Unsupervised Model Evaluation

Weijian Deng

Build a model that can see and generalize

Australian National University

ACM Multimedia 2024

Melbourne, Australia

Personal Information

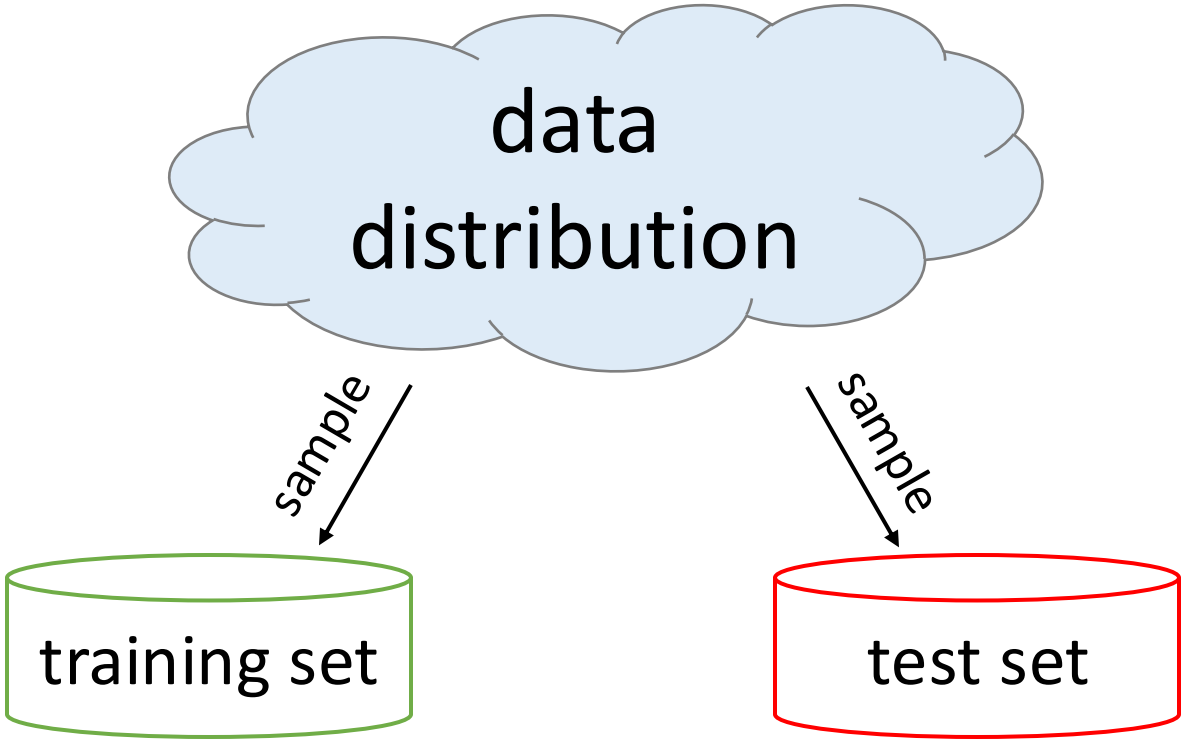
I am Actively Seeking Full-Time Research Roles

