— 7 —
Replication in machine learning

The preconditions for crisis exist in machine learning, too. And yet, the situation in machine learning is different. While accuracy numbers don't replicate, model rankings replicate to a significant degree.

Source: The Emerging Science of Machine Learning Benchmarks. M. Hardt, 2025. URL: https://mlbenchmarks.org. Compiled on 2025-11-02.

The previous chapter covered the replication crisis in the statistical sciences. Researcher degrees of freedom is one major culprit, especially in light of Goodhart's law. Researchers can always find creative, sometimes questionable ways around the statistical guardrails of the scientific publication process.

If statistical sciences broadly face a replication crisis, what does this imply for machine learning? In this chapter, we examine the empirical reality of replication in machine learning research.

7.1 The preconditions for crisis

The *preconditions* for crisis surely exist in machine learning as well. You might even argue they are stronger in some significant ways.

To say that machine learning research has a culture of rapid publication is an understatement; the community produces an unfathomable volume of papers each year. For example, the paper that invented residual networks picked up a quarter million citations in the last five years. The paper that introduced transformers raked in around 55000 citations in 2024 alone. Both papers presented marvelous inventions worthy of the fame, but who can read a quarter million papers?

Peer review has long struggled to keep up with the relentless onslaught of new papers. The field's largest conference, Neural Information Processing Systems (abbreviated NeurIPS), received around 16,000 paper submissions in 2024; a hierarchy of around 100 senior area chairs, 1000 area chairs, and well over 10000 reviewers are tasked with sorting out peer review.

The pressure on peer review was already apparent in 2014 when—at just around 1600 submissions—NeurIPS was still small compared with the behemoth that it's since grown into. In the *NeurIPS experiment* of 2014, the program chairs Corinna Cortes and Neil Lawrence randomly assigned 166 papers (or 10% of the submissions) to two independent reviewer panels. What they found is striking, although perhaps not surprising to some: Of the papers accepted by one of the panels, only about half of them would've been accepted by the other committee.[1] This was early experimental evidence of what has since become a widespread belief among researchers and a source of consolation after rejection notifications: The outcome of peer review is rather random. The NeurIPS experiment was a brilliant idea that has since inspired much valuable experimentation with peer review, aimed at

improving the process.

Even if peer review were a lot more thorough than it is, there's only so much it can do. Machine learning embraces an extreme level of researcher degrees of freedom. It's the *anything goes* principle that's non-negotiable in the field. Researchers are unconstrained by statistical rigor or protocol. In particular, there is no preregistration or similar mechanisms whatsoever. The mere idea of trying to do something like preregistration doesn't compile. It would require a complete rewiring of how the community works. Progress in machine learning is by choice highly incremental and adaptive in its nature. Popular test sets, like CIFAR-10 and ImageNet, have easily been used millions of times in data-dependent ways.

The only methodological commitment in the community is to report evaluations on shared test sets. But why should the average value of a loss function on a test set be any more reliable than a p-value? Both are statistics computed on a sample, to which Goodhart's law would seem to apply in the same way. In this sense, machine learning shares an Achilles heel with statistics. Moreover, whereas datasets in the empirical sciences are typically sampled specifically for a study, machine learning test sets sit on the internet for everyone to use freely.

As a result of these factors, many have worried about a major scientific crisis lurking in the shadows of machine learning research. Indeed, a 2018 article in *Science* proclaimed exactly that.[2] Nevertheless, we will see that the situation in machine learning is in some important ways different than in other fields. Toward developing this point, it's helpful to draw some distinctions about the different standards of replication.

7.2 Replication in machine learning

A minimal bar of replication is *code re-execution*. This means that you can download and run the code linked to from a paper. If the dataset is public, this means that with some reasonable amount of elbow grease you can re-execute the experiments from the paper and *reproduce* the claims and plots in the paper. Reproducibility is replication in essentially *identical conditions*. Other than the hardware used to run the experiment, nothing has changed.

Code re-execution standards have benefited in the last ten years from the impressive open source software ecosystem that's grown around machine learning. It's marvelous how simple it is to clone and run code these days

compared to what it was even ten years ago. The contributions of thousands of open source developers have undoubtedly transformed machine learning for the better.

That said, the state of reproducibility is far from perfect. In 2019, NeurIPS ran a reproducibility challenge, where researchers volunteered to reproduce the findings of machine learning papers. Volunteers claimed 173 papers for reproduction and documented the results of their efforts in public reports. These reports make it clear that reproducibility isn't a binary. Reproduction requires some varying amount of skill and effort. But most findings do seem to at least reproduce with effort.

Code re-execution is also an important bar to insist on, since it incentivizes open science. For re-execution to work, the data, the model, and the code have to be available. The rise of proprietary models in the language model era threatens the minimal bar of code re-execution. Academics shouldn't give up on it.

Rapid re-execution is an important goal for the scientific community, but it's necessarily only a minimal bar of scientific replication. Code re-execution says very little about the scientific validity of claims in a paper. A result wouldn't be interesting if it held only under one specific set of conditions. When scientists say ResNet-152 is better than Inception V3 on ImageNet, they don't mean to say that this comparison holds only on one specific set of 50,000 test points and nowhere else. We expect the claim to be true somewhat more robustly at least.

A stronger bar to clear is replication in *similar conditions*. Here, we ask that claims are robust under minor variations in experimental conditions. In machine learning work, it's reasonable to ask that things replicate if you were to draw a fresh dataset from the same population. This kind of replication rules out that your model is good on only one sample. Theoretically, this is reasonably well captured by generalization under repeated sampling. In practice, resampling isn't so easy. It's hard to draw a new sample from the same population. Think back to Chapter 2, where we saw that populations change over time and rarely are completely stationary. The most we can do is try our best to recreate a new test set using the exact same steps as in the original creation procedure. Sometimes we get lucky and at the time of creation there were additional data points set aside and forgotten that haven't been used as a test set before. This was the case with the MNIST *lost digits*[3]—more later! A prudent practitioner might also set aside an extra secret test set at the time of benchmark creation for use in a future replication

study. But if the new test set is created retroactively, it will be at least a bit different from the original one.

