BIG ARCHITECTURES AND OVERFITTING THE TEST SET

Alireza Shafaei

MACHINE LEARNING READING GROUP -- DECEMBER 2019

Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht*, Rebecca Roelofs, Ludwig Schmidt, Vaishaal Shankar

*Authors ordered alphabetically. Ben did none of the work.

What?

- We've made a substantial amount of progress on benchmark vision classification problems over the past decade.
- We are effectively tuning the hyper-params on the test sets.
- Are we actually making progress on these problems or just overfitting to the *specific test sets*?

How?

- Create new test sets for *ImageNet* and *CIFAR10* following the same procedures as the original work as much as possible.
- Evaluate all the previous models on the new test.

How?

- Create new test sets for *ImageNet* and *CIFAR10* following the same procedures as the original work as much as possible.
- Evaluate all the previous models on the new test.

- **CIFAR10** accuracy drops range from **3% to 15%**.
- **ImageNet** accuracy drops range from **11% to 14%**.
  - Amounts to about 5 years of progress.

Why? (Discuss)

- Are we overfitting to the test set?
- Is the new test set not similar enough in distribution to the original?

Why?

- Because the models have *adapted* to the specific samples in the test sets?
- **But**,
  - The relative order is almost exactly preserved!
  - For every 1% improvement on the original >1% improvement on the new set!
- Does adapting to a specific set necessarily contradict these observations?
  - If we don’t have these properties, we are adapting for sure.
  - But does having these properties mean we are not adapting?

Why?

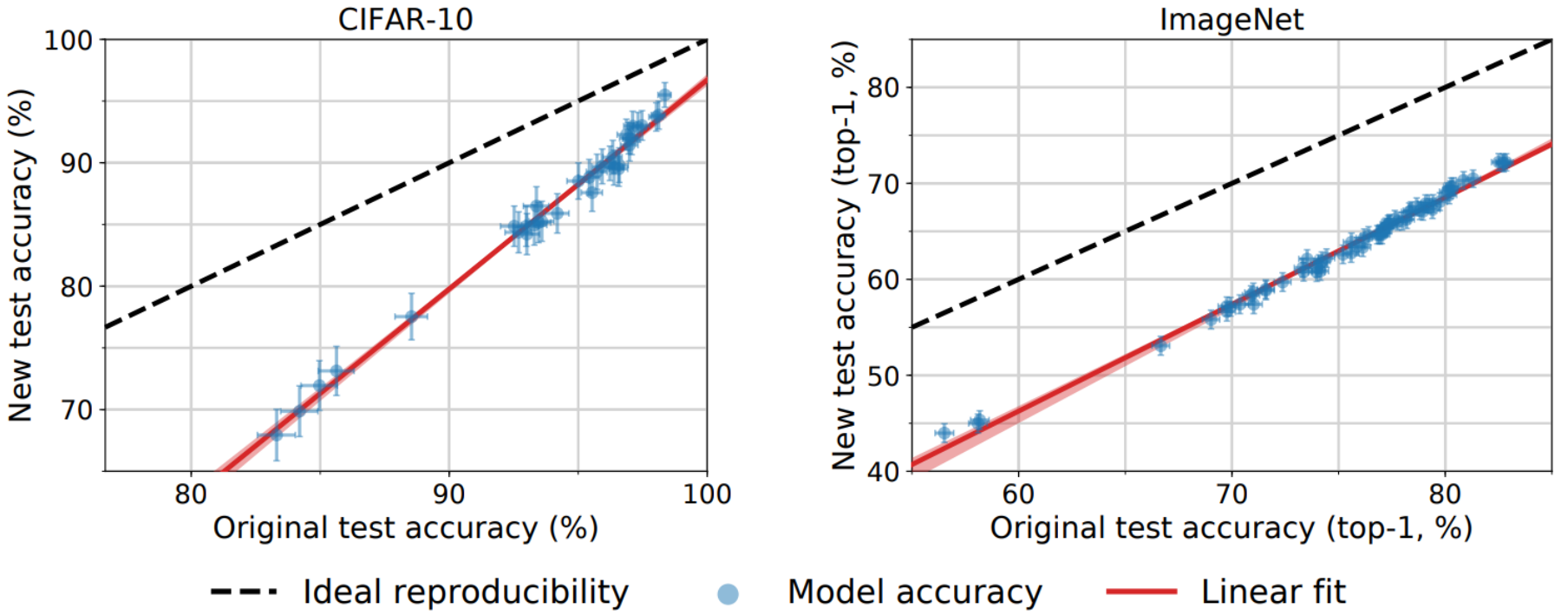


Figure 1: Model accuracy on the original test sets vs. our new test sets. Each data point corresponds to one model in our testbed (shown with 95% Clopper-Pearson confidence intervals). The plots reveal two main phenomena: (i) There is a significant drop in accuracy from the original to the new test sets. (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.