- PhD in Machine Learning
- Research Fellow at ANU

Research Interest

- Robust Foundational Vision-Language Models
- 3D Modelling & Generation

Pillars in Machine Learning

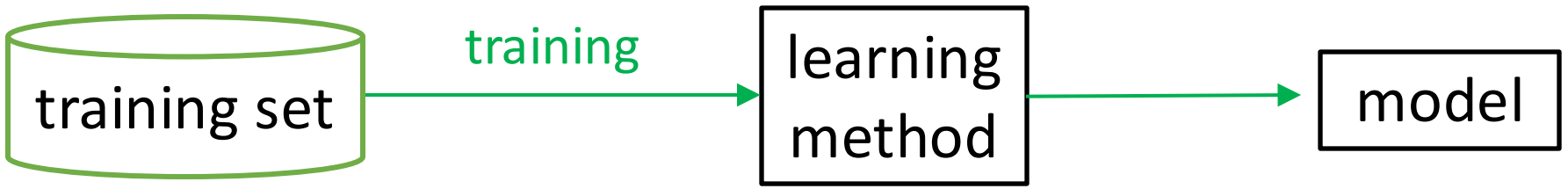


i.i.d. assumption

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

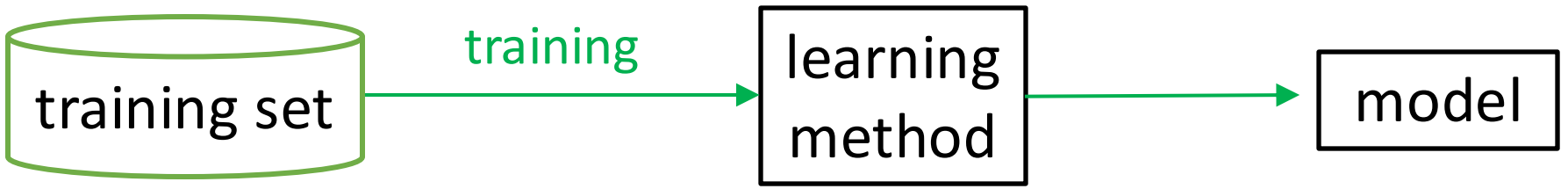
Pillars in Machine Learning

Training phase



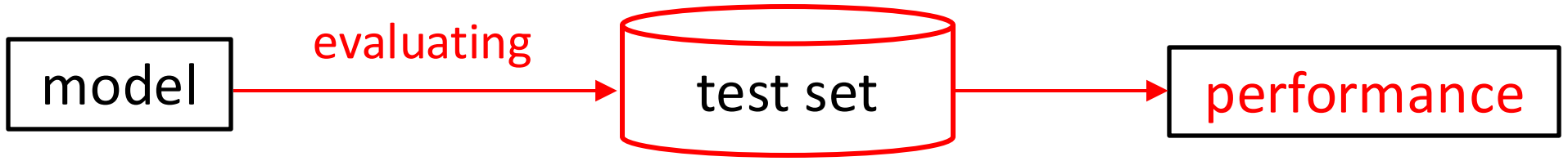
Pillars in Machine Learning

Training phase



Generalization evaluation

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



Supervised Evaluation in Textbook

Test set is fully annotated

Ground truths are provided

test image

#2

#4

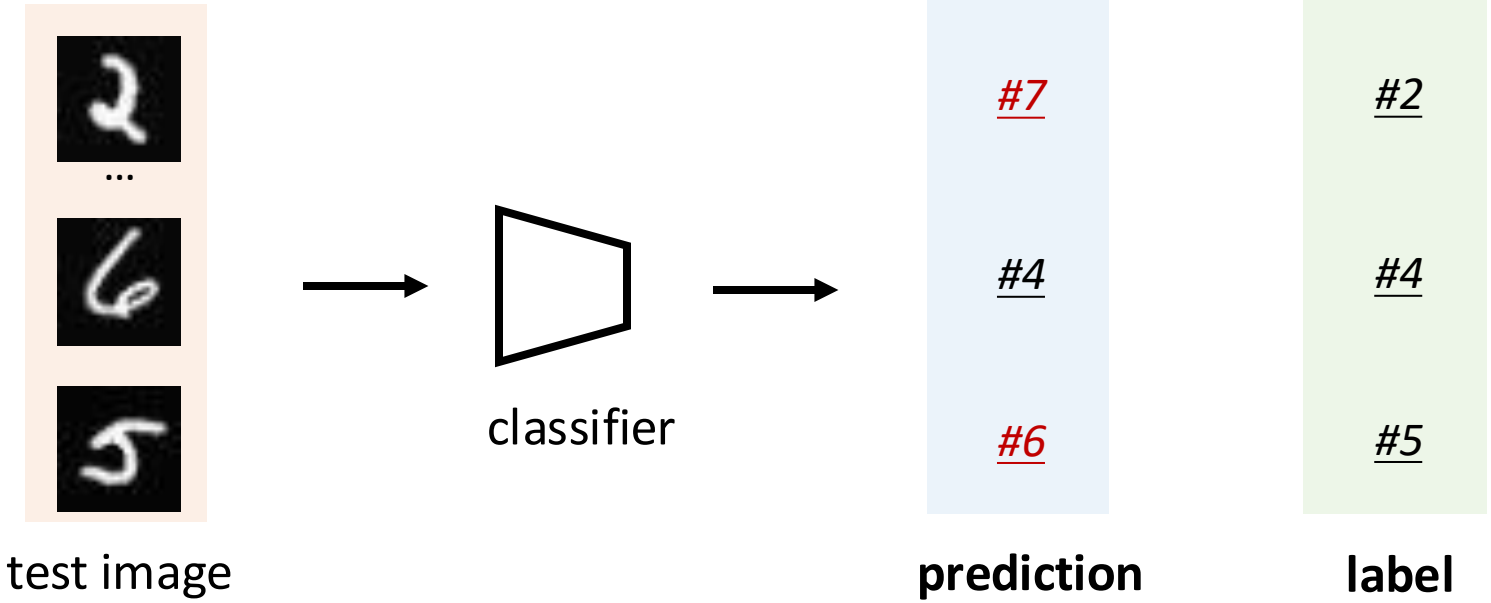
#5

label

Supervised Evaluation in Textbook

Test set is fully annotated

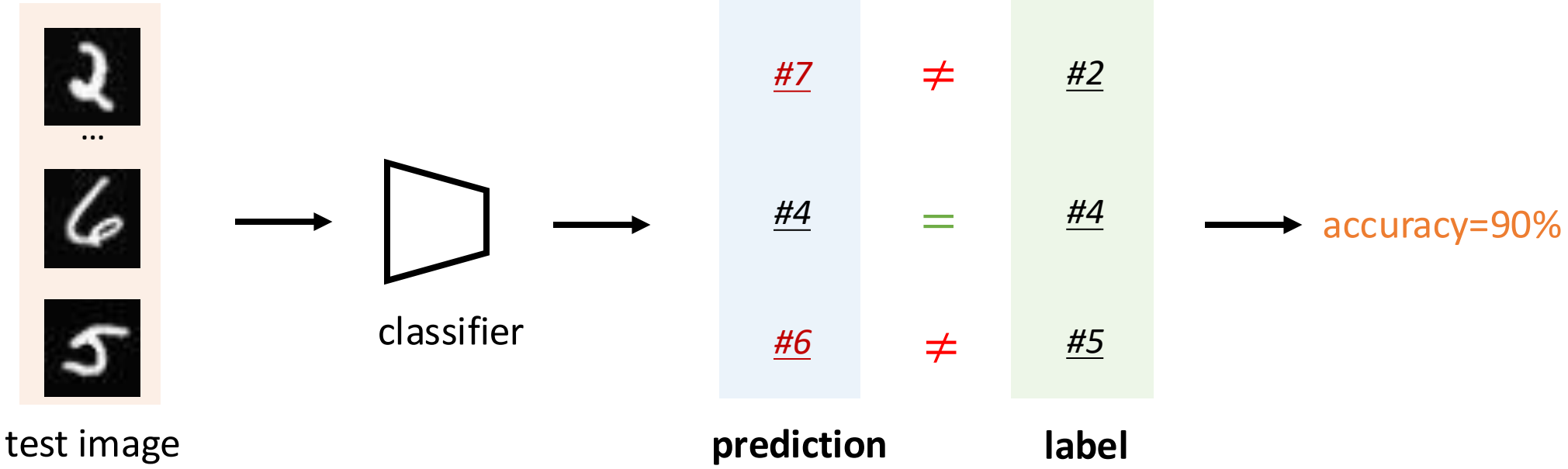
Ground truths are provided



Supervised Evaluation in Textbook

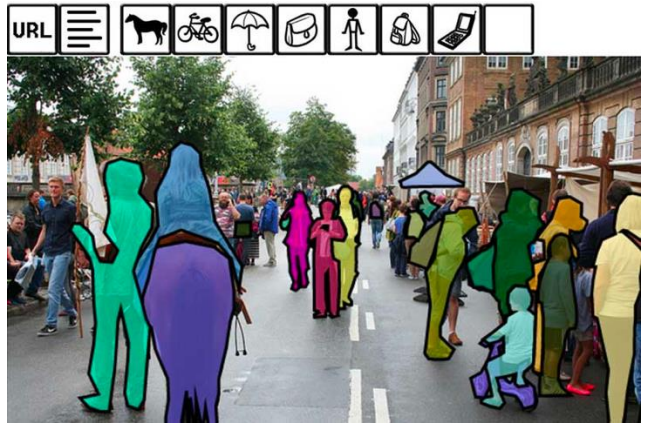
Test set is fully annotated

Ground truths are provided



In-distribution Benchmarks

ImageNet



MSCOCO

Cityscape

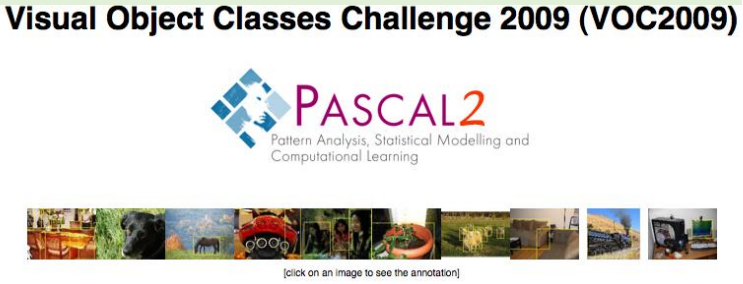
PASCAL

Is Supervised Evaluation Feasible?

Yes!

