Is IMAGENET Solved?
Evaluating Machine Accuracy

Becca Roelofs

December 10, 2021

Thank you to my collaborators

Ludwig Schmidt

Vaishaal Shankar

Horia Mania

Alex Fang

Ben Recht

Do ImageNet Classifiers Generalize to ImageNet?
Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, Vaishaal Shankar. ICML 2019

Evaluating Machine Accuracy on ImageNet.
Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, Ludwig Schmidt. ICML 2020

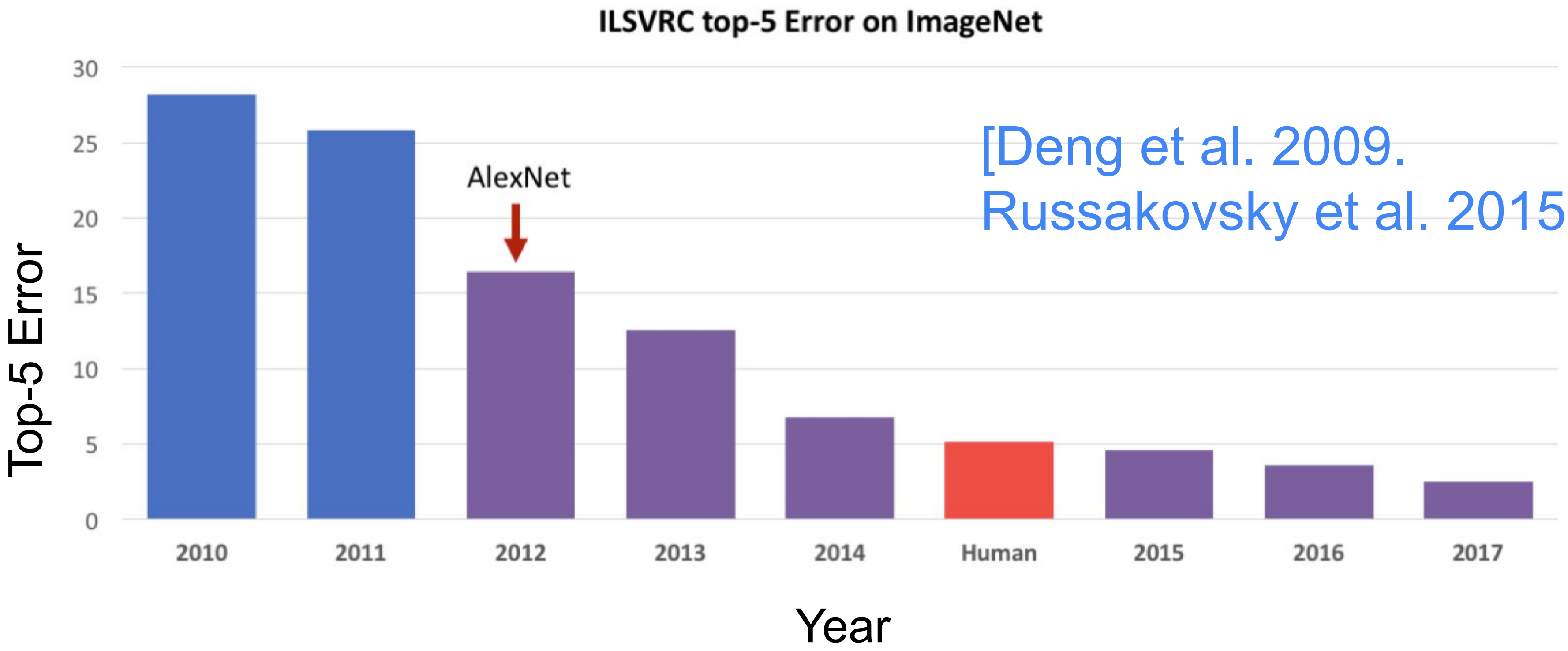
High level questions

1. How could we improve ImageNet evaluations?

2. How does model ImageNet compared to human performance?

3. How robust are ImageNet models compared to human performance?

ImageNet

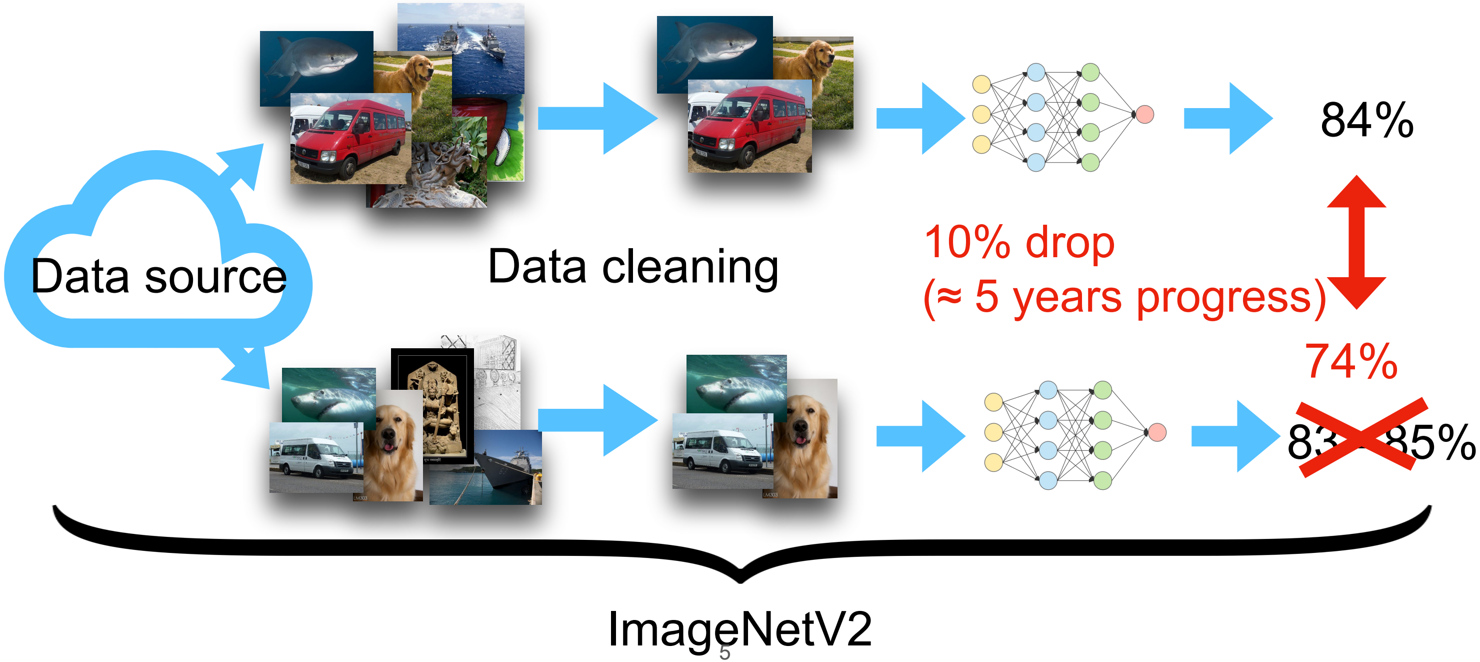


[Deng et al. 2009.
Russakovsky et al. 2015.]

