Australian National University

NEC Laboratories America

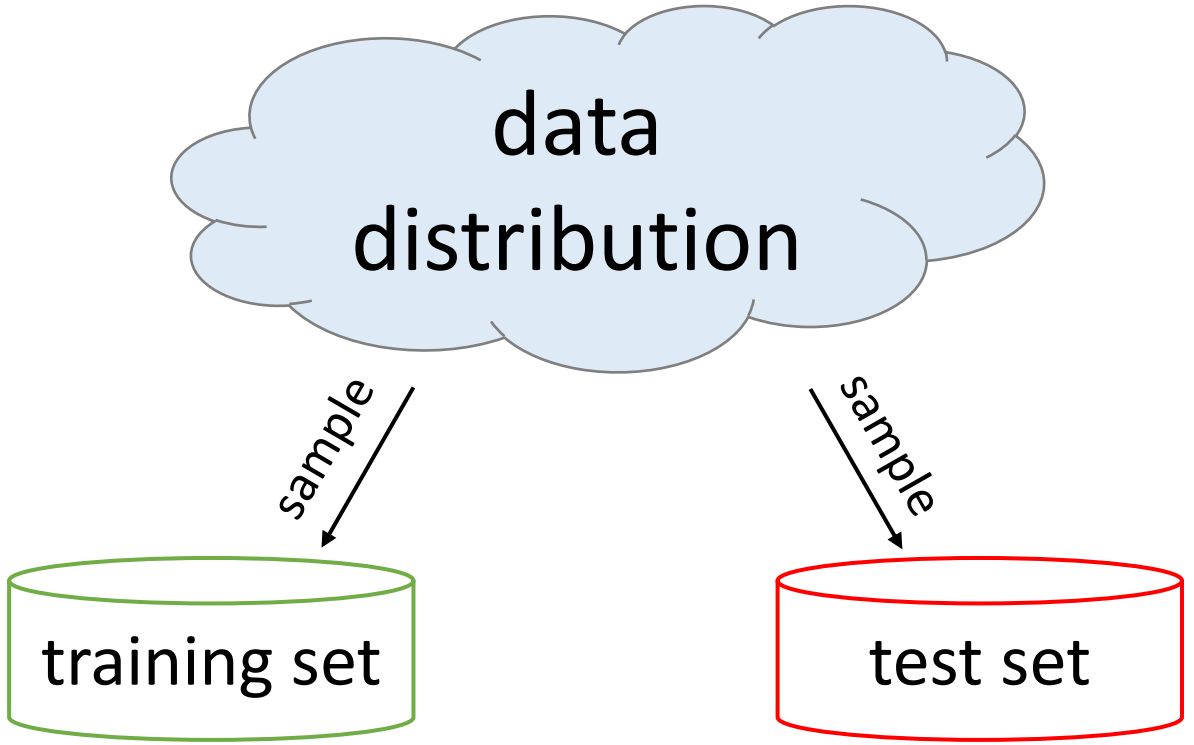
ICML International Conference On Machine Learning

Confidence and Dispersity Speak: Characterizing Prediction Matrix for Unsupervised Accuracy Estimation

Weijian Deng[1] Yumin Suh[2] Stephen Gould[1] Liang Zheng[1]

[1]Australian National University [2]NEC Labs America

Classical Model Evaluation

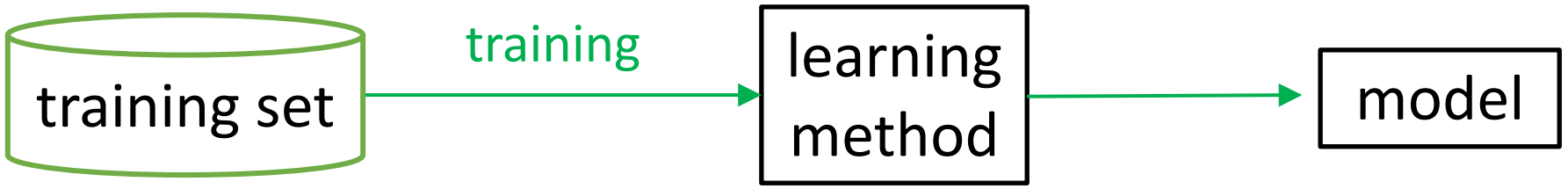


i.i.d. assumption

1) train set and test set are **independent** from each other;
2) they are **identically distributed**;