Why though?

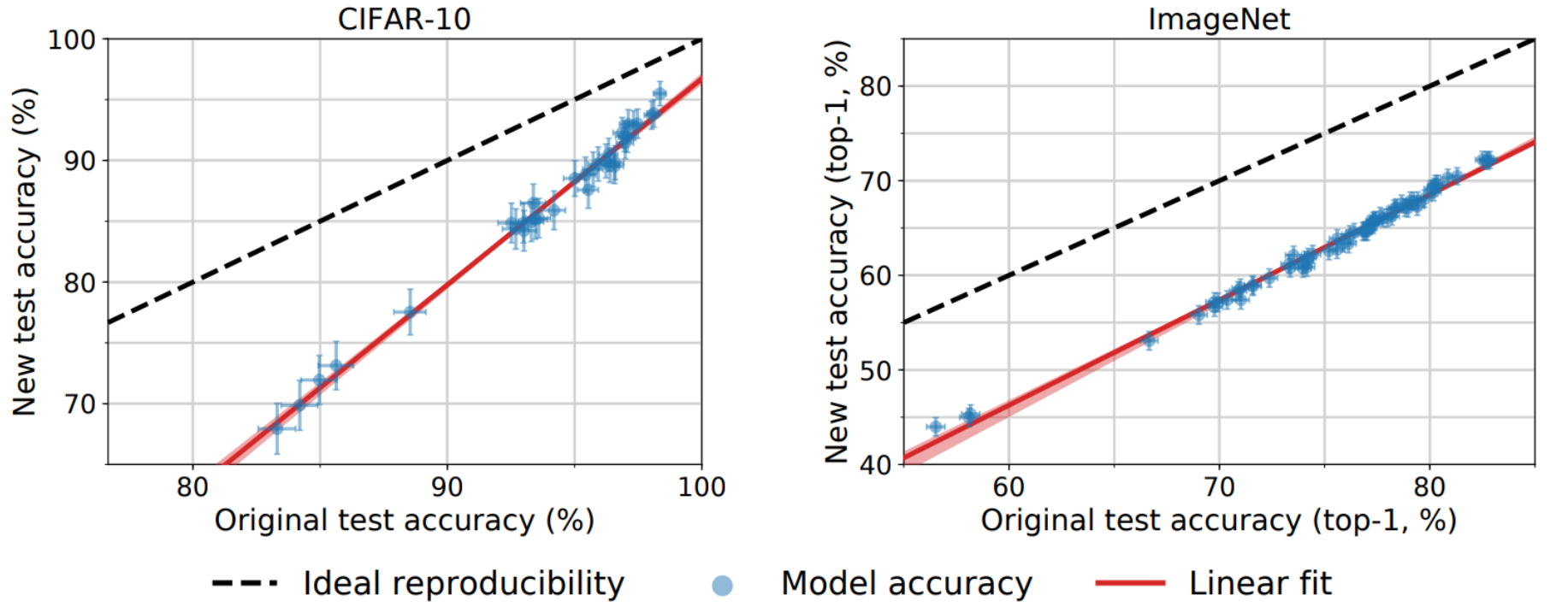


Figure 1: Model accuracy on the original test sets vs. our new test sets. Each data point corresponds to one model in our testbed (shown with 95% Clopper-Pearson confidence intervals). The plots reveal two main phenomena: (i) There is a significant drop in accuracy from the original to the new test sets. (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.

Why though?