Call this evidentiary bar of replication under similar conditions *internal validity*. Internal validity says that a claim is valid *within* a given setup and a fixed target population, e.g., ImageNet. Internal validity is a reasonable yardstick for good machine learning work. However, internal validity says nothing about how well a model developed in one setting might work elsewhere.

An even stronger bar to consider is that claims replicate in *different conditions*. Call this *external validity*. The claim stands even if the population changes. Applying a model to a different population almost always entails some necessary changes to the code, and some amount of engineering effort to get the code to run on the new setup. We'll apply this criterion within reason. After all, no machine learning method performs well everywhere. But external validity asks for more than just replication under resampling.

As we already touched on in Chapter 4, pinning down formal definitions of internal and external validity is challenging, as the concepts vary from one context to another and run into epistemological debates. We avoid a deeper philosophical discussion by taking a pragmatic route: Internal validity is about generalization to *more of the same* population. External validity is about generalization to *different populations.*

7.3 The trouble with absolute benchmark numbers

Imagine playing a fun little game for computer vision researchers: Given an image, classify which vision benchmark it comes from. This game is the starting point for the 2011 paper *An unbiased look at dataset bias* by Torralba and Efros.[4] Anecdotal evidence suggests that experts are actually pretty good at the game. Moreover, models trained on the classification task got high accuracy on the kind of benchmarks that would appear in computer vision papers in 2011. Even current computer vision datasets—proposed for model training not just evaluation—are still easy to tell apart. State-of-the-art vision models achieve 80%+ accuracy in a three way classification task involving three recent datasets all based on web crawled images.[5,6]

The fact that benchmark prediction is easy is a sign of *dataset biases* that identify each benchmark like a fingerprint. These biases include selection and filtering biases, capture biases (how an image is taken), and annotator

biases (how labels are assigned). No two ways of assembling a dataset are ever quite the same. A consequence of the resulting idiosyncrasies is that classifiers trained on one dataset don't do just as well on any other dataset. The accuracy drops when moving from one benchmark to another are typically in the double digits.

There's another corollary to the observation. Since all benchmarks are easy to tell apart, it can't be the case that two of them *both* reflect the *visual world*. Torralba and Efros contend that datasets ultimately are all trying to capture the same domain, the visual world. But apparently they don't succeed. All datasets are pretty far from whatever demands the real world poses. There's another way to look at this corollary. If benchmarks are useful, this must mean that a benchmark can be a useful benchmark without representing the real world.

By the way, a lesser known part of the paper proposes a market perspective on measuring dataset value. The question is: How many data points from one benchmark would you exchange for a single data point from another? This defines a kind of exchange rate between datasets, and by extension gives us a way to talk about the value of a dataset.

Torralba and Efros launched a new field of study. Their key findings about accuracy drops have since been confirmed in hundreds of papers. Accuracy numbers change even with slight changes to the population. The problem they brought attention to now falls under the umbrella term *distribution shift*. Distribution shift is a catch all for situations where model performance changes from one dataset to another.

Changing populations

Torralba and Efros demonstrated that model performance can drop sharply from one dataset to the other. While they considered fairly different computer vision benchmarks, the same is still true even if we move between seemingly very similar distributions.

Whoever first deployed a model on an online platform must've quickly learned that performance in the real world degrades over time. A recommender system trained on a week of clicks in January will have likely gotten much worse by February. But not only time degrades model performance. Small transformations like object rotations, partial occlusions, or image artifacts can significantly deteriorate the accuracy of a vision model. Style changes, grammatical transformation, and different prompts can throw off

language models.

Distribution shift or dataset shift refers to differences in the distribution on which a model was trained and the distribution on which it is tested. The word makes most sense in cases where we expect the distributions to be morally similar. You wouldn't call the difference between any two arbitrary distributions *shift*. When people say distribution shift, there's an underlying assumption that the two distributions should've been close in some intuitive sense.

A mathematical fact says that if two distributions are close in total variation distance, then model performance must also be close on the two distributions. This follows from the material in Chapter 2: Total variation distance bounds the difference of any bounded loss function on the two distributions. Unfortunately, two naturally occurring distributions are rarely close in total variation distance. If you were to take pictures with two different cameras—each filtering and encoding an image differently—the resulting distributions would almost certainly be far in total variation distance.

Machine learning excels at exploiting dataset artifacts. Whatever signal there is in the data, large enough models will use it for better predictions. Hundreds of papers have documented the fragility of machine learning models under distribution shift.

If distribution shift is a problem, what is the solution?

Researchers have actively studied this problem for decades. *Domain adaptation* refers to the challenge of training a model on one data distribution (the source domain) and applying it to another distribution (the target domain), typically under the assumption that the task remains the same. Usually, you have some data from the target domain, although it may be unlabeled. Domain adaptation is one of the tools for mitigating distribution shift. A closely related problem is *domain generalization*, where the goal is to train models on multiple source domains to improve generalization to an unseen target domain, without any target data available during training.

More broadly, transfer learning is a bag of techniques to use models from one domain to improve learning in another. Unlike domain adaptation, transfer learning may involve different tasks across domains. A common case of transfer learning is to take a deep neural net trained on ImageNet, remove its final classification layer, add a different classification layer on top, and fine-tune the last layer on a new dataset. ImageNet turned out to be quite useful for this kind of transfer learning.

There are also benchmarks for domain generalization and testing model performance across distribution shifts. For example, DomainBed tests the ability of models to generalize to new domains. DomainBed contains simple variants of basic benchmarks, such as *Colored MNIST*, or *Rotated MNIST*.[7] WILDS is a curated collection of 10 datasets that aim to capture real-world distribution shifts, for example, variation in location and time of satellite imaging data.[8]

There are two primary findings from these benchmarks:

1. Methods specifically designed for domain generalization or robustness to distribution shift do not consistently outperform empirical risk minimization. What works best is the *kitchen sink* approach: Train the best model you can on all the data you have and hope for the best.
2. Accuracy numbers, and by extension absolute benchmark metrics, don't have external validity. They change sharply under dataset variation.

