

# Predicting Generalization of Deep Models

Shreyas and Vihari

# Classic theory

Generalization gap is bounded by some complexity measure of the fitted function with high probability.

$$\boxed{\text{test-train}} \leq \sqrt{\frac{C}{m}} + \text{tiny term}$$

↓

Training data size

Generalization gap:  
test error - empirical train error

C: Complexity measure  
- A function of trained model and training data

# Deep nets and classic theory

Classic complexity measures cannot explain why deep models generalize.

Empirically explore complexity measures based on function and training data properties that explain the generalization gap in practice.

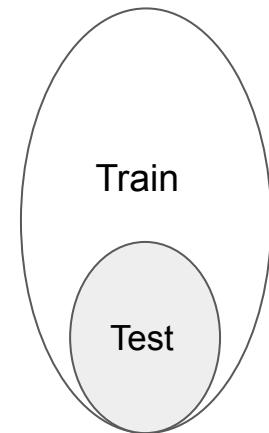
C -> empirical\_predictor(train\_data, trained\_model)

Utility:

- Motivates further theory.
- Improves ML application safety.

# Classic evaluation

- Traditionally, performance evaluated on a held-out test split of train data, called in-distribution (ID) evaluation.
- However, in-distribution test distribution need not represent real-world test distributions.
- Moreover, in-distribution performance is often inflated. I.e. average performance in practice is often lower than in-domain.
  - Ex: Dependence on metal token for predicting pneumonia



# OOD Evaluation: Beyond ID evaluation

We need an alternate evaluation:

- that informs how well a model performs in the real-world.
- that measures “true” progress, i.e. by not incentivising predictions due to incidental features.

But how can we possibly quantify performance of a model in the arbitrary, creative and complex real-world? 😞

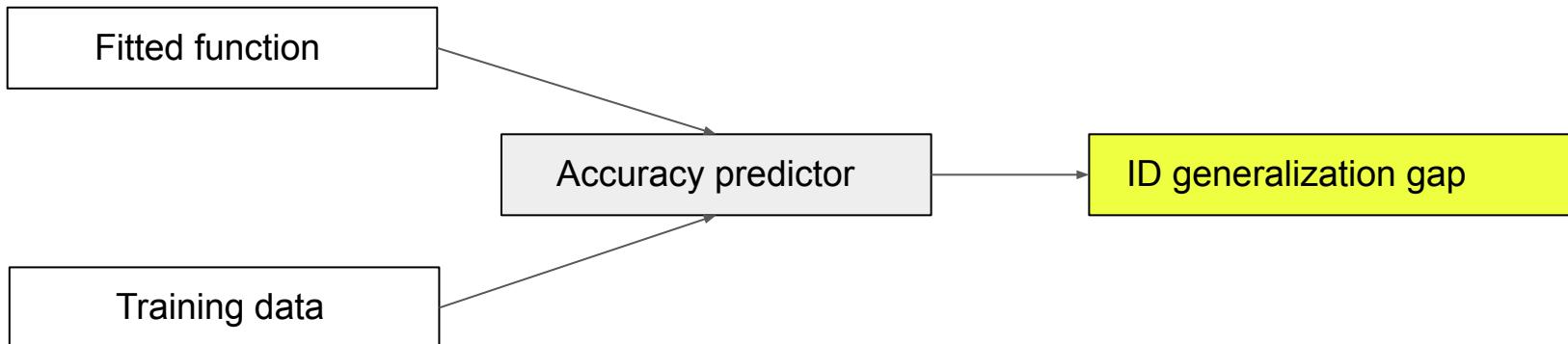
# Outline

- Part 1: Complexity measures for in-distribution generalization gap.
- Part 2: Complexity measures for out-of-distribution generalization gap.
- Part 3: Predicting accuracy on any dataset without labelled data.
- Takeaways and future work.

# Part I: In-domain Generalization

# This part...

We will look at empirical measures that predict real-world performance given the fitted function and in-domain distribution.



# Survey and overview

Published as a conference paper at ICLR 2020

---

## FANTASTIC GENERALIZATION MEASURES AND WHERE TO FIND THEM

**Yiding Jiang<sup>\*†</sup>, Behnam Neyshabur\*, Hossein Mobahi, Dilip Krishnan, Samy Bengio**  
Google Research

{ydjiang, neyshabur, hmobahi, dilipkay, bengio}@google.com

Summarizes and compares generalization measures existing to date on **Image classification** datasets: CIFAR-10, SVHN, using **ConvNet** architecture.

# Setup

Trained models are generated by setting 7 common hyperparameters (batch size, dropout, lr, etc.) to 3 values ( $3^7 = 2187$  models).

$$\theta := (\theta_1, \dots, \boxed{\theta_n}) \in \Theta,$$

n<sup>th</sup> Hyperparam

where  $\Theta := \Theta_1 \times \dots \times \Theta_n$

Let  $\mu(\theta)$  be the proposed complexity measure.

Let  $g(\theta)$  denote the true generalization gap computed using held-out test set.

Rank models  $(\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n)})$  using  $\mu$  or  $g$  and the ranked lists should be consistent for a good complexity measure  $\mu$ .

## Setup (continued) ...

Metric: Granulated Kendall's coefficient  $\Psi$

$$\psi_i \triangleq \frac{1}{m_i} \sum_{\theta_1 \in \Theta_1} \cdots \sum_{\theta_{i-1} \in \Theta_{i-1}} \sum_{\theta_{i+1} \in \Theta_{i+1}} \cdots \sum_{\theta_n \in \Theta_n} \tau(\cup_{\theta_i \in \Theta_i} \{(\mu(\theta), g(\theta))\})$$

Ranking distance  
High if rank assigned by  
mu and g align.

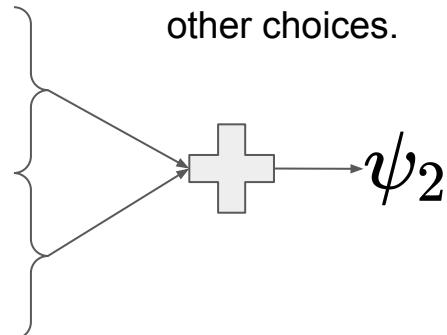
Per hyperparameter ranking comparison to incentivise measures that can predict the effect of any hyperparameter.

# Metric explained

$$\Psi \triangleq \frac{1}{n} \sum_{i=1}^n \psi_i$$

Avoids rewarding weaker measures.  
For eg. if measure captures depth of network well, then can rank models overall well without capturing effect of other choices.

Model Index	Dropout	Learning rate
1	0.3	0.1
2	0.3	0.05
3	0.7	0.1
4	0.7	0.05



# Baseline complexity measures

	batch size	dropout	learning rate	depth	optimizer	weight decay	width	overall $\tau$	$\Psi$	
Corr	vc dim 19	0.000	0.000	0.000	-0.909	0.000	0.000	-0.171	-0.251	-0.154
	# params 20	0.000	0.000	0.000	-0.909	0.000	0.000	-0.171	-0.175	-0.154
	1/ $\gamma$ (22)	0.312	-0.593	0.234	0.758	0.223	-0.211	0.125	0.124	0.121
	entropy 23	0.346	-0.529	0.251	0.632	0.220	-0.157	0.104	0.148	0.124
	cross-entropy 21	0.440	-0.402	0.140	0.390	0.149	0.232	0.080	0.149	0.147
	oracle 0.02	0.380	0.657	0.536	0.717	0.374	0.388	0.360	0.714	0.487
	oracle 0.05	0.172	0.375	0.305	0.384	0.165	0.184	0.204	0.438	0.256
	canonical ordering	0.652	0.969	0.733	0.909	-0.055	0.735	0.171	N/A	N/A

Oracle: True ranking of models + noise

Canonical: Simple ranking rules based on common wisdom

Starting observation: existing vc dimension based complexity measures do not explain generalization gap.