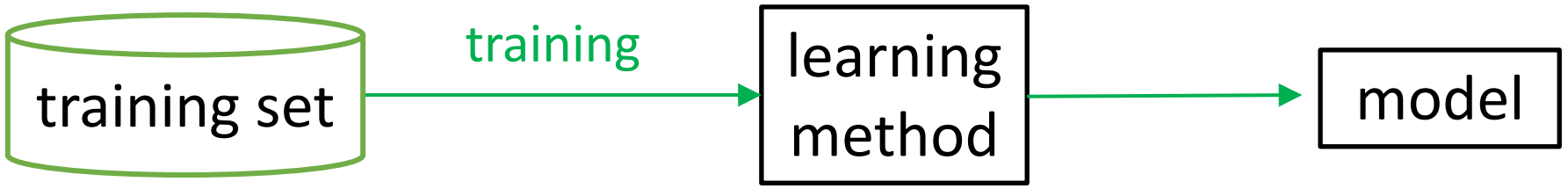
Classical Model Evaluation

Training phase



Classical Model Evaluation

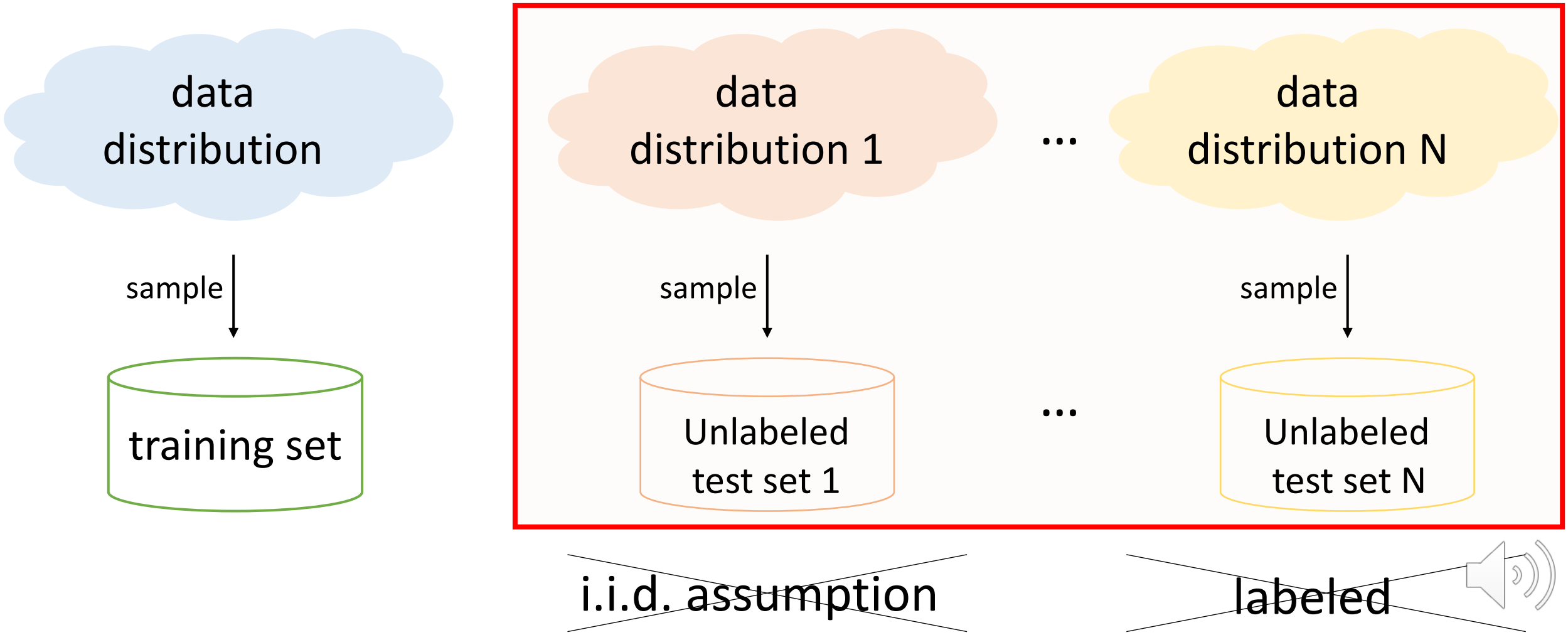
Training phase



Generalization evaluation

How well it performs well on new, previously unseen inputs?

Model Evaluation in Real-world Applications



Unsupervised Accuracy Estimation

ImageNet Training Set

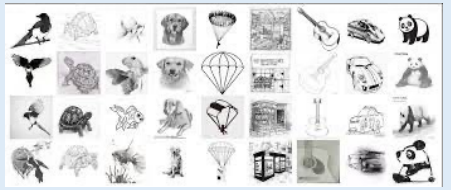
ViT

Accuracy?

Unlabeled Test Sets

ImageNet-R(endiction)

ImageNet-S(ketch)



...