The first finding is an instance of Sutton's *bitter lesson*:[9] Building bigger models on more data usually beats clever algorithm design. Although it's not our main focus in this chapter, the bitter lesson is an interesting meta fact about machine learning that has long contributed to the importance of empirical testing in the field.

The second finding is more directly about replication. Accuracy numbers evidently do not satisfy external validity. They are best interpreted literally: On this specific dataset, the model makes so many errors. We know from Hoeffding's bound that they satisfy the most narrow interpretation of internal validity. If we sample twice from the exact same distribution, the numbers are close up to statistical error. But we don't seem to get much more than this.

There are fundamental reasons why absolute accuracy numbers don't mean much on their own. Take any benchmark of your choice and replace 20% of the labels with randomly drawn labels from the set of C classes. The new benchmark is for all intents and purposes operationally equivalent. It gives you all the same model comparisons. Up to small statistical fluctuations, leaderboard climbing works the same way. However, an accuracy number A on the original benchmarks maps to $0.8A + 0.2/C$ on the new benchmark. Benchmarks may be "harder" or "easier" in terms of accuracy numbers, while being equivalent benchmarks. Accuracy numbers without additional context or domain knowledge therefore can't say much. Of course, labels aren't completely random in real benchmarks, but the same situation arises

naturally when using annotation sources of varying quality. This can strongly affect accuracy, but may not impact model comparisons as we explore further in Chapter 9.

7.4 Model rankings in the ImageNet era

ImageNet is best known as the benchmark that kicked off the deep learning revolution during the decade following 2012. When people say ImageNet, they usually refer to the dataset that was released for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. The competition was organized yearly from 2010 until 2017 to measure progress in computer vision.[10]

In 2012, a deep convolutional neural network called AlexNet achieved top-5 error 15.3% on the ILSVRC-2012 dataset, vastly outperforming the second best contender that came in at 26.2%.[11] Created by Krizhevsky, Sutskever, and Hinton, effectively using GPU hardware, the model that beat the competition demonstrated that end-to-end trained convolutional neural networks were the way to go for computer vision. The competition heated up quite a bit over the years to follow.

Around 2015, major companies like Google, Microsoft, and Baidu had deep learning teams working on taking first place. It was the first time the rivalry between the United States and China over AI development played out in such a public manner. In a now largely forgotten controversy, Baidu allegedly made too many queries to the test set to gain an advantage in the competition. The ILSVRC organizers disqualified Baidu from the 2015 competition and banned the company from competing in the 2016 rendition. Microsoft went on to win the 2015 competition with Kaiming He's deep residual networks (ResNets), getting the top-5 error down to 3.57%. ResNets were here to stay and would become one of the crown jewels of the ImageNet era.

The ILSVRC-2012 data features roughly 1.3 million labeled images from 1000 different classes. The test size contains 50000 instances, that means only 50 per class. It's a fairly idiosyncratic dataset. Of the 1,000 classes in ILSVRC-2012, just three are about people: groom, baseball player, and scuba diver. On the other hand, there 118 are dog breeds. The reasons for these choices aren't entirely clear, though they reflect the field's focus at the time on handling fine-grained classification tasks. Although ILSVRC-2012 is kind of quirky, researchers learned that representations trained on the

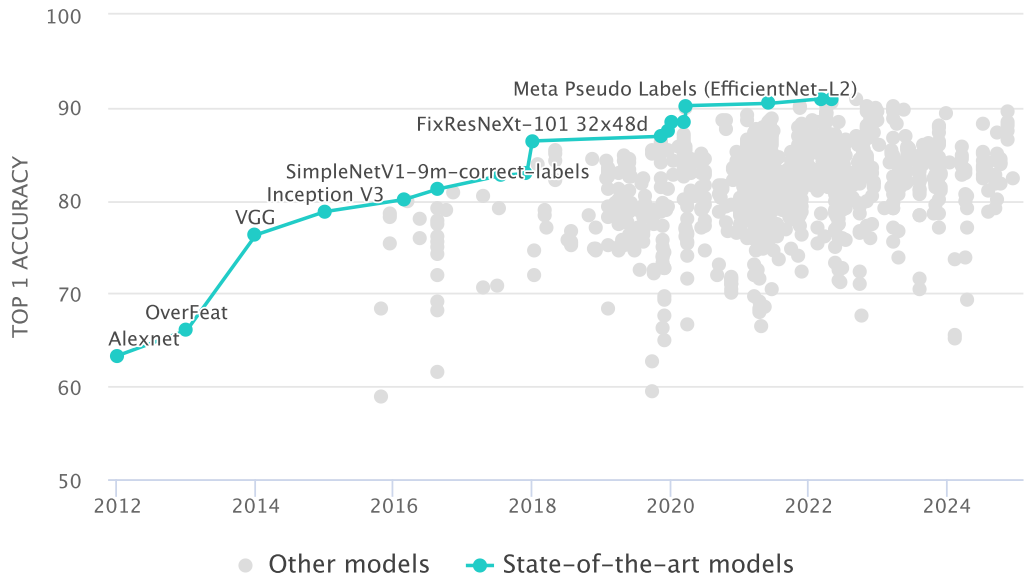

Figure 7.1: The ILSVRC2012 test set supported over a decade of active model development according to PapersWithCode

dataset make good features for other tasks.

Long after the official competition was over, the academic leaderboard on ImageNet remained the central benchmark in the field. When attention eventually moved elsewhere in 2022, ILSVRC-2012 had supported over a decade of intense model development.