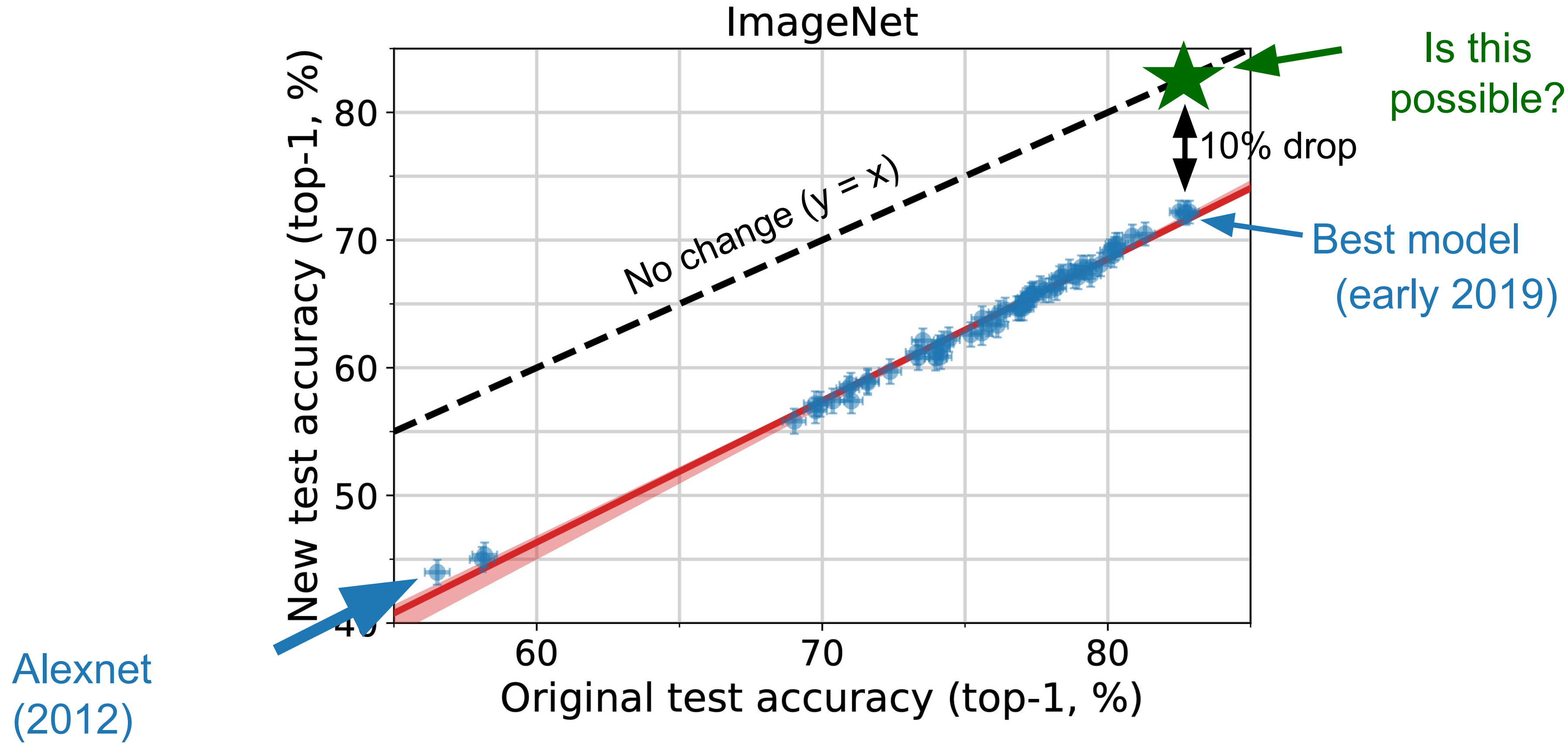
Are ImageNet performance measurements valid?

ImageNetV2

Generalization: At the very least, the models should perform just as well on new data from the same source.

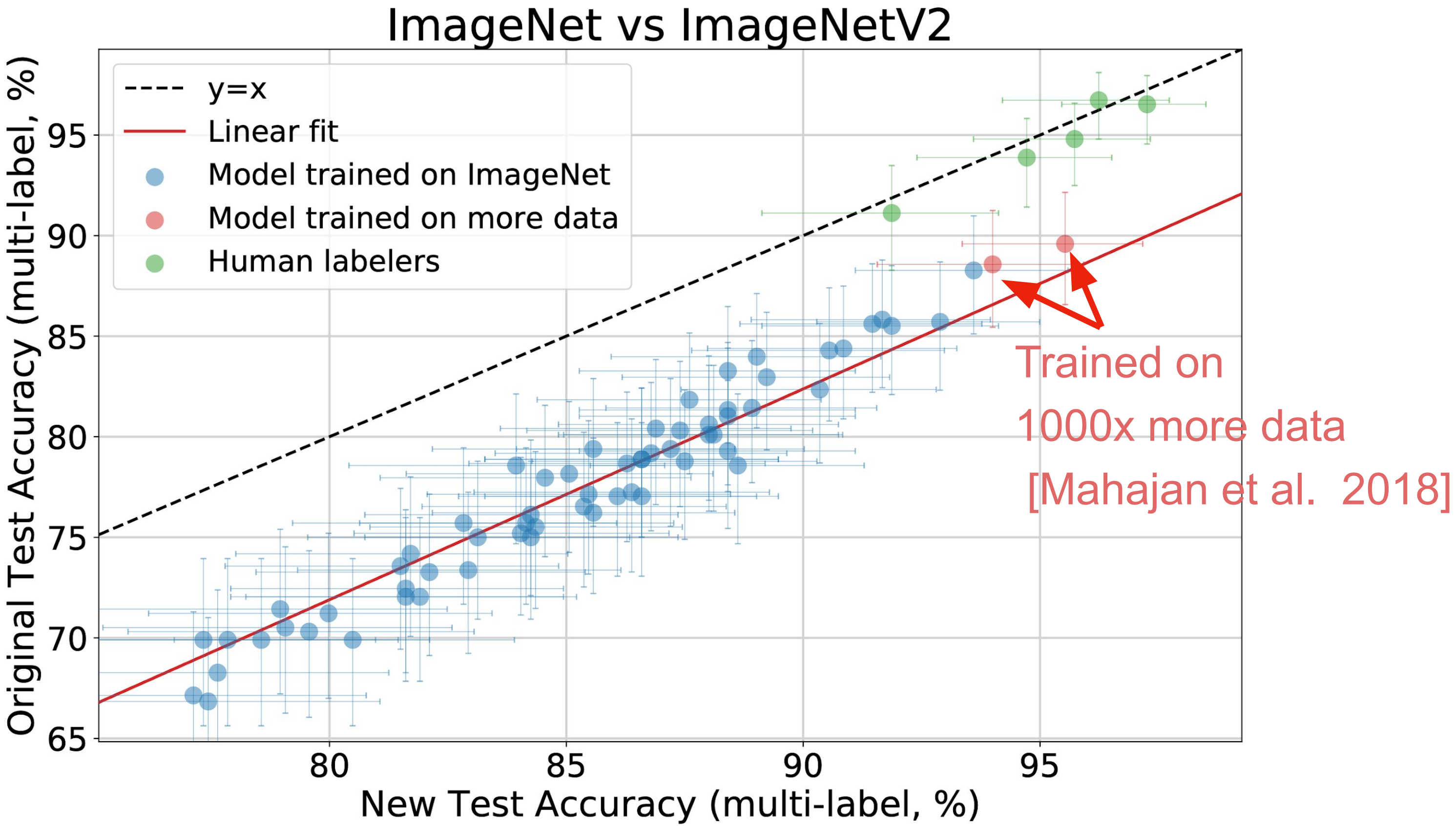


[Recht, Roelofs, Schmidt, Shankar ’19]



Is this accuracy drop from distribution shift avoidable?

[Recht, Roelofs, Schmidt, Shankar ’19]

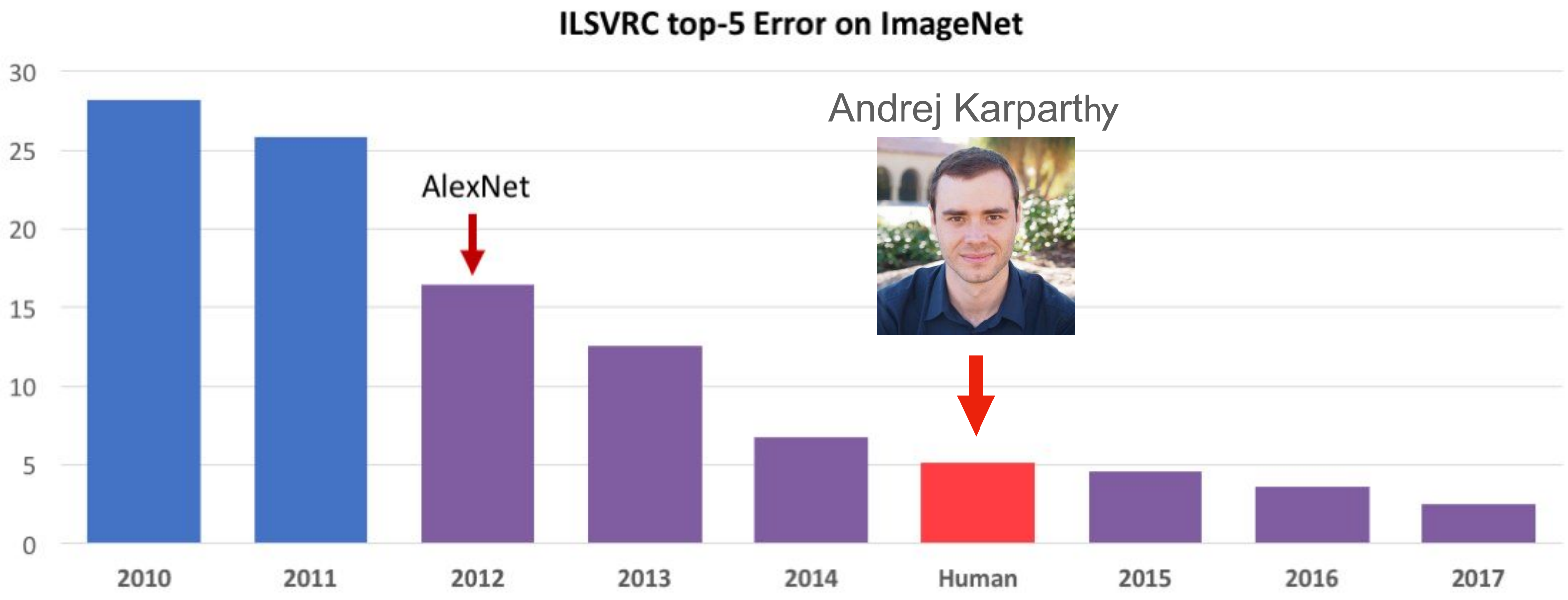


Humans are substantially more robust to distribution shift!

