Assessing the Historical Accuracy of ImageNet

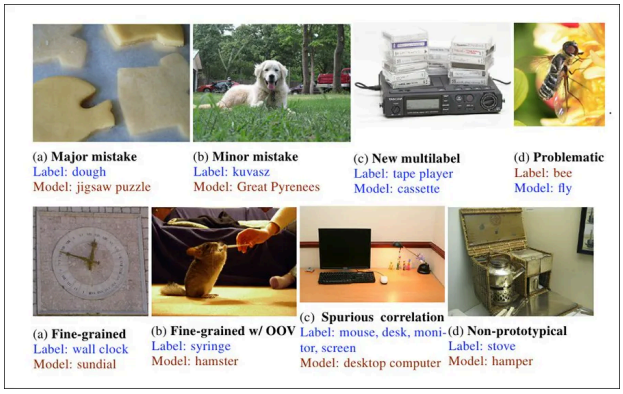
Originally published at https://www.unite.ai/assessing-the-historical-accuracy-of-imagenet/

A new study from Google Research and UC Berkeley adds to longstanding criticism regarding the computer vision (CV) research sector's reliance on the venerable ImageNet dataset, and its many derivatives. After a great deal of labor-intensive manual evaluation, the authors conclude that almost 50% of the supposed mistakes that the best models make on the multi-label subset evaluation of ImageNet (where current top-performing models achieve more than 97% top-1 accuracy) are not actually in error.

From the paper:

> *'Our analysis reveals that nearly half of the supposed mistakes are not mistakes at all, and we uncover new valid multi-labels, demonstrating that, without careful review, we are significantly underestimating the performance of these models.*
>
> *'On the other hand, we also find that today's best models still make a significant number of mistakes (40%) that are obviously wrong to human reviewers.'*

The extent to which the mislabeling of datasets – particularly by unskilled crowdsource workers – may be skewing the sector, was revealed by the study's painstaking approach to evaluation of the image/text pairings across a large swathe of the history of ImageNet.

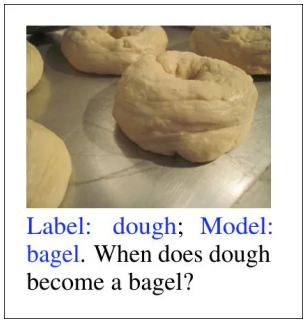


In the top row, examples of Mistake Severity: in the first two examples here, the new model simply gets the predicted label wrong; in the third example, the new model identifies a previously-missing multi-label (a label that addresses a novel categorization of the image); in the final image in the top row, the model's prediction is ambiguous, because the picture is a bee-fly and not a fly. However, the average bee belongs to the Diptera insect order, and so this exception would be almost impossible to spot, even for an expert annotator. In the row below are four mistake categories, with examples. Source: https://arxiv.org/pdf/2205.04596.pdf

The researchers employed a small number of dedicated evaluators to painstakingly review historical error records in ImageNet dataset evaluation, finding that a great many of the error judgements are themselves in error – a discovery that potentially revises some of the poor scoring that many projects have obtained on ImageNet benchmarks over the years.

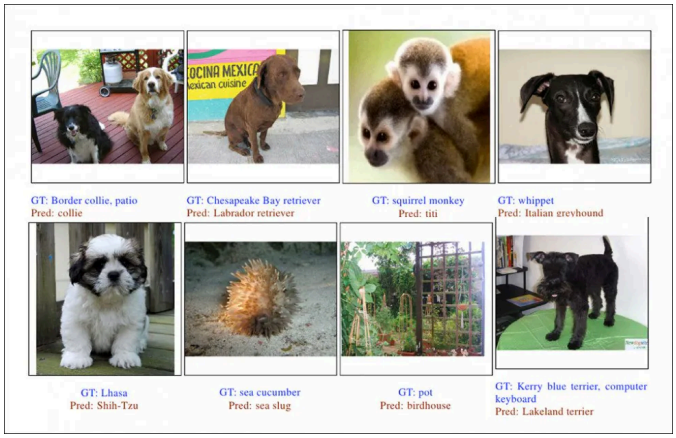
As ImageNet entrenches in the CV culture, the researchers contend that improvements in accuracy are thought to yield diminishing returns, and that new models which overstep established label accuracy, and which suggest new (i.e. additional) labels may be being punished, essentially, for non-conformity.

'For example,' the authors observe. *'should we penalize models for being the first to predict that a pre-baked bagel may be a bagel, as one of the model's we review in this work does?'*



From the paper, a newer model defies prior prediction that the object in the photo is dough, and suggests that the object is actually already a bagel).

From the point of view of a crowdsourced worker tasked with identifying such an object, this is a semantic and even philosophical quandary that can only be resolved by multi-labeling (as often occurs in later subsets and subsequent iterations of ImageNet); in the above case, the object is indeed both dough and at least a nascent bagel.



Major (above) and minor (below) mistakes that emerged when testing custom models in the research. Original ImageNet labels are the first images on the left.