The competition dataset is only a small part of the larger ImageNet database that is a vast collection of millions of human-annotated images. The class labels come from WordNet, a database of English nouns grouped into synonyms, or *synsets*. For instance, car and automobile belong to the same synset. WordNet also arranges these synsets into a hierarchy: chair falls under furniture, for example. WordNet existed before ImageNet and partly inspired its creation. Scale was an important aspect of ImageNet from the get go. Introduced in 2009 by Fei-Fei Li and her team at Princeton University, the first version of ImageNet contained around 5,000 synsets with an average of 600 images per category.[12] By 2011, ImageNet had grown to 32,000 categories.

ImageNet creation

Building ImageNet involved two key components: gathering candidate images for each class and using gig workers to label them. For the first step, creators used image search engines like Flickr. For labeling, they turned to Amazon's Mechanical Turk (MTurk), an online labor market where workers reviewed images and decided whether they matched the assigned category. It's worth taking a closer look at the ImageNet creation process. Annotating millions of images efficiently was a daunting test, carried out in roughly four steps:

1. Candidate set creation: For each synset noun, create a set of candidate images by searching for the noun on image search engines.
2. Present MTurk workers with a grid of candidate images for each class. Let them select the ones they think belong to the class.
3. Retain all images above a certain *selection frequency* corresponding to the fraction of turkers choosing each image for inclusion.
4. Remove near duplicates.

Note that workers didn't just label random images on the internet. They labeled images that were already selected as candidates. To be more precise, they didn't label them in the sense of coming up with a description or class name themselves. Workers confirmed if a candidate image belonged to a candidate category. Candidate images were then selected based on annotator agreement. This double selection process is what gives ImageNet its characteristic look.[13] Images typically represent the category front and center. Due to the many similar classes of dog breeds, however, the dataset is relatively hard for untrained human observers.

ImageNet test set replication

The intense use and competitive pressure that ImageNet experienced over the course of a decade begs a question: Can we trust the reported progress on ImageNet? Or did researchers led themselves astray by overfitting to ImageNet.

To gain insight into this question, researchers carefully created a new test set for ImageNet by re-executing the dataset creation process for the CIFAR-10 and ImageNet ILSVRC 2012 test sets.[14] With the new test sets at hand, the team of researchers evaluated a slew of ImageNet era models on the new test instances.

The two primary findings were:

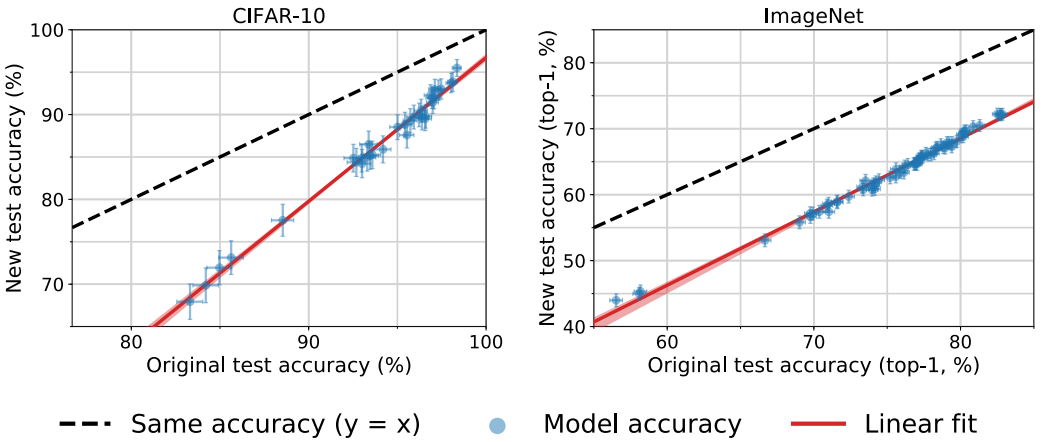

Figure 7.2: Scatter plot of model accuracies on the original test sets vs. new test sets for CIFAR-10 and ImageNet. Each data point corresponds to one model from a set of representative models. Source: Recht, Roelofs, Schmidt, Shankar (2019)

1. Accuracy numbers drop significantly between the old and the new test set.
2. Model rankings are largely preserved.

Even though the researchers took care to create a dataset in the same way that ImageNet ILSVRC 2012 came about, the differences were stark enough to cause accuracy numbers to drop sharply. This confirms that accuracy numbers don't even replicate under a serious attempt to sample from the same distribution twice.

The findings are best summarized in a scatter plot that shows for each model its performance on the old test set (on the x-axis) versus the new test set (on the y-axis). The main diagonal corresponds to equal performance. Points below the line perform worse on the new test set than the old one. We've seen such a scatter plot before at the beginning of Chapter 6, where we looked at replication efforts on psychology papers. This one looks quite different.

The scatter plot has a curious property: Models seem to cluster around a line. In fact, this *on the line* phenomenon appears in many different contexts.[15] It is not limited to test set reconstructions, but appears more generally when models trained on one distribution are evaluated in another. One important caveat is that models must all be trained on the same training data. The phenomenon breaks down if models were trained on different datasets.

The slope of the line fit has interesting implications:

- Slope > 0: The highest accuracy model on old data is also highest accuracy on new data. The relationship between in and out of domain accuracy is strictly monotone. This means model rankings are preserved!
- Slope > 1: The higher the model's accuracy, the less its accuracy drops on the new domain.

A positive slope occurs in almost all cases. This makes a powerful case for the external validity of model rankings. Whenever the *on the line* phenomenon occurs with a positively sloped line, model rankings must be preserved. Any monotone relationship imply the same. It so happens to be a line of positive slope.

The meaning of a slope greater than 1 is a bit more subtle. It doesn't always happen, but the slope for ImageNet is just a bit larger than one. A line of slope greater than one is evidence against *adaptive overfitting*. It means that newer models—that achieve higher accuracy and had more time to overfit—show a smaller difference between old and new test accuracy.