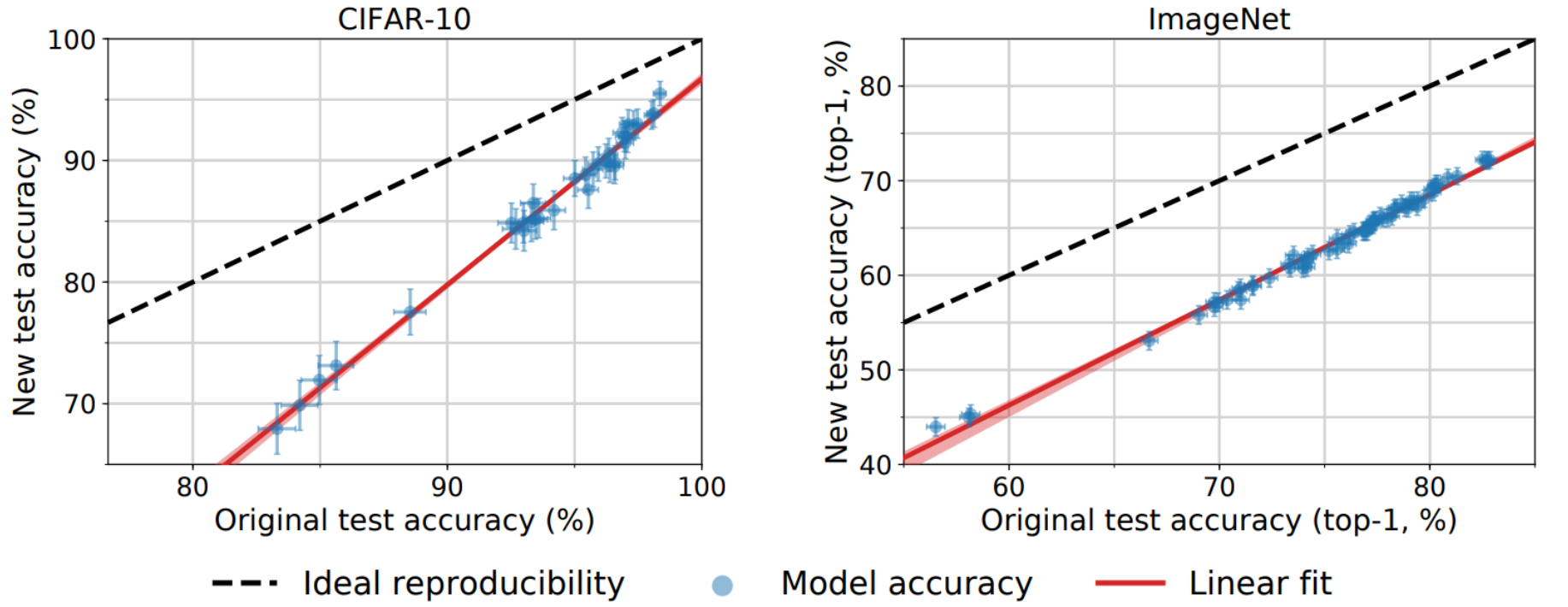


Figure 1: Model accuracy on the original test sets vs. our new test sets. Each data point corresponds to one model in our testbed (shown with 95% Clopper-Pearson confidence intervals). The plots reveal two main phenomena: (i) There is a significant drop in accuracy from the original to the new test sets. (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.

Why though?

- Because the models have *adapted* to the specific samples in the test sets?
- **But**,
  - The relative order is almost exactly preserved!
  - For every 1% improvement on the original >1% improvement on the new set!

- We can get almost the same results if evaluate on easy samples.
- It also shows that current classifiers *still do not generalize reliably even in the benign environment of a carefully controlled reproducibility experiment*.

Making sense of these results

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

Making sense of these results

$$L_{\mathcal{D}}(\hat{f}) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\Big[\mathbb{I}[\hat{f}(x) \neq y]\Big] \approx L_S(\hat{f}) = \frac{1}{|S|}\sum_{(x,y)\in S} \mathbb{I}[\hat{f}(x) \neq y]$$

Making sense of these results

$$L_{\mathcal{D}}(\hat{f}) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\Big[\mathbb{I}[\hat{f}(x) \neq y]\Big] \approx L_S(\hat{f}) = \frac{1}{|S|}\sum_{(x,y)\in S}\mathbb{I}[\hat{f}(x) \neq y]$$

Making sense of these results

$$L_{\mathcal{D}}(\hat{f}) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\Big[\mathbb{I}[\hat{f}(x) \neq y]\Big] \approx L_S(\hat{f}) = \frac{1}{|S|}\sum_{(x,y)\in S} \mathbb{I}[\hat{f}(x) \neq y]$$

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

Disentangling the losses.