Evaluating human accuracy on ImageNet

- Prior work
- ImageNet images have multiple correct labels
- Current accuracy metrics
- Our proposal: multi-label accuracy

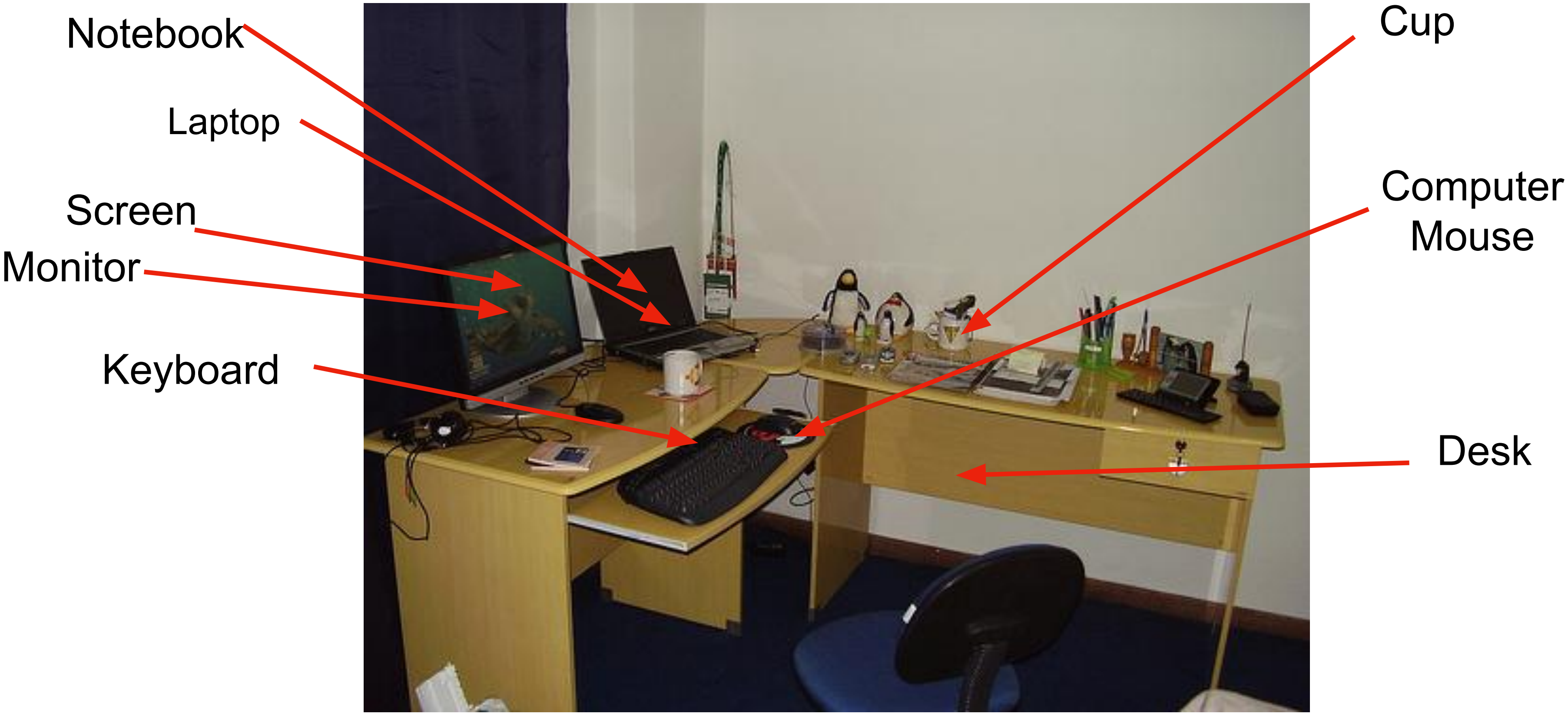
Previous human accuracy study



[Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, Berg, Li ’15]

Limitations of prior work

- Evaluated only one human subject
- Measured top-5 accuracy
- Did not evaluate robustness to distribution shift



ImageNet label: desk

Which of these labels should count as correct?

Current accuracy metrics

Top-1 Accuracy

Mushroom vs. Gyromitra

Desk, Laptop, Monitor

Tusker vs African elephant

Dock, Pier …

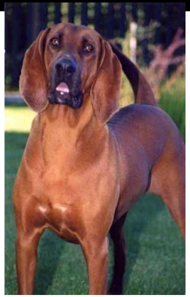
Subset Relationships

Crowded Images

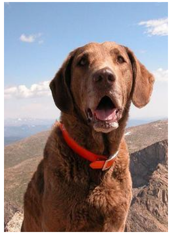
Too hard!

Top-5 Accuracy

Vizsla



Redbone



Chesapeake Bay Retriever

Rhodesian Ridgeback

Too easy!

Our proposal: multi-label accuracy

- Each classifier predicts one label per image.
- A label counts as correct if it is present in the image.

ImageNet label: Picket Fence

Multi-label annotations: Groom, Bowtie, Gown, Picket Fence

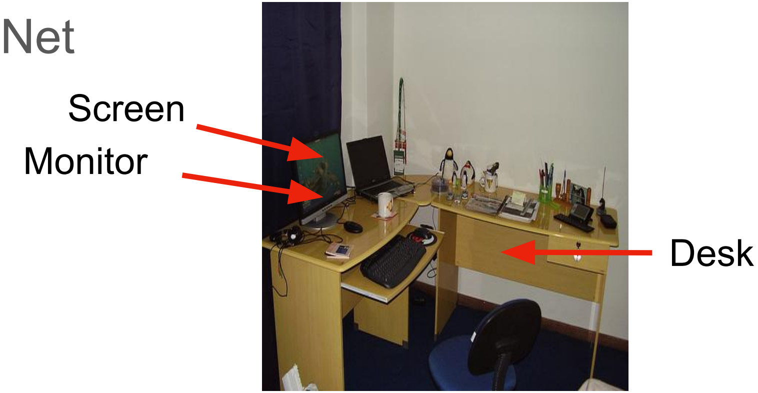
Multi-label annotations improve ImageNet evaluation

Multi-label is a **more meaningful** metric for ImageNet

Allows for comparison with **human performance**

Resolves issues caused by **ambiguous class boundaries**, including equivalent classes and subset relationships

Our multi-label accuracy evaluations also ignore images with **incorrect ground truth label**



Tusker vs African elephant

Collecting Multi-label annotations

1. Trained human experts in the ImageNet Class hierarchy

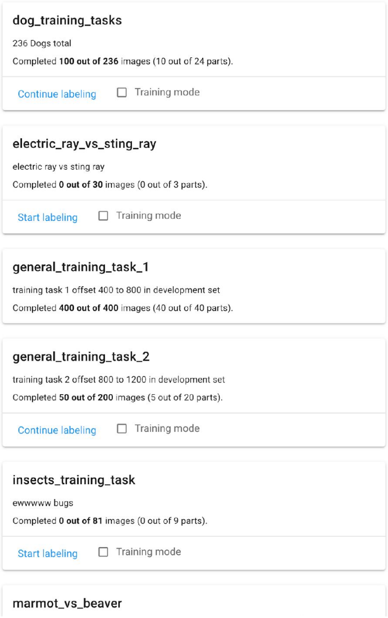
2. Built a Web UI for reviewing unique model predictions