# Norm-based measures

		batch size	dropout	learning rate	depth	optimizer	weight decay	width	overall $\tau$	$\Psi$
Corr	Frob distance 40	-0.317	-0.833	-0.718	0.526	-0.214	-0.669	-0.166	-0.263	-0.341
	Spectral orig 26	-0.262	-0.762	-0.665	-0.908	-0.131	-0.073	-0.240	-0.537	-0.434
	Parameter norm 42	0.236	-0.516	0.174	0.330	<b>0.187</b>	0.124	-0.170	0.073	0.052
	Path norm 44	0.252	<b>0.270</b>	0.049	<b>0.934</b>	0.153	0.338	<b>0.178</b>	<b>0.373</b>	<b>0.311</b>
	Fisher-Rao 45	<b>0.396</b>	0.147	<b>0.240</b>	-0.553	0.120	<b>0.551</b>	0.177	0.078	0.154
	oracle 0.02	0.380	0.657	0.536	0.717	0.374	0.388	0.360	0.714	0.487

Frob distance: distance from initialization

Spectral: measure based on spectral norm of parameters  
(both fail to predict)

Distance from initialization does not matter. Param norm better

Pathnorm: a simple scale invariant complexity measure. (element-wise) square all parameters and accumulate sum of outputs (k=number of classes/outputs) for all-ones input.

$$\mu_{\text{path-norm}} = \sum_{i=1}^k f_{w^2}(\mathbf{1})[i]$$

# Flatness-based measures

PAC-Bayesian measures

If the prior distribution is  $P$  and posterior (after training) is  $Q$  on  $w$ .

Then the expected generalization gap is bounded as below (McAllester, D. A. '99)

$$\mathbb{E}_{w \sim Q} [L(f_w)] \leq \mathbb{E}_{w \sim Q} [\hat{L}(f_w)] + \sqrt{\frac{KL(Q||P) + \log(\frac{m}{\delta})}{2(m-1)}} \text{ w.p } 1 - \delta$$

when  $P = \mathcal{N}(0, \sigma^2 I), Q = \mathcal{N}(w, \sigma^2 I)$ , then (Neyshabur et.al. 2017)

$$\mu_{\text{pac-bayes}} = \frac{\|w\|_2^2}{4\sigma^2} + \log\left(\frac{m}{\delta}\right)$$

$\sigma$  set to largest value s.t.

$$\mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 I)} \hat{L}(f_{w+u}) \leq 0.1$$

## Flatness-based measures (continued...)

Bound based on worst-case flatness (Keskar et.al. 2016).

The magnitude of  $\omega$  (number of parameters) length g.v. of variance  $\sigma^2$  is

$$\alpha = \sigma \sqrt{2 \log\left(\frac{2\omega}{\delta}\right)} \text{ w.p. } 1 - \delta/2$$

$$\mathbb{E}_{\mathbf{u} \sim \mathcal{N}(u, \sigma^2 I)} [L(f_{\mathbf{w}+\mathbf{u}})] \leq \max_{|u_i| \leq \alpha} \hat{L}(f_{\mathbf{w}+\mathbf{u}}) + \sqrt{\frac{\frac{\|\mathbf{w}-\mathbf{w}^0\|_2^2 \log(2\omega/\delta)}{2\alpha^2} + \log\left(\frac{2m}{\delta}\right) + 10}{m-1}}$$

$$\mu_{\text{sharpness}} = \frac{\|w\|_2^2 \log(2\omega)}{4\alpha^2} + \log\left(\frac{m}{\delta}\right)$$

where  $\alpha$  is to largest number such that  $\max_{|u_i| \leq \alpha} \hat{L}(f_{w+u}) \leq 0.1$

# Sharpness-based measures (continued...)

		batch size	dropout	learning rate	depth	optimizer	weight decay	width	overall $\tau$	$\Psi$
Corr	sharpness-orig 52	0.542	-0.359	0.716	0.816	0.297	0.591	0.185	0.400	0.398
	pacbayes-orig 49	0.526	-0.076	0.705	0.546	<b>0.341</b>	0.564	-0.086	0.293	0.360
	1/ $\alpha'$ sharpness mag 62	<b>0.570</b>	<b>0.148</b>	<b>0.762</b>	0.824	0.297	<b>0.741</b>	<b>0.269</b>	<b>0.484</b>	<b>0.516</b>
	1/ $\sigma'$ pacbayes mag 61	0.490	-0.215	0.505	<b>0.896</b>	0.186	0.147	0.195	0.365	0.315
	oracle 0.02	0.380	0.657	0.536	0.717	0.374	0.388	0.360	0.714	0.487

Sharpness magnitude  $\alpha$  is the most informative of generalization gap. (which is as good as it gets on these datasets when compared to oracle 0.02)

# A surprisingly simple baseline

Published as a conference paper at ICLR 2022

---

## ON PREDICTING GENERALIZATION USING GANs

Yi Zhang<sup>1,2</sup>, Arushi Gupta<sup>1</sup>, Nikunj Saunshi<sup>1</sup>, and Sanjeev Arora<sup>1</sup>

<sup>1</sup>Princeton University, Computer Science Department  
`{y.zhang, arushig, nsaunshi, arora}@cs.princeton.edu`  
<sup>2</sup>Microsoft Research

Use (conditional-)GANs to generate test data, and use it to predict generalization gap.

# Algorithm

---

**Algorithm 1** Predicting test performance

---

**Require:** target classifier  $f$ , training set  $S_{\text{train}}$ , GAN training algorithm  $\mathcal{A}$

1. Train a conditional GAN model using  $S_{\text{train}}$ :

$G, D = \mathcal{A}(S_{\text{train}})$  where  $G, D$  are the generator and discriminator networks.

2. Generate a synthetic dataset by sampling from the generator  $G$ :

$$S_{\text{syn}} = \{(\tilde{x}_1, \tilde{y}_1), \dots, (\tilde{x}_N, \tilde{y}_N)\} \text{ where } \tilde{x}_i, \tilde{y}_i = G(z_i, \tilde{y}_i).$$

The  $z_i$ 's are drawn i.i.d. from  $G$ 's default input distribution.  $N$  and  $\tilde{y}_i$ ' are chosen so as to match statistics of the training set.

**Output:** the synthetic accuracy  $\hat{g}(f) := \frac{1}{|S_{\text{syn}}|} \sum_{(\tilde{x}, \tilde{y}) \in S_{\text{syn}}} \mathbf{1}[f(\tilde{x}) = \tilde{y}]$  as the prediction

---

# Results

Task	No.1 team			No.2 team	No.3 team	<b>Ours</b>
	DBI*LWM	MM	AM	R2A	VPM	
1 : VGG on CIFAR-10	25.22	1.11	15.66	52.59	6.07	<b>62.62</b>
2 : NIN on SVHN	22.19	47.33	<b>48.34</b>	20.62	6.44	34.72
4 : AllConv on CINIC-10	31.79	43.22	47.22	<b>57.81</b>	15.42	52.80
5 : AllConv on CINIC-10	15.92	34.57	22.82	24.89	10.66	<b>53.56</b>
8 : VGG on F-MNIST	9.24	1.48	1.28	13.79	16.23	<b>30.25</b>
9 : NIN on CIFAR-10	25.86	20.78	15.25	11.30	2.28	<b>33.51</b>

# Takeaways

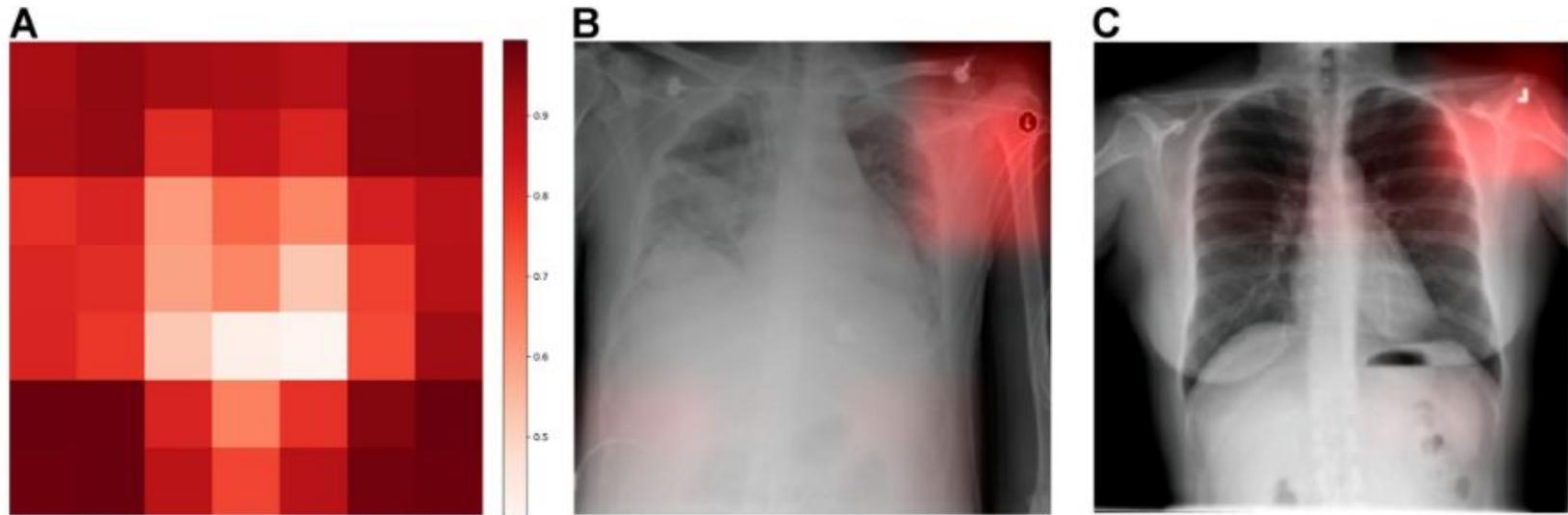
- We have had some success empirically developing measures that explain ID generalization.
- There is scope for predicting even better though.
- Limitation: Effectiveness of synthetic data generated by GANs is unknown.