In a similar study, Yadav and Bottou's *Cold Case: The Lost MNIST Digits* revisited the MNIST digits benchmark by recovering and testing on previously discarded samples from the original NIST dataset used to construct MNIST.[3] It's hard to imagine a benchmark that's considered more overused than MNIST. Researchers have long considered MNIST a solved problem that is now little more than a "unit test" to see if software runs. Nevertheless, the scatter plot for the lost digits has the same positively sloped linear trend as for CIFAR and ImageNet. In particular, model rankings are roughly preserved.

To give one more example, a similar observation is true for Kaggle competitions. In a meta study of numerous Kaggle competitions, scores on the public leaderboard exhibited a strong positive correlation with scores on the private leaderboard. Participants in Kaggle competitions routinely worry about overfitting to the public leaderboard. Problems with competitions certainly occur. But these problems generally seem to be more about failures with the competition setup—data splits, target variable definition, loss functions, etc.—than with overfitting to the leaderboard.

What has emerged from these empirical findings is a certain internal validity of iron rule: Competitive testing yields model rankings that routinely replicate in sufficiently similar conditions.

ImageNot and external validity

We can go a step further and ask if model rankings transfer from ImageNet to other datasets. The answer turns out to be, yes, again. Kornblith, Shlens, and Le[16] showed that progress on ImageNet transfers to other computer vision benchmarks, such as CIFAR-10 and CIFAR-100, Pascal Visual Object Classes (VOC) benchmark, SUN397, and Caltech-101. Models transfer particularly well when trained on ImageNet and fine-tuned on other datasets. But even training from scratch yields a positive correlation between ImageNet accuracy and accuracy on other datasets.

In another test of external validity, researchers created ObjectNet[17], an object detection benchmark consisting of 50000 carefully selected test images. Like ImageNet, the images show an object front and center, but unlike ImageNet the images feature transformations like rotations, occlusions, and variations in lighting. This results in a large accuracy drop going from ImageNet to ObjectNet. Still, the model rankings are preserved.

On further thought, you might register an objection. These datasets are still all alike in one important characteristic. These were all reasonable computer vision benchmarks, carefully created by experts. Perhaps things break down on a less reasonable, less curated dataset.

To test this possibility, researchers created a toy dataset called ImageNot that matches the scale and class diversity of ImageNet but differs in every other regard. Whereas ImageNet was carefully curated with multiple annotators per image, ImageNot is based on a quick and dirty web crawl. ImageNot is not only a test set, but comes with one million training points matching the size of the ImageNet ILSVRC 2012 training set.

In some more detail, the steps in creating ImageNot were:

1. Pick 1000 arbitrary classes while avoiding all ImageNet synsets in the WordNet hierarchy.
2. Images come from the LAION-5B image database that features nearly six billion image-caption pairs crawled from the internet. Select images based on the similarity between the class name and the caption, according to a RoBERTa text embedding.
3. Implement some additional safety filters.

Importantly, no annotators filtered the images or cleaned the labels. As a result, the images in ImageNot are all over the place.

Retraining several key ImageNet era model architectures on ImageNot reveals that the model rankings are preserved. The model rankings on

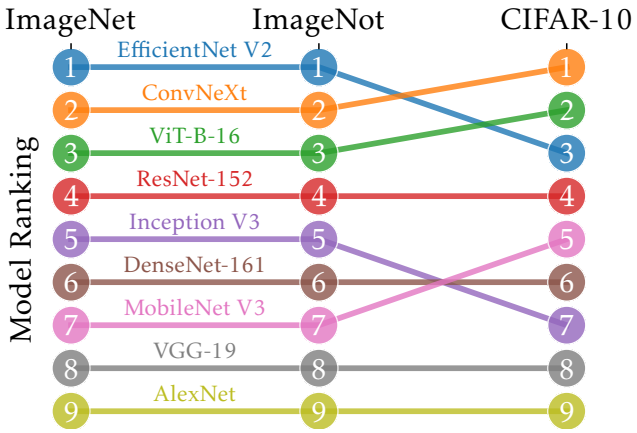

Figure 7.3: Model rankings are preserved between ImageNet and ImageNot

ImageNet and ImageNot turn out to be exactly the same: 1. EfficientNet[18] (2019), 2. ConvNeXt[19] (2022), 3. ViT-B-16 Vision Transformers[20] (2020), 4. Residual Networks[21] (2016), 5. Inception V3[22] (2015), 6. DenseNet[23] (2017), 7. MobileNet V3 Large[24] (2019), 8. VGG[25] (2014), 9. AlexNet[11] (2012).

What's perhaps more striking is that the *relative* improvement from one model to the next are also roughly the same. This adds a quantitative dimension to the stability of model rankings. Relative improvements from one model to the next also replicate.

The surprising transfer ability of model rankings and relative improvements suggests that there is a certain external validity of the iron rule: If you beat the previous best under sufficiently general conditions, it will likely replicate elsewhere. The ImageNot experiment tests how far we can stretch the external validity of model rankings. But apart from running such empirical tests, there is little we know about the external validity of model rankings.

7.5 Lessons learned

The statistical scientific crisis is real and it affects machine learning, too. But something a bit different has been happening in machine learning. We can summarize the key empirical findings in a few points:

1. Absolute benchmark numbers, such as accuracy numbers, may satisfy the minimal bar of code re-execution. They also satisfy the theoretical

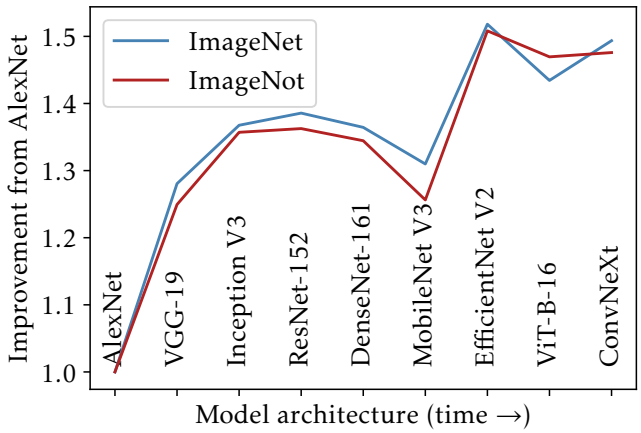

Figure 7.4: Relative accuracy improvements from model to the next are approximately the same on ImageNet and ImageNot.

guarantee of replication under i.i.d.-sampling (Chapter 4). But they appear to give little more than that. Even seemingly benign forms of distribution shift change absolute numbers significantly.