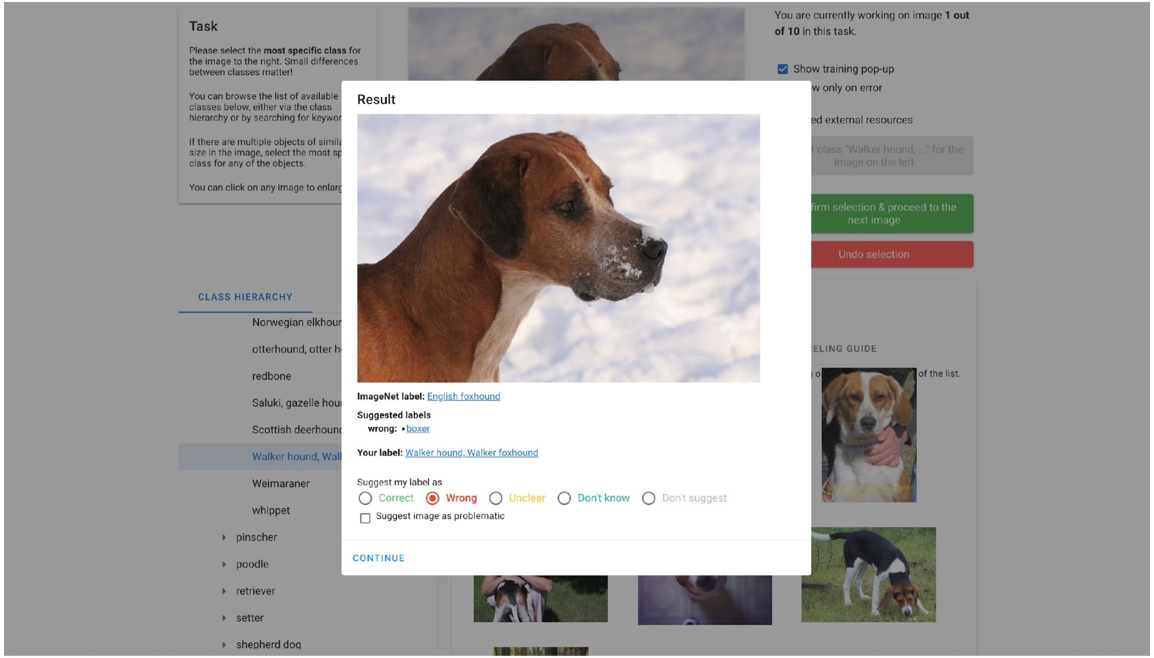
Collecting Multi-label annotations

1. **Trained human experts in the ImageNet Class hierarchy**

2. Built a Web UI for reviewing unique model predictions

Training humans experts





Humans completed training tasks designed to cover difficult class distinctions and received feedback on their predictions

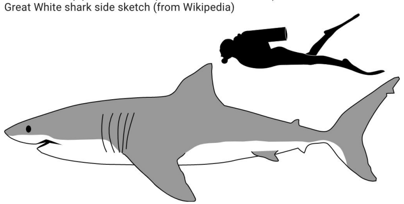
Training human experts

We created a **labeling guide**:

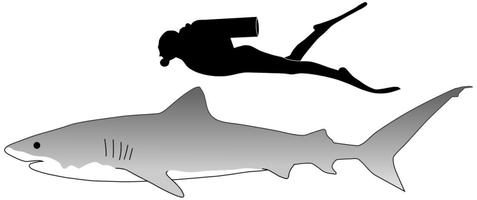
Potentially confused with **Tiger shark** and **Hammerhead shark**.
Hammerhead sharks are usually easy to identify based on their distinctive head.
The distinction with **Tiger shark** is more complicated.

Great white shark vs Tiger shark

- Points to compare
  - Stripes
    - Tiger sharks have vertical stripes
  - Thickness of the main body
    - Great White sharks are thicker
  - Head
    - Tiger sharks seem to have a more wedge-like / pointy shape
  - Fins on the underside
    - Tiger sharks have larger fins on the underside
  - Shape of tail
    - The top part of the tail (see the sketches below)
- Great White shark side sketch (from Wikipedia)



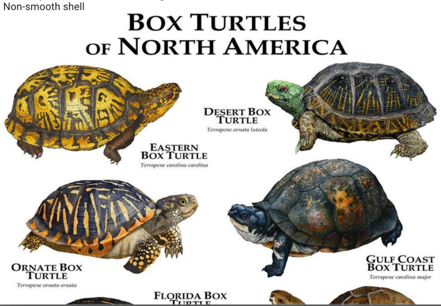
- Tiger shark side sketch (from Wikipedia). Note for instance the larger fins on the underside towards the tail end.



- More information on https://fishingbooker.com/blog/tiger-shark-vs-great-white-shark/

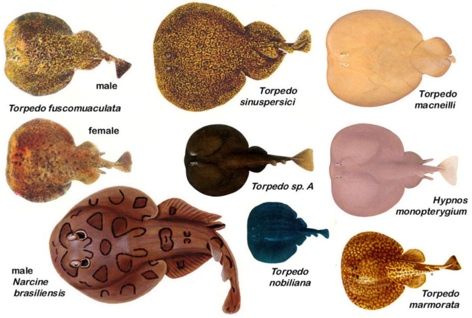
- Box turtle
  - Highly domed carapace
  - Dark colored shells with orange to yellow patterning (color varies widely)
  - Males have red eyes, while females have yellow and brown eyes
  - Hinged plastron

  - Can retract completely into the shell
  - Fully terrestrial
  - Found in forests and fields
  - Feet elephant-like, without webbing between toes
  - Non-smooth shell



Stingray vs. electric ray

- This is a hard class distinction.
- Some training images are incorrect.
- **Electric rays** tend to have a fin at the end of their tail, for instance (source biophysics.sbg.ac.at)



- The tails of **electric rays** also tend to be wider and shorter than those of a **stingray**.
- **Stingrays** look more like this (source unknown, via zazzle.com)

Sharks

Turtles

Stingrays

Collecting Multi-label annotations

1. Trained human experts in the ImageNet Class hierarchy

2. **Built a Web UI for reviewing unique model predictions**

Collecting multi-label annotations

n03461385 grocery store, grocery, food market, market

a marketplace where groceries are sold; "the grocery store included a meat market"

Correct Wrong Unclear Don't know Unreviewed

n07717556 butternut squash

buff-colored squash with a long usually straight neck and sweet orange flesh

Correct Wrong Unclear Don't know Unreviewed

n07716906 spaghetti squash

medium-sized oval squash with flesh in the form of strings that resemble spaghetti

Correct Wrong Unclear Don't know Unreviewed

n07717410 acorn squash

small dark green or yellow ribbed squash with yellow to orange flesh

Correct Wrong Unclear Don't know Unreviewed

set all unreviewed to wrong

set assigned wnid to correct

toggle image name

Problematic

Collecting multi-label annotations

toggle image name ☐ Problematic

n04152593 screen, CRT screen
the display that is electronically created on the surface of the large end of a cathode-ray tube
◉ Correct ○ Wrong ○ Unclear ○ Don't know ○ Unreviewed

n03179701 desk
a piece of furniture with a writing surface and usually drawers or other compartments
◉ Correct ○ Wrong ○ Unclear ○ Don't know ○ Unreviewed

n03180011 desktop computer
a personal computer small enough to fit conveniently in an individual workspace
◉ Correct ○ Wrong ○ Unclear ○ Don't know ○ Unreviewed

n03793489 mouse, computer mouse
a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad; "a mouse takes much more room than a trackball"
◉ Correct ○ Wrong ○ Unclear ○ Don't know ○ Unreviewed

n03782006 monitor
electronic equipment that is used to check the quality or content of electronic transmissions
◉ Correct ○ Wrong ○ Unclear ○ Don't know ○ Unreviewed

n03529860 home theater, home theatre
television and video equipment designed to reproduce in the home the experience of being in a movie theater
○ Correct ◉ Wrong ○ Unclear ○ Don't know ○ Unreviewed

set all unreviewed to wrong | set assigned wnid to correct

Multi-label statistics

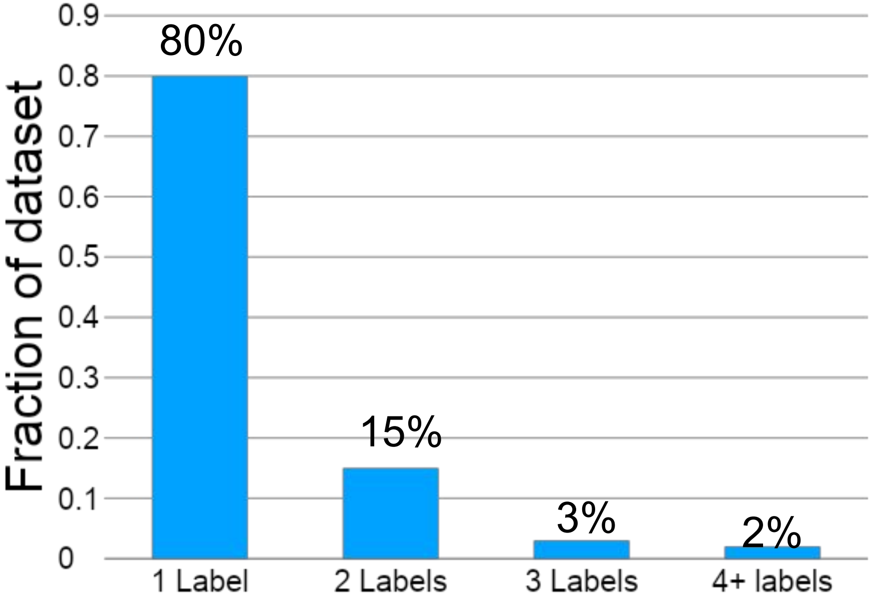
20,000 images annotated from ImageNet and 20,683 from ImageNetV2.
182,597 unique model predictions reviewed.

Multi-label statistics

20,000 images annotated from ImageNet and 20,683 from ImageNetV2.
182,597 unique model predictions reviewed.

1. **How many ImageNet images have more than one correct label?**

Fraction of ImageNet validation images with multiple correct labels



Multi-label statistics

20,000 images annotated from ImageNet and 20,683 from ImageNetV2.
182,597 unique model predictions reviewed.

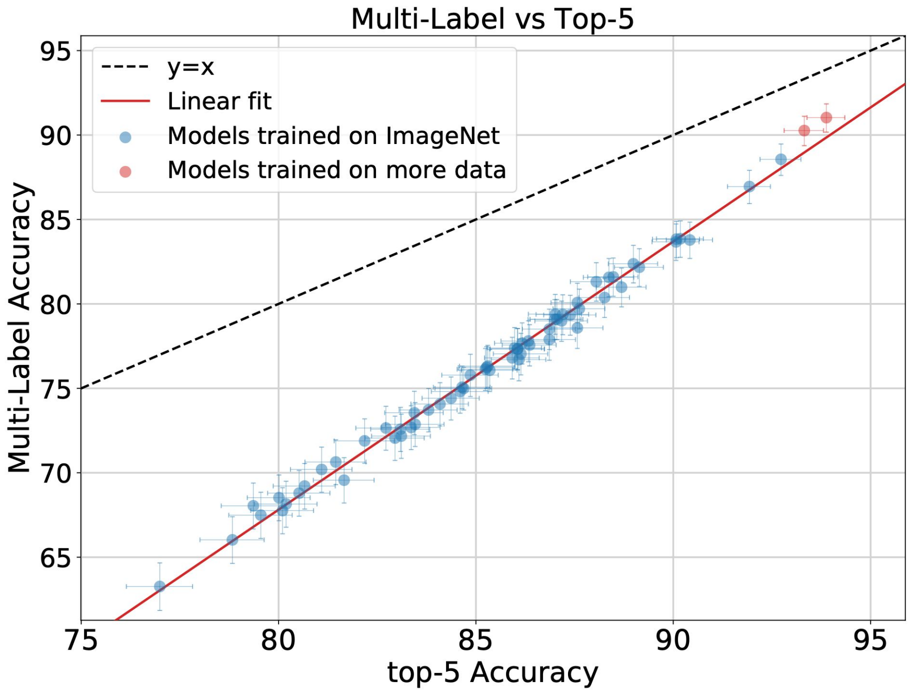
1. How many ImageNet images have more than one correct label?

2. **How do multi-label metrics compare to top-1 / top-5 accuracy?**

Multi-label versus Top-1 or Top-5 Accuracy

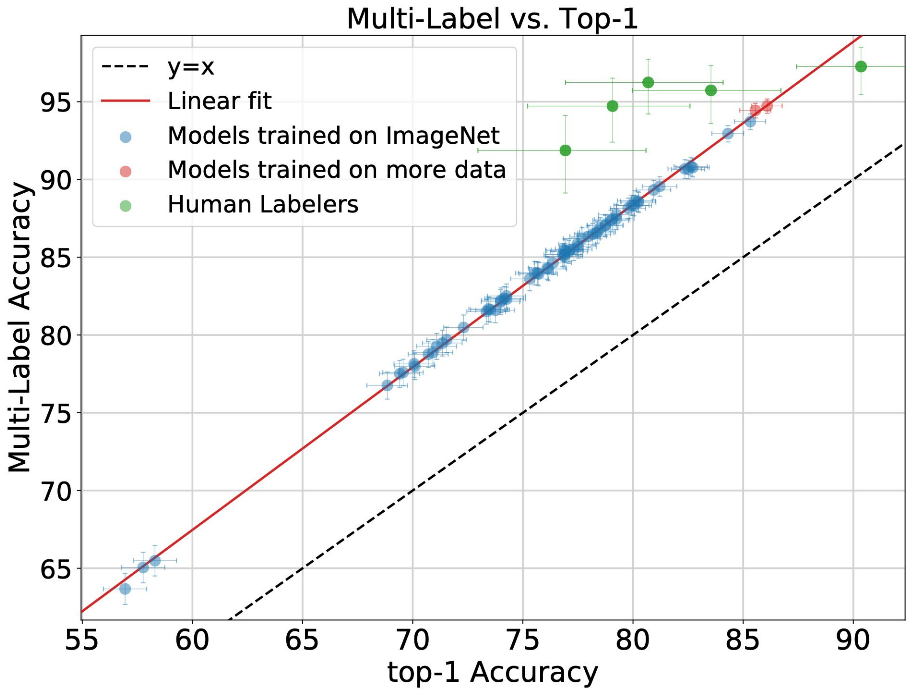
Model (in testbed)	Top1 Accuracy	Top5 Accuracy	Multi-Label Accuracy
Best	86%	97%	95%
Worst	57%	79%	64%
Median	77%	93%	85%

Multi-Label vs Top-5 Accuracy



Multi-label accuracy is harder (lower) than top-5 accuracy.
Improving top-5 accuracy improves multi-label accuracy.

Multi-label vs Top-1 accuracy



Multi-label accuracy is easier (higher) than top-1 accuracy
For models, improving top-1 accuracy improves multi-label accuracy.

Multi-label statistics

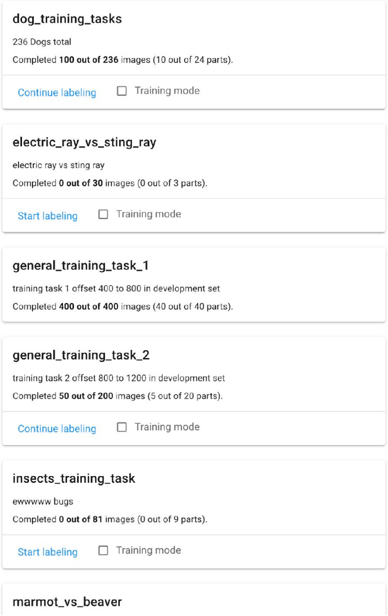
20,000 images annotated from ImageNet and 20,683 from ImageNetV2.
182,597 unique model predictions reviewed.

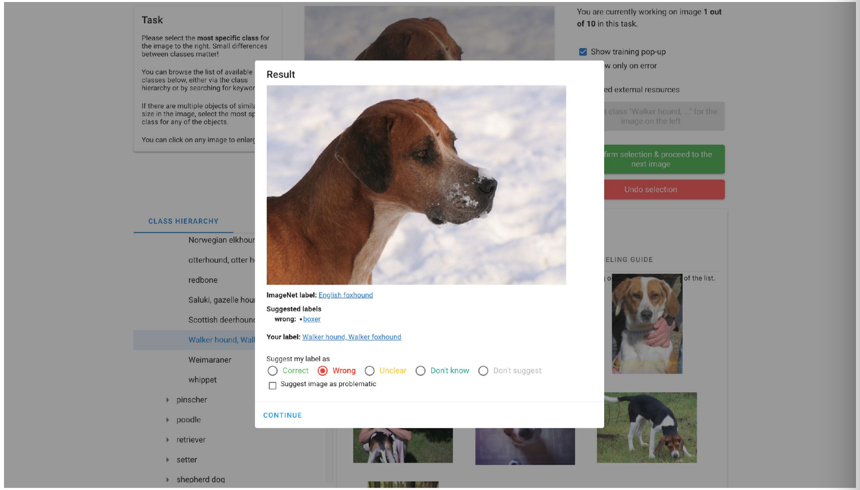
1. How many ImageNet images have more than one correct label?

2. How do multi-label metrics compare to top-1 / top-5 accuracy?

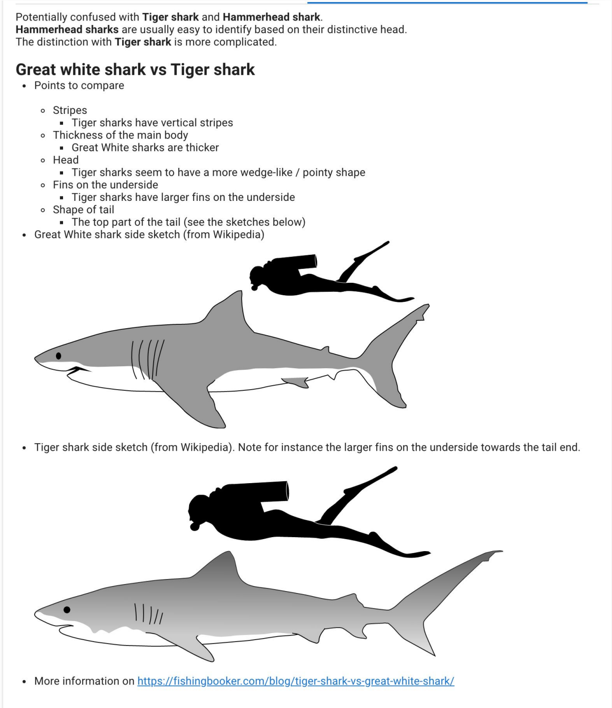
3. **How do humans perform on multi-label metrics compared to machines?**

Training humans for high performance



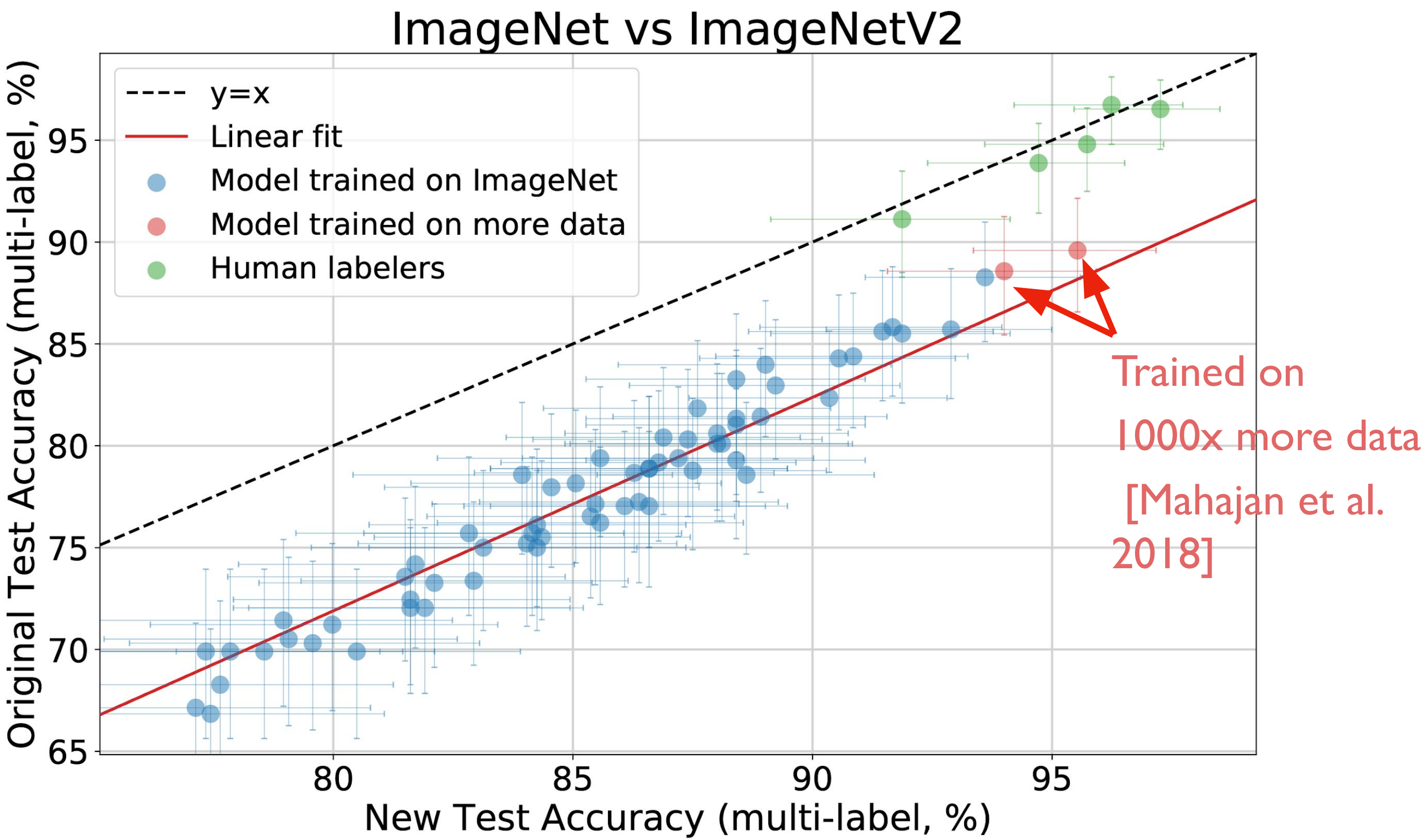


Training Tasks



Labeling Guide

All 5 humans trained for 3 months on 1000+ images



Humans are more accurate and substantially more robust than models

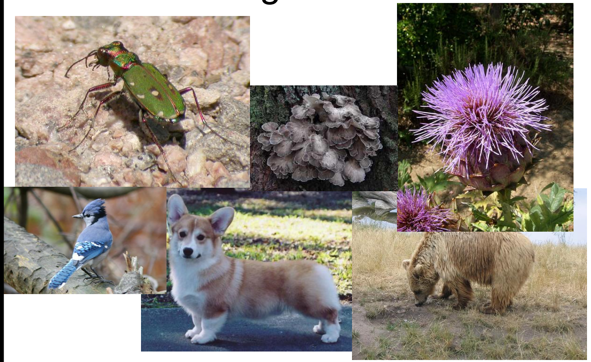
Does human robustness and performance vary across class subsets of ImageNet?

Best Model accuracy: 96%

Best Model Accuracy: 96%

Best Model Accuracy: 95%

Organisms



Objects

Best model accuracy: 90% (-6%)
Best human accuracy: 97% (+0.5%)

Accuracy difference between ImageNet and ImageNetV2

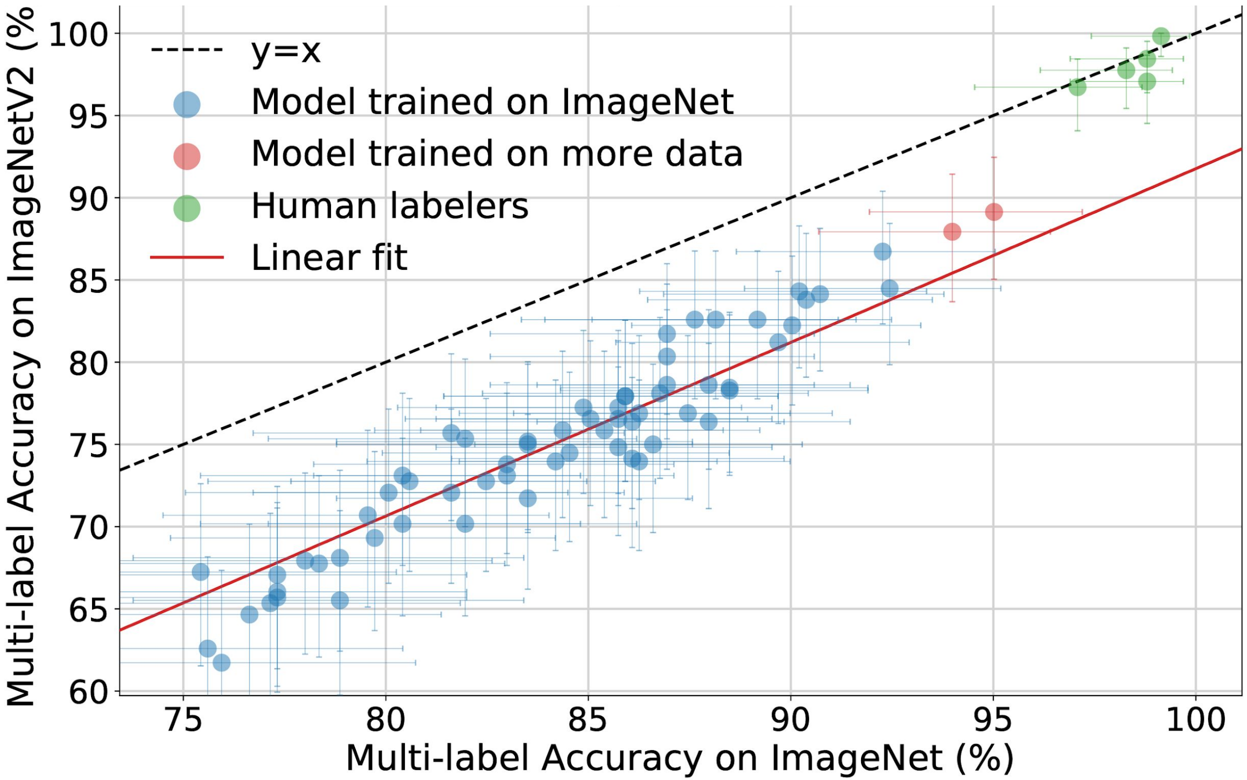
Best model accuracy: 90% (-6.3%)
Best human accuracy: 93%(+0.2%)

