CENTER FOR BIOSYSTEMS AND BIOTECH DATA SCIENCE

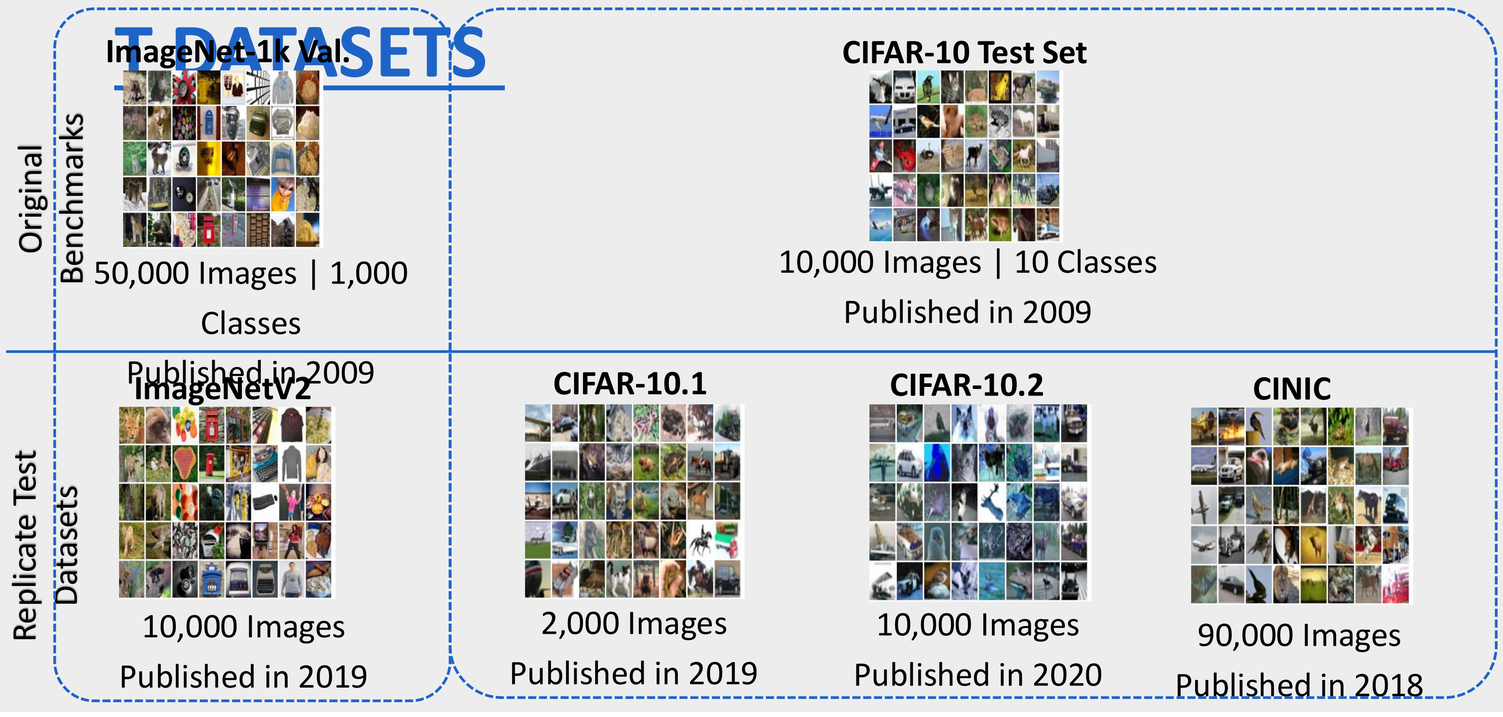
Esla Timothy Anzaku, Haohan Wang, Ajiboye Babalola, Arnout Van Messem, and Wesley De Neve

RE-ASSESSING ACCURACY DEGRADATION:

UNDERSTANDING DNN BEHAVIOR ON SIMILAR-BUT-NON-IDENTICAL TES T DATASETS

How do we diagnose AI model failure on near-identical test datasets?

