External Validity and Machine Learning

Lecture Scribe: Kevin Ma —— Instructor: Professor Ben Recht

1 Introduction

Lets first start off with some very high level background of the idea of generalization.

Given: independent identically distributed sample $S = z_1, ..., z_n$ from dist D

The **goal** is to find a good predictor function f, given the generalization error equation below:

$$\text{Generalization Error} = R[f] - R_S[f]$$

where $R[f]$ is the population risk, which is an unknown, and $R_S[f]$ is the empirical risk, which can be minimized using your favorite gradient descent algorithm. Our goal is to find $R[f]$ such that it minimizes the generalization error. Thus, a model that minimizes the generalization error is a model that is likely to perform well on unseen data as it does training data.

So how do we address this issue of minimizing generalization error? One of the old conventional machine learning wisdom is to address the tension between the model's complexity and its ability to learn from the original data and generalize to new data. This principle can be seen as relating to the bias-variance trade off where if a model has high bias (which means the model is too simple), it may be unable to learn complex patterns in the data, leading to high generalization error. However, if a model has high variance (meaning the model is too complex), it may over-fit to the data and perform poorly on new, unseen data, which will also lead to high generalization error. Essentially, in the case of high variance, we may be prone to 'training on the testing data'. This idea is conveyed visually in Figure 1.

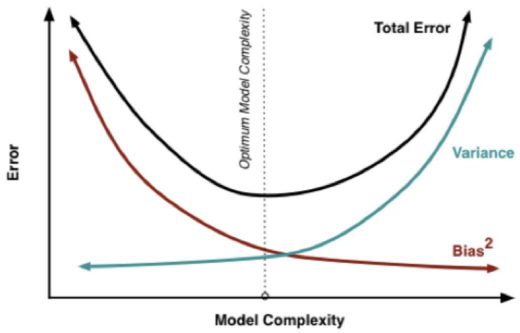

Figure 1: Bias-Variance Trade off Graph

However, according to Ben, this graphical representation of the trend does not tell the whole story. A study in 2019 investigated this phenomenon [3]. In this paper, they studied whether or not higher test accuracy on models carry over to new dataset generated using the same method for ImageNet.

Conventional wisdom would have you suggest that if you tune a model to adapt to specific images in the original test set, this model should generalize poorly to images in a new test set from the same source. However, the models with the highest accuracy on the original test sets were still the models with the highest accuracy on the new test set. This is depicted visually in Figure 2.

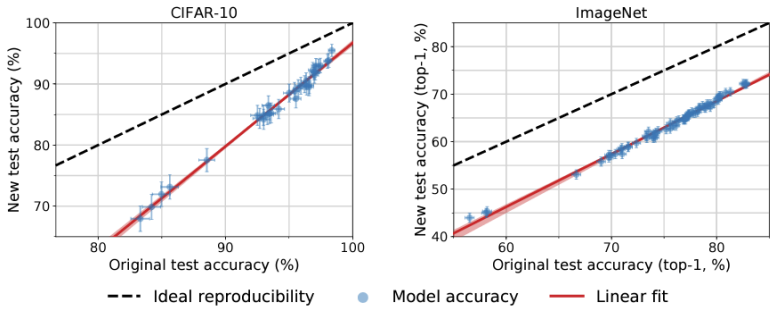

Figure 2: Model accuracy on the original test sets vs their new test set. Each data point corresponds to one model in the test bed. The plot shows two phenomena. There is a significant drop in accuracy from the original test set to the new test set. However, the model's accuracy corresponds well with the original test set vs the new test set.

The primary root cause of this stems in how we collect our data. Throughout the rest of the scribe note, I will provide both Ben's opinions and inject a bit of my opinions. I will denote when an opinion is Ben's and when it is mine throughout the scribe note.

2 A Tour of the Creation of ImageNet and CIFAR-10

In that same paper [3], ImageNet and CIFAR10 was recreated to test whether this idea of a bias-variance trade off is true. To do this, the authors of the paper needed to recreate a new dataset using the same procedure as the original ImageNet and CIFAR10. In the section below, we discuss how the authors originally created each dataset.