# Part II: OOD Generalization

# Why OOD?

- So far, we have seen measures that can predict in-distribution generalization gap well.
- But, in-distribution test error often over-estimates real-world performance.
- Out-of-distribution error is a better proxy for real-world performance.

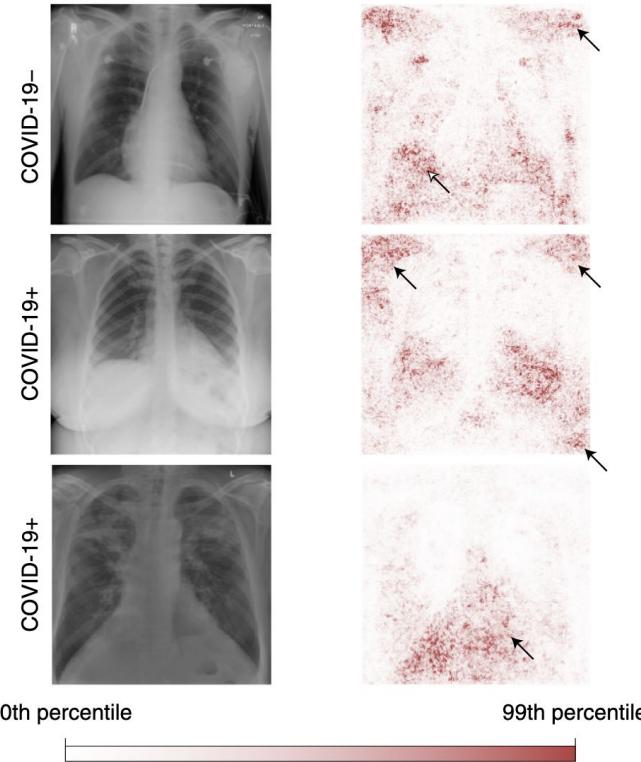
# What's wrong with ID error? Example 1



Metal token placement revealed the hospital system which further informed disease statistics leading to non-trivial accuracy on in-domain (seen hospitals).

Figure and finding: Zech, John R., et al. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study." *PLoS medicine* 15.11 (2018): e1002683.

# What's wrong with ID error? Example 2



Prediction of negative and positive based on superficial/irrelevant features (image edges, cardiac silhouette, diaphragm)

DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*. 2021 Jul;3(7):610-9.

# IID is a myth.

---

## **Do ImageNet Classifiers Generalize to ImageNet?**

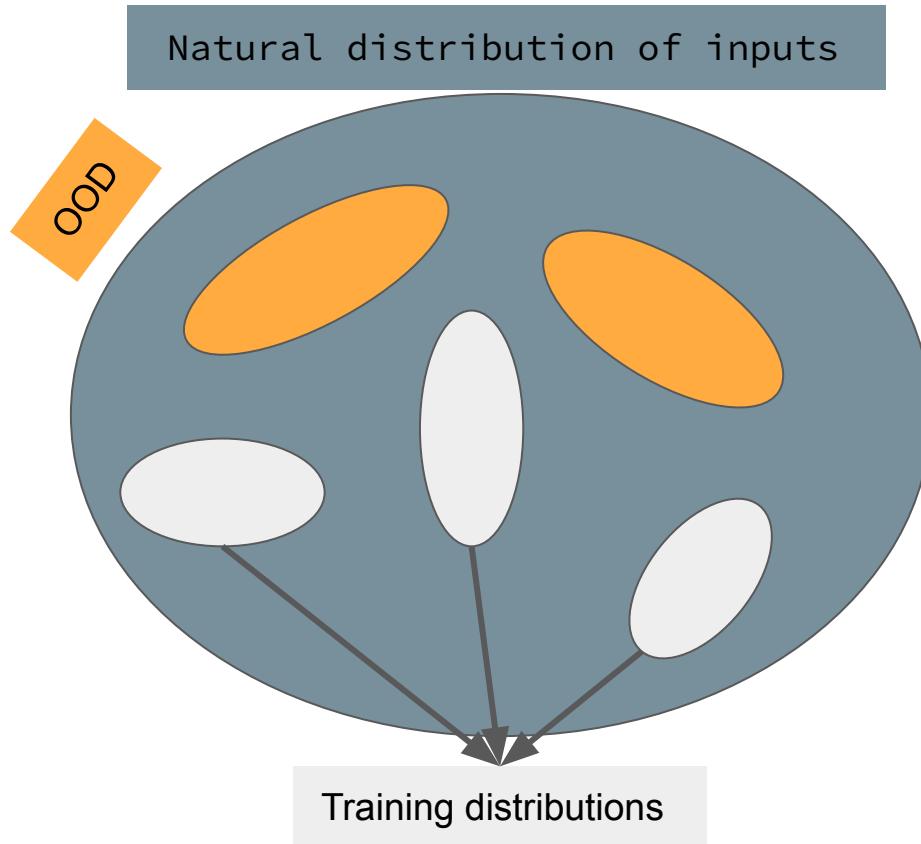
---

**Benjamin Recht<sup>\*</sup>** <sup>1</sup> **Rebecca Roelofs** <sup>1</sup> **Ludwig Schmidt** <sup>1</sup> **Vaishaal Shankar** <sup>1</sup>

- Closely replicates the data generation of CIFAR-10 and Imagenet to create new test sets.
- Yet, the performance dropped by 3% – 15% on CIFAR-10 and 11% – 14% on ImageNet, which they attribute to distribution shift. (New test sets contains more harder examples.)
- It is hard to replicate iid even if we want to.

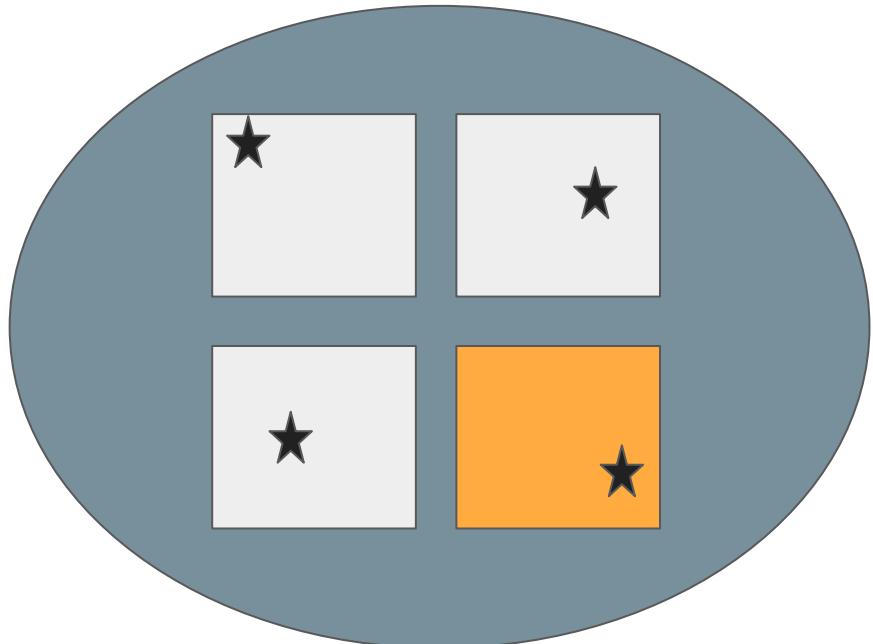
# What is OOD?

- There exists an underlying latent natural distribution of inputs we care about.
- Training distributions are a small sample from the underlying natural distribution.
- OOD evaluation concerns evaluation on examples sampled from the underlying natural distribution instead of in-distribution.