- Deep Neural Networks (DNNs) power AI systems, and their reliability is crucial
- DNNs perform well on standard benchmarks like ImageNet and CIFAR-10
- Yet, their accuracy drops substantially when tested on replicated test datasets
- Are these accuracy drops due to true model failure, or are we misinterpreting the effectiveness of DNNs?
- This work offers a new lens on how to evaluate the reliability of DNNs under minimal dataset changes



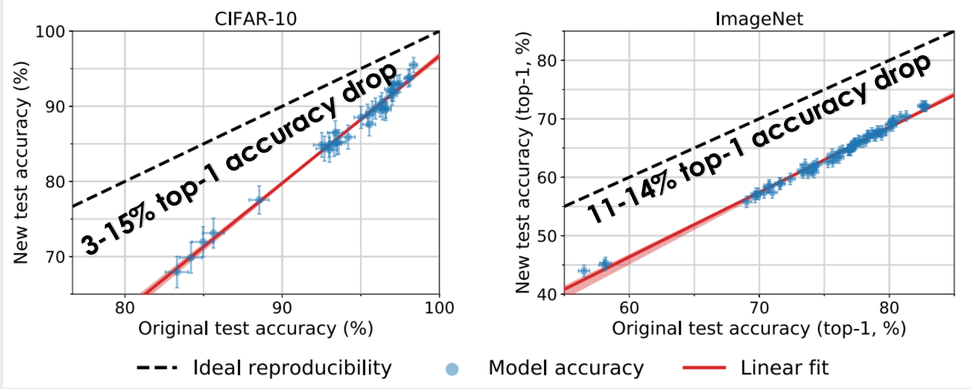
1. Conventional Evaluation Approach

Methodology

- The accuracy across all datapoints in the original and replicated test datasets is evaluated for comparison
- This ignores how confident DNNs are about their predictions

Results

- Unexpected and largely unexplained accuracy gaps



Conclusions

- Unexpected accuracy drops are interpreted as evidence of poor generalization
- DNNs are seen as less trustworthy, even under small data changes

Implications

- DNN reliability is questioned on replicated datasets
- Both researchers and practitioners could be misled about model behavior

Recht et al., "Do ImageNet Classifiers Generalize to ImageNet?," ICML, 2019.

2. Proposed Evaluation Framework

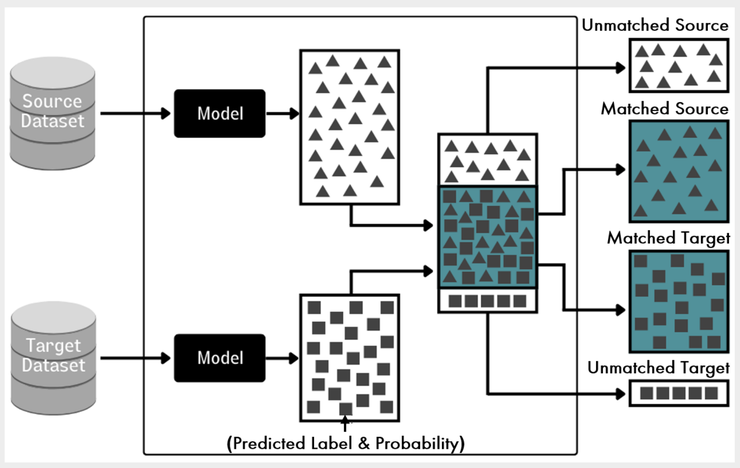
Our Approach

- We integrate model uncertainty into the evaluation of DNN effectiveness
- We assess not just what DNNs predict, but also how confidently they do so

2. Proposed Evaluation Framework (Cont.)

Methodology

- Split the test datasets into subsets with matching predictive uncertainty profiles and assess them as illustrated below