2.1 Creation of CIFAR10

CIFAR10 is a labelled image dataset created by Alex Krizhevsky [2], which is a labelled subset of the 80 million tiny images dataset collected by Alex Torralba [4]. The creators of the CIFAR-10 dataset used a two-stage process to create it. First, they used an automatic process to extract candidate images from the 80 million tiny image dataset. Then, they presented these candidate images to student labelers who reviewed and labelled these images to create the final dataset. CIFAR10 researchers would then remove the unsuitable images and remove near duplicates. Afterwards, the dataset was randomly split into class-balanced train and test sets. More details can be found in the original paper. The authors in [3] recreated this test dataset using a very similar procedure; however, apparently on a much smaller scale.

2.2 Creation of ImageNet

ImageNet was another very popular labelled image dataset created by a team at Stanford [1]. The creators first used WordNet to define a list of object categories hierarchically. Images were collected online by querying various search engines to gather a list of URLs of images that are labeled with the corresponding words. They then utilized Amazon Mechanical Turk workers to manually verify and label the images. A recreation of the web interface and task that is assigned to the workers can be seen in Figure 3 below.

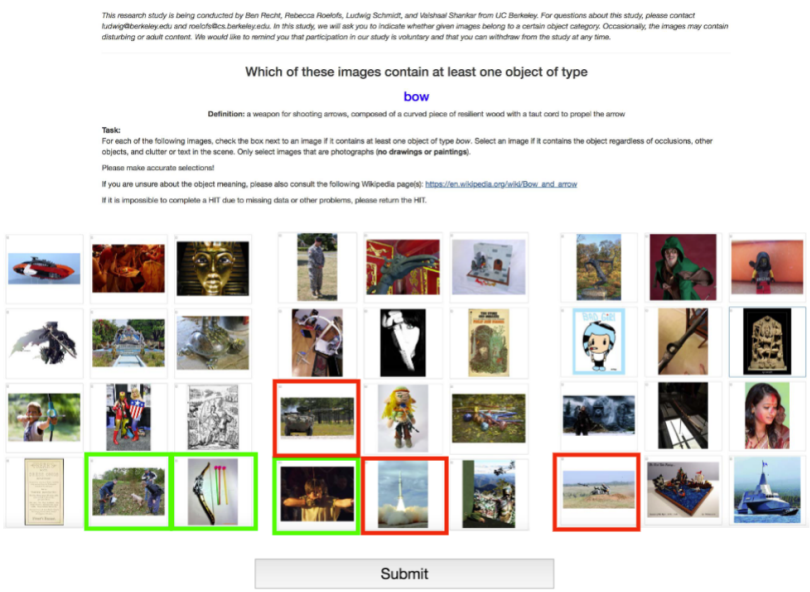

Figure 3: A recreation of the Amazon Turk workers interface for ImageNet.

In order to ensure consistency, the same image may be annotated by multiple workers, and only images that were annotated consistently across different workers were used. When recreating the new ImageNet dataset, the authors in [3] also ensured that there were marginal differences in the percentage of consistency across Amazon Turk Workers.

2.3 Discussion

We reflect back on the results in Figure 2. Recall that each of those datapoints are models from the state-of-the-art machine learning models from 2009 and onwards. As you can probably infer, the less accurate models were models from earlier years and the more accurate models are models from more recent innovations. As we said previously, we would generally expect that a model with super high test accuracy on the original dataset would have lower test accuracy on a new dataset, but this was refuted in the study from [3].

Here are a few key things that Ben mentioned in class that surprised him from these results. (1) Models with higher test accuracy on the original dataset were more resilient to drops when tested on the new dataset, which can be seen in Figure 4.

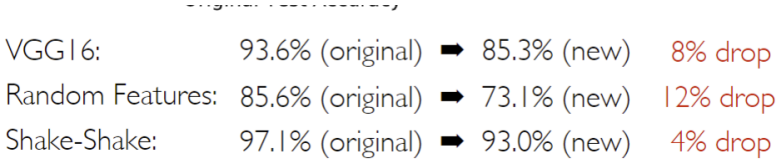

Figure 4: Comparison of the accuracy between different models trained on CIFAR10

(2) The slope was greater than 1, and (3) there was always a decrease in accuracy from the original dataset to the new dataset.

