Understanding the Behavior of DNNs on Replicated Datasets

Esla T. Anzaku[1], Haohan Wang[2], Ajiboye Babalola[3], Seyed A. Mousavi[1], Wesley De Neve[1], Arnout Van Messem[4]

[1]Ghent University Global Campus, South Korea

[2]University of Illinois, USA

[3]George Mason Korea, South Korea

[4]Université de Liège, Belgium

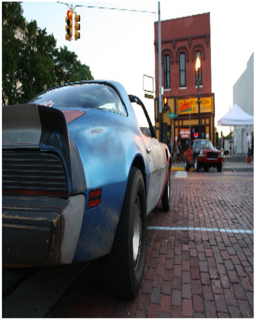
DSSV - July 9, 2025

Introduction

- **DNNs**: remarkable performance during model creation
- **Image recognition**: CIFAR-10 and ImageNet
- **Generalization** is crucial
- **Emerging trend**: use replicated test dataset
  - Created by closely following methodology and procedures of original dataset
- **Challenges**:
  1. Unexpected accuracy drop on similar test datasets
     - Not entirely explained by generalization shortcomings or dataset disparities
     - Introduce new evaluation framework leveraging uncertainty estimates generated by models under study
  2. Inherent single-label assumption in image recognition
     - Can this help explain the accuracy drop?
     - Propose new evaluation metric taking the multi-label nature of images into account

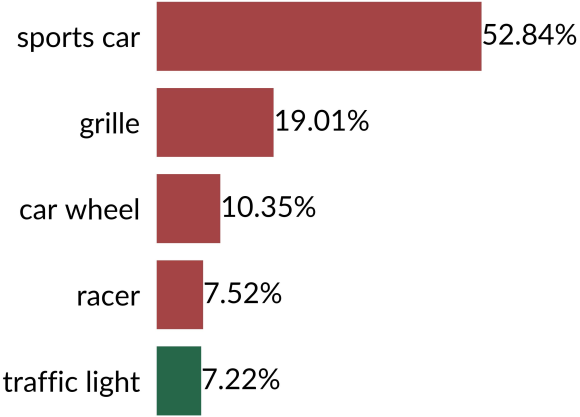
Image recognition

Input Image



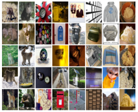
Ground Truth: traffic light

Predictions



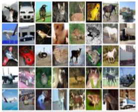
Replicated datasets

ImageNet-1k Val. Set[1]



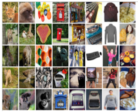
50,000 images, 1,000 classes
Published in 2009 [4]

CIFAR-10 Test Set



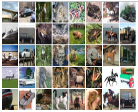
10,000 images, 10 classes
Published in 2009 [5]

ImageNetV2



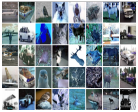
10,000 images
Published in 2019 [7]

CIFAR 10.1



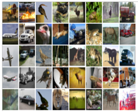
2,000 images
Published in 2019 [7]

CIFAR 10.2



10,000 images
Published in 2020 [6]

CINIC

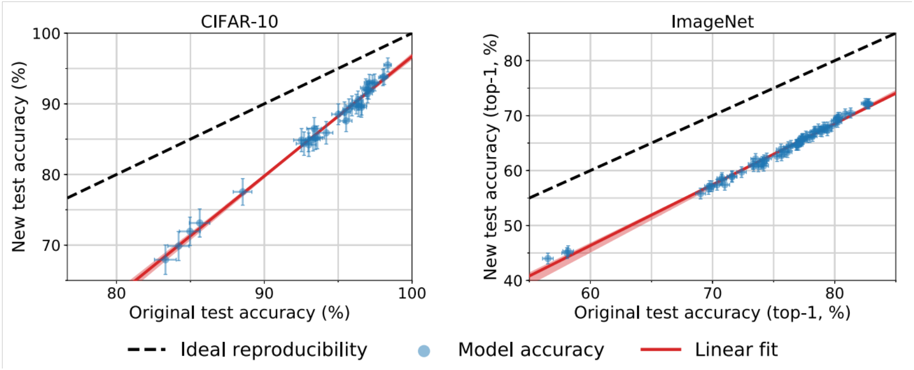


90,000 images
Published in 2018 [3]

[1] ImageNetV1

Accuracy Degradation [2]