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

- Is it the generalization gap?
  - The magnitude of difference is larger than the likely range of fluctuations due to the sampling process.
- Is it the distribution gap?
  - Hard to quantify.
  - Went to great length to minimize these differences.
  - It is the same for all the models.
- Is it the adaptivity gap?

Disentangling the losses.

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

- Is it the adaptivity gap?
  - Later models *probably* should have a larger adaptivity gap because of *successive hyper-param tuning*.
  - But the improvements on the original test set translates into a higher improvement on the new test set!

Disentangling the losses.

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

- Is it the adaptivity gap?
  - Later models probably should have a larger adaptivity gap because of successive hyper-param tuning.
  - But the improvements on the original test set translates into a higher improvement on the new test set!

- It is most likely the distribution gap!

Experiments

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_aug	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

Experiments

with a linear function. On CIFAR-10, the new accuracy of a model is approximately given by the following formula:

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

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

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

Experiments

- Tested **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.
- Conjecture that the accuracy drops stem from **small variations in the human annotation process**.

Experiments – Labelling ImageNet

- Each turker is presented with *48 images* and *a label*.
- Must pick the images that belong to that label.
- 20 turkers per label.

- Some images will be picked more often than others for each label.

Experiments – Labelling ImageNet

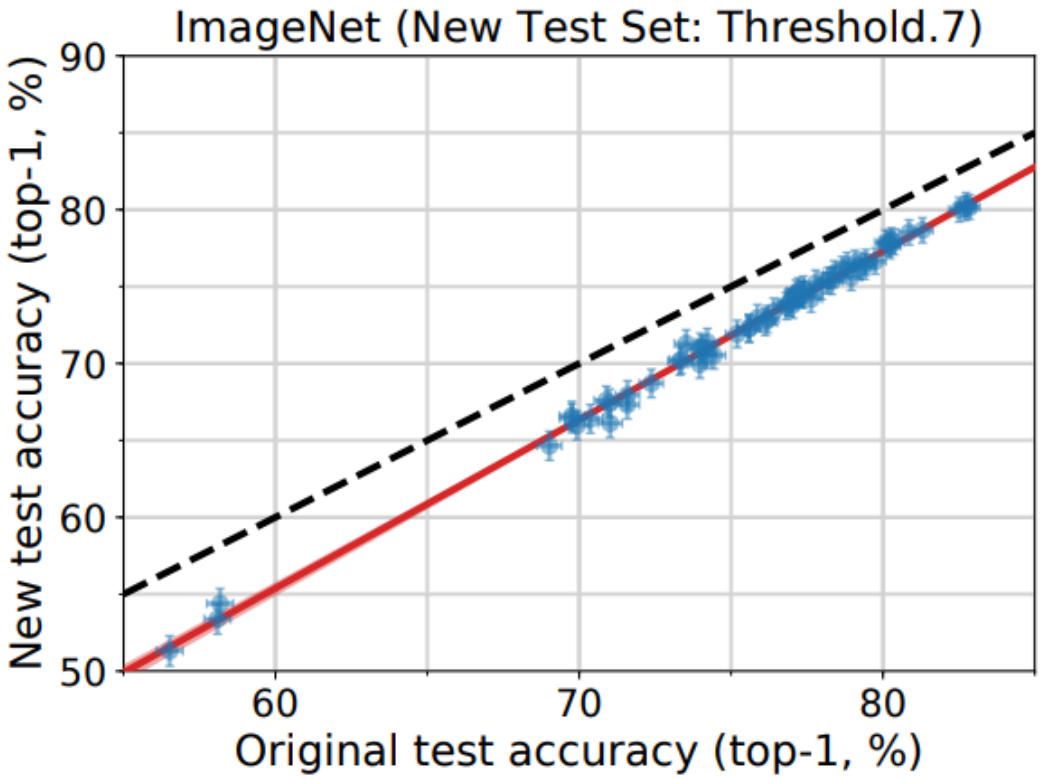
Sampling Strategies. In order to understand how the MTurk selection frequency affects the model accuracies, we explored three sampling strategies.

