 93.0 [91.8, 94.1] | 4.1 | 5 | 1 |
| 16 | wide_resnet_28_10 | 95.9 [95.5, 96.3] | 89.7 [88.3, 91.0] | 6.2 | 14 | 2 |
| 23 | resnet_basic_110 | 93.5 [93.0, 93.9] | 85.2 [83.5, 86.7] | 8.3 | 24 | -1 |
| 27 | vgg_15_BN_64 | 93.0 [92.5, 93.5] | 84.9 [83.2, 86.4] | 8.1 | 27 | 0 |
| 30 | cudaconvnet | 88.5 [87.9, 89.2] | 77.5 [75.7, 79.3] | 11.0 | 30 | 0 |
| 31 | random_features_256k_auc | 85.6 [84.9, 86.3] | 73.1 [71.1, 75.1] | 12.5 | 31 | 0 |

**ImageNet Top-1**

| Orig. Rank | Model | Orig. Accuracy | New Accuracy | Gap | New Rank | Δ Rank |
|---|---|---|---|---|---|---|
| 1 | pnasnet_large_tf | 82.9 [82.5, 83.2] | 72.2 [71.3, 73.1] | 10.7 | 3 | -2 |
| 4 | nasnetalarge | 82.5 [82.2, 82.8] | 72.2 [71.3, 73.1] | 10.3 | 1 | 3 |
| 21 | resnet152 | 78.3 [77.9, 78.7] | 67.0 [66.1, 67.9] | 11.3 | 21 | 0 |
| 23 | inception_v3_tf | 78.0 [77.6, 78.3] | 66.1 [65.1, 67.0] | 11.9 | 24 | -1 |
| 30 | densenet161 | 77.1 [76.8, 77.5] | 65.3 [64.4, 66.2] | 11.8 | 30 | 0 |
| 43 | vgg19_bn | 74.2 [73.8, 74.6] | 61.9 [60.9, 62.8] | 12.3 | 44 | -1 |
| 64 | alexnet | 56.5 [56.1, 57.0] | 44.0 [43.0, 45.0] | 12.5 | 64 | 0 |
| 65 | fv_64k | 35.1 [34.7, 35.5] | 24.1 [23.2, 24.9] | 11.0 | 65 | 0 |

*Table 1.* Model accuracies on the original CIFAR-10 test set, the original ImageNet validation set, and our new test sets. Δ Rank is the relative difference in the ranking from the original test set to the new test set in the full ordering of all models (see Appendices C.3.3 and D.4.4). For example, $\Delta\text{Rank} = -2$ means that a model dropped by two places on the new test set compared to the original test set. The confidence intervals are 95% Clopper-Pearson intervals. Due to space constraints, references for the models can be found in Appendices C.3.2 and D.4.3.

two main trends and discuss the results further in Section 5.

**A Significant Drop in Accuracy.** All models see a large drop in accuracy from the original test sets to our new test sets. For widely used architectures such as VGG (Simonyan and Zisserman, 2014) and ResNet (He et al., 2016a), the drop is 8% on CIFAR-10 and 11% on ImageNet. On CIFAR-10, the state of the art (Cubuk et al., 2018) is more robust and only drops by 3% from 98.4% to 95.5%. In contrast, the best model on ImageNet (Liu et al., 2018) sees an 11% drop from 83% to 72% in top-1 accuracy and a 6% drop from 96% to 90% in top-5 accuracy. So the top-1 drop on ImageNet is larger than what we observed on CIFAR-10.

To put these accuracy numbers into perspective, we note that the best model in the ILSVRC[6] 2013 competition achieved 89% top-5 accuracy, and the best model from ILSVRC 2014 achieved 93% top-5 accuracy. So the 6% drop in top-5 accuracy from the 2018 state-of-the-art corresponds to approximately five years of progress in a very active period of machine learning research.

[6]ILSVRC is the ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015).