- Test set is fully annotated
- Training and test sets are usually from the same distribution

Cityscape



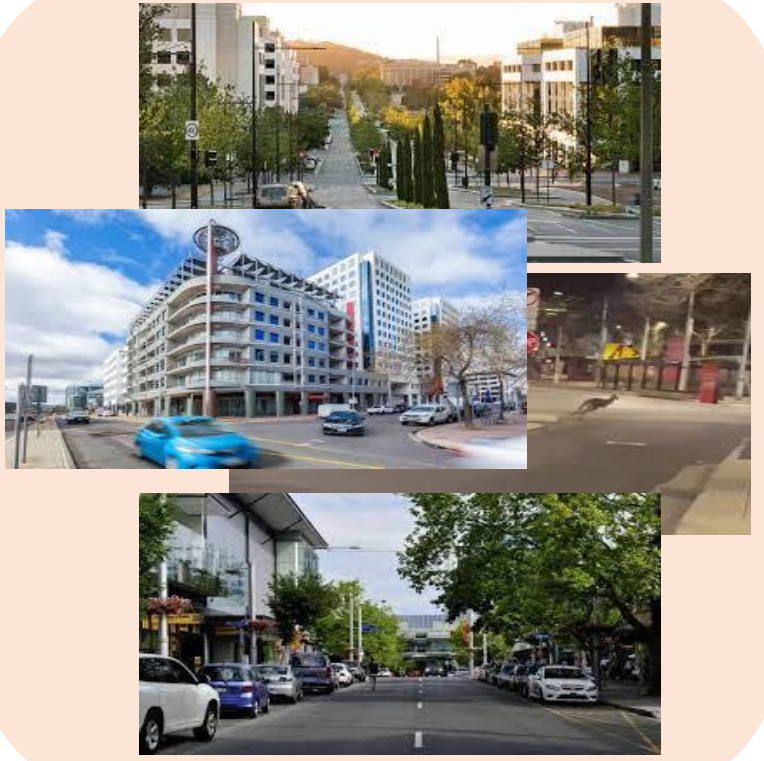
PASCAL

When Deploying a Self-Driving System?

Bremen city

Deploying

Canberra city



training

Self-Driving System

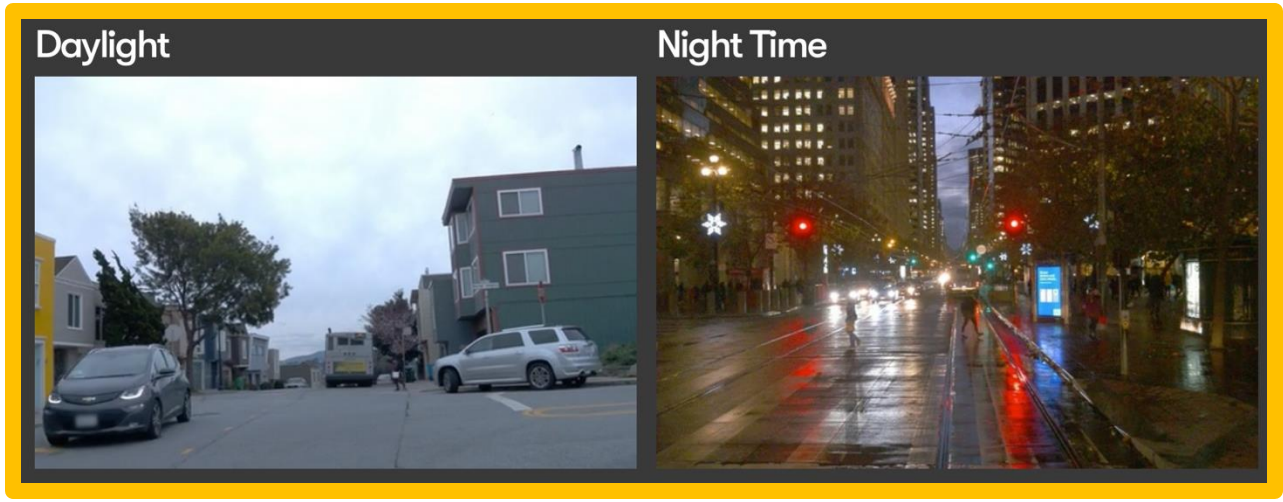
Geiger, Andreas, et al. "Vision meets robotics: The kitti dataset." The International Journal of Robotics Research 32.11 (2013): 1231-1237. AP
Cordts, Marius, et al. "The cityscapes dataset for semantic urban scene understanding." In CVPR, 2016

When Deploying a Self-Driving System…

Out-of-distribution test set

Distribution shift:

- Lighting condition (daylight vs. night time)
- Location (city vs. suburban)
- Environments (weather / construction)
- …



Sun, Pei, et al. "Scalability in perception for autonomous driving: Waymo open dataset." In CVPR, 2020

When Deploying a Self-Driving System…

Out-of-distribution test set

Distribution shift:

No!

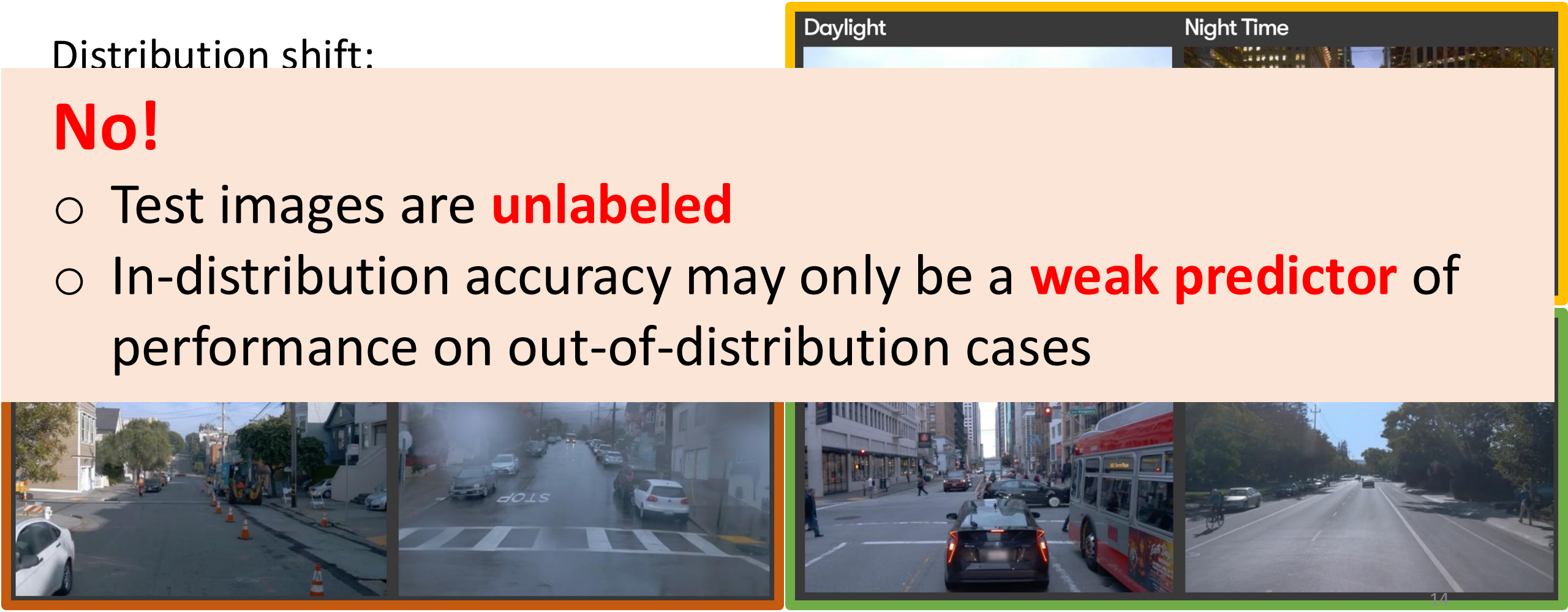
- Test images are **unlabeled**

Sun, Pei, et al. "Scalability in perception for autonomous driving: Waymo open dataset." In CVPR, 2020