Given a trained model, the goal is to **predict its performance on unlabelled test sets**

Deng, Weijian "Are Labels Necessary for Classifier Accuracy Evaluation?", In CVPR, 2021; TPAMI 2022

Accuracy Estimation with Classifier Prediction

- **SoftMax scores of classifiers on unlabelled data are informative**

Average Confidence (AC)

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In ICLR, 2017

Average Thresholded Confidence (ATC)

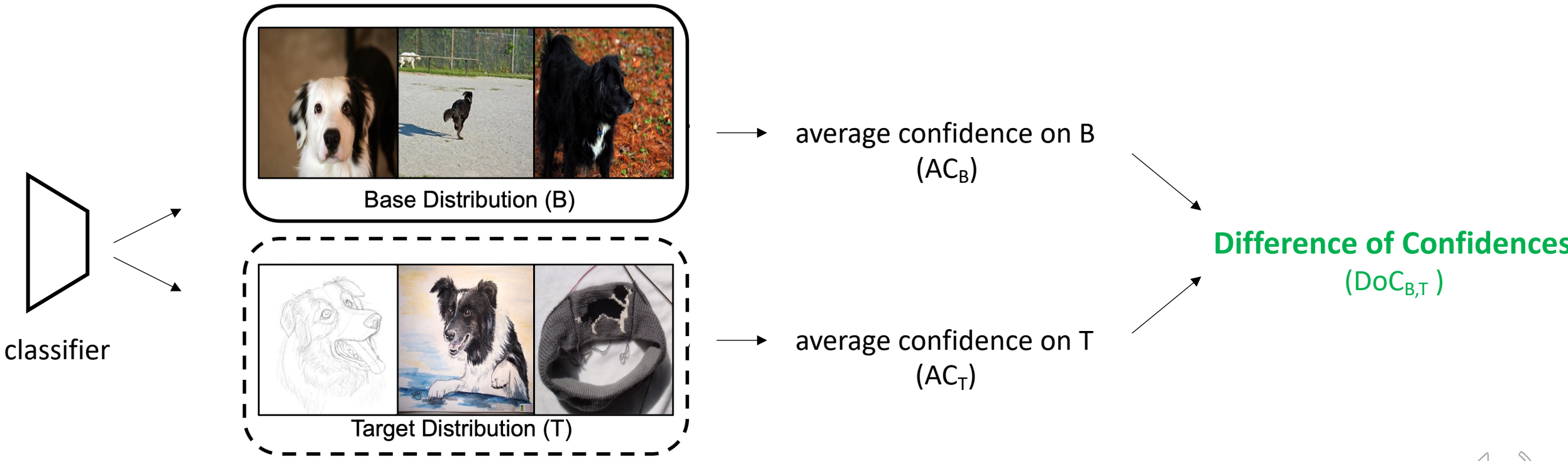
Garg, Saurabh, et al. "Leveraging unlabeled data to predict out-of-distribution performance." In ICLR, 2022

Difference of Confidence (DoC)

Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., & Schmidt. "Predicting with confidence on unseen distributions", In ICCV, 2021

Accuracy Estimation with Classifier Prediction

- **Difference of Confidences encodes useful information about distribution shift**

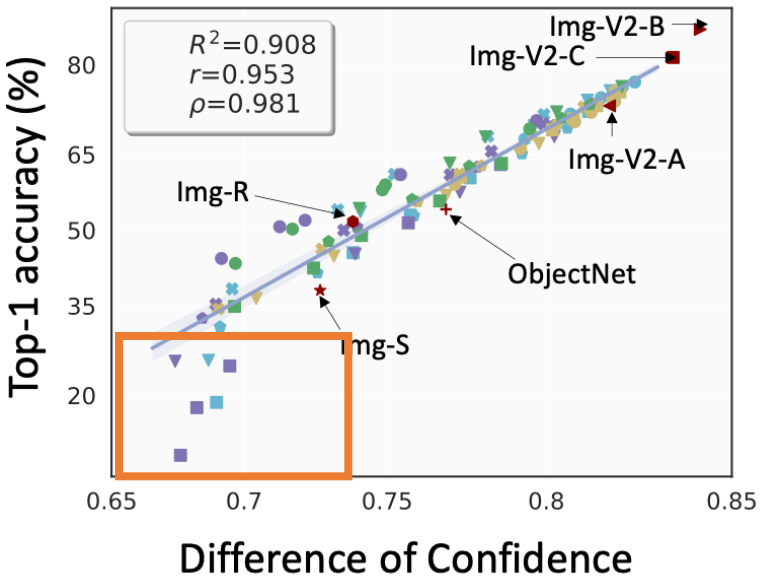


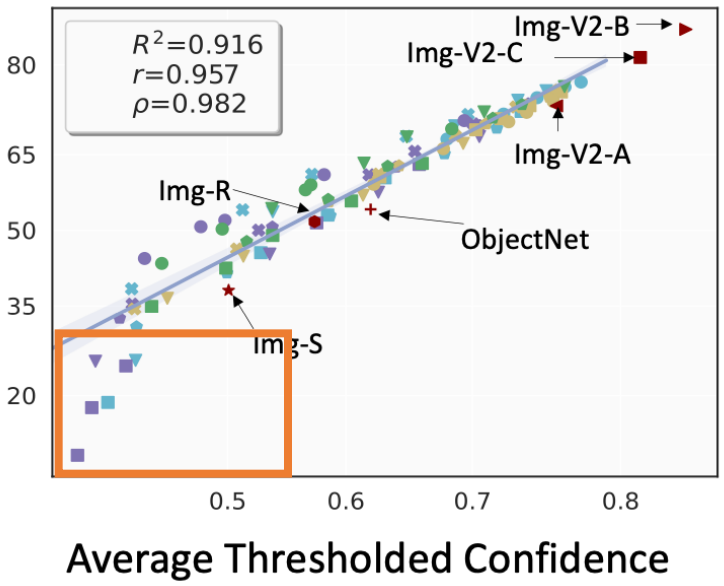
Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., & Schmidt. "Predicting with confidence on unseen distributions", In ICCV, 2021