# OOD Evaluation: Toy Example

- Different distribution of inputs are defined by different positioning of the object in the image.
- OOD Evaluation is about evaluating on inputs with arbitrarily positioned object.



Natural distribution of inputs  
:= object appearing anywhere

# OOD Evaluation: Real Examples

- A medical diagnosis application trained on a handful of hospitals, but we wish to generalize to any hospital.

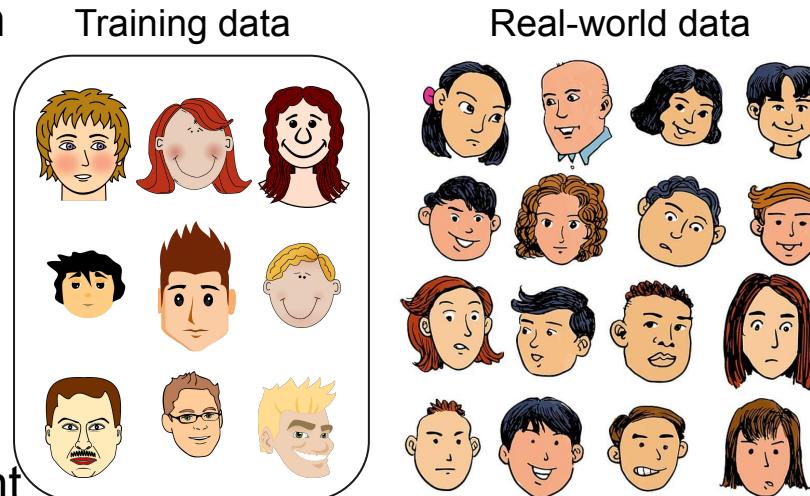
ID: training hospitals

OOD: any hospital

- A speech recognition application that is trained on a handful of speakers should generalize to any speaker (with any accent or ethnicity) in the world.

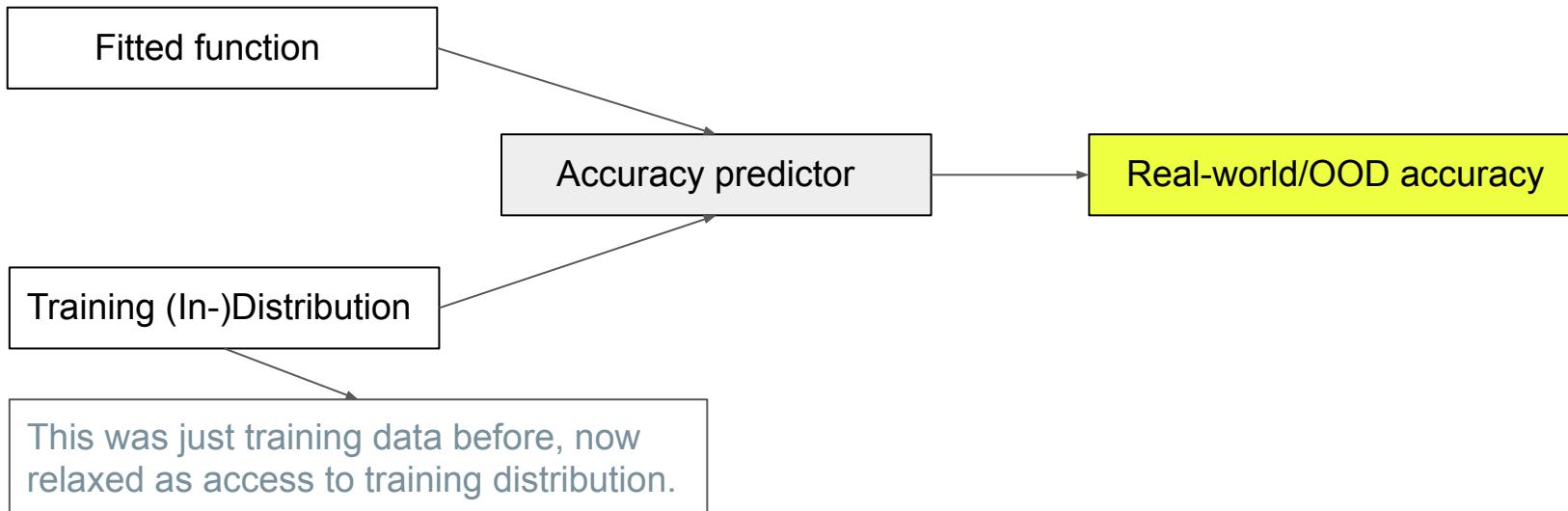
ID: training speakers

OOD: any speaker



# This part...

We will look at empirical measures that predict real-world performance given the fitted function and in-domain distribution.



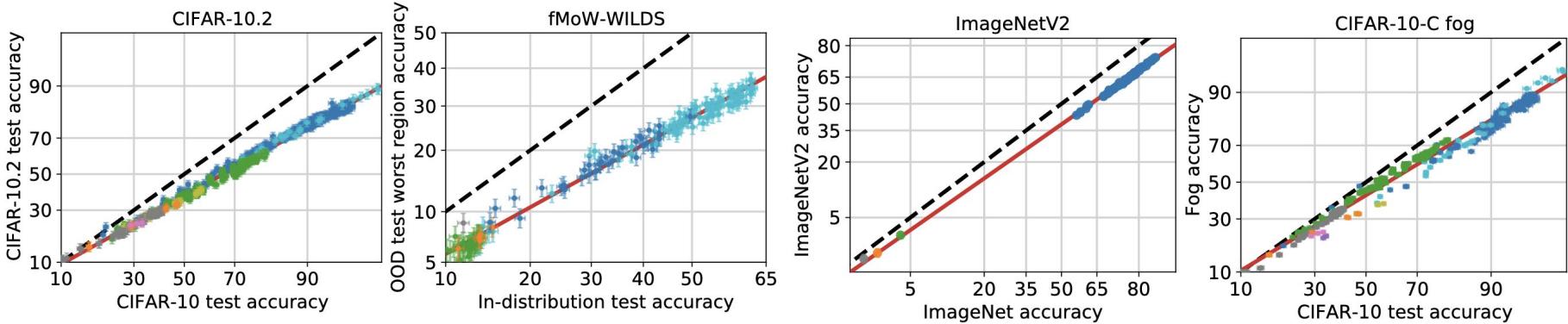
# ID performance is a surprisingly strong metric!

---

## Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization

---

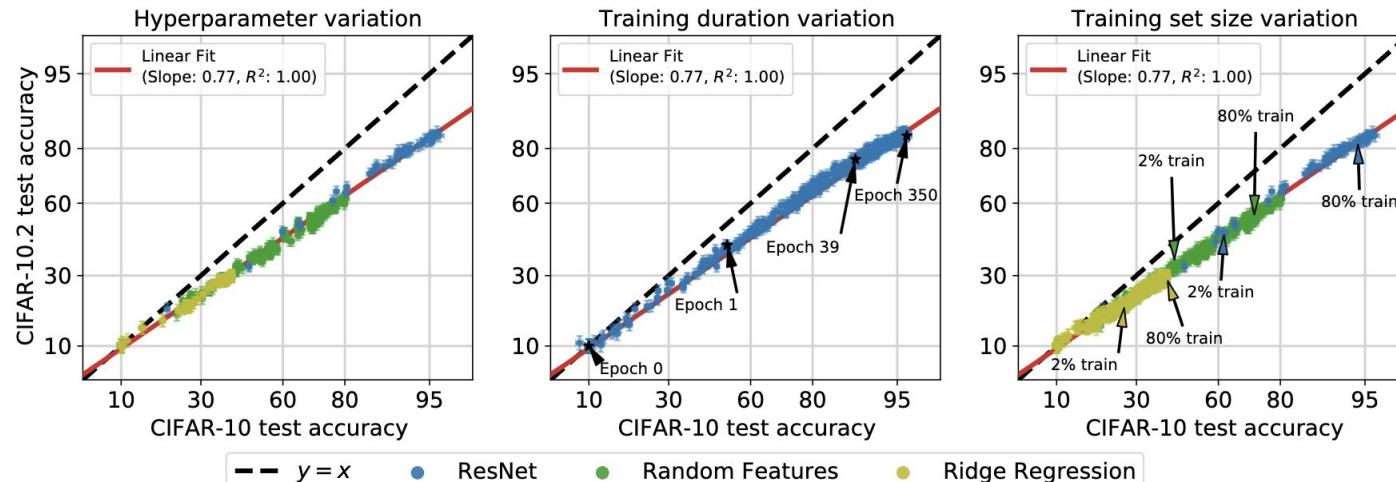
**John Miller<sup>1</sup> Rohan Taori<sup>2</sup> Aditi Raghunathan<sup>2</sup> Shiori Sagawa<sup>2</sup> Pang Wei Koh<sup>2</sup> Vaishaal Shankar<sup>1</sup>**  
**Percy Liang<sup>2</sup> Yair Carmon<sup>3</sup> Ludwig Schmidt<sup>4</sup>**