When Deploying a Self-Driving System…

Out-of-distribution test set

Distribution shift:

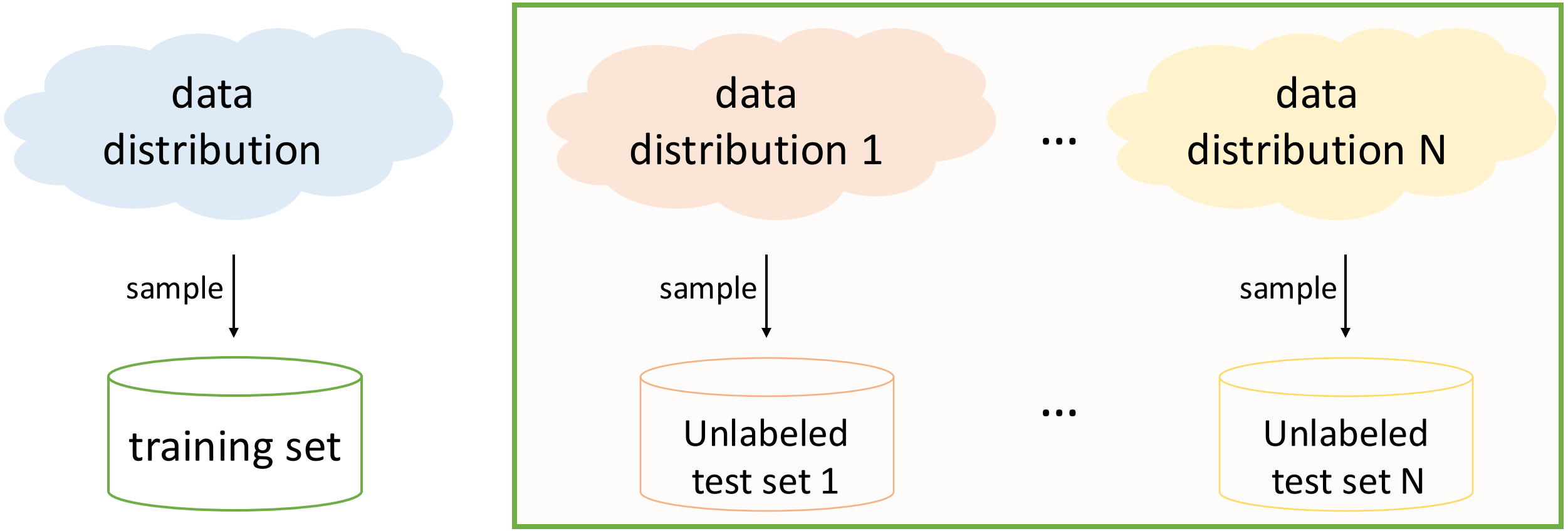


No!

- Test images are **unlabeled**
- In-distribution accuracy may only be a **weak predictor** of performance on out-of-distribution cases

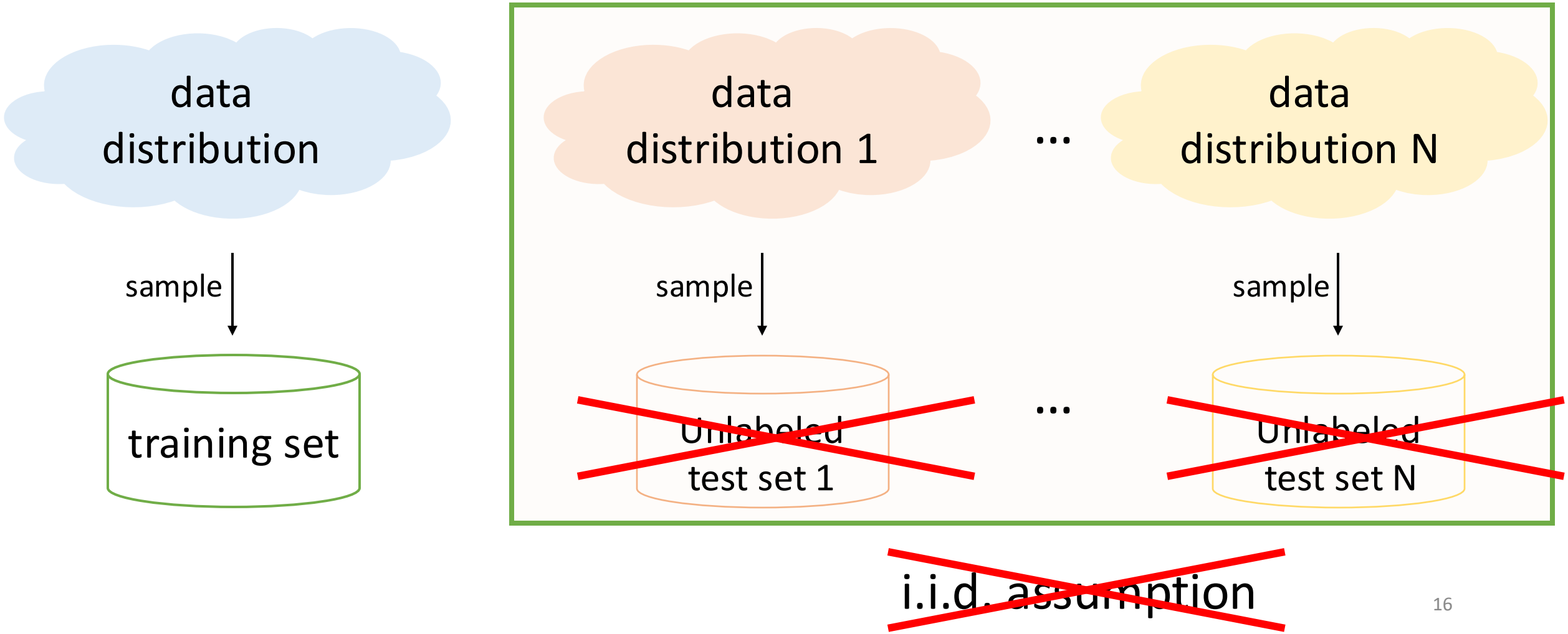
Sun, Pei, et al. "Scalability in perception for autonomous driving: Waymo open dataset." In CVPR, 2020