1. Get model predictions

2. Match predictions & generate test subsets

3. Assess the test subsets

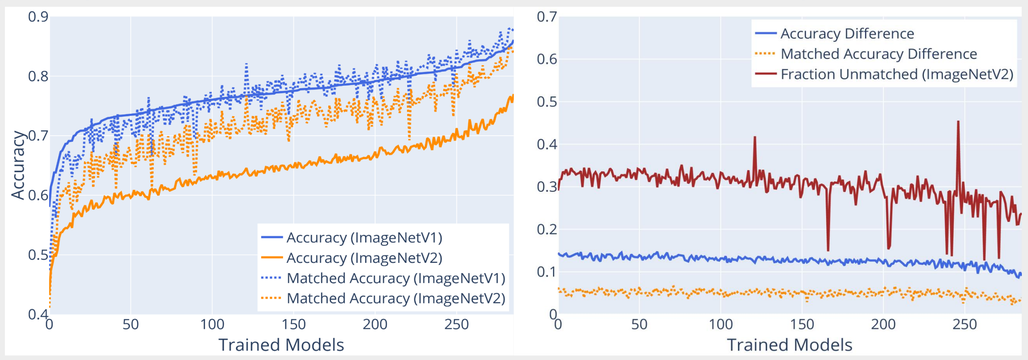
Model behavior is similar on source and target datasets

IF

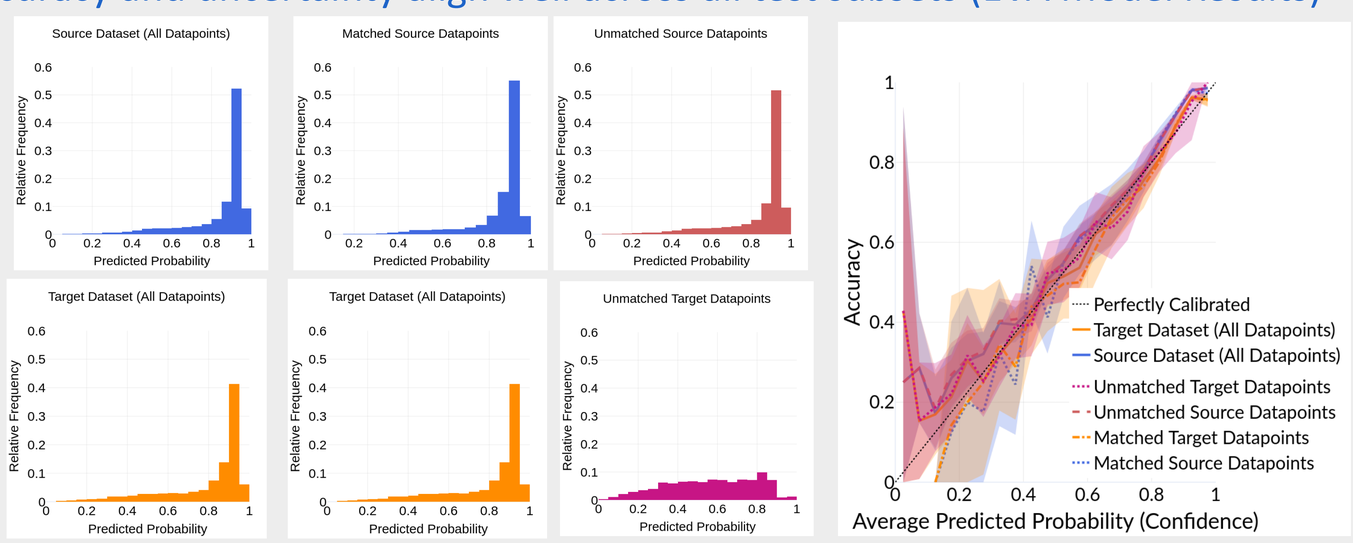
- ❑ Accuracy gap on matched subsets is substantially smaller
- ❑ All subsets have similar accuracy vs. uncertainty relationship

Findings

1. Accuracy drops are not as substantial if the uncertainty of DNNs is leveraged



2. Accuracy and uncertainty align well across all test subsets (EVA Model Results)



Implications

- Better conclusions about DNN effectiveness result in a better understanding of DNN behavior

For more details, check our paper: "Re- assessing Accuracy Degradation: Understanding DNN behavior on Similar-but-non-identical Test Datasets," Machine Learning Journal, 2025.

@ EslaTimothy.Anzaku@ghent.ac

research

GHENT UNIVERSITY

GHENT UNIVERSITY GLOBAL CAMPUS

LIÈGE université Mathematics

MASON KOREA George Mason University

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN