From ImageNet to Image Classification:

Contextualizing Progress on Benchmarks

5th July, 2020
PR12 Paper Review
JinWon Lee
Samsung Electronics

Dimitris Tsipras, et al., "From ImageNet to Image Classification: Contextualizing Progress on Benchmarks"

Related Papers

- Lucas Beyer, et al., "Are we done with ImageNet?"
  - https://arxiv.org/abs/2006.07159
- Kai Xiao, et al., "Noise or Signal: The Role of Image Backgrounds in Object Recognition"
  - https://arxiv.org/abs/2006.09994

Datasets for Image Classification

- *How aligned are existing benchmarks with their motivating real-world tasks?*

- The sheer size of machine learning datasets makes meticulous data curation virtually impossible. Dataset creators thus resort to scalable methods such as automated data retrieval and crowd-sourced annotation.

- As a result, the dataset and its corresponding annotations can sometimes be ambiguous, incorrect, or otherwise misaligned with ground truth.

Correct Labels?

~~(a) missile~~ ~~(b) stage~~ ~~(c) monastery~~ ~~(d) Staffordshire bull terrier~~

ImageNet Labels

projectile **acoustic guitar** **church** **American Staffordshire terrier**

ImageNet Creation Pipeline

- Image and label collection
  - The ImageNet creators first selected a set of classes using the WordNet hierarchy.
  - Then, for each class, they sourced images by querying several search engines with the WordNet synonyms of the class, augmented with synonyms of its WordNet parent node(s).
  - To further expand the image pool, the dataset creators also performed queries in several languages.

ImageNet Creation Pipeline

- Image validation via the CONTAINS task
  - The ImageNet creators employed annotators via the Mechanical Turk (MTurk) crowd-sourcing platform.
  - Annotators were presented with its description (along with links to the relevant Wikipedia pages) and a grid of candidate images.
  - Their task was then to select all images in that grid that contained an object of that class.

Revisiting the ImageNet Labels

- The root cause for many of these errors is that the image validation stage (i.e., the CONTAINS task) asks annotators only to verify if a specific proposed label.

- Crucially, annotators are never asked to choose among different possible labels for the image and, in fact, have no knowledge of what the other classes even are.

Images with Multiple Objects

- Annotators are instructed to ignore the presence of other objects when validating a particular ImageNet label for an image.

IN label:
Model T

Objects: 3
car wheel
Model T
grille

IN label:
Afghan hound

Objects: 2
malinois
Afghan hound

IN label:
cab

Objects: 3
cab
car wheel
palace

IN label:
pot

Objects: 2
bucket
pot

IN label:
toyshop

Objects: 2
teddy
toyshop

IN label:
maypole

Objects: 2
maypole
pole

IN label:
Siamese cat

Objects: 2
quilt
Siamese cat

IN label:
castle

Objects: 2
castle
seashore

IN label:
hamper

Objects: 2
hamper
wine bottle

IN label:
china cabinet

Objects: 4
goblet
vase
china cabinet
tray

IN label:
Model T

Objects: 4
car wheel
Model T
grille
convertible

IN label:
airliner

Objects: 2
airliner
wing

Biases in Image Filtering

- Since annotators have no knowledge of what the other classes are, they do not have a sense of the granularity of image features they should pay attention to.
- Moreover, the task itself does not necessary account for their expertise, e.g., between all the 24 terrier breeds that are present in ImageNet.

From Label Validation to Image Classification

- Authors would like them to classify the image, selecting all the relevant labels for it. However, asking (untrained) annotators to choose from among all 1,000 ImageNet classes is infeasible.
- First, they obtain a small set of (potentially) relevant candidate labels for each image.
- Then, they present these labels to annotators and ask them to select one of them for each distinct object using what they call the CLASSIFY task.
- They use 10,000 images from ImageNet validation set – 10 randomly selected images per class.

Obtaining Candidate Labels

- Potential labels for each image by simply combining the top-5 predictions of 10 models from different parts of the accuracy spectrum with the existing ImageNet label.

- Asking annotators whether an image contains a particular class but for all potential labels. The outcome of this experiment is a selection frequency for each image-label pair, i.e., the fraction of annotators that selected the image as containing the corresponding label

Obtaining Candidate Labels

Which of these images contain at least one object of type

Lhasa or Lhasa apso

Definition: a breed of terrier having a long heavy coat raised in Tibet as watchdogs

If you are unsure about the object meaning, please also consult the following Wikipedia page(s): https://en.wikipedia.org/wiki/Lhasa_Apso

Task:

For each of the following images, check the box next to an image if it contains at least one object of type *Lhasa or Lhasa apso*.

Select an image if it contains the object **regardless of occlusions, other objects, and clutter or text** in the scene. Only select images that are photographs (**no drawings or paintings**).

Please make accurate selections!

If it is impossible to complete a HIT due to missing data or other problems, please return the HIT. Blatantly incorrect answers might cause the HIT to be rejected.