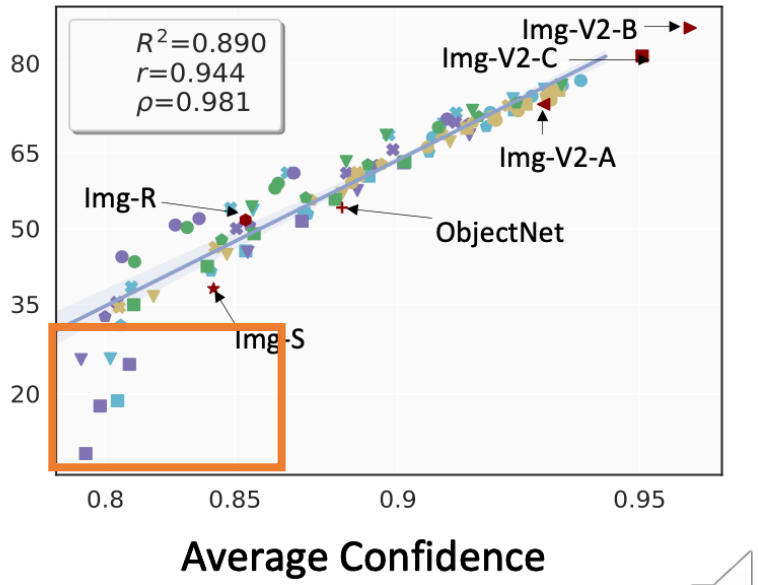
Prediction Alone Is Not Sufficient

Predication score-based method **cannot well capture** the test sets in the **low-accuracy region**

Each point denote one test set from ImageNet-C



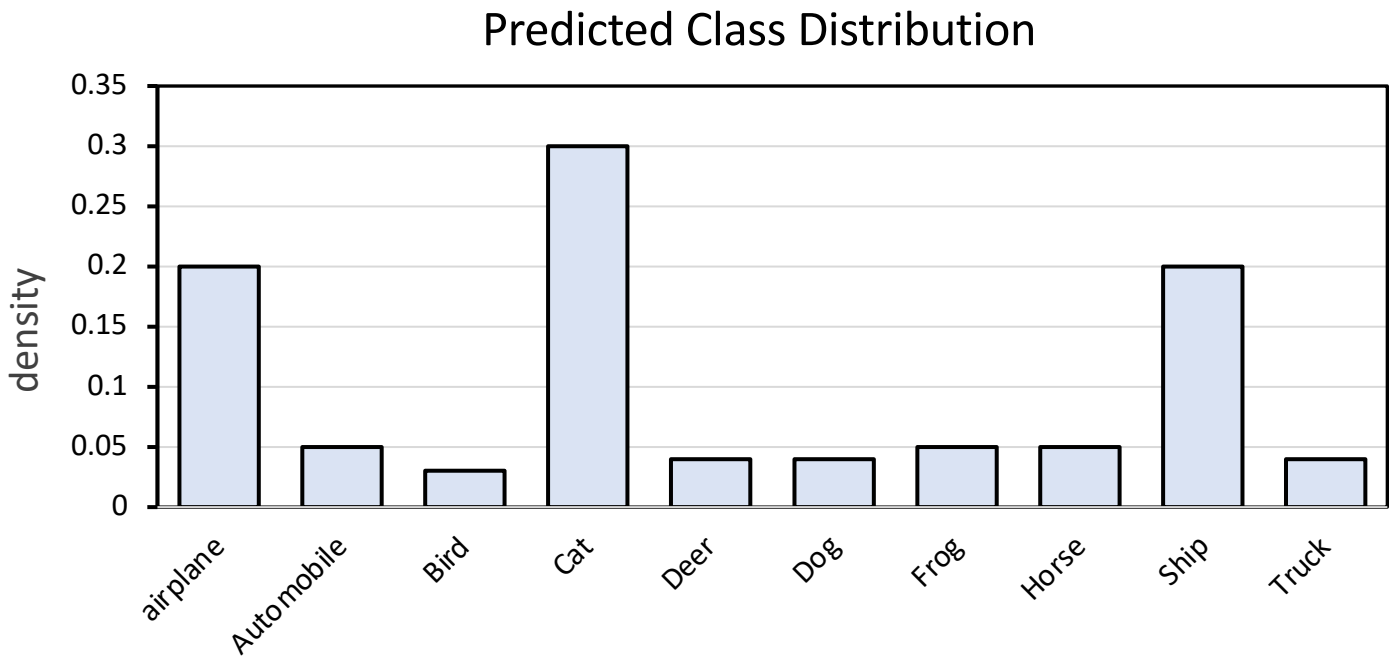




Prediction Alone Is Not Sufficient

- **Classifiers tend to classify data into specific categories**

CIFAR-10: predictions are mainly assigned to "cat", "boat", and "airplane".



Prediction Dispersity Is Also Informative

- **Confidence** reflects whether the individual prediction is certain
- **Dispersity** indicates how the predictions are distributed across all categories

Our key insight:

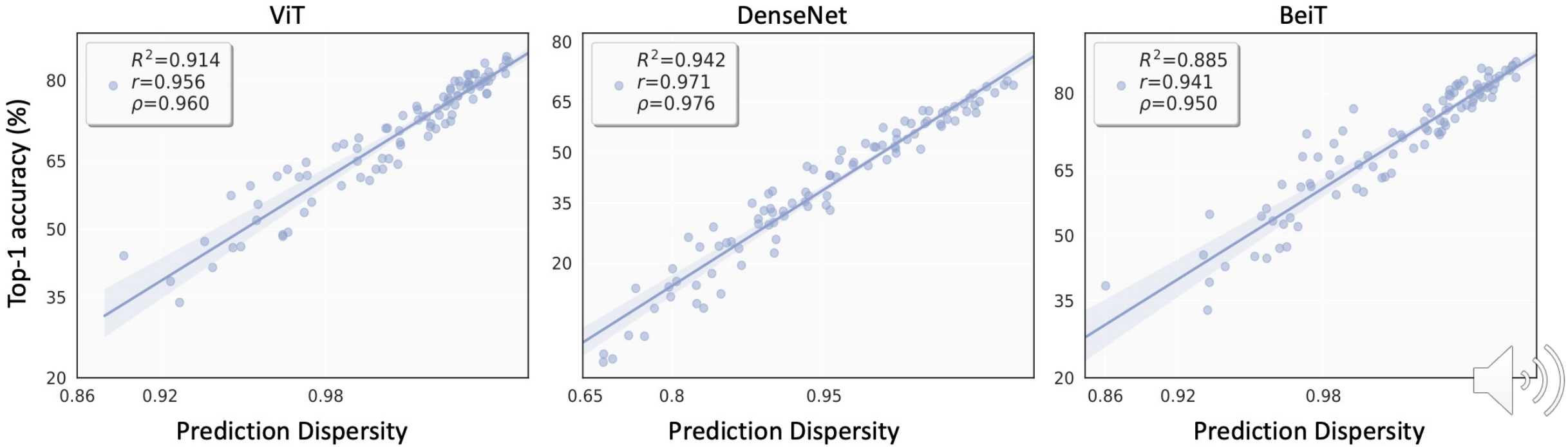
A well-performing model should give predictions with **high confidence and dispersity**

Prediction Dispersity Is Also Informative

Prediction Dispersity Score:

We first calculate the histogram of the number of the predicted class and then use entropy to measure the degree of balance

Each point denote one test set from ImageNet-C



Characterizing Confidence and Dispersity

Nuclear norm has been shown to be effective in characterizing both properties

Cui, Shuhao, et al. "Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations." In CVPR, 2020