The two obvious solutions are to assign more resources to labeling (which is a challenge, within the budgetary constraints of most computer vision research projects); and, as the authors emphasize, to regularly update datasets and label evaluation sub-sets (which, among other obstacles, risks to break 'like for like' historical continuity of benchmarks, and to litter new research papers with qualifications and disclaimers regarding equivalence).

As a step to remedying the situation, the researchers have developed a new sub-dataset of ImageNet called *ImageNet-Major* (ImageNet-M), which they describe as *'a 68-example "major error" slice of the obvious mistakes made by today's top models—a slice where models should achieve near perfection, but today are far from doing so.'*

The paper is titled *When does dough become a bagel? Analyzing the remaining mistakes on ImageNet*, and is written by four authors from Google Research, together with Sara Fridovich-Keil of UC Berkeley.

Technical Debt

The findings are important because the remaining errors identified (or misidentified) in ImageNet, in the 16 years since its inception, the central study of the research, can represent the difference between a deployable model and one that's error-prone enough that it can't be let loose on live data. As ever, the last mile is critical.

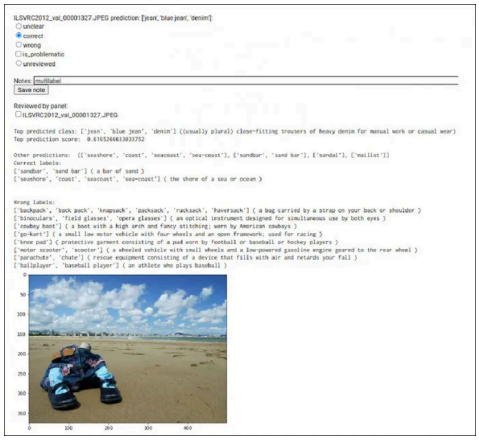
The computer vision and image synthesis research sector has effectively 'auto-selected' ImageNet as a benchmark metric, for a number of reasons — not least because a rash of early adopters, at a time where high-volume and well-labeled datasets were rarer than they are now, produced so many research initiatives that testing against ImageNet quickly became the only broadly applicable historical 'standard' for benchmarking new frameworks.

Method

Seeking out the 'remaining mistakes' in ImageNet, the researchers used a standard ViT model (capable of achieving an accuracy of 89.5%) with 3 billion parameters, *Vit-3B*, pretrained on JFT-3B and fine-tuned on ImageNet-1K.

Using the ImageNet2012_multilabel dataset, the researchers recorded the initial multi-label accuracy (MLA) of ViT-3B as 96.3%, during which the model made 676 apparent mistakes. It was these mistakes (and also mistakes produced by a Greedy Soups model) that the authors sought to investigate.

To evaluate the remaining 676 mistakes, the authors avoided crowdworkers, observing that mistakes of this type can be difficult for average annotators to spot, but assembled a panel of five expert reviewers, and created a dedicated tool to allow each reviewer to see at a glance the predicted class; the predicted score; the ground truth labels; and the image itself.



The UI built for the project.

In some cases, further research was necessary to resolve disputes among the panel, and Google Image search was used as an adjunct tool.

> *'[In] one interesting but not isolated case, a prediction of a taxi cab (with no obvious taxi cab indicators beyond yellow color) was present in the image; we determined the prediction to be correctly a taxi cab and not just a standard vehicle by identifying a landmark bridge in the background in order to localize the city, and a subsequent image search for taxis in that city yielded the images of the same taxi model and license plate design, validating the model's actually correct prediction.'*

After initial review of the mistakes found over several phases of the research, the authors formulated four novel mistake types: *fine-grained error*, where the predicted class is similar to a ground-truth label; *fine-grained with out-of-vocabulary (OOV)*, where the model identifies an object whose class is correct but not present in ImageNet; *spurious correlation*, where the predicted label is read out of context of the image; and *non-prototypical*, where the ground truth object is a specious example of the class that bears resemblance to the predicted label.

In certain cases, the ground truth was not itself 'true':

> *'After review of the original 676 mistakes [found in ImageNet], we found that 298 were either correct or unclear, or determined the original groundtruth incorrect or problematic.'*

After an exhaustive and complex round of experiments across a range of datasets, subsets and validation sets, the authors found that the two models under study were actually deemed correct (by the human reviewers) for half of the 'mistakes' they made under conventional techniques.

The paper concludes:

> *'In this paper, we analyzed every remaining mistake that the ViT-3B and Greedy Soups models make on the ImageNet multi-label validation set.*
>
> *'Overall, we found that: 1) when a large, high-accuracy model makes a novel prediction not made by other models, it ends up being a correct new multi-label almost half of the time; 2) higher accuracy models do not demonstrate an obvious pattern in our categories and severities of mistakes they solve; 3) SOTA models today are largely matching or beating the performance of the best expert human on the human-evaluated multi-label subset; 4) noisy training data and under-specified classes may be a factor limiting the effective measurement of improvements in image classification.'*

First published 15th May 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More