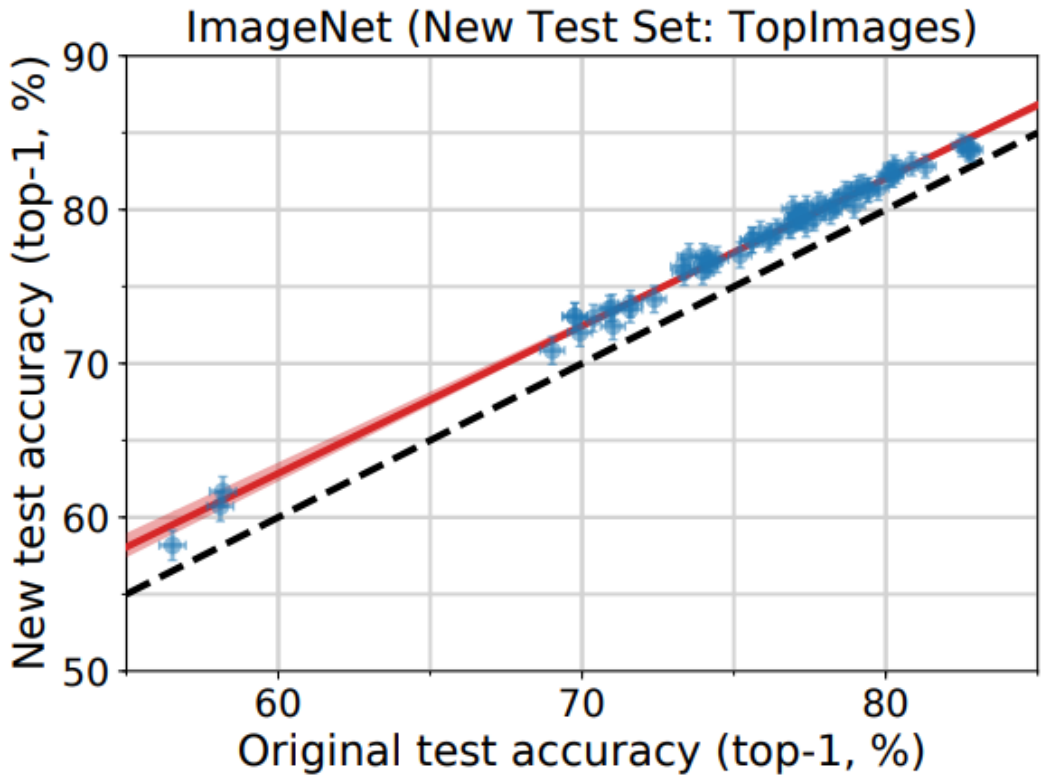
- **MatchedFrequency:** First, we estimated the selection frequency distribution for each class from the annotated original validation images. We then sampled ten images from our candidate pool for each class according to these class-specific distributions (see Appendix C.1.2 for details).
- **Threshold0.7:**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

Experiments – Labelling ImageNet

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%

Experiments – Labelling ImageNet





Ideal reproducibility Model accuracy Linear fit

Discussion

- At least on CIFAR-10 and ImageNet, multiple years of competitive test set adaptivity ***did not lead to diminishing accuracy*** numbers.
- The lack of adaptive overfitting contradicts conventional wisdom in machine learning. Maybe:
  - **The Ladder Mechanism***. An algorithm that protects the evaluation from adaptive overfitting.
  - **Limited Model Class**. Low-variance estimation*.

* The Ladder: A Reliable Leaderboard for Machine Learning Competitions

* Model Similarity Mitigates Test Set Overuse. arXiv 1905.12580 May 2019.

Discussion

- The distribution gap is the leading hypothesis.
- It is surprisingly hard to accurately replicate the distribution of current image classification datasets.
- The difficulty of 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.
- We could create a new correct test set with even lower model accuracies.
- ImageNet models still have difficulty generalizing from “easy” to “hard” images.

Future Work

- Adaptive overfitting: do other domains produce the same results?
- Distribution gap: what makes the new data harder?
- Robust models: can they eliminate the gap?
- Do humans fail to the same extent?
  - Preliminary result: no!
- More test sets could help us understand better.

Discussion

- We already knew:
  - Out-of-distribution samples lead to unreliable predictions.
  - Testing for sample-level in-distribution vs. out-of-distribution is not easy.
  - Adversarial samples break models completely.
  - Robust classification is not easy.
- Now:
  - “In-distribution samples” can also break predictions, in the sense that we can’t practically have a reasonable bound on the error even at set level.

Thank you!