Prediction Matrix $P \in \mathbb{R}^{N_t \times K}$

N_t test samples, and K classes

Nuclear Norm: the sum of singular values of prediction matrix

Nuclear Norm Is Effective

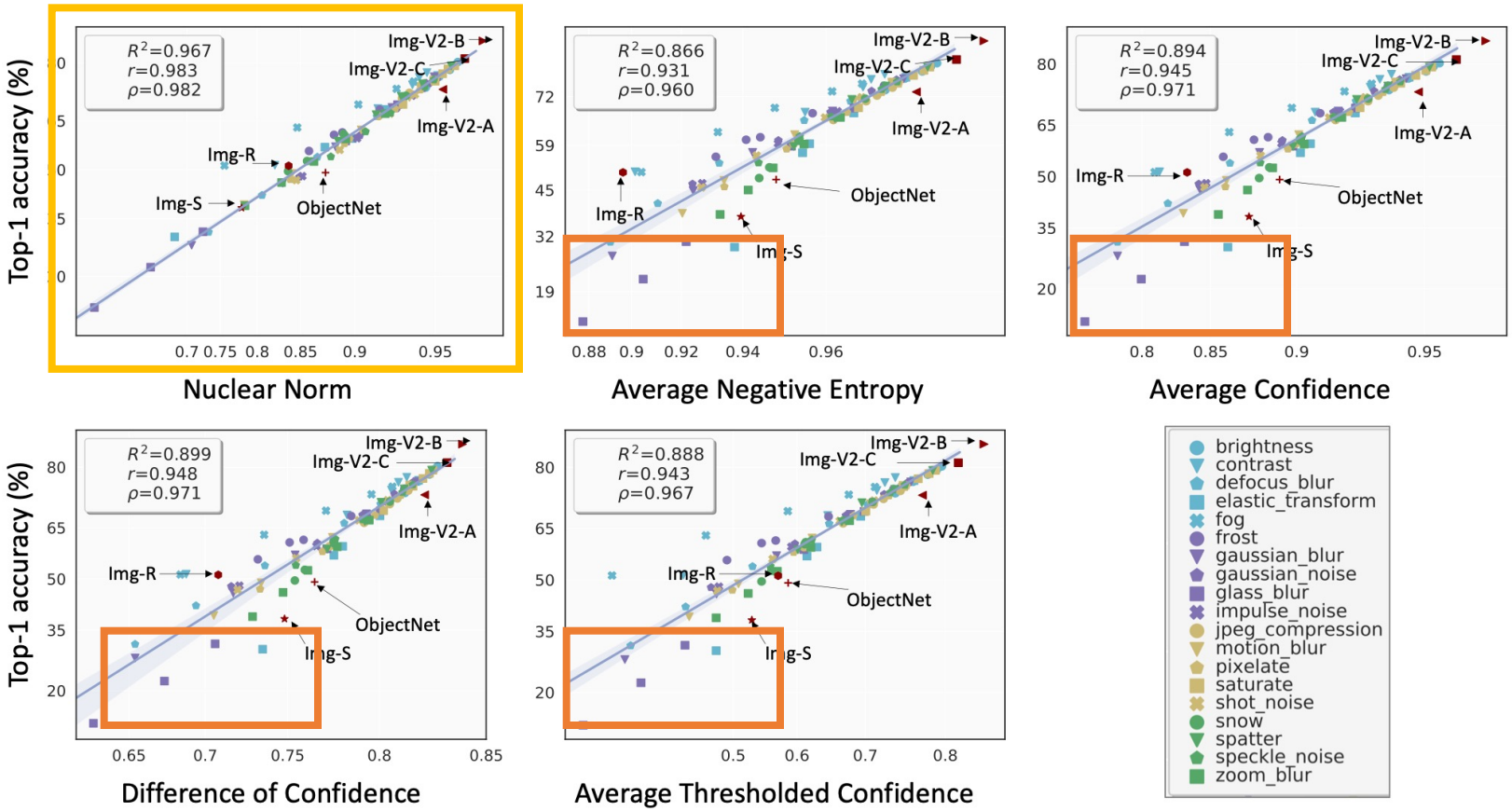
Nuclear norm exhibits the highest correlation strength (R^2 and ρ) with OOD accuracy across three setups

Setup	Model	AC		ANE		ATC		DoC		Nuclear Norm	
		R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ
ImageNet	ViT	0.970	0.990	0.964	0.988	0.978	0.990	0.961	0.990	**0.991**	**0.995**
	BeiT	0.977	0.994	0.964	0.989	0.985	0.995	0.979	0.994	**0.988**	**0.996**
	Swin	0.794	0.929	0.732	0.909	0.815	0.935	0.791	0.929	**0.949**	**0.961**
	DenseNet	0.938	0.984	0.929	0.979	0.961	0.989	0.937	0.984	**0.995**	**0.997**
	Res152-BiT	0.891	0.981	0.877	0.979	0.916	0.982	0.908	0.981	**0.981**	**0.991**
	ConvNeXt	0.894	0.971	0.866	0.960	0.888	0.967	0.899	0.971	**0.967**	**0.982**
	Average	0.911	0.975	0.889	0.968	0.924	0.976	0.911	0.975	**0.979**	**0.989**
CIFAR-10	ResNet-20	0.916	0.991	0.916	0.991	0.934	0.992	0.937	0.991	**0.989**	**0.995**
	RepVGG-A0	0.811	0.982	0.806	0.981	0.841	0.985	0.824	0.982	**0.992**	**0.996**
	VGG-11	0.973	0.994	0.973	0.995	0.984	**0.996**	0.964	0.994	**0.988**	**0.996**
	Average	0.900	0.989	0.900	0.988	0.920	0.991	0.908	0.989	**0.990**	**0.995**
CUB-200	ResNet-50	0.836	0.942	0.839	0.939	0.855	0.957	0.818	0.942	**0.989**	**0.997**
	ResNet-101	0.303	0.734	0.319	0.739	0.351	0.775	0.308	0.734	**0.987**	**0.998**
	PMG	0.892	0.979	0.893	0.977	0.977	0.991	0.903	0.979	**0.990**	**0.998**
	Average	0.677	0.885	0.684	0.885	0.727	0.908	0.677	0.885	**0.989**	**0.997**

Nuclear Norm Is Effective

Nuclear norm exhibits the highest correlation strength (R^2 and ρ) with OOD accuracy across three setups