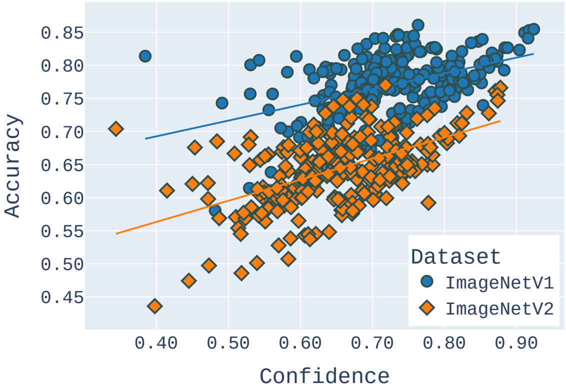
Accuracy degradation



Accuracy drop [7]

Unexplained and unexpected top-1 accuracy drop of 3-15% for CIFAR and 11-15% for ImageNet on replicated test datasets.

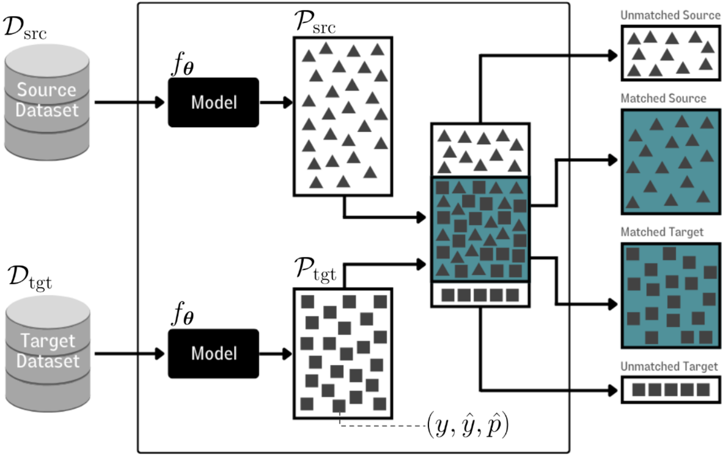
Accuracy vs. uncertainty relationship



Observation

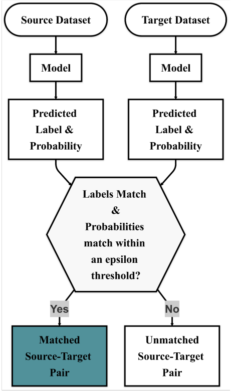
Models tend to be less confident and less accurate on ImageNetV2.

Proposed framework



Proposed framework

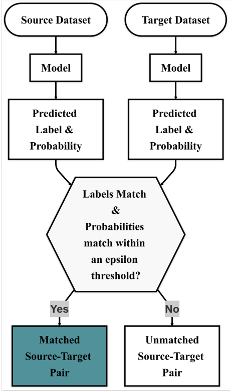
Idea: leverage DNN uncertainty in model assessment



1. Obtain model predictions
2. Match predictions and make subsets
3. Assess test subsets
   Model behavior is similar on source and target dataset if
   - Accuracy gap on matched subsets is substantially smaller
   - All subsets have similar accuracy versus uncertainty relationship

Proposed framework

Idea: leverage DNN uncertainty in model assessment



Conventional accuracy assessment

- Uses all datapoints
- Treats all predictions equally
- Ignores model uncertainty
- Assumes dataset characteristics are same

Proposed evaluation framework

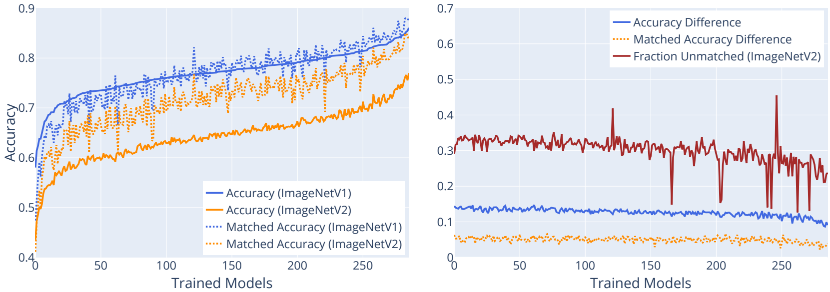
- Matches similar predictions
- Creates fair comparison subsets
- Leverages model uncertainty
- Accounts for differences in dataset characteristics

Experimental setup

- ImageNetV1 vs. ImageNetV2[2]
- 286 pre-trained ImageNet models
  - Architectures: ResNet, EficientNet, MobileNet, ConvNeXt v2, ViTs, ...