# Linear Trend preserved over many factors

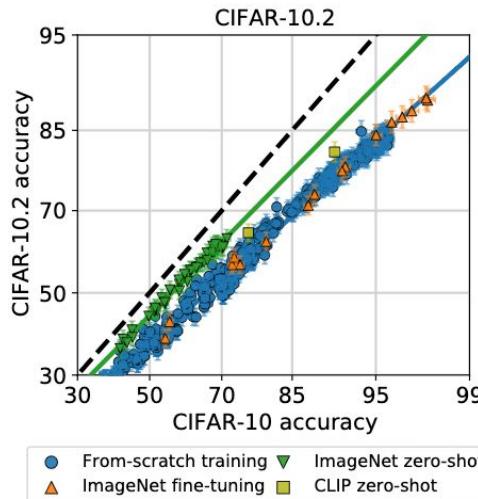
- The linear dependence is preserved over:

- Model hyperparameters
- Training dataset size
- Training duration

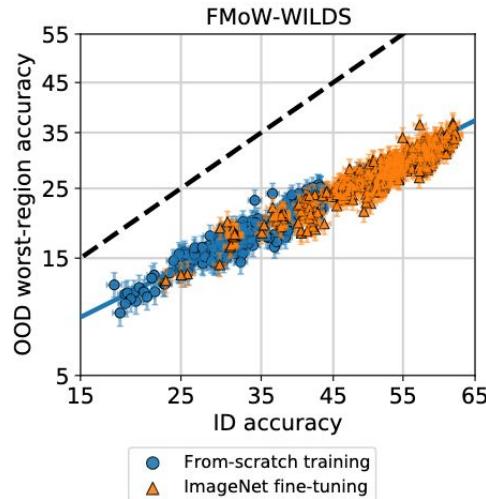


# Interesting Behaviour in other settings

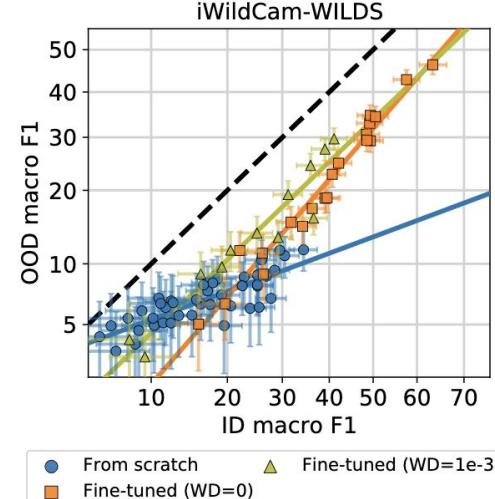
- Pre-training
  - Zero-shot behaviour is better generalised than fine-tuned behaviour
  - Pre-training helps improve generalisation to OOD characteristics if not present in training data



OOD datasets contain similar viewpoints as ID

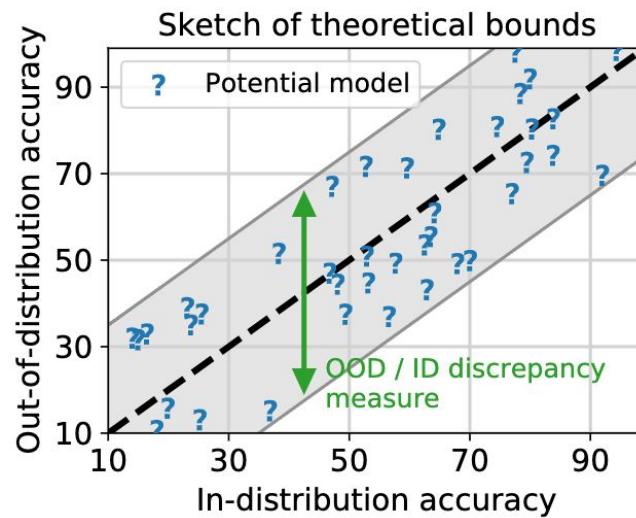


OOD datasets contain very different viewpoints from ID



# Explaining Linear Trend with Gaussian Shift model

- Classic Generalisation Theory [Mansour et al. 2009] says
  - For a model  $f$  trained on a distribution  $D$ , relate accuracy on  $D$  to accuracy on an OOD distribution  $D'$  as:  $|\text{acc}_{D'}(f) - \text{acc}_D(f)| \leq d(D, D')$



# Explaining Linear Trend with Gaussian Shift model

- [Miller et al.] show a stronger linear trend than this

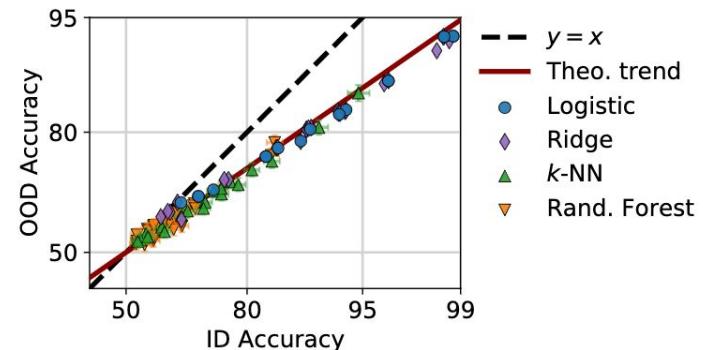
- Consider  $D$  such that  $x|y \sim \mathcal{N}(\mu \cdot y; \sigma^2 I_{d \times d})$ ,  
random noise on  $\mathbb{R}^d$  sphere
- Consider  $D'$  shifted as  $\mu' = \alpha \cdot \mu + \beta \cdot \Delta$  and  $\sigma' = \gamma \cdot \sigma$
- Consider linear classifier  $x \mapsto \text{sign}(\theta^\top x)$

**Theorem 1.** In the setting described above where  $\Delta$  is independent of  $\theta$ , let  $\delta \in (0, 1)$ . With probability at least  $1 - \delta$ , we have

$$\left| \Phi^{-1}(\text{acc}_{D'}(\theta)) - \frac{\alpha}{\gamma} \Phi^{-1}(\text{acc}_D(\theta)) \right| \leq \frac{\beta}{\gamma \sigma} \sqrt{\frac{2 \log 2/\delta}{d}}.$$



vanishes in high-d



# A Larger scale evaluation of Metric Correlation

---

## Assaying Out-Of-Distribution Generalization in Transfer Learning

---

**Florian Wenzel<sup>\*1</sup>**

**Carl-Johann Simon-Gabriel<sup>1</sup>**

**David Kernert<sup>1</sup>**

**Bernt Schiele<sup>1</sup>**

**Andrea Dittadi<sup>† 2</sup>**

**Max Horn<sup>1</sup>**

**Chris Russell<sup>1</sup>**

**Bernhard Schölkopf<sup>1</sup>**

**Peter Gehler<sup>1</sup>**

**Dominik Zietlow<sup>1</sup>**

**Thomas Brox<sup>1</sup>**

**Francesco Locatello<sup>1</sup>**

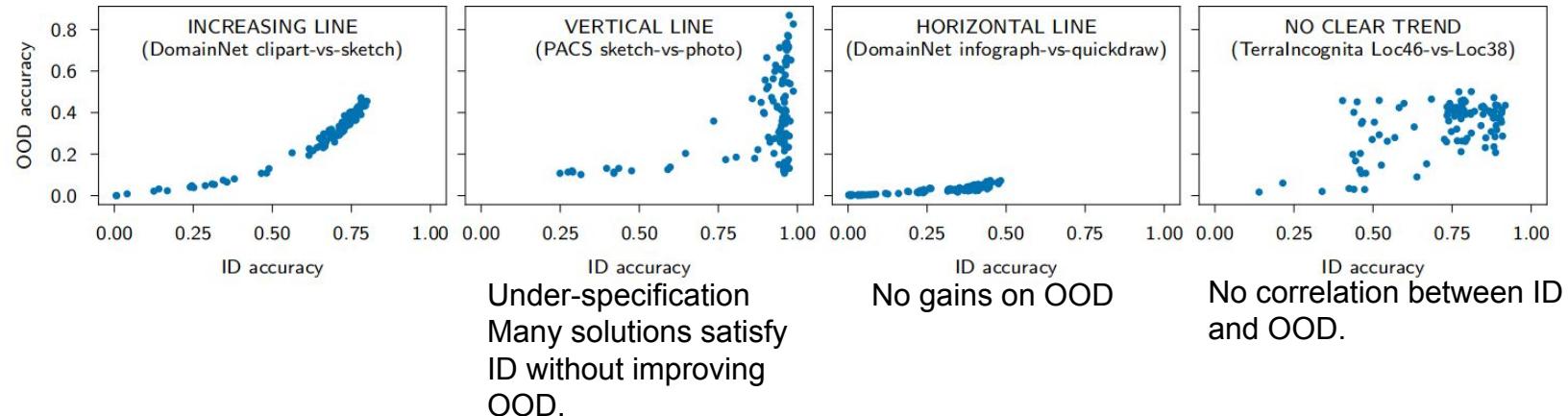
<sup>1</sup> AWS Tübingen <sup>2</sup> Technical University of Denmark

They evaluate the ID-OOD accuracy observation more rigorously.