Now **disclaimer**, these are my personal thoughts as to why this is happening and why I do not think these results, while surprising, are bad things. First, I want to address the issue of the decrease in test accuracy. There are a few plausible reasons as to why the test accuracy decreased when trained on the

new dataset, and I think the most plausible one is the fact that there are misaligned labels between two images in the test set and the train set. The reason this happened, and this was briefly addressed in lecture and also in the original paper [3], is because humans are creating these data, and humans are extremely inconsistent. Even if we try to ensure reliability by only keeping annotated images between consistent workers, there is still going to be human errors because of the following reasons:

- **Annotator Bias**: We cannot control the bias of annotators, and we cannot control who the annotators are when it comes to Amazon Turk workers. If a group of annotators happen to share a common bias, the fact that they agree on the same label will mean that an image will be incorrectly labelled.
- **Subjectivity**: Categories are subjective and open to interpretation. Different annotators have different opinions on what a certain category means, and these inconsistencies can propogate into our quality control of the dataset. If several people disagree what a bow means and your description of it, but only one of them has it right, would that not lead to bias especially if our quality control is simply just agreement across multiple AmazonTurk workers?

Again, **disclaimer**, these are just my opinions as to why I think this is happening. The reality is that nobody knows for sure, but we can only make educated guesses. Now for the second part, I want to bring attention back to the fitted straight line of all the trained models on Figure 4. This was a surprising result and broke some conventional wisdom about machine learning theory. However, these results may not be that bad it shows that machine learning models are resilient to variability in data and will generally work well on new data, which can be desirable in many cases.

3 Additional Examples

Ben brought up a few more examples of how the dataset can impact the performance of your prediction along with more examples of why he thinks generalization is not dependent on overfitting or the bias-variance trade off.

3.1 Subimages across timed videos

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) hosted a competition where video datasets curated by ImageNet team in 2015 were annotated across 30 classes, 1314 validation videos, and 4000 training videos (presented as 1M jpeg frames). Something surprising from the results is that small incremental changes in frame can lead to drastic changes in the prediction, which can be seen in Figure 5.

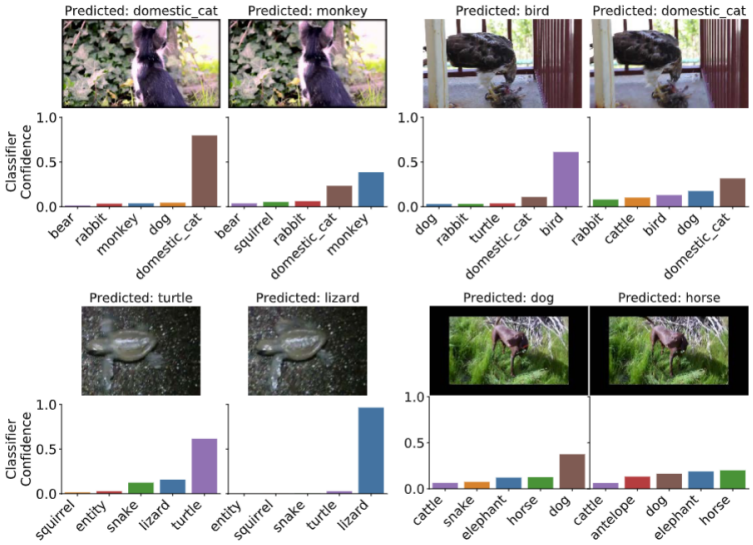

Figure 5: Prediction comparison across frames

3.2 Kaggle

Just as a background, Kaggle is a competition where Kaggle computes two scores, one for the private leaderboard and one for the public leaderboard. Therefore, people who do this competition will tune their model to maximize the accuracy of a public test set, but the winners are the ones who get the best accuracy on a private test set. Figure 6 compares the public score and the private score on a graph.