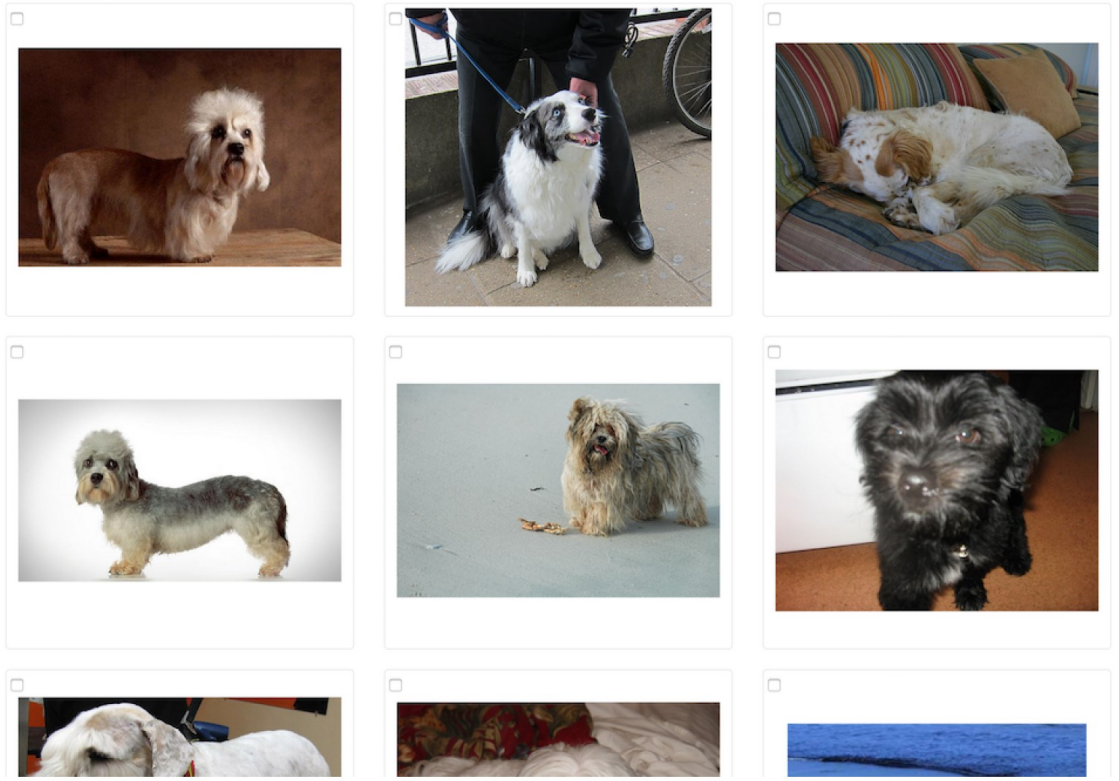


Image Classification via the CLASSIFY Task

- Asking annotators to identify
  - All labels that correspond to objects in the image.
  - The label for the main object (according to their judgment).
  - Select only one label per distinct object except mutually exclusive objects.

- Identifying the main label and number of objects
  - From each annotator's response, authors learn what they consider to be the label of the main object, as well as how many objects they think are present in the image

Image Classification via the CLASSIFY Task

Select which labels appear in the image

(Please read the instructions carefully as they are somewhat unusual)

Task:

1. Valid labels: select ALL labels that correspond to DISTINCT objects in the image. If for a single thing you cannot decide between multiple labels (which cannot all be true at the same time---e.g, different animal breeds), select the one that seems most likely.

> **Example:** Select ONLY ONE dog breed for a single dog and ONE shoe type for a single shoe (even if you are unsure about the correct breed/type---just pick one). BUT select BOTH "car" and "car wheel" for a car with visible wheels, BOTH "fur" and "coat" for a fur coat, BOTH "grocery store" and "orange" for oranges inside a grocery store (as these correspond to distinct things in the image).

2. Main object: from the chosen labels, select the one corresponding to the MAIN OBJECT in the image by clicking the appropriate radio button. (If you cannot decide which object is the main one, pick your best guess.)

If unsure about what a label means, you can consult the corresponding Wikipedia pages.

Examples:

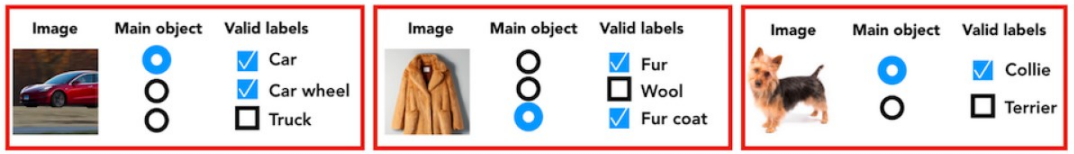


Image | Main object | Valid Labels

- **pedestal or plinth or footstall**
  Definition: an architectural support or base (as for a column or statue)
  Wikipedia: https://en.wikipedia.org/wiki/Pedestal
- **obelisk**
  Definition: a stone pillar having a rectangular cross section tapering towards a pyramidal top
  Wikipedia: https://en.wikipedia.org/wiki/Obelisk
- **pirate or pirate ship**
  Definition: a ship that is manned by pirates
  Wikipedia: https://en.wikipedia.org/wiki/Piracy

Submit

Quantifying the Benchmark-Task Alignment of ImageNet

- Multi-object images
  - More than 1/5 contains at least two objects.
  - Models perform significantly worse on multi-label images based on top-1 accuracy (measured w.r.t. ImageNet labels): accuracy drops by more than 10% across all models.
  - There are pairs of classes which consistently co-occur.
  - Model accuracy is especially low on certain classes that systematically co-occur.
  - In light of this, a more natural notion of accuracy for multi-object images would be to consider a model prediction to be correct if it matches the label of any object in the image.