# Setup

- 10 tasks, 36 datasets, 172 (ID, OOD) pairs
- Models: models pretrained on ImageNet and finetuned on ID datasets
- Metrics: classification error, NLL, expected calibration error (ECE), adversarial classification error.
- Metrics are evaluated on ID test set, held-out OOD test set, and corrupted test set. (corruptions are 1 of 17 synthetic image distortions, for eg. low-pass or high-pass filters)
- Best hyperparameters: number of training steps and learning rate, augmentation strategy picked using held-out (ID-OOD) experiments.

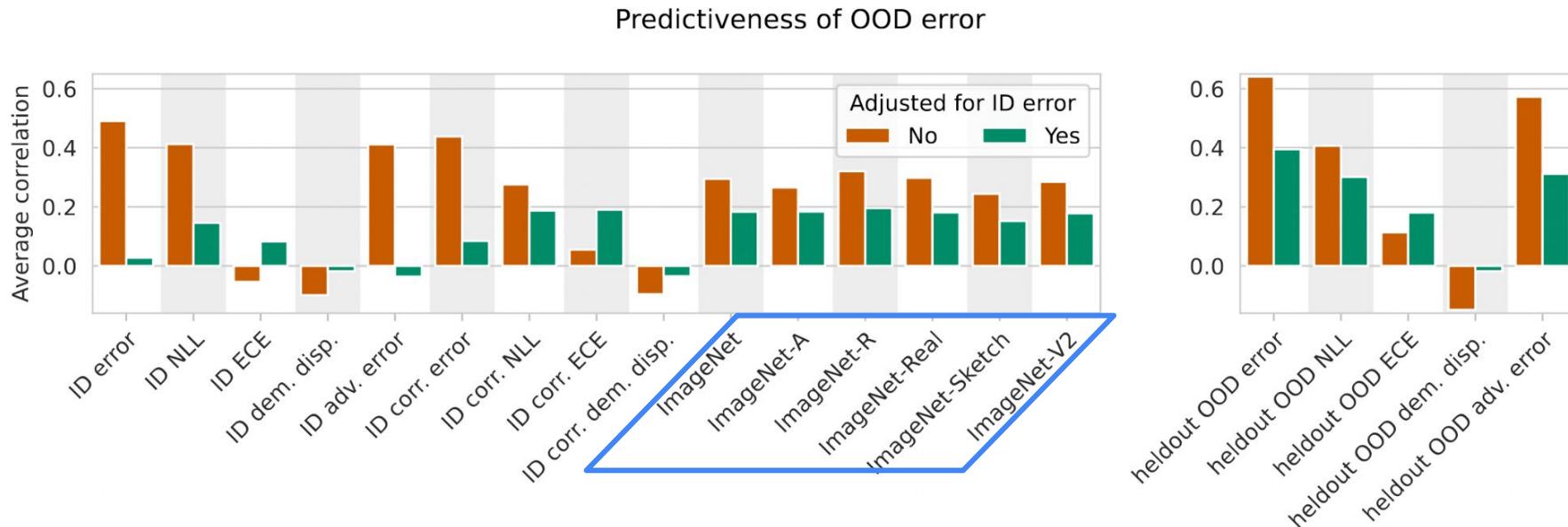
# There is more to ID-OOD correlation



But a trend of decreasing OOD with increasing ID accuracy is never observed.  
(overfitting to spurious features not hurting robustness?)

**So, a good strategy to improve OOD is to improve ID.**

# Which metric is most informative?



- ID error is unsurprisingly the most predictive of the OOD error after heldout OOD error.
- Focus on green bars for how well the metric explains residuals from ID-OOD linear fit.
- Accuracy and robustness of pretrained model translates to robustness of downstream model.

# Augmentations and architecture

		Full dataset				Few-shot-100			Few-shot-10		
		Error	Adv err	Corr err	ECE	Error	Adv err	ECE	Error	Adv err	ECE
ID	OOD	0.11 (1e-1)	0.44 (7e-3)	6.63 (8e-8)	0.43 (3e-3)	-0.22 (3e-1)	0.71 (1e-3)	0.58 (8e-2)	2.10 (2e-4)	3.36 (3e-6)	1.33 (8e-3)
	OOG	1.99 (2e-5)	2.13 (2e-5)		2.52 (4e-5)	1.34 (1e-3)	1.88 (3e-5)	2.12 (2e-4)	2.34 (2e-5)	2.96 (2e-6)	1.12 (2e-2)

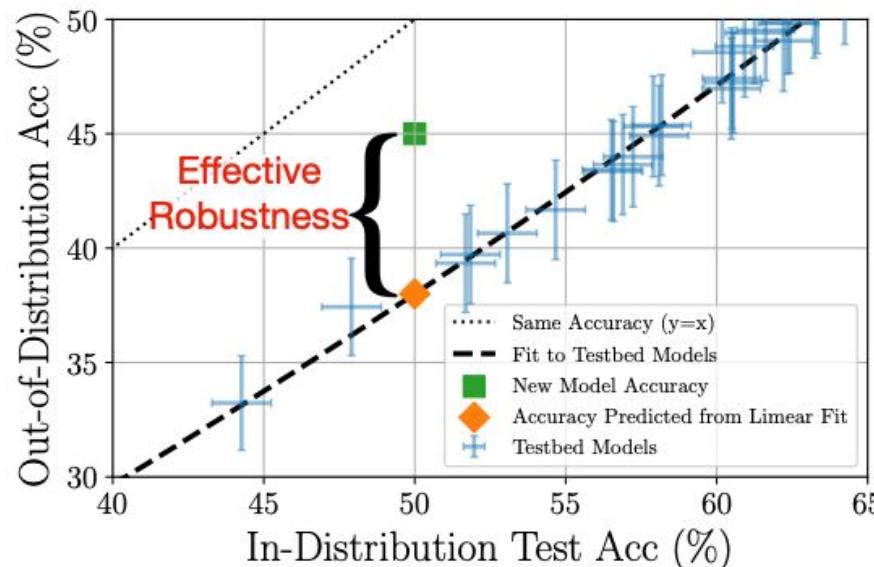
Performance improvement over not using Aug. (p-value)  
Black: insignificant.

Data augmentations (RandAugment or AugMix) improve accuracy for any OOD type (natural, corrupted or adversarial).

Architecture matters and vision transformers are most robust across the board.

# Models that have high “effective robustness”

- Models that lie above the linear fit are said to have “effective robustness”
- Exceedingly rare: zero-shot CLIP models, very few pre-trained Imagenet models



# How do we identify effective robustness?

- One approach: Explanation through spectral properties of both models and OOD data influence the corresponding effective robustness
- Models that are robust to high-frequency data features generalise better

---

## Models Out Of Line: A Fourier Lens On Distribution Shift Robustness

---

Sara Fridovich-Keil<sup>†,\*</sup>, Brian R. Bartoldson<sup>‡</sup>, James Diffenderfer<sup>‡</sup>,

Bhavya Kailkhura<sup>‡</sup>, Peer-Timo Bremer<sup>‡</sup>

<sup>†</sup>UC Berkeley, <sup>‡</sup>Lawrence Livermore National Laboratory

# Effective Robustness correlates with Spectral Properties

- CLIP Models are uncharacteristically robust to semantic content-preserving corruptions
- Higher frequency features ~ semantically meaningful content
- Pre-training and data augmentation encode more high-frequency features

Low Frequency Interpolation - semantic content-preserving corruption



High Frequency Interpolation - semantic content-destroying corruption



# Takeaways

- Unlike for ID error, no one measure can explain OOD error.
- Deep networks are surprisingly resilient to overfitting. Not only do they avoid overfitting to seen examples, but also avoid overfitting to seen training distributions (in-distributions).
- Future work should focus on characterizing the distribution shift better.