ImageNet Setup

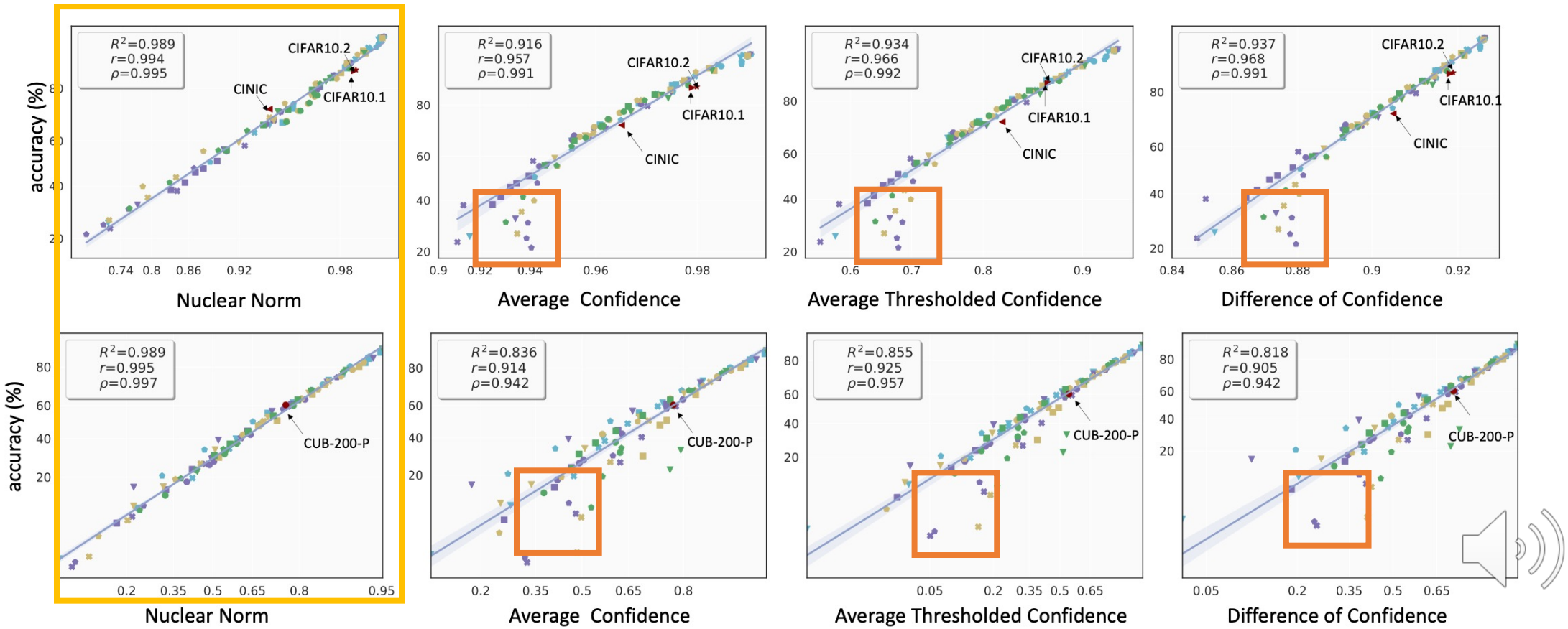


Nuclear Norm Is Effective

Nuclear norm exhibits the highest correlation strength (R^2 and ρ) with OOD accuracy across three setups

CIFAR-10 Setup

CUB-200 Setup



Discussion: Class Imbalance

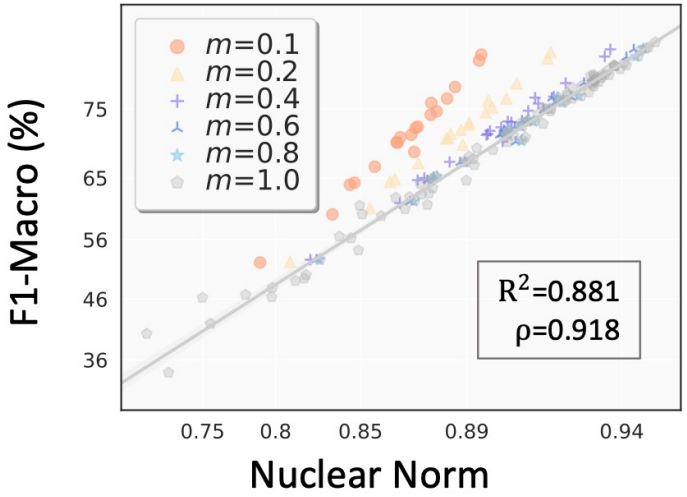
Point1: Nuclear Norm assumes the **class distribution is roughly uniform**

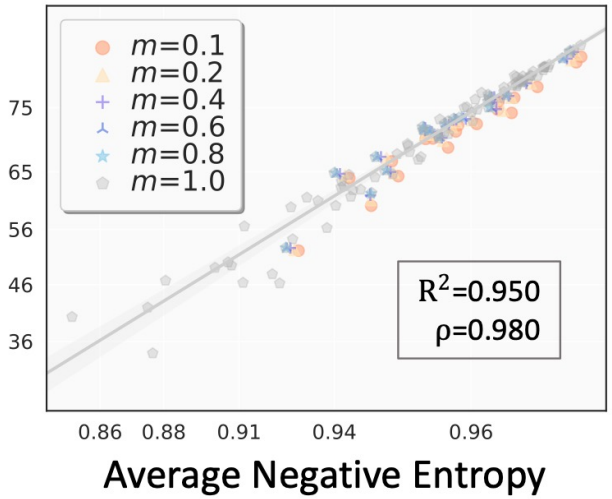
Observations:

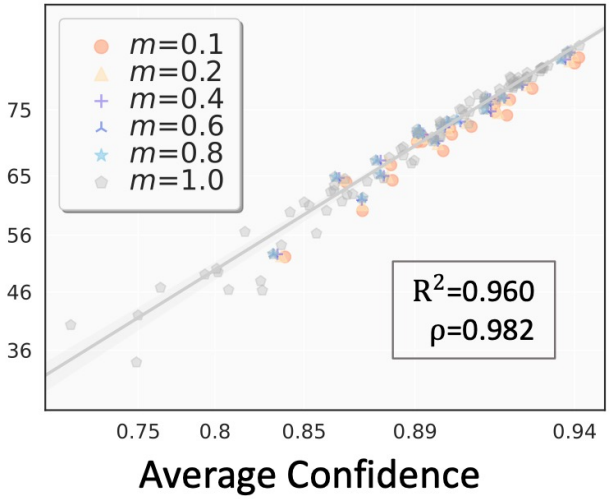
1) Other methods are stable under class imbalance;

2) Nuclear Norm **is resistant to moderate class imbalance;**

3) Nuclear Norm **is less effective under severe class imbalance**;





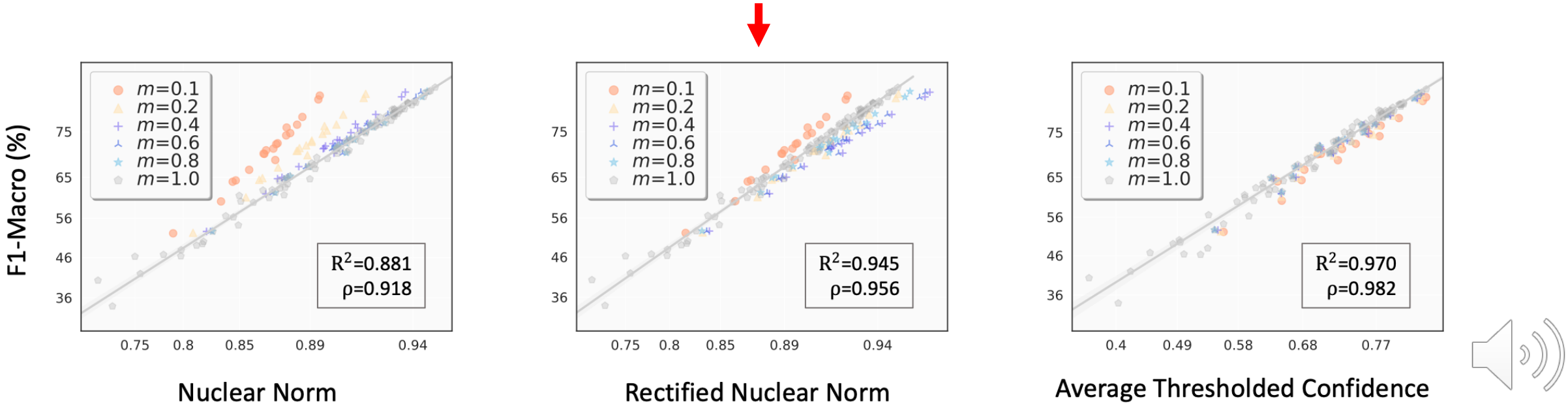


Discussion: Class Imbalance

Point2: Estimate class distribution using BBSE

Lipton, Z., Wang, Y.-X., and Smola, A. Detecting and correcting for label shift with black box predictors. In ICML, 2018

Make Nuclear Norm focus on major classes instead of all classes: it helps



Discussion: Class Imbalance

Point3: Accurately characterize the class-specific prediction dispersity

If we have **prior knowledge** about the imbalanced class distribution, we can expect class predictions to follow it rather than a uniform one.

Discussion: Class Imbalance

Point3: Accurately characterize the class-specific prediction dispersity

If we have **prior knowledge** about the imbalanced class distribution, we can expect class predictions to follow it rather than a uniform one.

Mutual Information Maximizing

$$H\left(\frac{1}{n_t}\sum_{i=1}^{n_t}\boldsymbol{P}_{i,:}\right) - \frac{1}{n_t}\sum_{i=1}^{n_t}H(\boldsymbol{P}_{i,:})$$

$\boldsymbol{P} \in \mathbb{R}^{n_t \times k}$ is the prediction matrix

By default, this term encourages the predictions to be globally balanced

Potential Improvements: revise it to align with the known class distribution

Summary

- **Prediction dispersity is a useful** property that correlates strongly with classifier accuracy on various test sets

- **Considering both prediction confidence and dispersity** using **nuclear norm** to achieve more **accurate accuracy estimations**

- Nuclear Norm **is resistant to moderate class imbalance**

Future work:

estimating class distribution for better characterizing prediction dispersity

Thank you!

For more information, please refer to https://weijiandeng.xyz