Multi-Object Images

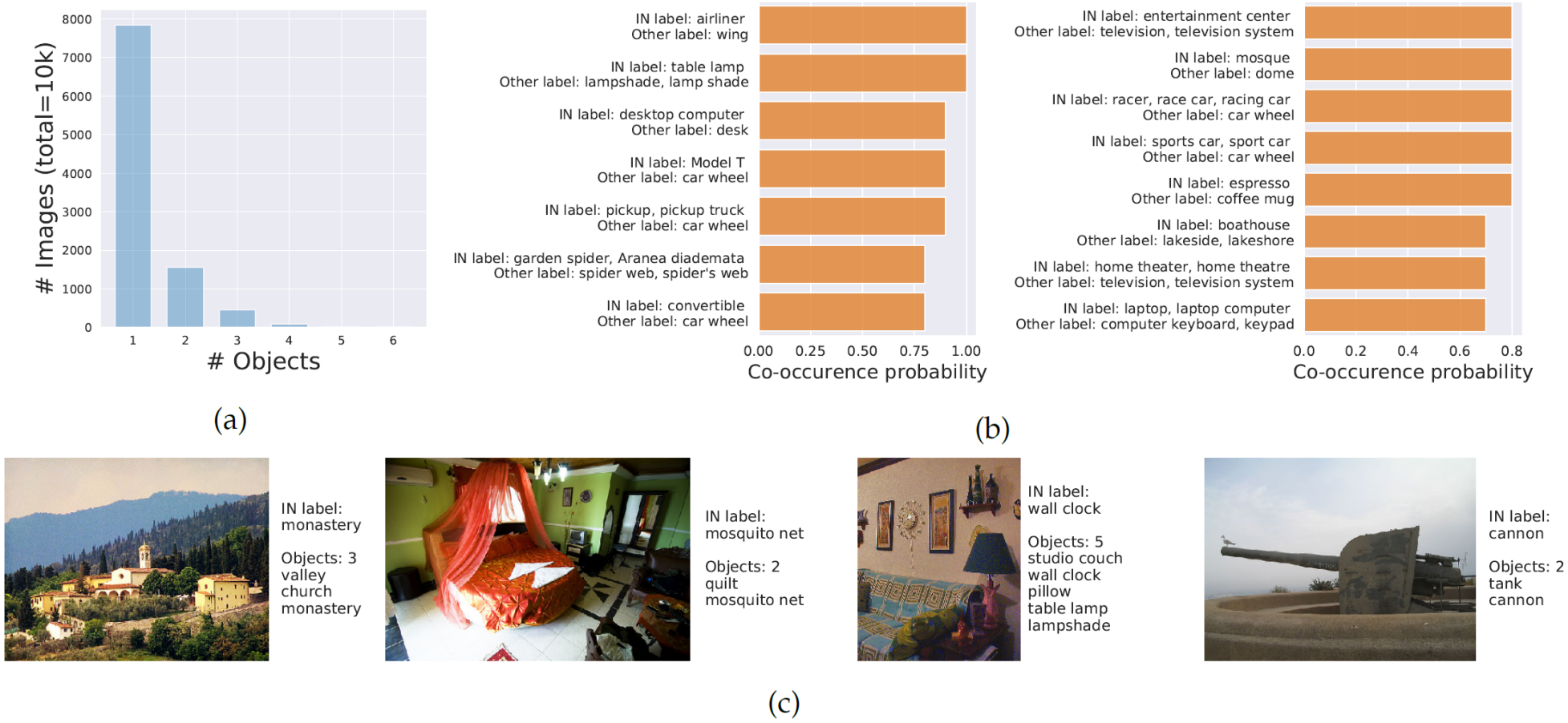


Figure 3: (a) Number of objects per image—more than a fifth of the images contains two or more objects from ImageNet classes. (b) Pairs of classes which consistently co-occur as distinct objects. Here, we visualize the top 15 ImageNet classes based on how often their images contain another *fixed* object ("Other label"). (c) Random examples of multi-label ImageNet images (cf. Appendix Figure 17 for additional samples).

Multi-Object Images

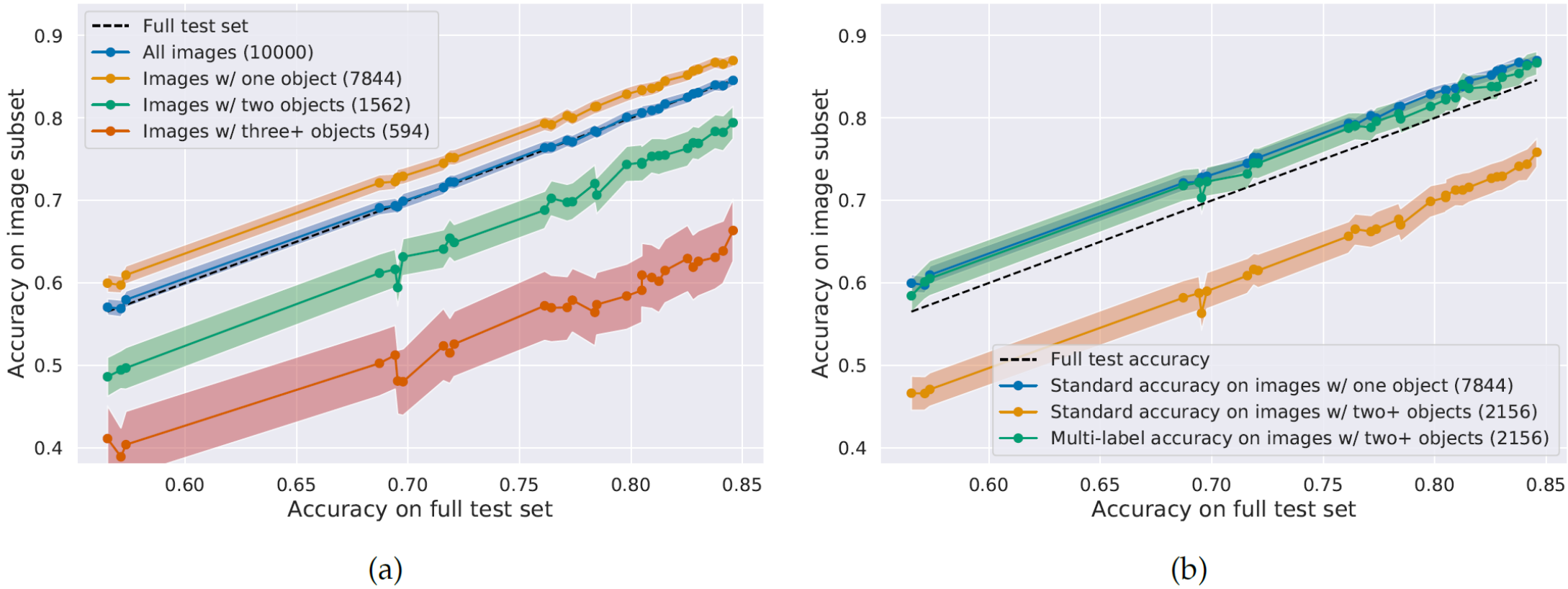


Figure 4: (a) Top-1 model accuracy on multi-object images (as a function of overall test accuracy). Accuracy drops by roughly 10% across all models. (b) Evaluating multi-label accuracy on ImageNet: the fraction of images where the model predicts the label of *any* object in the image. Based on this metric, the performance gap between single- and multi-object images virtually vanishes. Confidence intervals: 95% via bootstrap.

Multi-Object Images

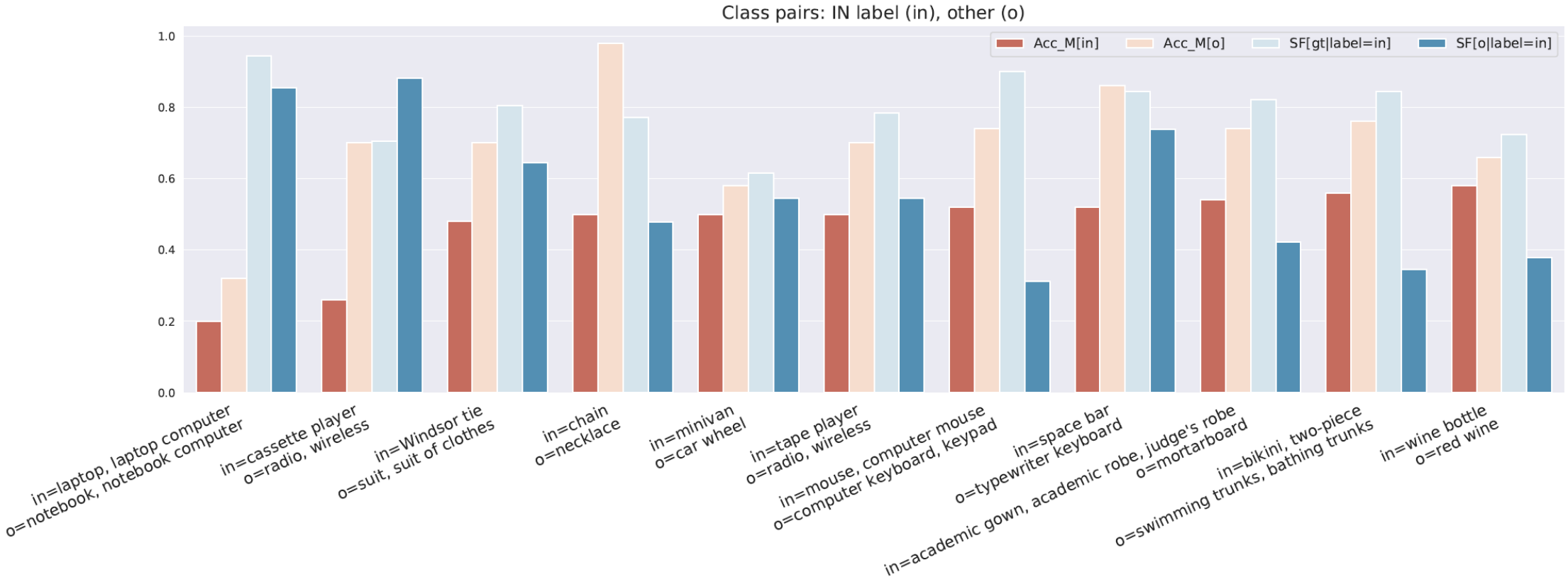
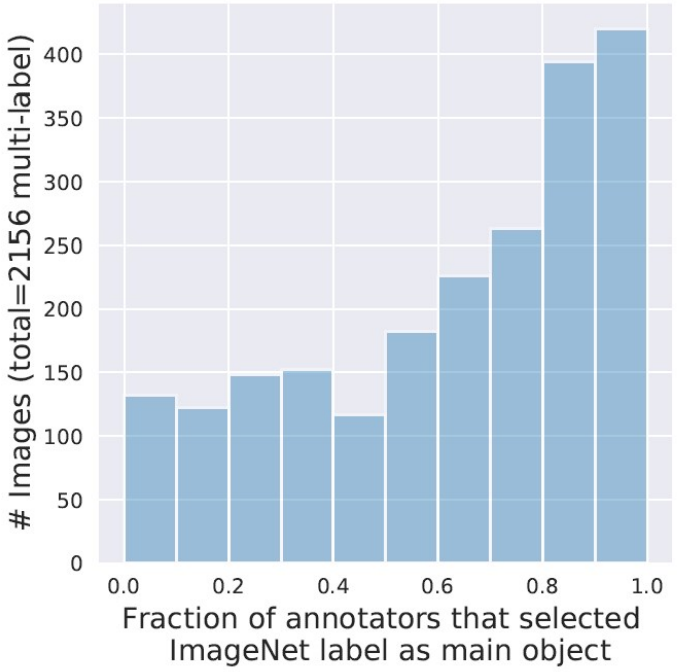


Figure 21: Classes where model accuracy is consistently low due to frequent object co-occurences: in these cases, an object from the ImageNet class frequently co-occurs with (or is a sub-part of) objects from another class. Here, models seem to be unable to disambiguate the two classes completely, and thus perform poorly on one/both classes. Note that human selection frequency for the ImageNet class is high, indicating that an object from that class is present in the image.

Multi-Object Images

- Human-label disagreement
  - Although models suffer a sizeable accuracy drop on multi-object images, they are still relatively good at predicting the ImageNet label.
  - This bias could be justified whenever there is a distinct main object in the image and it corresponds to the ImageNet label.
  - However, for nearly a third of the multi-object images, the ImageNet label does not denote the most likely main object as judged by human annotators.

Human-Label disagreement



(a)

ImageNet label:
feather boa

Main label:
sunglasses

Objects: 3
sunglasses
sunglass
feather boa

ImageNet label:
wooden spoon

Main label:
mixing bowl

Objects: 2
wooden spoon
mixing bowl

ImageNet label:
ladle

Main label:
fountain

Objects: 2
ladle
fountain

ImageNet label:
bulletproof vest

Main label:
assault rifle

Objects: 2
assault rifle
bulletproof vest

(b)

Figure 5: (a) Fraction of annotators that selected the ImageNet label as denoting the main image object. For 650 images (out of 2156 multi-object images), the majority of annotators select a label *other* than the ImageNet one as the main object. (b) Random examples of images where the main label as per annotators contradicts the ImageNet label (cf. Appendix Figure 19 for additional samples).

Human-Label disagreement

IN label:
caldron

Main label:
street sign

Objects: 2
street sign
caldron

IN label:
assault rifle

Main label:
military uniform

Objects: 3
assault rifle
military uniform
rifle

IN label:
Windsor tie

Main label:
suit

Objects: 2
Windsor tie
suit

IN label:
cleaver

Main label:
gown

Objects: 2
cleaver
gown

IN label:
bearskin

Main label:
trombone

Objects: 2
bearskin
trombone

IN label:
water bottle

Main label:
restaurant

Objects: 3
goblet
restaurant
water bottle

IN label:
plate

Main label:
soup bowl

Objects: 2
plate
soup bowl

IN label:
coffeepot

Main label:
teapot

Objects: 3
teapot
cup
pot

IN label:
bonnet

Main label:
wool

Objects: 2
wool
bonnet

IN label:
consomme

Main label:
plate

Objects: 2
consomme
soup bowl

IN label:
banana

Main label:
orange

Objects: 2
orange
banana

IN label:
vault

Main label:
church

Objects: 2
altar
church

IN label:
picket fence

Main label:
seashore

Objects: 2
picket fence
seashore

IN label:
shoji

Main label:
sliding door

Objects: 4
sliding door
shoji
dining table
window shade

IN label:
sock

Main label:
Loafer

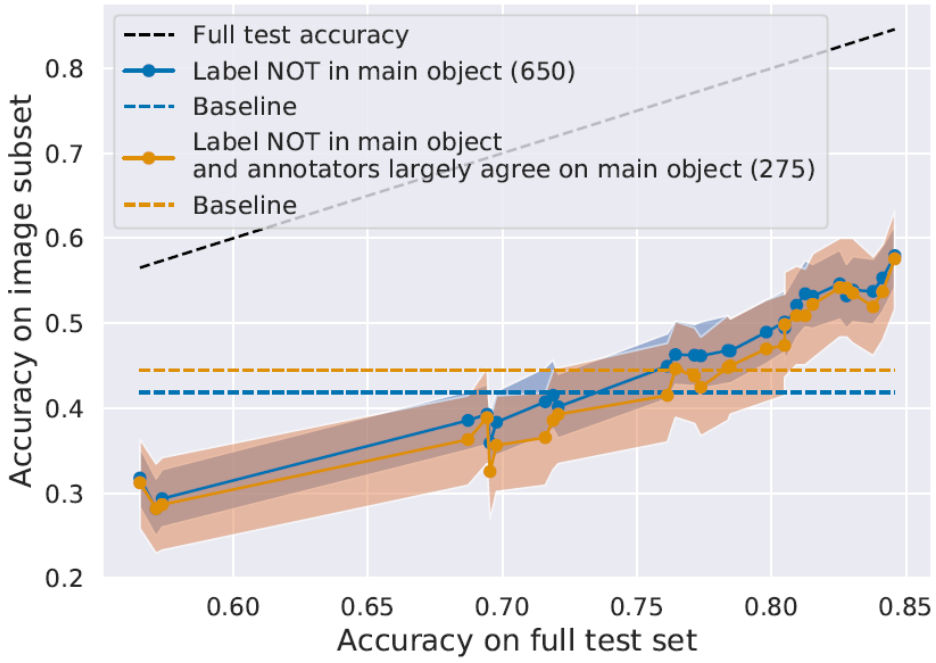
Objects: 2
sock
Loafer

IN label:
ping-pong ball

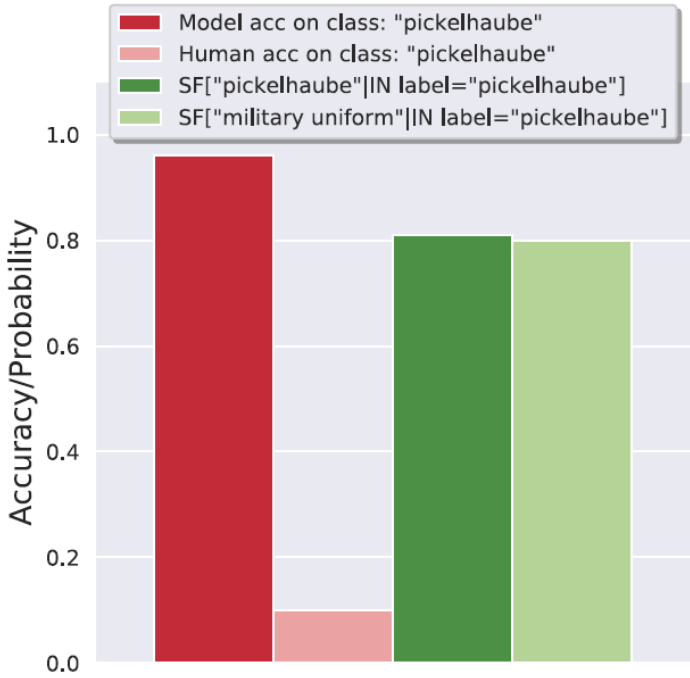
Main label:
upright

Objects: 2
ping-pong ball
upright

Human-Label disagreement



(a)



ImageNet label: *pickelhaube*

Main label: *military uniform*

Objects: *2 (military uniform, pickelhaube)*

(b)

Figure 6: (a) Model accuracy on images where the annotator-selected main object does not match the ImageNet label. Models perform much better than the baseline of randomly choosing one of the objects in the image (dashed line)—potentially by exploiting dataset biases. (b) Example of a class where humans disagree with the label as to the main object, yet models still predict the ImageNet label. Here, for images of that class, we plot both model and annotator accuracy as well as the selection frequency (SF) of both labels.

Quantifying the Benchmark-Task Alignment of ImageNet

- Bias in label validation
  - Recall that annotators are asked a somewhat leading question, of whether a specific label is present in the image, making them prone to answering positively even for images of a different, yet similar, class.
  - Under the original task setup (i.e., the CONTAINS task), annotators consistently select multiple labels, for nearly 40% of the images, another label is selected at least as often as the ImageNet label.
  - Even when annotators perceive a single object, they often select as many as 10 classes.
  - If instead of asking annotators to judge the validity of a specific label(s) in isolation, asking them to choose the main object among several possible labels simultaneously (i.e., via the CLASSIFY task), they select substantially fewer labels.