[2]Similar experiments and results are available for CIFAR-10.

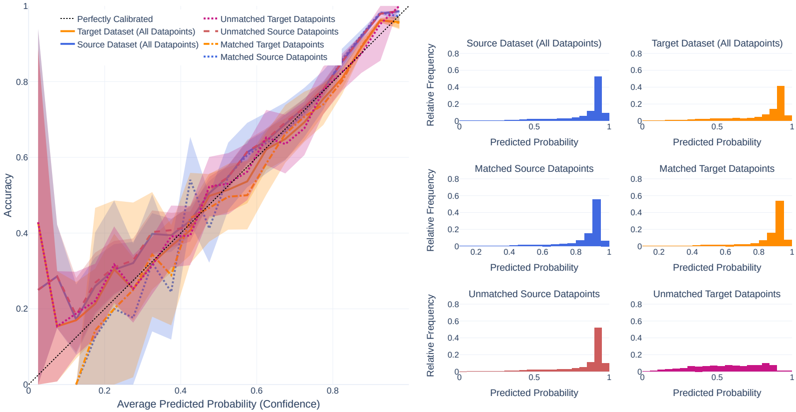
Results



Observation

Leveraging uncertainty leads to significantly lower accuracy gap

Results



Observations

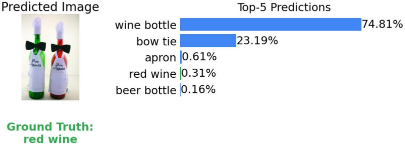
- Different test subsets with different accuracies and uncertainty distributions
- Yet similar accuracy-uncertainty relationship

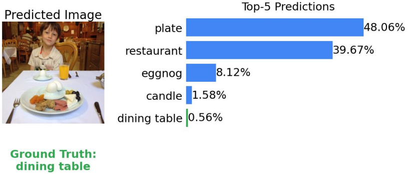
Conclusions

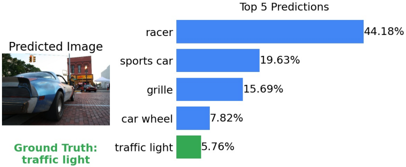
- Top-1 accuracy gaps are substantially lower than earlier reported.
- Accuracy-uncertainty profiles are consistent across matched and unmatched subsets.
- DNNs demonstrate better robustness on replicated datasets than earlier reported.
- Test and replicated datasets differ in subtle ways that need further investigation.

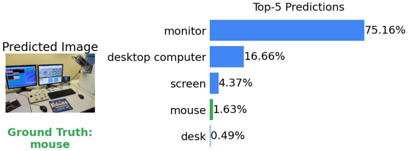
Single-label Assumption [1]

Single-label assumption









Single-label assumption vs. multi-label nature

Since standard evaluation metrics are constrained to a single ground-thruth label, conventional top-1 metrics will often underestimate model performance.

Alternative evaluation methods

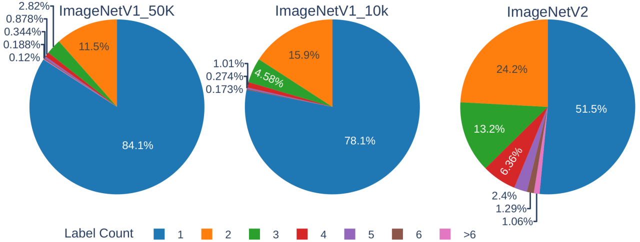
- **Top-5 accuracy:** verifies whether at least one of the 5 highest-ranked predictions matches the ground-truth label but does not evaluate whether all relevant categories are identified.
- **ReaL accuracy:** expands the ground-truth label set but considers only the top-ranked prediction.

Proposed selection mechanism

- C: number of classes
- Dataset: $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$
- Corresponding softmax output: $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \dots, \hat{\mathbf{y}}_N\}$ with $\hat{\mathbf{y}}_i \in \mathbb{R}^C$
- k_i: number of ground-truth classes for i^{th} image[3]

Variable top-k selection mechanism

For each datapoint $\mathbf{x}_i$, the top-k_i predictions are obtained by selecting the indices corresponding to the highest k_i values in $\hat{\mathbf{y}}_i$.



[3]There have been attempts to assign multiple labels to ImageNet using ReaL.