2. Absolute benchmark numbers, in particular, do *not* have external validity. They do not replicate under significant variation in testing conditions.
3. Relative model comparisons and model rankings in the ImageNet era robustly satisfy internal validity. Model rankings replicate under reasonable attempts to recreate original testing conditions.
4. Model rankings are sometimes stable even under major dataset variation. In particular, they show signs of external validity. The scope of external validity is not fully known.

Measurement versus ranking

The empirical reality highlights the distinction between quantitative measurement of model abilities versus model rankings. The first is measurement on a *cardinal* scale; the other is measurement on an *ordinal* scale. Absolute benchmark numbers, construed as cardinal measurements of a model's latent abilities, lack validity. Model rankings, on the other hand, fare much better.

We don't fully understand why. But one important aspect is that measurement and ranking behave differently under competitive pressure. Goodhart's law applies to cardinal measurement. Statistical measurement on a cardinal

scale fails under competition. This is the reason the anthropologist Marilyn Strathern paraphrased Goodhart's law as: *When a measure becomes a target, it ceases to be a good measure.* This is true for cardinal measurement.

Rankings, in contrast, are robust to competition. That's sort of the point of ranking and it's not hard to make this formal. Suppose each participant in a competition has some true quality. What we observe when we attempt to measure quality, however, is the true quality in addition to the effort the participant put into beating the competition. If each participant puts in the same effort, then the ranking reveals the correct ordering by underlying quality. The numbers are off, but the ranking is fine.

Benchmarks make a virtue out of gaming the metric. Rather than fighting Goodhart's law, they lean into it. This has the consequence that absolute numbers on benchmarks are meaningless. But on the flip side, if everyone exhaustively competes over the leaderboard, things can still work out: Benchmarks can correctly identify the best model at any point in time. The next chapter continues on this thread.

Notes

Liu and He repeated the Torralba and Efros game on modern computer vision benchmarks in 2024.[5] The main takeaway is the same. It's still feasible to predict which dataset an image comes from with high accuracy. Using state-of-the-art neural networks, they achieved over 80% accuracy in distinguishing images from three large collections (YFCC100M, CC12M, and the DataComp 1B dataset). This is even higher than the 2011 study's results, despite these newer datasets being far larger and more varied. It confirms that each dataset—even lightly curated ones scraped from the web—still has a unique "signature." Zeng, Yin, and Liu continue the investigation arguing that the classifiers find meaningful differences in these datasets.[6] Mansour and Heckel report similar findings for text datasets.[26]

For a book covering early work on dataset shift and domain generalization, see *Dataset shift in machine learning*.[27] There are many classical theory results in this area.[28–30]

The fact that nothing beats plain empirical risk minimization for domain generalization is broadly true in different contexts, including the DomainBed[7] benchmark by Gulrajani and Lopez-Paz and the WILDS benchmark by Koh et al.[8] Wild-time is a related benchmark that tests resilience to

temporal distribution shift.[31] Causal methods that were specifically designed to mitigate distribution shift also do not outperform standard empirical minimization.[32,33]

Recht, Roelofs, Schmidt, and Shankar conducted the ImageNet replication study.[14] In particular, they found that a selection frequency of 0.73 most closely resembled the original test set. The new test set became known as ImageNet V2 and has been used in many papers on distribution shift since. There's a lot of related work on ImageNet, in particular. Tsipras et al.[34] discuss the alignment of the ImageNet benchmark with real-world image classification, focusing on the difference between selecting one class label per image and annotating all objects present in the image. Fang, Kornblith, and Schmidt ask whether progress on ImageNet transfers to the real-world.[35] *Are we done with ImageNet?*, Beyer et al. ask, and collect new labels for ImageNet to see how much accuracy drops.[36] Feuer et al. compare different data curation strategies for ImageNet.[37]

Yadav and Bottou[3] use the MNIST lost digits for a similar replication study on MNIST. Surprisingly, even on MNIST, a much smaller and older benchmark, the findings are similar. Roelofs et al.[38] conduct a meta study of overfitting in Kaggle competitions, showing a strong correlation between public leaderboard and the final private leaderboard. Miller et al. conduct another such replication study on the Stanford Question Answering Dataset (SQuAD), a population natural language processing benchmark, and confirm the findings in this domain. Miller's dissertation has a few of these results and provides additional background.[39] LeJeune, Liu, and Heckel prove the monotonicity of risk relationships under distribution shift in a theoretical model.[40]

The linear relationship between in-domain and out-of-domain performance breaks down when models were trained on different training data. This first became apparent when the CLIP model was released that appeared off "off the line" out-of-domain performance. This advantage is due to greater diversity in the training data.[41,42] We'll return to this topic in Chapter 10 in the context of generative models.

Taori et al. measure robustness to natural distribution shifts in image classification.[43] They found that most specialized robustness improvements (like training on corrupted images, or adversarial defenses) did *not* transfer better to these natural shifts. Moreover, methods that help with one type of shift (say, noise corruption) often don't help with others. The one factor that consistently improved robustness was using more data: models pretrained

on massive, diverse datasets tended to maintain higher accuracy under shifts. Geirhos et al. compare the robustness of humans with computer vision models on out-of-domain vision tasks.[44] Ramanugan et al. study the connection between pre-training data diversity and fine-tuning robustness.[44]

Huh, Agrawal, and Efros evaluated what makes ImageNet good for transfer learning.[45] Kornblith, Shlens, Le[16] demonstrated the value of ImageNet as a basis for transfer learning. Salaudeen and Hardt created ImageNot building on the data-generating process of Shirali and Hardt,[13] who studied the question what makes ImageNet different from LAION.[46] Kataoka et al. demonstrate that pre-training image classification models doesn't necessarily need *natural* images.[47]

The argument in this chapter is not that there aren't any problems with replication and reproducibility in machine learning. Indeed, several works have pointed out serious problems. A 2018 article in *Science* argued that AI faces a reproducibility crisis, noting that AI studies often come without published code or with incomplete implementation details, making it "hard to verify" many claims.[2] Henderson et al. examined deep reinforcement learning algorithms and showed their performance can vary wildly due to seemingly minor choices—random seeds, environment stochasticity, or evaluation protocols—to the point that conclusions in some RL papers were unreliable.[48] They argued for higher standards, like running multiple random seeds and reporting variance.