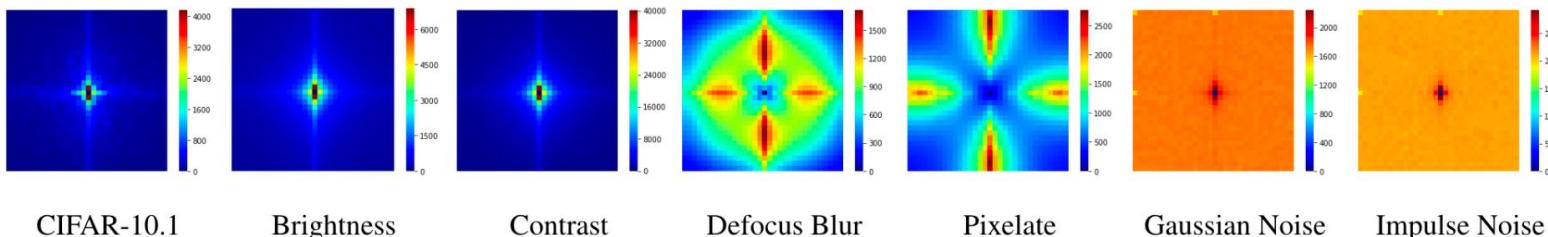


Figure 2: Power spectral densities for a selection of low (CIFAR-10.1, brightness, contrast), mid (defocus blur, pixelate), and high (gaussian noise, impulse noise) frequency shifts w.r.t. CIFAR-10.

# Part III: Accuracy Prediction on Target Domain

# Beyond simple correlation...

- How do we look beyond metrics that simply correlate with OOD accuracy?
- Can we predict test error on *any* dataset with access to just unlabelled test data?
- Can we rank models on how well they would perform on a dataset with just unlabelled data?

## ASSESSING GENERALIZATION VIA DISAGREEMENT

**Yiding Jiang**\*  
Carnegie Mellon University  
ydjiang@cmu.edu

**Vaishnavh Nagarajan**\*†  
Google Research  
vaishnavh@google.com

**Christina Baek, J. Zico Kolter**  
Carnegie Mellon University  
{kbaek, zkoltter}@cs.cmu.edu

## LEVERAGING UNLABELED DATA TO PREDICT OUT-OF-DISTRIBUTION PERFORMANCE

**Saurabh Garg**\*  
Carnegie Mellon University  
sgarg2@andrew.cmu.edu

**Sivaraman Balakrishnan**  
Carnegie Mellon University  
sbalakri@andrew.cmu.edu

**Zachary C. Lipton**  
Carnegie Mellon University  
zlipton@andrew.cmu.edu

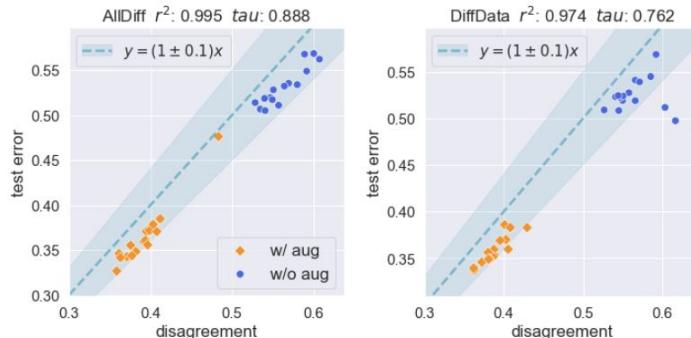
**Behnam Neyshabur**  
Google Research, Blueshift team  
neyshabur@google.com

**Hanie Sedghi**  
Google Research, Brain team  
hsedghi@google.com

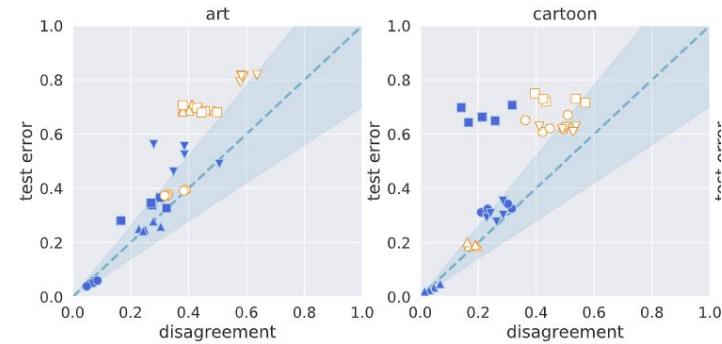
- By training separate classifiers and measuring disagreement *on the same unlabelled data*

# Disagreement Tracks Generalisation Error

- Assume we have two hypotheses  $h, h'$  from the same training procedure (same optimiser, LR, data-batching, hardware) on a data distribution  $\mathcal{D}$
- $\text{TestErr}_{\mathcal{D}}(h) \triangleq \mathbb{E}_{\mathcal{D}} [\mathbb{1}[h(X) \neq Y]]$  and  $\text{Dis}_{\mathcal{D}}(h, h') \triangleq \mathbb{E}_{\mathcal{D}} [\mathbb{1}[h(X) \neq h'(X)]]$ .
- Also easy to show  $|\text{TestErr} - \text{Dis}| < \text{Calibration Error}$ 
  - Issue: This correlation only exists for calibrated models, circular issue sometimes!
  - Issue: Calibration degrades very fast under shift



In-Distribution (ResNet18 on CIFAR100)



Distribution Shift (PACS Dataset)

# Issue: Calibration Degrades under Shift

Published in Transactions on Machine Learning Research (10/2022)

---

## A Note on “Assessing Generalization of SGD via Disagreement”

Andreas Kirsch

Yarin Gal

OATML, Department of Computer Science  
University of Oxford

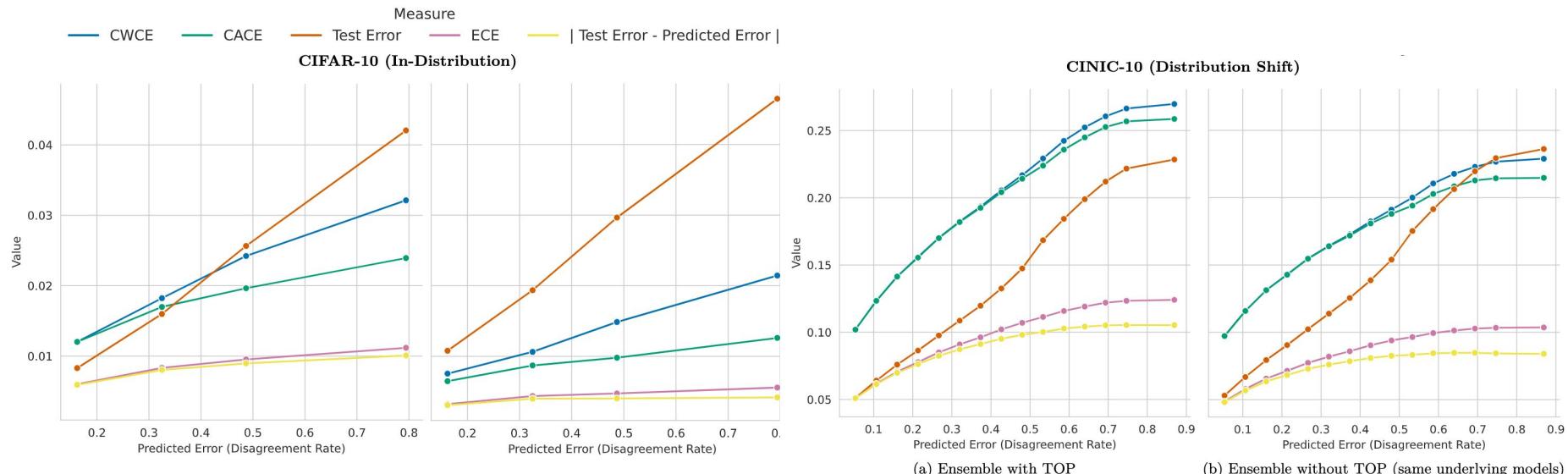
*andreas.kirsch@cs.ox.ac.uk*

*yarin.gal@cs.ox.ac.uk*

- They show that Theorem 1 does not hold under distribution shift
- “we can only hope to trust model calibration for in-distribution data, while under distribution shift, the calibration ought to deteriorate”

# Issue: Calibration Degrades under Shift

- We can't successfully trust disagreement  $\sim$  test error under high miscalibration



# Takeaways

- Predicting ID generalization gap is somewhat easier and measures based on sharpness or evaluation using synthetic examples (GAN) perform reasonably well. Although they are not perfect and can be improved further.
- OOD generalization gap OTH is understandably multi-faceted and no one measure can explain it well.
- ID generalization is surprisingly well-correlated with OOD generalization. Besides, negative effect of ID performance on OOD is never observed (although could be possible).
- Best strategy for improving real-world performance therefore is to continue to improve ID performance.