**Few Changes in the Relative Order.** When sorting the models in order of their original and new accuracy, there are few changes in the respective rankings. Models with comparable original accuracy tend to see a similar decrease in performance. In fact, Figure 1 shows that the original accuracy is highly predictive of the new accuracy and that the relationship can be summarized well with a linear function. On CIFAR-10, the new accuracy of a model is approximately given by the following formula:

$$\text{acc}_{\text{new}} \;=\; 1.69 \cdot \text{acc}_{\text{orig}} - 72.7\% \;.$$

On ImageNet, the top-1 accuracy of a model is given by

$$\text{acc}_{\text{new}} \;=\; 1.11 \cdot \text{acc}_{\text{orig}} - 20.2\% \;.$$

Computing a 95% confidence interval from 100,000 bootstrap samples gives $[1.63, 1.76]$ for the slope and $[-78.6, -67.5]$ for the offset on CIFAR-10, and $[1.07, 1.19]$ and $[-26.0, -17.8]$ respectively for ImageNet.

On both datasets, the slope of the linear fit is *greater* than 1. So models with higher original accuracy see a smaller drop on the new test sets. In other words, model robustness *improves* with increasing accuracy. This effect is less pronounced on ImageNet (slope 1.1) than on CIFAR-10 (slope

1.7). In contrast to a scenario with strong adaptive overfitting, neither dataset sees diminishing returns in accuracy scores when going from the original to the new test sets.

### 3.4. Experiments to Test Follow-Up Hypotheses

Since the drop from original to new accuracies is concerningly large, we investigated multiple hypotheses for explaining this drop. Appendices C.2 and D.3 list a range of follow-up experiments we conducted, e.g., re-tuning hyperparameters, training on part of our new test set, or performing cross-validation. However, none of these effects can explain the size of the drop. We conjecture that the accuracy drops stem from small variations in the human annotation process. As we will see in the next section, the resulting changes in the test sets can significantly affect model accuracies.

## 4. Understanding the Impact of Data Cleaning on ImageNet

A crucial aspect of ImageNet is the use of MTurk. There is a broad range of design choices for the MTurk tasks and how the resulting annotations determine the final dataset. To better understand the impact of these design choices, we assembled three different test sets for ImageNet. All of these test sets consist of images from the same Flickr candidate pool, are correctly labeled, and selected by more than 70% of the MTurk workers on average. Nevertheless, the resulting model accuracies vary by 14%. To put these numbers in context, we first describe our MTurk annotation pipeline.

**MTurk Tasks.** We designed our MTurk tasks and user interface to closely resemble those originally used for ImageNet. As in ImageNet, each MTurk task contained a grid of 48 candidate images for a given target class. The task description was derived from the original ImageNet instructions and included the definition of the target class with a link to a corresponding Wikipedia page. We asked the MTurk workers to select images belonging to the target class regardless of "occlusions, other objects, and clutter or text in the scene" and to avoid drawings or paintings (both as in ImageNet). Appendix D.4.1 shows a screenshot of our UI and a screenshot of the original UI for comparison.

For quality control, we embedded at least six randomly selected images from the original validation set in each MTurk task (three from the same class, three from a class that is nearby in the WordNet hierarchy). These images appeared in random locations of the image grid for each task. In total, we collected sufficient MTurk annotations so that we have at least 20 annotated validation images for each class.

The main outcome of the MTurk tasks is a *selection frequency* for each image, i.e., what fraction of MTurk workers selected the image in a task for its target class. We recruited at least ten MTurk workers for each task (and hence for each image), which is similar to ImageNet. Since each task contained original validation images, we could also estimate how often images from the original dataset were selected by our MTurk workers.

**Sampling Strategies.** In order to understand how the MTurk selection frequency affects the model accuracies, we explored three sampling strategies.

- **MatchedFrequency:** First, we estimated the selection frequency distribution for each class from the annotated original validation images. We then sampled ten images from our candidate pool for each class according to these class-specific distributions (see Appendix D.1.2 for details).
- **Threshold0.7:** For each class, we sampled ten images with selection frequency at least 0.7.
- **TopImages:** For each class, we chose the ten images with highest selection frequency.

In order to minimize labeling errors, we manually reviewed each dataset and removed incorrect images. The average selection frequencies of the three final datasets range from 0.93 for TopImages over 0.85 for Threshold0.7 to 0.73 for MatchedFrequency. For comparison, the original validation set has an average selection frequency of 0.71 in our experiments. Hence all three of our new test sets have higher selection frequencies than the original ImageNet validation set. In the preceding sections, we presented results on MatchedFrequency for ImageNet since it is closest to the validation set in terms of selection frequencies.