Evaluation Beyond Textbook:
Out-of-distribution and Unlabelled Evaluation

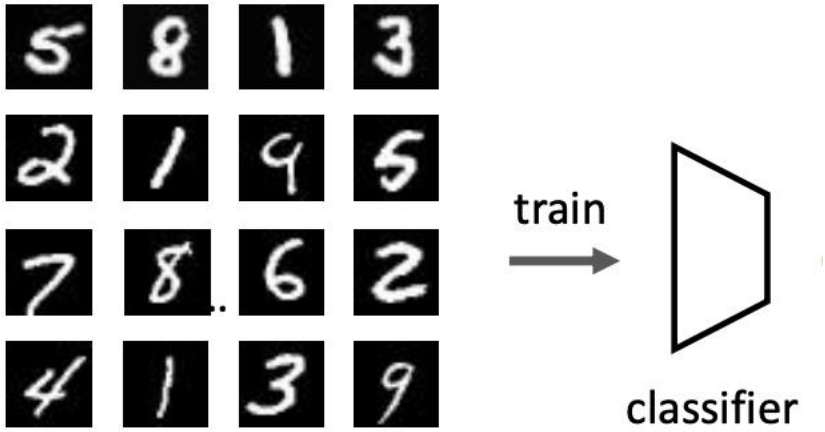


i.i.d. assumption

Evaluation Beyond Textbook:
Out-of-distribution and Unlabelled Evaluation



Unsupervised Evaluation: Problem Definition

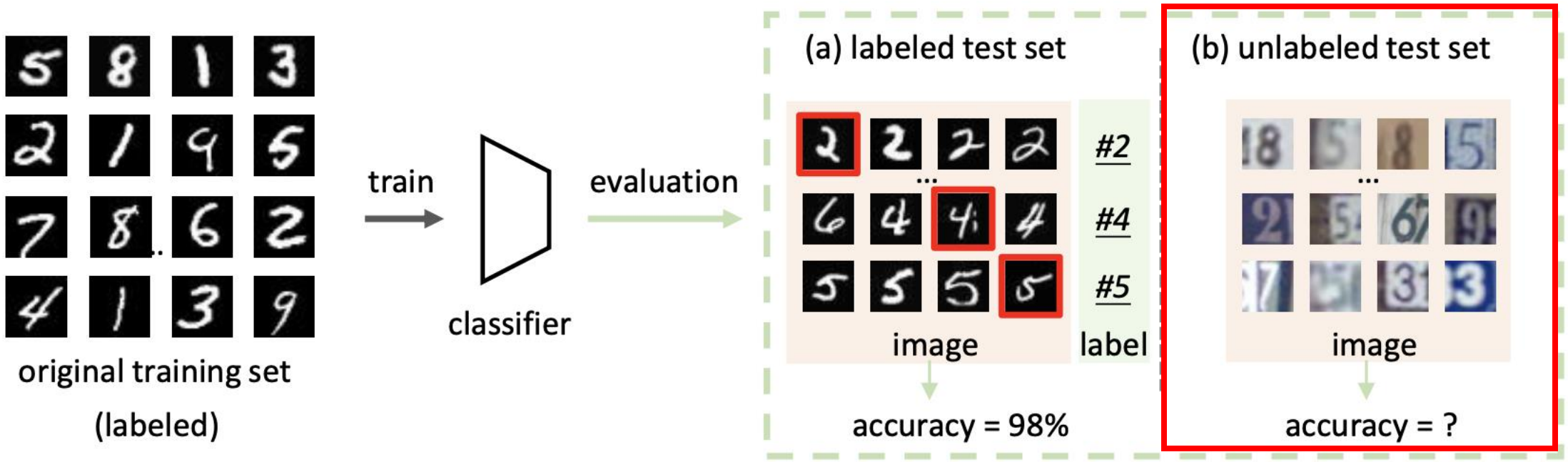


Given

- A training dataset
- A classifier trained on this dataset
- A test set without labels

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

Unsupervised Evaluation: Problem Definition



Given

- A training dataset
- A classifier trained on this dataset
- A test set without labels

We want to *estimate*:
accuracy on the unlabelled test set

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

Perspectives

In collaboration with other researchers, we have contributed three perspectives

1. Weijian Deng, Liang Zheng: **AutoEval: Are Labels Always Necessary for Classifier Accuracy Evaluation?** (**TPAMI 2021**)
2. Weijian Deng, Liang Zheng: **Are Labels Always Necessary for Classifier Accuracy Evaluation?** (**CVPR 2021**).
3. Weijie Tu, Weijian Deng, Tom Gedeon, Liang Zheng: **A Bag-of-Prototypes Representation for Dataset-Level Applications** (**CVPR 2023**).
4. Weijian Deng, Stephen Gould, Liang Zheng: **What Does Rotation Prediction Tell Us About Classifier Accuracy Under Varying Testing Environments?** (**ICML 2021)**
5. Xiaoxiao Sun, Yunzhong Hou, Weijian Deng, Hongdong Li, Liang Zheng: **Ranking Models in Unlabeled New Environments** (**ICCV 2021**).
6. Yuli Zou*, Weijian Deng*, Liang Zheng (*Equal Contribution): **Adaptive Calibrator Ensemble: Navigating Test Set Difficulty in Out-of-Distribution Scenarios** (**ICCV 2023**).
7. Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, Tom Gedeon: **An Empirical Study Into What Matters for Calibrating Vision-Language Models** (**ICML 2024**).
8. Weijie Tu, Weijian Deng, Tom Gedeon: **A Closer Look at the Robustness of Contrastive Language-Image Pre-Training (CLIP)** (**NeurIPS 2023**).
9. Weijian Deng, Stephen Gould, Liang Zheng: **On the Strong Correlation Between Model Invariance and Generalization** (**NeurIPS 2022**).
10. Renchunzi Xie, Ambroise Odonnat, Vasilii Feofanov, Weijian Deng, Jianfeng Zhang, Bo An: **MANO: Exploiting Matrix Norm for Unsupervised Accuracy Estimation Under Distribution Shifts** (**NeurIPS 2024**).
11. Weijie Tu, Weijian Deng, Tom Gedeon, Liang Zheng: **What Does Softmax Probability Tell Us About Classifiers Ranking Across Diverse Test Conditions?** (**TMLR 2024**).
12. Weijian Deng, Yumin Suh, Stephen Gould, Liang Zheng: **Confidence and Dispersity Speak: Characterizing Prediction Matrix for Unsupervised Accuracy Estimation** (**ICML 2023**).