# Takeaways (continued) ...

- Disagreement between calibrated ensemble of classifiers is a reliable estimate for accuracy even on OOD datasets.

## Future Work/Limitations

- Empirically found successful generalization measures need to be weaved in to a convincing theory.
- Predicting OOD robustness has still a long way to go. ID is the best predictor of OOD but far from ideal.

## Takeaways (continued) ...

- Distribution shift or digression from training dataset needs to be better understood. Currently, very crude characterization: natural, adversarial or artificial shift.
- Population with worst generalization is more important than expected generalization that we looked at in this session. No existing problem formulation captures that.
- Creating calibrated ensembles for accuracy prediction on any unlabeled dataset. Calibration breaks easier than accuracy on distribution shifts. How to create calibrated ensembles? What is the uncertainty on accuracy estimate, which could depend on the nature of distribution shift.

Thanks!

# Issues with Standard OOD Evaluation?

- Dependence on specific OOD datasets and corruption choices
- Standard OOD benchmarks may not translate to real-world deployment
- Disentangling intrinsic model robustness vs specific training schedules
- Spurious metrics for measurement - calibration? ID accuracy? empirical risk?
- Hard to understand causes of either success or failure.

# This session

- Can we quantify real-world performance without OOD datasets?
- What model properties inform distributional robustness?
- Dataset independent way of measuring OOD robustness.

# Large Scale Experiments on OOD

- Miller et al 2021
  - Established a roughly linear trend b/w ID and OOD accuracy
  - Justifications using Gaussian distribution shift model
  - Does not work well with different covariance shifts, pre-training, adversar

# Assaying Out-Of-Distribution Generalization in Transfer Learning

# Models Out Of Line: A Fourier Lens On Distribution Shift Robustness

# Models out of Line: A Fourier Lens

- Other papers noticed a linear dependence b/w ID and OOD performance
- Models out of this line have “remarkable” robustness, how do we identify?
- How do spectral properties of models and OOD data affect robustness?
- Perturb image’s Fourier amplitude while keeping phase constant - preserving semantic information
- Perturb image’s Fourier phase while keeping amplitude constant - destroying semantic information
- Measure - High Frequency Fraction (HFF), Consistent Distance (CF)
- Three tunable "knobs" that are available to the neural network practitioner and have been shown to impact OOD robustness: pruning [7], data augmentation [15], and weight ensembling [34].
- Main Takeaway: If model is robust to high frequency perturbation, then it is robust to semantics similar to a human.

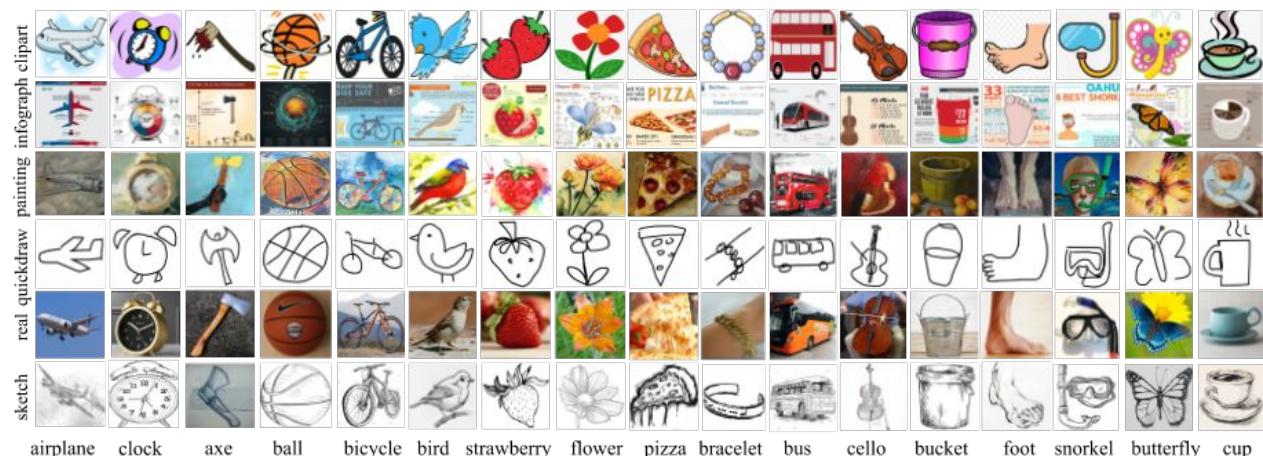
# Theoretical Metrics

Gannon et al

# Standard OOD Evaluation

Measure the strength of an algorithm through testing on datasets with distribution shift.

- PACS, VLCS, DomainNet



# Standard OOD Evaluation

Measure the strength of an algorithm through testing on datasets with distribution shift.

- PACS, VLCS, DomainNet
- WILDS

Camelyon17

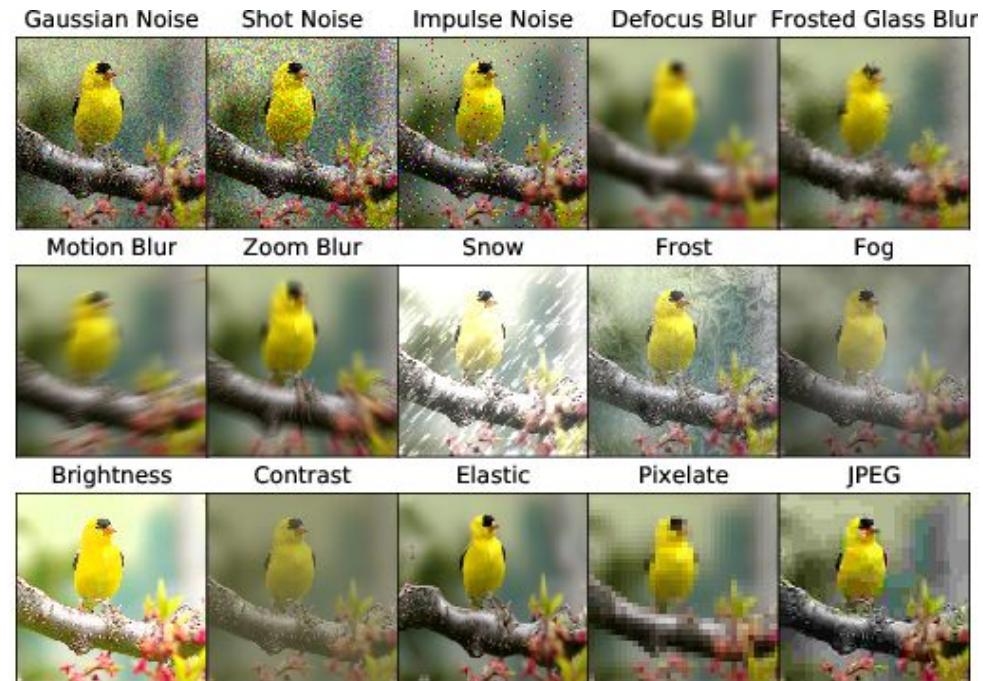
Train			Val (OOD)	Test (OOD)	
y = Normal	d = Hospital 1	d = Hospital 2	d = Hospital 3	d = Hospital 4	d = Hospital 5

<https://wilds.stanford.edu/datasets/>

# Standard OOD Evaluation

Measure the strength of an algorithm through testing on datasets with distribution shift.

- PACS, VLCS, DomainNet
- WILDS
- Corruption datasets



# Standard OOD Evaluation

Measure the strength of an algorithm through testing on datasets with distribution shift.

- PACS, VLCS, DomainNet
- WILDS
- Corruption datasets
- Spurious correlations

		label: object	
		waterbird	landbird
spurious attribute: background	water background		
	land background		
	majority	majority	minority
	minority	minority	majority

Image source: GroupDRO slides at ICLR 2020