Proposed evaluation metric

- Define G subgroups based on the number of ground-truth labels g, containing N_g datapoints
- Datapoints: $\mathbf{x}_{g,i}$ with ground-truth labels $\mathbf{y}^{\text{gt}}_{g,i} \in \{0,1\}^C$
- Predictions $\hat{\mathbf{y}}_{g,i} \in \{0,1\}^C$
- Subgroup accuracy

$$A_g = \frac{1}{N_g} \sum_{i=1}^{N_g} \frac{1}{C} \sum_{c=1}^{C} \mathbb{I}(y^{\text{gt}}_{g,i,c} = \hat{y}_{g,i,c})$$

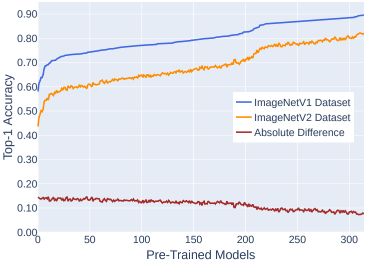
Average Subgroup Multi-label Accuracy (ASMA)

$$\text{ASMA} = \frac{1}{G} \sum_{g=1}^{G} A_g$$

Experimental setup

- ImageNetV1 vs. ImageNetV2
- 350 pre-trained ImageNet models
  - 100 top performing models (based on top-1 accuracy)
  - 250 randomly selected models covering a wide range of architectures: ResNet, EfficientNet, MobileNet, ConvNeXt, ViTs, ...
- Three evaluation metrics:
  - Top-1 accuracy: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i^{\mathrm{gt}})$
  - ReaL accuracy: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i \in \mathbf{y}_i^{\mathrm{plaus}})$ with $\mathbf{y}_i^{\mathrm{plaus}}$ set of plausible labels
  - ASMA

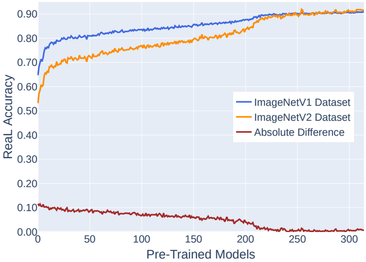
Results – Top-1 accuracy



Observations

- Performance on ImageNetV2 consistently lower
- Accuracy gap: 6-14%

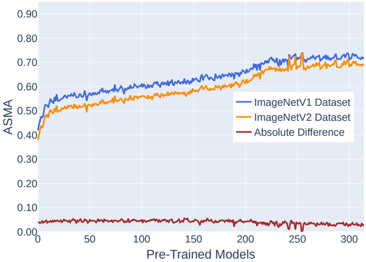
Results – ReaL accuracy



Observations

- Difference between ImageNetV1 and ImageNetV2 lowers noticeably
- For 78 models: gap $< 1\%$
- Accuracy gap: 0-11%

Results – ASMA



Observations

- Difference decreases further
- For 4 models: gap $< 1\%$
- Accuracy gap: 0-6%

Conclusions

- Top-1 accuracy overestimates DNN performance gaps
- This overestimation is (partially) due to ignoring the multi-label nature of images
- Top-1 accuracy masks DNNs with desirable multi-label class prediction properties

To conclude

References

[1] Esla Timothy Anzaku, Seyed Amir Mousavi, Arnout Van Messem, and Wesley De Neve. The Impact of the Single-Label Assumption in Image Recognition Benchmarking. *arXiv*, arXiv:2412.18409, 2025.

[2] Esla Timothy. Anzaku, Haohan Wang, Ajiboye Babalola, Arnout Van Messem, and Wesley De Neve. Re-assessing accuracy degradation: a framework for understanding DNN behavior on similar-but-non-identical test datasets. *Machine Learning*, 114(84), 2025.

[3] Luke N. Darlow, Elliot J. Crowley, Antreas Antoniou, and Amos J. Storkey. CINIC-10 is not ImageNet or CIFAR-10. *arXiv*, arXiv:1810.03505, 2018.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A Large-scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[5] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. CIFAR-10 (Canadian Institute for Advanced Research). 2009.

[6] Shangyun Lu, Bradley Nott, Aaron Olson, Alberto Todeschini, Puya Vahabi, Carmon Yair, and Ludwig Schmidt. Harder or Different? A Closer Look at Distribution Shift in Dataset Reproduction. In *Uncertainty and Robustness in Deep Learning Workshop (UDL), ICML*, 2020.

[7] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5389–5400, 2019.