Bias in Label Validation

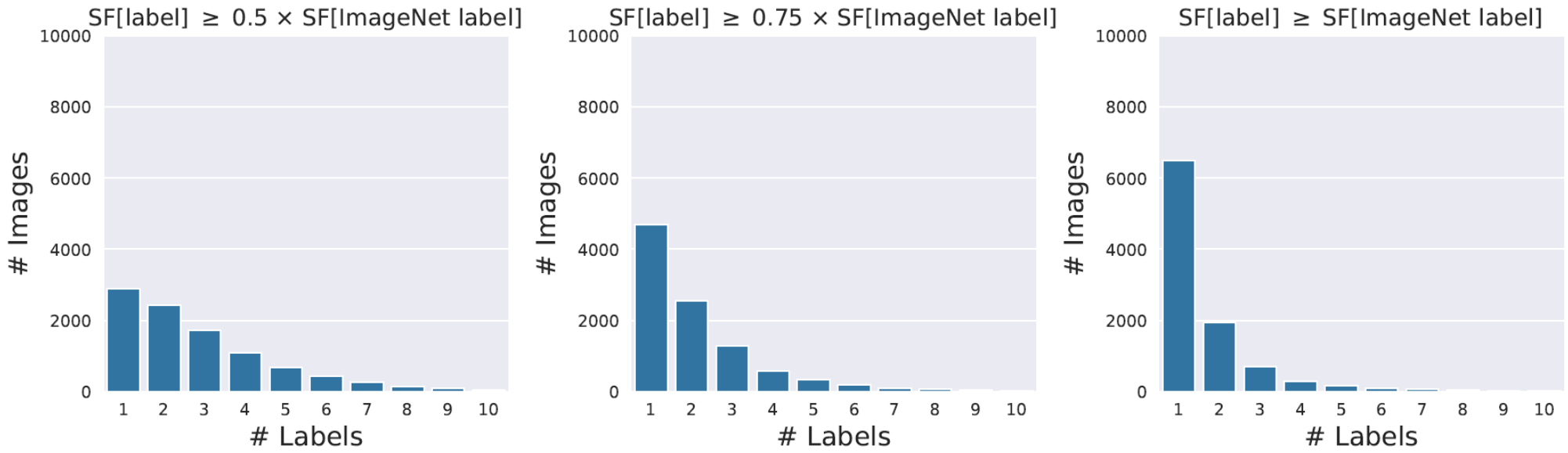


Figure 7: Number of labels per image that annotators selected as valid in isolation (determined by the selection frequency of the label relative to that of the ImageNet label). For more than 70% of images, annotators select another label at least half as often as they select the ImageNet label (*leftmost*).