**Results.** Table 2 shows that the MTurk selection frequency has significant impact on both top-1 and top-5 accuracy. In particular, TopImages has the highest average MTurk selection frequency and sees a small *increase* of about 2% in both average top-1 and top-5 accuracy compared to the original validation set. This is in stark contrast to MatchedFrequency, which has the lowest average selection frequency and exhibits a significant drop of 12% and 8%, respectively. The Threshold0.7 dataset is in the middle and sees a small decrease of 3% in top-1 and 1% in top-5 accuracy.

In total, going from TopImages to MatchedFrequency decreases the accuracies by about 14% (top-1) and 10% (top-5). For comparison, note that after excluding AlexNet (and the SqueezeNet models tuned to match AlexNet (Iandola et al., 2016)), the range of accuracies spanned by all remaining convolutional networks is roughly 14% (top-1) and 8% (top-5). So the variation in accuracy caused by the three sampling strategies is larger than the variation in accuracy among all post-AlexNet models we tested.

Figure 2 plots the new vs. original top-1 accuracies on Threshold0.7 and TopImages, similar to Figure 1 for MatchedFrequency before. For easy comparison of top-1

| Sampling Strategy | Average MTurk Selection Freq. | Average Top-1 Accuracy Change | Average Top-5 Accuracy Change |
|---|---|---|---|
| MatchedFrequency | 0.73 | -11.8% | -8.2% |
| Threshold0.7 | 0.85 | -3.2% | -1.2% |
| TopImages | 0.93 | +2.1% | +1.8% |

*Table 2.* Impact of the three sampling strategies for our ImageNet test sets. The table shows the average MTurk selection frequency in the resulting datasets and the average changes in model accuracy compared to the original validation set. We refer the reader to Section 4 for a description of the three sampling strategies. All three test sets have an average selection frequency of more than 0.7, yet the model accuracies still vary widely. For comparison, the original ImageNet validation set has an average selection frequency of 0.71 in our MTurk experiments. The changes in average accuracy span 14% and 10% in top-1 and top-5, respectively. This shows that details of the sampling strategy have large influence on the resulting accuracies.

and top-5 accuracy plots on all three datasets, we refer the reader to Figure 1 in Appendix D.4.4. All three plots show a good linear fit.

## 5. Discussion

Due to space constraints, we defer a discussion of related work to Appendix A. Furthermore, Appendix B contains a theoretical model for the accurate linear fit observed in Figures 1 and 2. Here, we return to the main question from Section 2: *What causes the accuracy drops?* As before, we distinguish between two possible mechanisms.

### 5.1. Adaptivity Gap

In its prototypical form, *adaptive* overfitting would manifest itself in diminishing returns observed on the new test set (see Section 2.1). However, we do not observe this pattern on either CIFAR-10 or ImageNet. On both datasets, the slope of the linear fit is *greater* than 1, i.e., each point of accuracy improvement on the original test set translates to more than 1% on the new test set. This is the opposite of the standard overfitting scenario. So at least on CIFAR-10 and ImageNet, multiple years of competitive test set adaptivity did not lead to diminishing accuracy numbers.

While our experiments rule out the most dangerous form of adaptive overfitting, we remark that they do not exclude all variants. For instance, it could be that any test set adaptivity leads to a roughly constant drop in accuracy. Then all models are affected equally and we would see no diminishing returns since later models could still be better. Testing for this form of adaptive overfitting likely requires a new test set that is truly i.i.d. and not the result of a separate data collection effort. Finding a suitable dataset for such an experiment is an interesting direction for future research.

The lack of adaptive overfitting contradicts conventional wisdom in machine learning. We now describe two mechanisms that could have prevented adaptive overfitting:

**The Ladder Mechanism.** Blum and Hardt introduced the Ladder algorithm to protect machine learning competitions against adaptive overfitting (Blum and Hardt, 2015). The core idea is that constrained interaction with the test set can allow a large number of model evaluations to succeed, even if the models are chosen adaptively. Due to the natural form of their algorithm, the authors point out that it can also be seen as a mechanism that the machine learning community *implicitly* follows.