Several *meta-studies* have tried to quantify the scope of the problem. Gundersen & Kjensmo (2018) reviewed dozens of papers from top AI conferences and scored them against reproducibility criteria.[49] They found many papers left out crucial details like hyperparameters or even a clear problem statement. Only a single-digit percentage of papers explicitly stated which software versions they used, and few provided enough information for exact re-execution. Pineau et al. (2020) report on the NeurIPS Reproducibility Challenge, in which volunteers attempted to replicate results of submitted papers.[50] While many results do reproduce (with effort), there are also cases where reproductions fall short of reported performance, or where key baselines were missing. An annual venue called MLRC continues the effort. Researchers at OpenAI attempted to use LLM agents to replicate 20 machine learning research papers from ICML 2024.[51] The best agents still struggled with the task; machine learning researchers did better with effort.

Kapoor and Narayanan[52] add evidence that machine learning is not immune to the replication crisis. They focus on the issue of data leakage from

test sets intro training sets, and show how it invalidates numerous findings that apply machine learning to scientific problems. In a similar vein, Leech et al. report on questionable practices in machine learning.[53] The study focuses on the categories of contamination, cherry picking, and misreporting, discussing several of the themes from the previous chapter, such as researchers degrees of freedom.

Bouthillier, Laurent, Vincent draw a distinction between the reproducibility of numerical results and the reproducibility of scientific claims under different sources of variation. While the former kind may have improved, the latter not necessarily.[54] Liao et al.[55] discuss the distinction between internal and external validity in the context of machine learning benchmarks, providing a taxonomy of different aspects the two notions of validity. In a similar vein, Liao, Taori, and Schmidt argue why external validity matters for machine learning research.[56]

The study of Goodhart's law in the context of machine learning goes back to work on *strategic classification*.[57] See Rosenfeld's tutorial for pointers.[58] In economics, related problems fall under the umbrella of principal-agent problems.

Bibliography

1. Cortes, C. & Lawrence, N. D. Inconsistency in conference peer review: Revisiting the 2014 NeurIPS experiment. *arXiv:2109.09774* (2021) (↑ 2).
2. Hutson, M. Artificial intelligence faces reproducibility crisis. *Science* **359,** 725–726 (2018) (↑ 3, 19).
3. Yadav, C. & Bottou, L. *Cold case: The lost MNIST digits* in *Neural Information Processing Systems (NeurIPS)* (2019) (↑ 4, 13, 18).
4. Torralba, A. & Efros, A. A. *Unbiased look at dataset bias* in *Conference on Computer Vision and Pattern Recognition (CVPR)* (2011), 1521–1528 (↑ 5).
5. Liu, Z. & He, K. A Decade's Battle on Dataset Bias: Are We There Yet? *arXiv:2403.08632* (2024) (↑ 5, 17).
6. Zeng, B., Yin, Y. & Liu, Z. Understanding Bias in Large-Scale Visual Datasets. *arXiv:2412.01876* (2024) (↑ 5, 17).
7. Gulrajani, I. & Lopez-Paz, D. In search of lost domain generalization. *arXiv:2007.01434* (2020) (↑ 8, 17).
8. Koh, P. W. *et al. Wilds: A benchmark of in-the-wild distribution shifts* in *International Conference on Machine Learning (ICML)* (2021), 5637–5664 (↑ 8, 17).
9. Sutton, R. The bitter lesson. *Incomplete Ideas (blog)* **13,** 38 (2019) (↑ 8).
10. Russakovsky, O. *et al.* Imagenet large scale visual recognition challenge. *International journal of computer vision* **115,** 211–252 (2015) (↑ 9).
11. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *Imagenet classification with deep convolutional neural networks* in *Neural Information Processing Systems (NeurIPS)* (2012) (↑ 9, 15).
12. Deng, J. *et al. ImageNet: A large-scale hierarchical image database* in *Conference on Computer Vision and Pattern Recognition (CVPR)* (2009), 248–255 (↑ 10).
13. Shirali, A. & Hardt, M. What makes ImageNet look unlike LAION? *arXiv:2306.15769* (2023) (↑ 11, 19).
14. Recht, B., Roelofs, R., Schmidt, L. & Shankar, V. *Do imagenet classifiers generalize to imagenet?* in *International Conference on Machine Learning (ICML)* (2019), 5389–5400 (↑ 11, 18).
15. Miller, J. P. *et al. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization* in *International Conference on Machine Learning (ICML)* (2021), 7721–7735 (↑ 12).
16. Kornblith, S., Shlens, J. & Le, Q. V. *Do better imagenet models transfer better?* in *Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 2661–2671 (↑ 14, 19).