Organisms

Best model accuracy: 89% (-5.9%)
Best human accuracy: 99.8% (+0.7%)

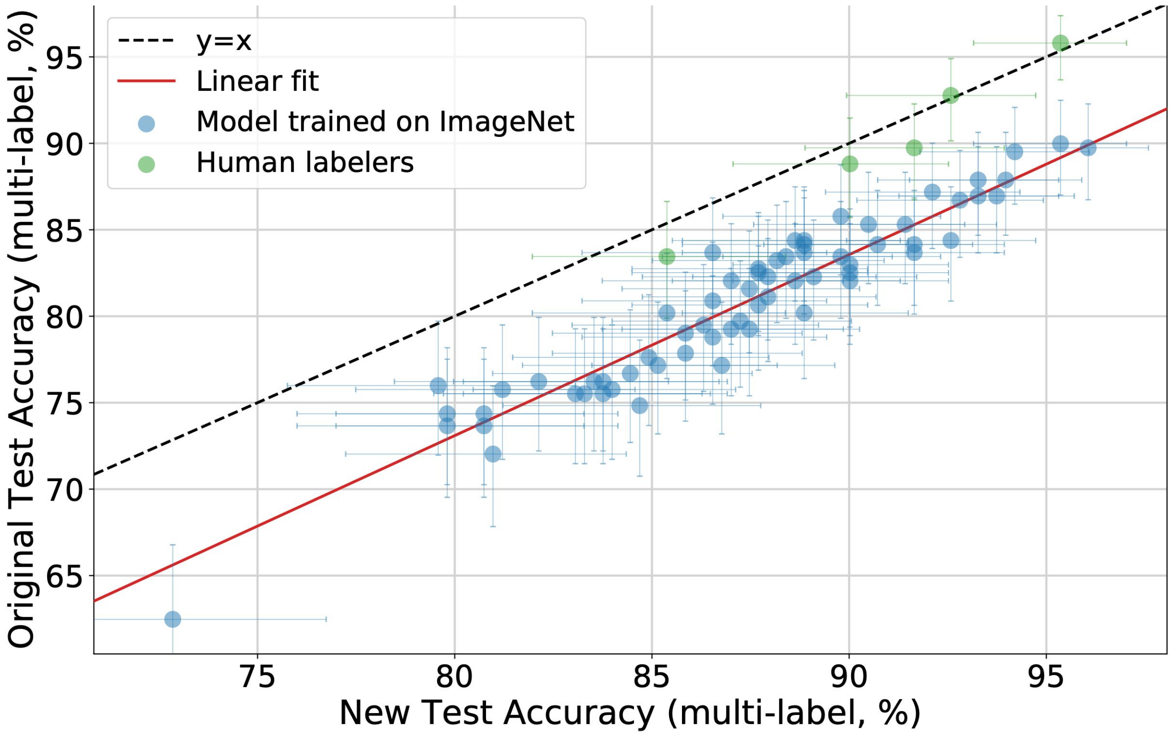
Objects

Only objects



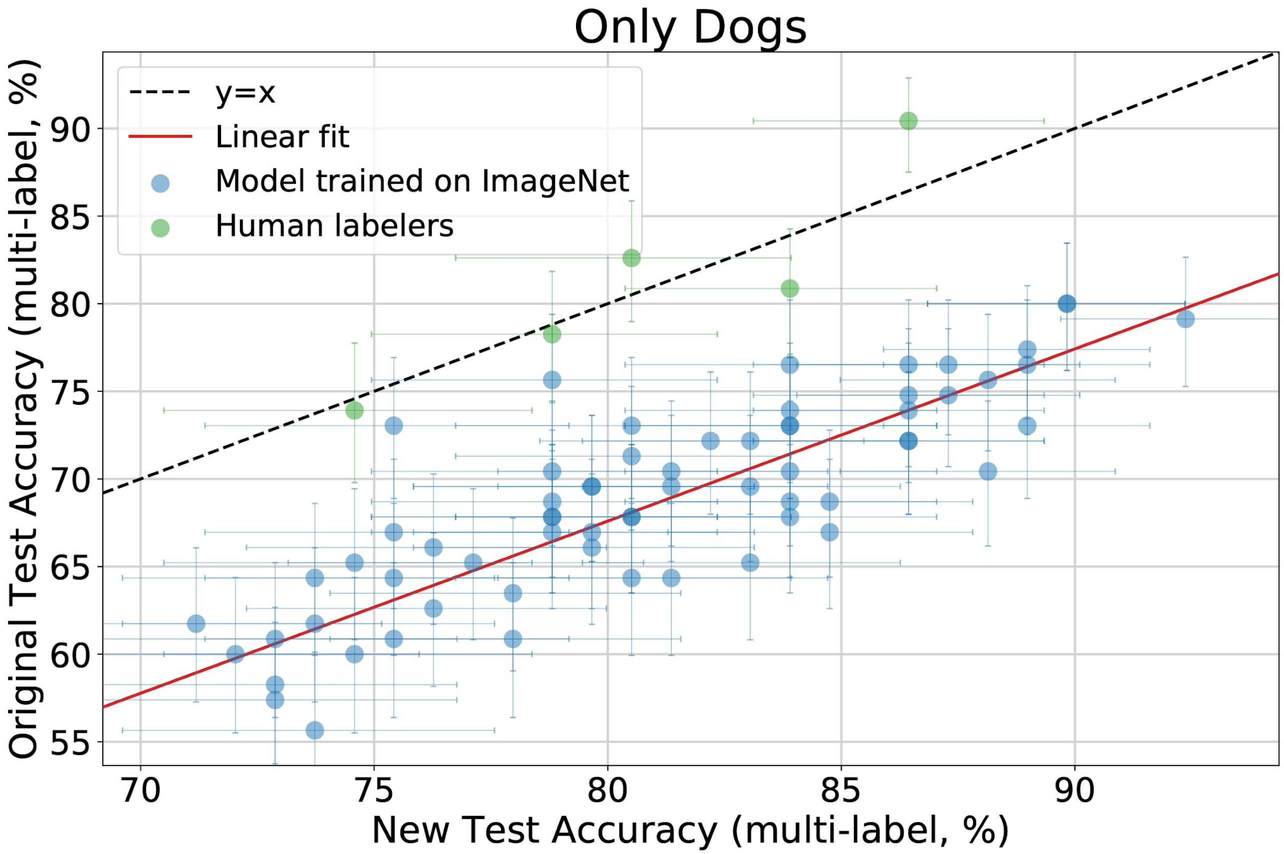
Humans are more accurate and more robust on objects

Only organisms



Humans are less accurate but more robust on organisms

Only dogs



Humans are substantially **less** accurate but **more** robust on dogs

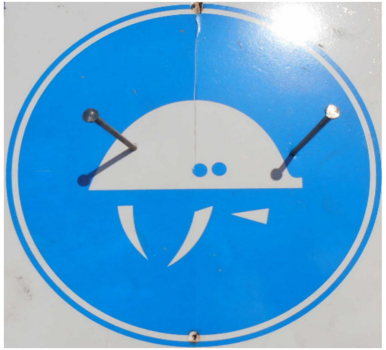
Mistake analysis

Humans: 10 images misclassified by all human labelers (1 monkey, 9 dogs)

Models: 27 images misclassified by all models (19 objects, 8 organisms)
Example of model mistakes:

Cup

Yawl



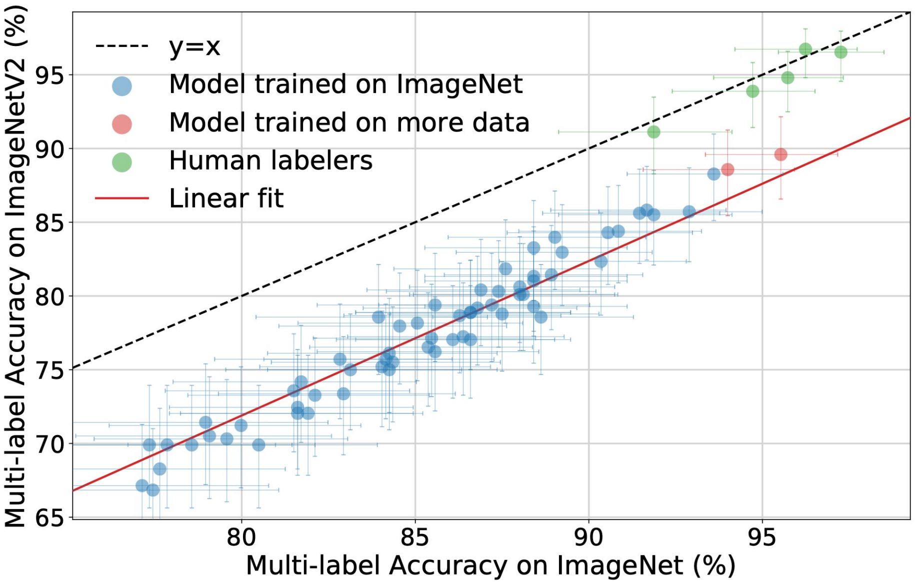
Nail

Spotlight

Majority of model mistakes are objects
Majority of human mistakes are dogs

Is ImageNet Solved?

- The best human labeler has higher accuracy than the best model on ImageNet, especially on the object subset
- Humans are **more robust** than models to ImageNet/ImageNetV2 distribution shift.

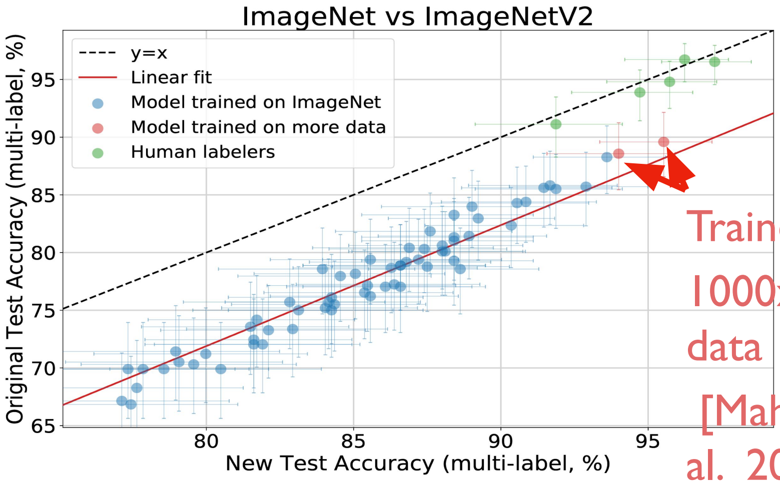


There is still room for improvement on ImageNet.

Recommendations for better ImageNet evaluations

1. Measure multi-label accuracy

2. Report performance on dogs, organisms, and inanimate objects separately.



Trained on 1000x more data

[Mahajan et al. 2018]

3. Evaluate performance to distribution shift.

https://www.tensorflow.org/datasets/catalog/imagenet2012_multilabel

https://github.com/modestyachts/evaluating_machine_accuracy_on_imagenet

Model (in testbed)	Top1 Accuracy	Top5 Accuracy	Multi-Label Accuracy
Best	86%	97%	95%
Worst	57%	79%	64%
Median	77%	93%	85%

Our proposal: multi-label accuracy

Prediction is correct if **any** of the correct labels is

ImageNet label: `Tusker`

Correct Labels:

`African Elephant, Tusker`

ImageNet label: `Picket Fence`

Labels:

`Groom, Bowtie, Gown, Picket Fence`

Multi-label annotations

ImageNet Label: `Picket Fence`
Additional Labels: `Groom, Bowtie`

ImageNet Label: `Tusker`
Additional Labels: `African Elephant`

ImageNet Inconsistencies

Mushroom

ILSVRC2012_val_00023237.JPEG

Subset Relationships

Wood Rabbit

Problematic Images

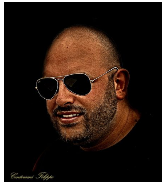
n02641379 gar, garfish, garpike, billfish, Lepisosteus osseus

Gloss: primitive predaceous North American fish covered with hard scales and having long jaws with needlelike teeth

Synsets are not synonyms

Sunglass

a convex lens that focuses the rays of the sun; used to start a fire



ILSVRC2012_val_00030556.JPEG

Redefined Classes

Magpie

ILSVRC2012_val_00035348.JPEG

Drawings or Paintings

ILSVRC2012_val_00033112.JPEG

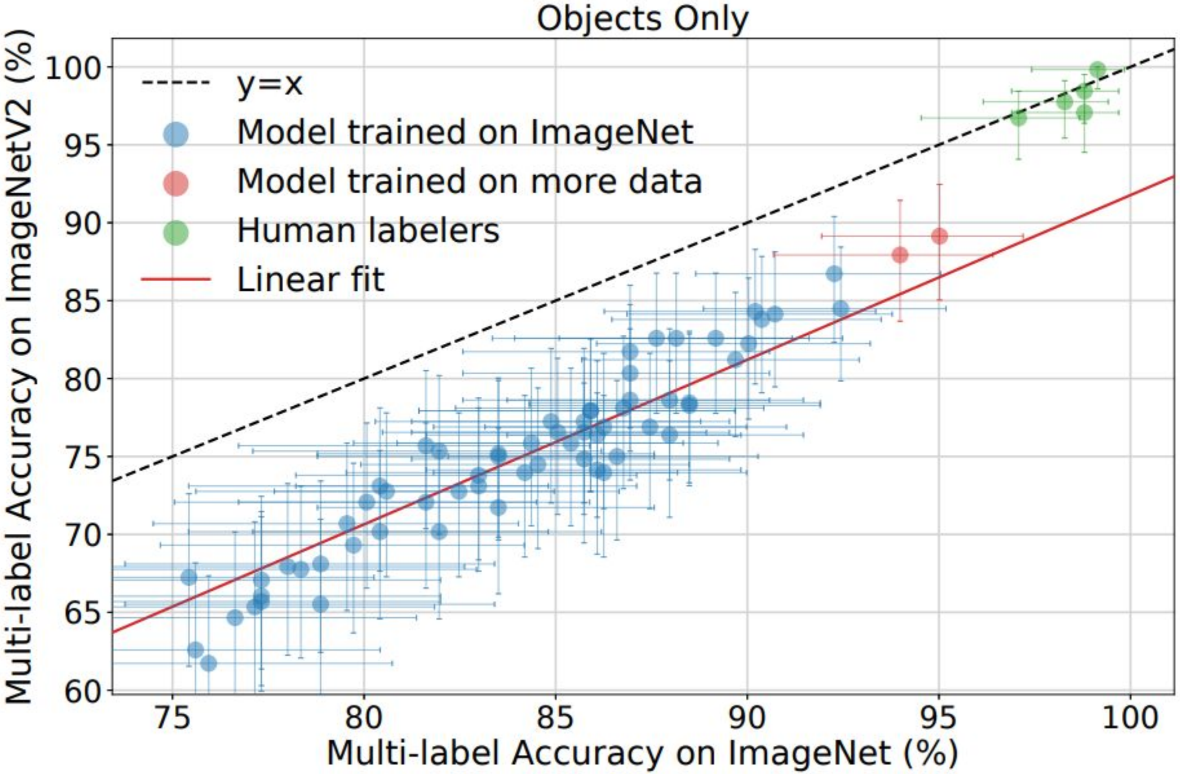
ILSVRC2012_val_00029666.JPEG

Near Duplicates

Sensitivity of the model to selection frequency

Illustrates that labeling biases play a large role in ImageNet model accuracy

Multi-label accuracies on OBJECTS only



Images that are difficult for humans

the potential insight into the failure modes of image classification models. To have a point of comparison let us start with the human labelers. There were 10 images which were misclassified by all human labelers. These images consisted of one image of a monkey and nine images of dogs. On the other hand, there were 27 images misclassified by all 72 models considered by us. Interestingly, 19 out of these images correspond to object classes and 8 correspond to organism classes. We note that there are only two images that were misclassified by all models and human labelers, both of them containing dogs. Four of the 27 images which were difficult for the models are displayed in Figure 5. It is interesting that the failure cases of the models consist of many images of objects while the failure cases of human labelers are exclusively images of animals.

Recommendation for Future work

1. Measure multi-label accuracy. While top-1 accuracy is still a good predictor of multi-label accuracy for models, this is not guaranteed for the future. Moreover, multi-label accuracy is a more meaningful metric for the ImageNet classification task. 2. Report performance on dogs, other animals, and inanimate objects separately. Label noise and ambiguities are a smaller concern on the 590 object classes where human labelers can achieve 99%+ accuracy. 3. Evaluate performance to distribution shift.