**Limited Model Class.** Adaptivity is only a problem if we can choose among models for which the test set accuracy differs significantly from the population accuracy. Importantly, this argument does not rely on the number of *all* possible models (e.g., all parameter settings of a neural network), but only on those models that could actually be evaluated on the test set. For instance, the standard deep learning workflow only produces models trained with SGD-style algorithms on a fixed training set, and requires that the models achieve high training accuracy (otherwise we would not consider the corresponding hyperparameters). Hence the number of different models arising from the current methodology may be small enough so that uniform convergence holds.

Our experiments offer little evidence for favoring one explanation over the other. One observation is that the convolutional networks shared many errors on CIFAR-10, which could be an indicator that the models are rather similar. But to gain a deeper understanding into adaptive overfitting, it is likely necessary to gather further data from more machine learning benchmarks, especially in scenarios where adaptive overfitting *does* occur naturally.

### 5.2. Distribution Gap

The lack of diminishing returns in our experiments points towards the distribution gap as the primary reason for the accuracy drops. Moreover, our results on ImageNet show that changes in the sampling strategy can indeed affect model accuracies by a large amount, even if the data source and other parts of the dataset creation process stay the same.

So in spite of our efforts to match the original dataset creation process, the distribution gap is still our leading hypothesis for the accuracy drops. This demonstrates that it

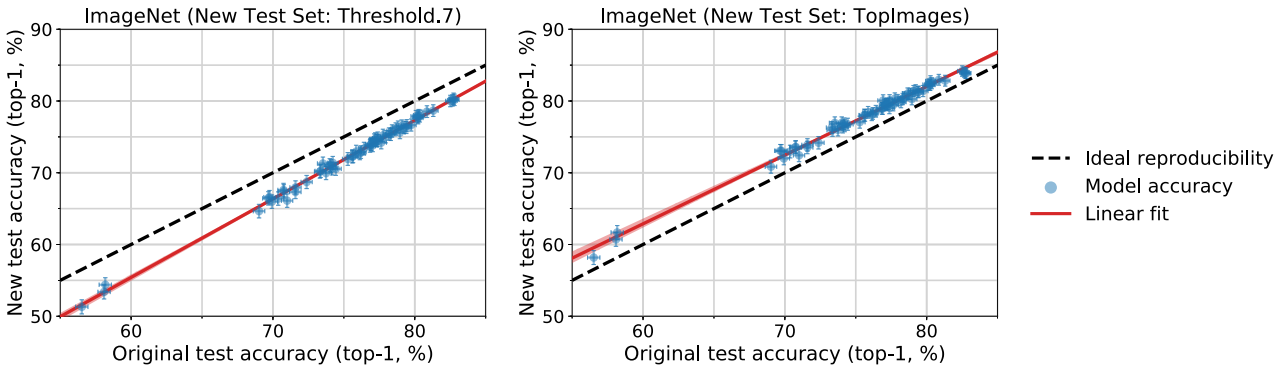


*Figure 2.* Model accuracy on the original ImageNet validation set vs. accuracy on two variants of our new test set. We refer the reader to Section 4 for a description of these test sets. Each data point corresponds to one model in our testbed (shown with 95% Clopper-Pearson confidence intervals). On Threshold0.7, the model accuracies are 3% lower than on the original test set. On TopImages, which contains the images most frequently selected by MTurk workers, the models perform 2% *better* than on the original test set. The accuracies on both datasets closely follow a linear function, similar to MatchedFrequency in Figure 1. The red shaded region is a 95% confidence region for the linear fit from 100,000 bootstrap samples.

is surprisingly hard to accurately replicate the distribution of current image classification datasets. The main difficulty likely is the subjective nature of the human annotation step. There are many parameters that can affect the quality of human labels such as the annotator population (MTurk vs. students, qualifications, location & time, etc.), the exact task format, and compensation. Moreover, there are no exact definitions for many classes in ImageNet (e.g., see Appendix D.4.8). Understanding these aspects in more detail is an important direction for designing future datasets that contain challenging images while still being labeled correctly.