17. Barbu, A. *et al. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models* in *Neural Information Processing Systems (NeurIPS)* (2019) (↑ 14).
18. Tan, M. & Le, Q. *Efficientnet: Rethinking model scaling for convolutional neural networks* in *International Conference on Machine Learning (ICML)* (2019), 6105–6114 (↑ 15).
19. Liu, Z. *et al. A convnet for the 2020s* in *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 11976–11986 (↑ 15).
20. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929* (2020) (↑ 15).
21. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition* in *Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 770–778 (↑ 15).
22. Szegedy, C. *et al. Going deeper with convolutions* in *Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 1–9 (↑ 15).
23. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. *Densely connected convolutional networks* in *Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 4700–4708 (↑ 15).
24. Howard, A. *et al. Searching for MobileNetv3* in *Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 1314–1324 (↑ 15).
25. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556* (2014) (↑ 15).
26. Mansour, Y. & Heckel, R. Measuring Bias of Web-filtered Text Datasets and Bias Propagation Through Training. *arXiv:2412.02857* (2024) (↑ 17).
27. Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A. & Lawrence, N. D. *Dataset shift in machine learning* (MIT Press, 2009) (↑ 17).
28. Ben-David, S., Blitzer, J., Crammer, K. & Pereira, F. *Analysis of representations for domain adaptation* in *Neural Information Processing Systems (NeurIPS)* (2006) (↑ 17).
29. Mansour, Y., Mohri, M. & Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. *arXiv:0902.3430* (2009) (↑ 17).
30. Ben-David, S. *et al.* A theory of learning from different domains. *Machine learning* **79,** 151–175 (2010) (↑ 17).
31. Yao, H. *et al. Wild-time: A benchmark of in-the-wild distribution shift over time* in *Neural Information Processing Systems (NeurIPS)* (2022), 10309–10324 (↑ 18).
32. Rosenfeld, E., Ravikumar, P. & Risteski, A. The risks of invariant risk minimization. *arXiv:2010.05761* (2020) (↑ 18).
33. Nastl, V. Y. & Hardt, M. *Do causal predictors generalize better to new domains?* in *Neural Information Processing Systems (NeurIPS)* (2024) (↑ 18).
34. Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A. & Madry, A. *From ImageNet to image classification: Contextualizing progress on benchmarks* in *International Conference on Machine Learning (ICML)* (2020), 9625–9635 (↑ 18).
35. Fang, A., Kornblith, S. & Schmidt, L. *Does progress on ImageNet transfer to real-world datasets?* in *Neural Information Processing Systems (NeurIPS)* (2024) (↑ 18).
36. Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X. & Oord, A. v. d. Are we done with ImageNet? *arXiv:2006.07159* (2020) (↑ 18).
37. Feuer, B. *et al. SELECT: A Large-Scale Benchmark of Data Curation Strategies for Image Classification* in *Neural Information Processing Systems (NeurIPS)* (2024), 136620–136645 (↑ 18).
38. Roelofs, R. *et al. A meta-analysis of overfitting in machine learning* in *Neural Information Processing Systems (NeurIPS)* (2019) (↑ 18).

39. Miller, J. P. *Validity Challenges in Machine Learning Benchmarks* (University of California, Berkeley, 2022) (↑ 18).
40. LeJeune, D., Liu, J. & Heckel, R. Monotonic risk relationships under distribution shifts for regularized risk minimization. *Journal of Machine Learning Research* **25,** 1–37 (2024) (↑ 18).
41. Fang, A. *et al. Data determines distributional robustness in contrastive language image pre-training (clip)* in *International Conference on Machine Learning (ICML)* (2022), 6216–6234 (↑ 18).
42. Ramanujan, V., Nguyen, T., Oh, S., Farhadi, A. & Schmidt, L. *On the connection between pre-training data diversity and fine-tuning robustness* in *Neural Information Processing Systems (NeurIPS)* (2023), 66426–66437 (↑ 18).
43. Taori, R. *et al. Measuring robustness to natural distribution shifts in image classification* in *Neural Information Processing Systems (NeurIPS)* (2020), 18583–18599 (↑ 18).
44. Geirhos, R. *et al. Partial success in closing the gap between human and machine vision* in *Neural Information Processing Systems (NeurIPS)* (2021), 23885–23899 (↑ 19).
45. Huh, M., Agrawal, P. & Efros, A. A. What makes ImageNet good for transfer learning? *arXiv:1608.08614* (2016) (↑ 19).
46. Salaudeen, O. & Hardt, M. ImageNot: A contrast with ImageNet preserves model rankings. *arXiv:2404.02112* (2024) (↑ 19).
47. Kataoka, H. *et al. Pre-training without natural images* in *Proc. Asian Conference on Computer Vision* (2020) (↑ 19).
48. Henderson, P. *et al. Deep Reinforcement Learning that Matters* in *AAAI Conference on Artificial Intelligence* (2018), 3207–3214 (↑ 19).
49. Gundersen, O. E. & Kjensmo, S. *State of the Art: Reproducibility in Artificial Intelligence* in *AAAI Conference on Artificial Intelligence* (2018), 1644–1651 (↑ 19).
50. Pineau, J. *et al.* Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research* **22,** 1–20 (2021) (↑ 19).
51. Starace, G. *et al.* PaperBench: Evaluating AI's Ability to Replicate AI Research. *arXiv:2504.01848* (2025) (↑ 19).
52. Kapoor, S. & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **4** (2023) (↑ 19).
53. Leech, G., Vazquez, J. J., Kupper, N., Yagudin, M. & Aitchison, L. Questionable practices in machine learning. *arXiv:2407.12220* (2024) (↑ 20).
54. Bouthillier, X., Laurent, C. & Vincent, P. *Unreproducible research is reproducible* in *International Conference on Machine Learning (ICML)* (2019), 725–734 (↑ 20).
55. Liao, T., Taori, R., Raji, I. D. & Schmidt, L. *Are we learning yet? A meta review of evaluation failures across machine learning* in *NeurIPS Datasets and Benchmarks Track* (2021) (↑ 20).
56. Liao, T. I., Taori, R. & Schmidt, L. *Why external validity matters for machine learning evaluation: Motivation and open problems* 2022 (↑ 20).
57. Hardt, M., Megiddo, N., Papadimitriou, C. & Wootters, M. *Strategic classification* in *Innovations in Theoretical Computer Science (ITCS)* (2016), 111–122 (↑ 20).
58. Rosenfeld, N. *Strategic ML: How to Learn With Data That 'Behaves'* in *ACM International Conference on Web Search and Data Mining* (2024), 1128–1131 (↑ 20).