Perspectives

In collaboration with other researchers, we have contributed three perspectives

Dataset-Dataset Distance

1. Weijian Deng, Liang Zheng: **AutoEval: Are Labels Always Necessary for Classifier Accuracy Evaluation?** (**TPAMI 2021**)
2. Weijian Deng, Liang Zheng: **Are Labels Always Necessary for Classifier Accuracy Evaluation?** (**CVPR 2021**).
3. Weijie Tu, Weijian Deng, Tom Gedeon, Liang Zheng: **A Closer Look at the Intra-Dataset-Level Applications** (**CVPR 2023**).
4. Weijian Deng, Stephen Gould, Liang Zheng: **What Does Rotation Prediction Tell Us About Classifier Accuracy Under Varying Testing Environments?** (**ICML 2021**)
5. Xiaoxiao Sun, Yunzhong Hou, Weijian Deng, Hongdong Li, Liang Zheng: **Ranking Models in Unlabeled New Environments** (**ICCV 2021**).
6. Yuli Zou*, Weijian Deng*, Liang Zheng (*Equal Contribution): **Adaptive Calibrator Ensemble: Navigating Test Set Difficulty in Out-of-Distribution Scenarios** (**ICCV 2023**).
7. Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, Tom Gedeon: **An Empirical Study Into What Matters for Calibrating Vision-Language Models** (**ICML 2024**).
8. Weijie Tu, Weijian Deng, Tom Gedeon: **A Closer Look at the Robustness of Contrastive Language-Image Pre-Training (CLIP)** (**NeurIPS 2023**).
9. Weijian Deng, Stephen Gould, Liang Zheng: **On the Strong Correlation Between Model Invariance and Generalization** (**NeurIPS 2022**).
10. Renchunzi Xie, Ambroise Odonnat, Vasilii Feofanov, Weijian Deng, Jianfeng Zhang, Bo An: **MANO: Exploiting Matrix Norm for Unsupervised Accuracy Estimation Under Distribution Shifts** (**NeurIPS 2024**).
11. Weijie Tu, Weijian Deng, Tom Gedeon, Liang Zheng: **What Does Softmax Probability Tell Us About Classifiers Ranking Across Diverse Test Conditions?** (**TMLR 2024**).
12. Weijian Deng, Yumin Suh, Stephen Gould, Liang Zheng: **Confidence and Dispersity Speak: Characterizing Prediction Matrix for Unsupervised Accuracy Estimation** (**ICML 2023**).

Perspectives

In collaboration with other researchers, we have contributed three perspectives

Dataset-Dataset Distance

Proxy Task/ Dataset

1. Weijian Deng, Liang Zheng: **AutoEval: Are Labels Always Necessary for Classifier Accuracy Evaluation?** (**TPAMI 2021**)
2. Weijian Deng, Liang Zheng: **Are Labels Always Necessary for Classifier Accuracy Evaluation?** (**CVPR 2021**).
3. Weijie Tu, Weijian Deng, Tom Gedeon, Liang Zheng: **A Closer Look at the Intra-Dataset-Level Applications** (**CVPR 2023**).
4. Weijian Deng, Stephen Gould, Liang Zheng: **What Does Rotation Prediction Tell Us About Classifier Accuracy Under Varying Testing Environments?** (**ICML 2021**)
5. Xiaoxiao Sun, Yunzhong Hou, Weijian Deng, Hongdong Li, Liang Zheng: **Ranking Models in Unlabeled New Environments** (**ICCV 2021**).
6. Yuli Zou*, Weijian Deng*, Liang Zheng (*Equal Contribution): **Adaptive Calibrator Ensemble: Navigating Test Set Difficulty in Out-of-Distribution Scenarios** (**ICCV 2023**).
7. Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, Tom Gedeon: **An Empirical Study Into What Matters for Calibrating Vision-Language Models** (**ICML 2024**).
8. Weijie Tu, Weijian Deng, Tom Gedeon: **A Closer Look at the Robustness of Contrastive Language-Image Pre-Training (CLIP)** (**NeurIPS 2023**).
9. Weijian Deng, Stephen Gould, Liang Zheng: **On the Strong Correlation Between Model Invariance and Generalization** (**NeurIPS 2022**).
10. Renchunzi Xie, Ambroise Odonnat, Vasilii Feofanov, Weijian Deng, Jianfeng Zhang, Bo An: **MANO: Exploiting Matrix Norm for Unsupervised Accuracy Estimation Under Distribution Shifts** (**NeurIPS 2024**).
11. Weijie Tu, Weijian Deng, Tom Gedeon, Liang Zheng: **What Does Softmax Probability Tell Us About Classifiers Ranking Across Diverse Test Conditions?** (**TMLR 2024**).
12. Weijian Deng, Yumin Suh, Stephen Gould, Liang Zheng: **Confidence and Dispersity Speak: Characterizing Prediction Matrix for Unsupervised Accuracy Estimation** (**ICML 2023**).

Perspectives

In collaboration with other researchers, we have contributed three perspectives

Dataset-Dataset Distance

1. Weijian Deng, Liang Zheng: **AutoEval: Are Labels Always Necessary for Classifier Accuracy Evaluation?** (**TPAMI 2021**)
2. Weijian Deng, Liang Zheng: **Are Labels Always Necessary for Classifier Accuracy Evaluation?** (**CVPR 2021**).
3. Weijie Tu, Weijian Deng, Tom Gedeon, Liang Zheng: **A Closer Look at the Intra-Dataset-Level Applications** (**CVPR 2023**).
4. Weijian Deng, Stephen Gould, Liang Zheng: **What Does Rotation Prediction Tell Us About Classifier Accuracy Under Varying Testing Environments?** (**ICML 2021**)

Proxy Task/ Dataset

5. Xiaoxiao Sun, Yunzhong Hou, Weijian Deng, Hongdong Li, Liang Zheng: **Ranking Models in Unlabeled New Environments** (**ICCV 2021**).
6. Yuli Zou*, Weijian Deng*, Liang Zheng (*Equal Contribution): **Adaptive Calibrator Ensemble: Navigating Test Set Difficulty in Out-of-Distribution Scenarios** (**ICCV 2023**).
7. Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, Tom Gedeon: **An Empirical Study Into What Matters for Calibrating Vision-Language Models** (**ICML 2024**).
8. Weijie Tu, Weijian Deng, Tom Gedeon: **A Closer Look at the Robustness of Contrastive Language-Image Pre-Training (CLIP)** (**NeurIPS 2023**).

Model Output

9. Weijian Deng, Stephen Gould, Liang Zheng: **On the Strong Correlation Between Model Invariance and Generalization** (**NeurIPS 2022**).
10. Renchunzi Xie, Ambroise Odonnat, Vasilii Feofanov, Weijian Deng, Jianfeng Zhang, Bo An: **MANO: Exploiting Matrix Norm for Unsupervised Accuracy Estimation Under Distribution Shifts** (**NeurIPS 2024**).
11. Weijie Tu, Weijian Deng, Tom Gedeon, Liang Zheng: **A Closer Look at the Calibration Stability Tell Us About Classifiers Ranking Across Diverse Test Conditions?** (**TMLR 2024**).
12. Weijian Deng, Yumin Suh, Stephen Gould, Liang Zheng: **Confidence and Dispersity Speak: Characterizing Prediction Matrix for Unsupervised Accuracy Estimation** (**ICML 2023**).

Model Outputs Are Already Informative

Model Outputs Are Already Informative

- Model predictions are already informative

model

1) Predicted class

$\hat{y}$

"dog"

2) Prediction confidence

$\hat{p}$

Maximum Softmax output

Model Outputs Are Already Informative

- Model predictions are already informative

model

1) Predicted class

$\hat{y}$

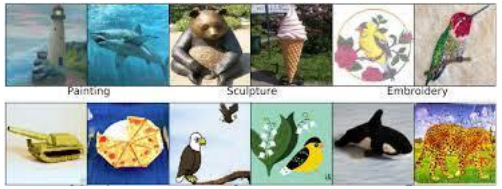
"dog"

2) Prediction confidence

$\hat{p}$

Maximum Softmax output

Test set 1



Average confidence (80%)

...

Test set N

Average confidence (30%)

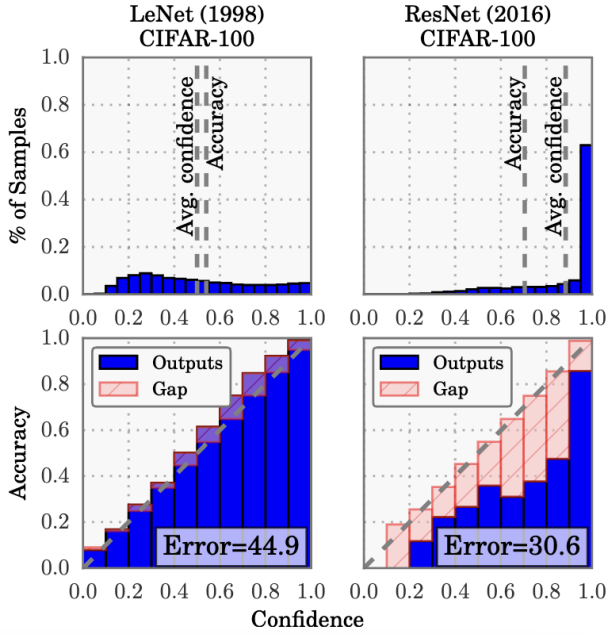
Models Tend to be poorly-calibrated

On Calibration of Modern Neural Networks

Chuan Guo [*1] **Geoff Pleiss** [*1] **Yu Sun** [*1] **Kilian Q. Weinberger** [1]

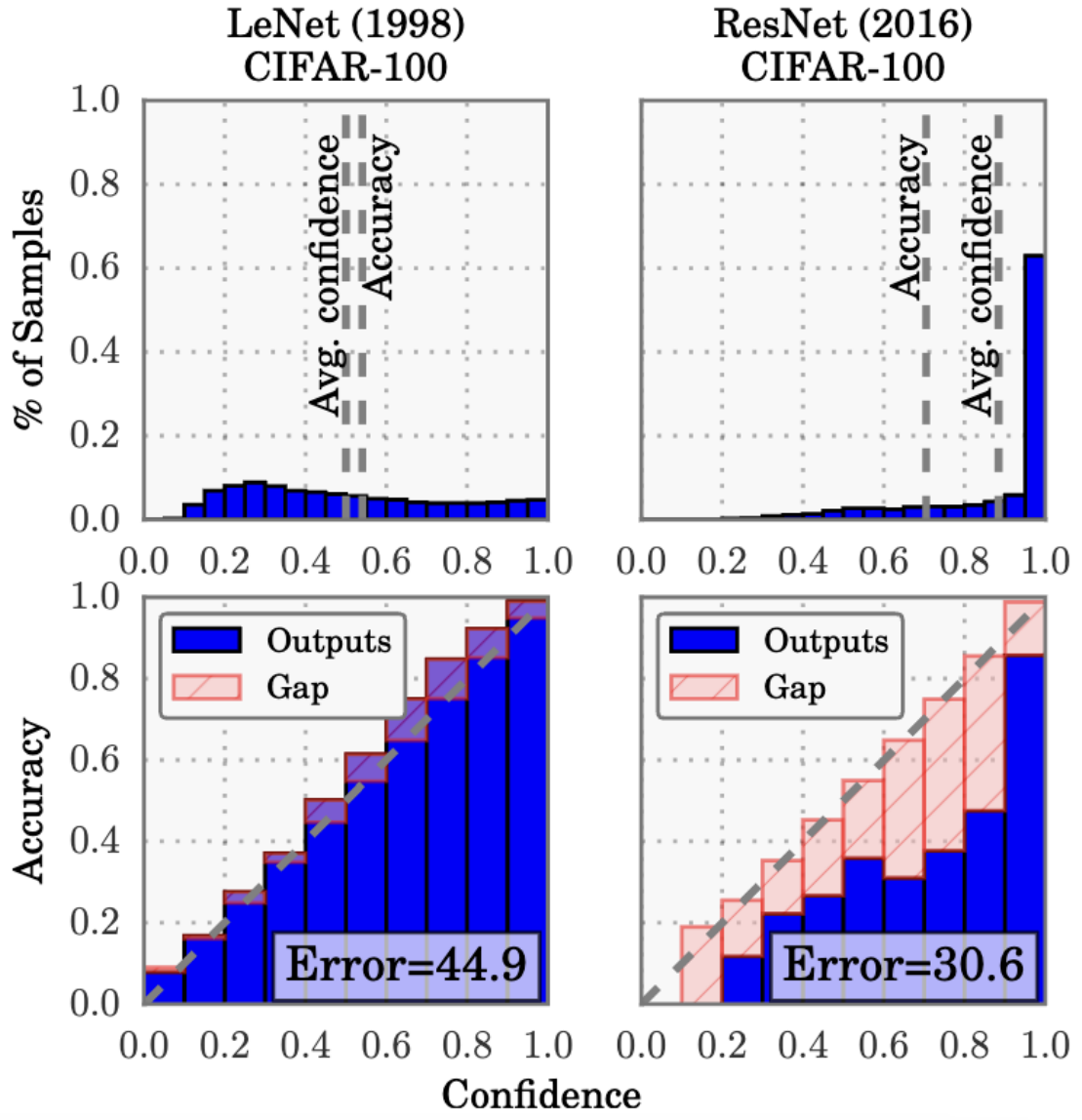
Abstract

Confidence calibration – the problem of predicting probability estimates representative of the true correctness likelihood – is important for classification models in many applications. We discover that modern neural networks, unlike those from a decade ago, are poorly calibrated. Through extensive experiments, we observe that depth, width, weight decay, and Batch Normalization are important factors influencing calibration. We evaluate the performance of various post-processing calibration methods on state-of-the-art architectures with image and document classification datasets. Our analysis and experiments not only offer insights into neural network learning, but also provide a simple and straightforward recipe for practical settings: on most datasets, *temperature scaling* – a single-parameter variant of Platt Scaling – is surprisingly effective at calibrating predictions.



On Calibration of Modern Neural Networks. In ICML 2017

Models Tend to be poorly-calibrated



On Calibration of Modern Neural Networks. In ICML 2017

Models Still Tend to be poorly-calibrated

Revisiting the Calibration of Modern Neural Networks

Matthias Minderer **Josip Djolonga** **Rob Romijnders** **Frances Hubis**
Xiaohua Zhai **Neil Houlsby** **Dustin Tran** **Mario Lucic**
Google Research, Brain Team
{mjlm, lucic}@google.com

Abstract

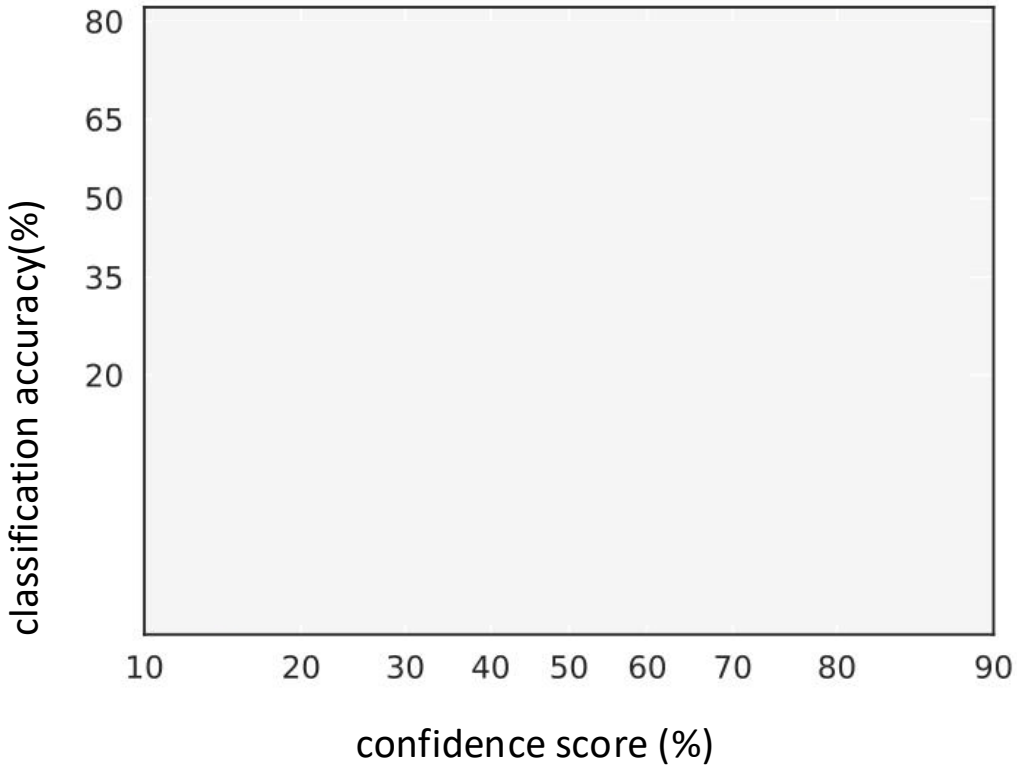
Accurate estimation of predictive uncertainty (model calibration) is essential for the safe application of neural networks. Many instances of miscalibration in modern neural networks have been reported, suggesting a trend that newer, more accurate models produce poorly calibrated predictions. Here, we revisit this question for recent state-of-the-art image classification models. We systematically relate model calibration and accuracy, and find that the most recent models, notably those not using convolutions, are among the best calibrated. Trends observed in prior model generations, such as decay of calibration with distribution shift or model size, are less pronounced in recent architectures. We also show that model size and amount of pretraining do not fully explain these differences, suggesting that architecture is a major determinant of calibration properties.

Revisiting the Calibration of Modern Neural Networks. In NeurIPS 2021

Model Outputs Are Already Informative

- Prediction confidence is indicative for unsupervised model evaluation

Strong Correlation

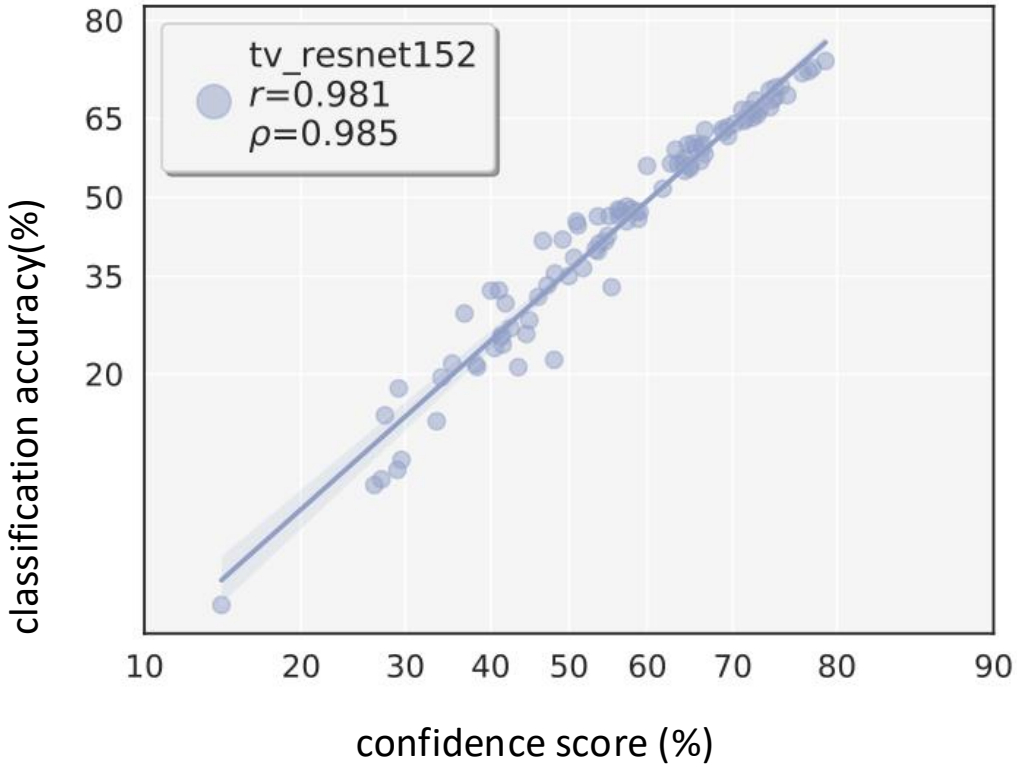


Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., & Schmidt. "Predicting with confidence on unseen distributions", In ICCV, 2021

Deng, W., Suh, Y., Gould, S. and Zheng, L . Confidence and Dispersity Speak: Characterising Prediction Matrix for Unsupervised Accuracy Estimation. In ICML 2024

Model Outputs Are Already Informative

- Prediction confidence is indicative for unsupervised model evaluation



Every point is a dataset

Strong Correlation

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

Deng, W., Suh, Y., Gould, S. and Zheng, L . Confidence and Dispersity Speak: Characterising Prediction Matrix for Unsupervised Accuracy Estimation. In ICML 2024

Thank You!