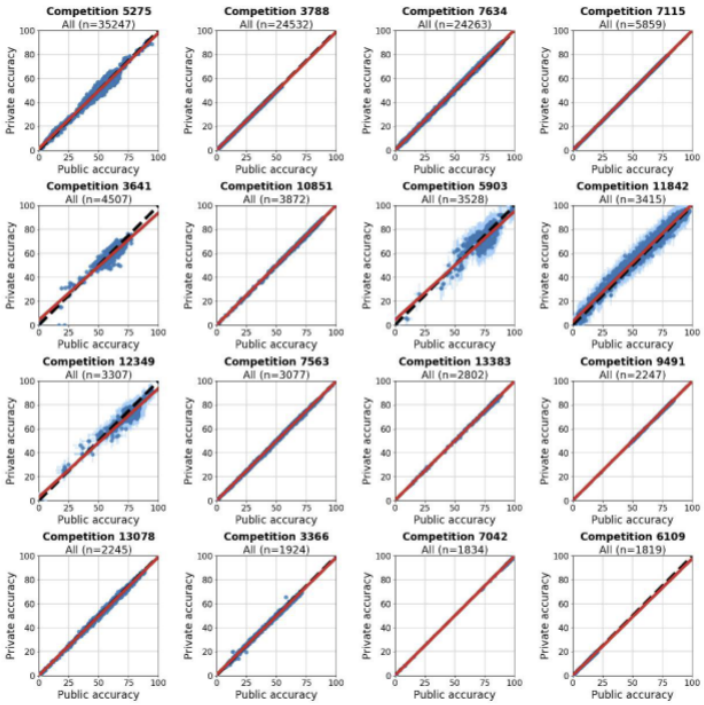

Figure 6: Aggregate of several Kaggle competition plots of the public score versus the private score

What the results show us is that the public scores and private scores have an almost direct correlation with each other, and this contradicts the idea that higher variance models with higher complexity lead to worse generalization.

3.3 Even more examples

Ben brought up even more examples that disproved the bias-variance trade off within the domain of generalization in machine learning. There are examples in MRI Reconstruction, Pose estimation, object detection, natural langauge processing, and more. Please refer to the lecture slides for more information.

4 Distribution Shift

Another question now arises. What if we changed the input data distribution that a model was trained on to something that it was not trained on. In machine learning, this is called distribution shift, and this is a problematic issue because machine learning models assume the training and test data comes from the same distribution. When this assumption is violated, the model's performance degrades because it is trying to apply patterns learned from one context to another context. In general, more data gives more robustness to this issue, which we can see in Figure 7.

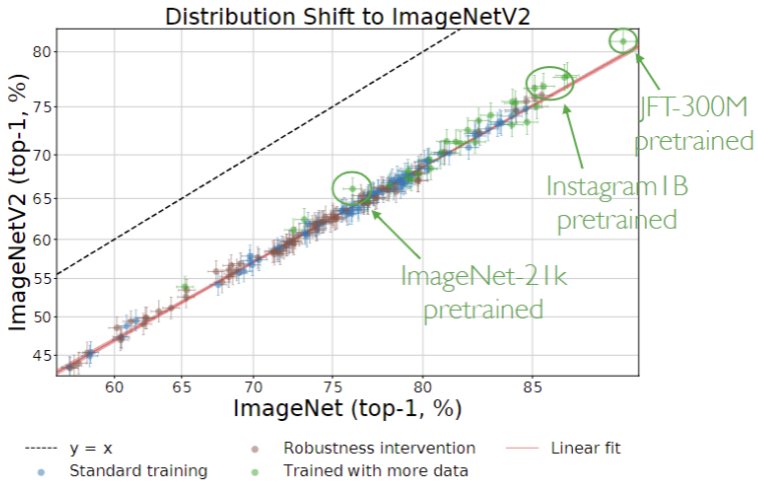

Figure 7: Distribution shift to ImageNetV2 from ImageNet, with some labels on how much data the models are trained on.

So what does this tell us? According to Ben, it basically shows us everything we have seen so far:

- Interpolating your training data is fine.
- Training on your test set is fine.
- Making models huge does not hurt.
- Making models huge does not help very much.
- As a result, bigger models have diminishing returns and is very wasteful
- Overall, distribution shift is real and very dangerous.

The final takeaway is that distribution shift is very dangerous and does happen, and this is most evident in critical safety situations, such as medicine and automotive. As stated in lecture regarding distribution shift: "accuracy numbers from test sets are notably brittle and susceptible to even minute natural variations in the data distribution". In his final slide, he states we need to reorient about how we talk about machine learning before and figure out a better way forward. I will add on top of that

statement by saying that we need to always ask ourselves, especially in safety-critical issues, "why are we using machine learning"? and "are there better more deterministic alternatives"?. In my opinion, if you cannot answer the first one or there are better deterministic solutions, you may not want to jump to the conclusion that machine learning is the best solution to your problem.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[2] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[3] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[4] Antonio Torralba, Rob Fergus, and William T Freeman. Tiny images. 2007.