The difficulty of clearly defining the data distribution, combined with the brittle behavior of the tested models, calls into question whether the black-box and i.i.d. framework of learning can produce reliable classifiers. Our analysis of selection frequencies in Figure 15 (Appendix D.4.7) shows that we could create a new test set with even lower model accuracies. The images in this hypothetical dataset would still be correct, from Flickr, and selected by more than half of the MTurk labelers on average. So in spite of the impressive accuracy scores on the original validation set, current ImageNet models still have difficulty generalizing from "easy" to "hard" images.

## 6. Conclusion & Future Work

The expansive growth of machine learning rests on the aspiration to deploy trained systems in a variety of challenging environments. Common examples include autonomous vehicles, content moderation, and medicine. In order to use machine learning in these areas responsibly, it is important that we can both train models with sufficient generalization abilities, and also reliably measure their performance. As our results show, these goals still pose significant hurdles even in a benign environment.

Our experiments are only a first step in addressing this reliability challenge. One important question is whether other machine learning tasks are also resilient to adaptive overfitting, but similarly brittle under natural variations in the data. Another direction is developing methods for more comprehensive yet still realistic evaluations of machine learning systems. Of course, the overarching goal is to develop learning algorithms that generalize reliably. While this is often a vague goal, our new test sets offer a well-defined instantiation of this challenge that is beyond the reach of current methods. Generalizing from our "easy" to slightly "harder" images will hopefully serve as a starting point towards a future generation of more reliable models.

## Acknowledgements

We would like to thank Tudor Achim, Alex Berg, Orianna DeMasi, Jia Deng, Alexei Efros, David Fouhey, Moritz Hardt, Piotr Indyk, Esther Rolf, and Olga Russakovsky for helpful discussions while working on this paper. Moritz Hardt has been particularly helpful in all stages of this project and – among other invaluable advice – suggested the title of this paper and a precursor to the data model in Section B. We also thank the participants of our human accuracy experiment in Appendix C.2.5 (whose names we keep anonymous following our IRB protocol).

This research was generously supported in part by ONR awards N00014-17-1-2191, N00014-17-1-2401, and N00014-18-1-2833, the DARPA Assured Autonomy (FA8750-18-C-0101) and Lagrange (W911NF-16-1-0552) programs, an Amazon AWS AI Research Award, and a gift from Microsoft Research. In addition, LS was supported by a Google PhD fellowship and a Microsoft Research Fellowship at the Simons Institute for the Theory of Computing.

## References


Alex Berg. Personal communication, 2018.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 2018. https://arxiv.org/abs/1712.03141.

Avrim Blum and Moritz Hardt. The Ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning (ICML)*, 2015. http://arxiv.org/abs/1502.04585.

Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. In *Neural Information Processing Systems (NIPS)*, 2017. https://arxiv.org/abs/1707.01629.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1610.02357.

Stephane Clinchant, Gabriela Csurka, Florent Perronnin, and Jean-Michel Renders. XRCE's participation to ImagEval. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.6670&rep=rep1&type=pdf, 2007.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011. http://proceedings.mlr.press/v15/coates11a.html.

Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. AutoAugment: Learning augmentation policies from data. https://arxiv.org/abs/1805.09501, 2018.

Jia Deng. *Large Scale Visual Recognition*. PhD thesis, Princeton University, 2012. ftp://ftp.cs.princeton.edu/techreports/2012/923.pdf.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. http://www.image-net.org/papers/imagenet_cvpr09.pdf.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with Cutout. https://arxiv.org/abs/1708.04552, 2017.

Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. http://arxiv.org/abs/1712.02779, 2017.

Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 2010. http://dx.doi.org/10.1007/s11263-009-0275-4.

Alhussein Fawzi and Pascal Frossard. Manitest: Are classifiers really invariant? In *British Machine Vision Conference (BMVC)*, 2015. https://arxiv.org/abs/1507.06535.

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. *Computer Vision and Image Understanding*, 2007. http://dx.doi.org/10.1016/j.cviu.2005.09.012.

Xavier Gastaldi. Shake-shake regularization. https://arxiv.org/abs/1705.07485, 2017.

Robert Geirhos, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2018. http://papers.nips.cc/paper/7982-generalisation-in-humans-and-deep-neural-networks.pdf.

Ben Hamner. Popular datasets over time. https://www.kaggle.com/benhamner/popular-datasets-over-time/data, 2017.

Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1610.02915.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *International Conference on Computer Vision (ICCV)*, 2015. https://arxiv.org/abs/1502.01852.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016a. https://arxiv.org/abs/1512.03385.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, 2016b. https://arxiv.org/abs/1603.05027.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning*

*Representations (ICLR)*, 2019. https://arxiv.org/abs/1807.01697.

Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. https://arxiv.org/abs/1804.00499.

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. https://arxiv.org/abs/1704.04861, 2017.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1709.01507.

Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1608.06993.

Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. https://arxiv.org/abs/1602.07360, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015. https://arxiv.org/abs/1502.03167.

Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Geometric robustness of deep networks: Analysis and improvement. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1711.09115.

Andrej Karpathy. Lessons learned from manually classifying CIFAR-10. http://karpathy.github.io/2011/04/27/manually-classifying-cifar10/, 2011.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. https://arxiv.org/abs/1710.05468, 2017.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better ImageNet models transfer better? http://arxiv.org/abs/1805.08974, 2018.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. https://storage.googleapis.com/openimages/web/index.html, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

Fei-Fei Li and Jia Deng. ImageNet: Where have we been? where are we going? http://image-net.org/challenges/talks_2017/imagenet_ilsvrc2017_v1.0.pdf, 2017.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. https://arxiv.org/abs/1405.0312.

Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *European Conference on Computer Vision (ECCV)*, 2018. https://arxiv.org/abs/1712.00559.

Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *Asian Conference on Pattern Recognition (ACPR)*, 2015. https://ieeexplore.ieee.org/document/7486599/.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. https://arxiv.org/abs/1706.06083.

Jitendra Malik. Technical perspective: What led computer vision to deep learning? *Communications of the ACM*, 2017. http://doi.acm.org/10.1145/3065384.

George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 1995. URL http://doi.acm.org/10.1145/219717.219748.

Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the Fisher kernel for large-scale image classification. In *European Conference on Computer Vision (ECCV)*, 2010. URL https://www.robots.ox.ac.uk/~vgg/rg/papers/peronnin_etal_ECCV10.pdf.

Jean Ponce, Tamara L. Berg, Mark Everingham, David A. Forsyth, Martial Hebert, Sveltana Lazebnik, Marcin Marszalek, Cordelia Schmid, Bryan C. Russell, Antionio Torralba, Chris. K. I. Williams, Jianguo Zhang, and Andrew Zisserman. *Dataset issues in object recognition*. 2006. https://link.springer.com/chapter/10.1007/11957959_2.

Ali Rahimi and Benjamin Recht. Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2009. https://papers.nips.cc/paper/3495-weighted-sums-of-random-kitchen-sinks-replacing-minimization-with-randomization-in-learning.

Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. Regularized evolution for image classifier architecture search. http://arxiv.org/abs/1802.01548, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Fei-Fei Li. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015. https://arxiv.org/abs/1409.0575.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2013. http://arxiv.org/abs/1312.6199.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. https://arxiv.org/abs/1409.4842v1.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. https://arxiv.org/abs/1512.00567.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, Inception-Resnet and the impact of residual connections on learning. In *Conference On Artificial Intelligence (AAAI)*, 2017. https://arxiv.org/abs/1602.07261.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. http://people.csail.mit.edu/torralba/publications/datasets_cvpr11.pdf.

Antonio Torralba, Rob Fergus, and William. T. Freeman. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008. https://ieeexplore.ieee.org/document/4531741/.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. http://www.cns.nyu.edu/pub/lcv/wang03-preprint.pdf.

Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. https://arxiv.org/abs/1801.02612.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1611.05431.

Yoshihiro Yamada, Masakazu Iwamura, and Koichi Kise. Shakedrop regularization. https://arxiv.org/abs/1802.02375, 2018.

Benjamin Z. Yao, Xiong Yang, and Song-Chun Zhu. Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks. In *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, 2007. https://link.springer.com/chapter/10.1007/978-3-540-74198-5_14.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016. https://arxiv.org/abs/1605.07146.

Xingcheng Zhang, Zhizhong Li, Chen Change Loy, and Dahua Lin. Polynet: A pursuit of structural diversity in very deep networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1611.05725.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1707.07012.