Bias in Label Validation

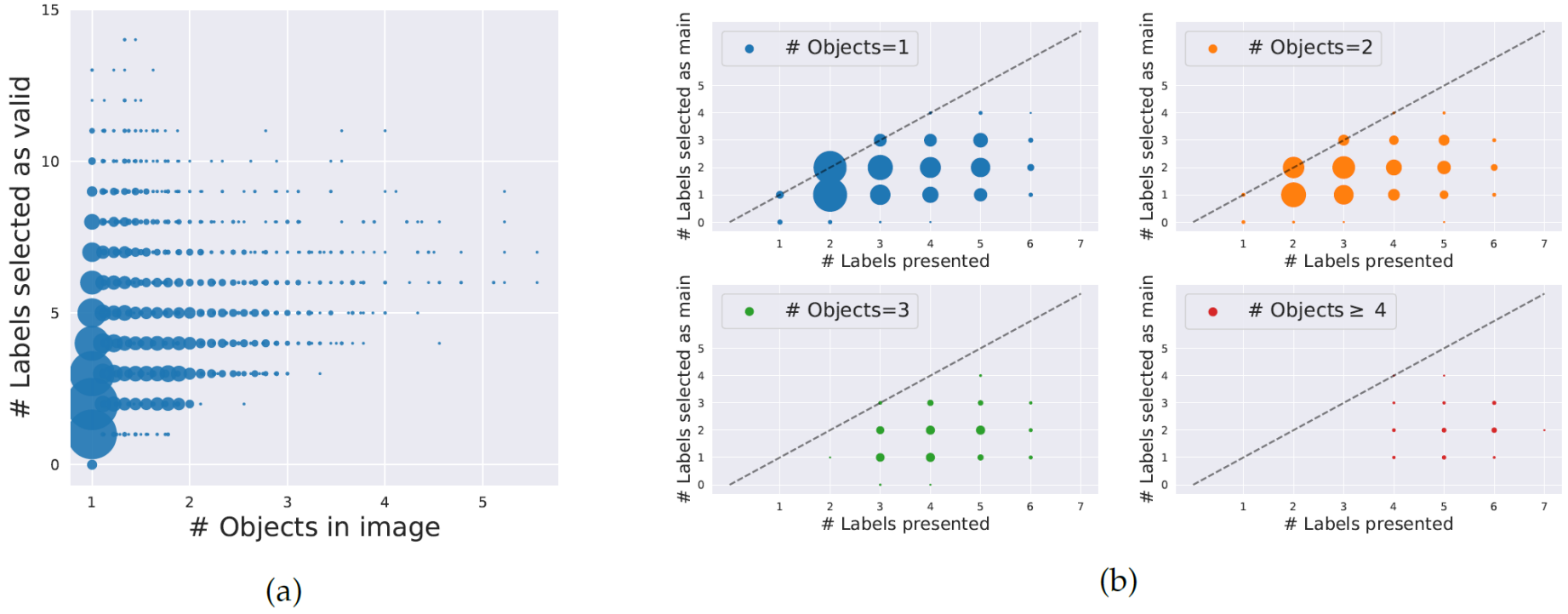


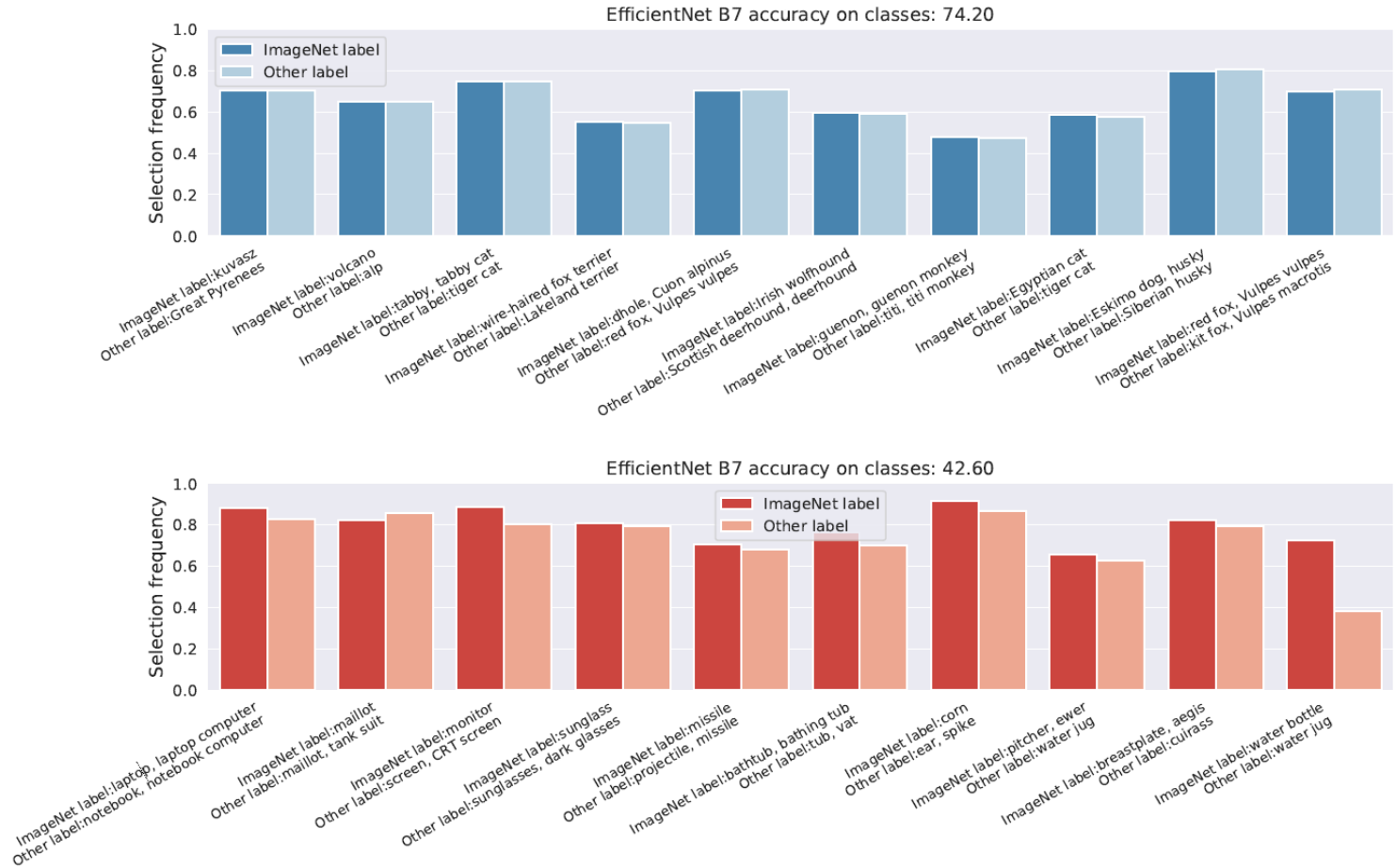
Figure 8: (a) Number of labels selected in the CONTAINS task: the y-axis measures the number of labels that were selected by at least two annotators; the x-axis measures the average number of objects indicated to be present in the image during the CLASSIFY task; the dot size represents the number of images in each 2D bin. Even when annotators perceive an images as depicting a single object, they often select multiple labels as valid. (b) Number of labels that at least two annotators selected for the main image object (in the CLASSIFY task; cf. Section 3.2) as a function of the number of labels presented to them. Annotator confusion decreases significantly when the task setup explicitly involves choosing between multiple labels *simultaneously*.

Bias in Label Validation

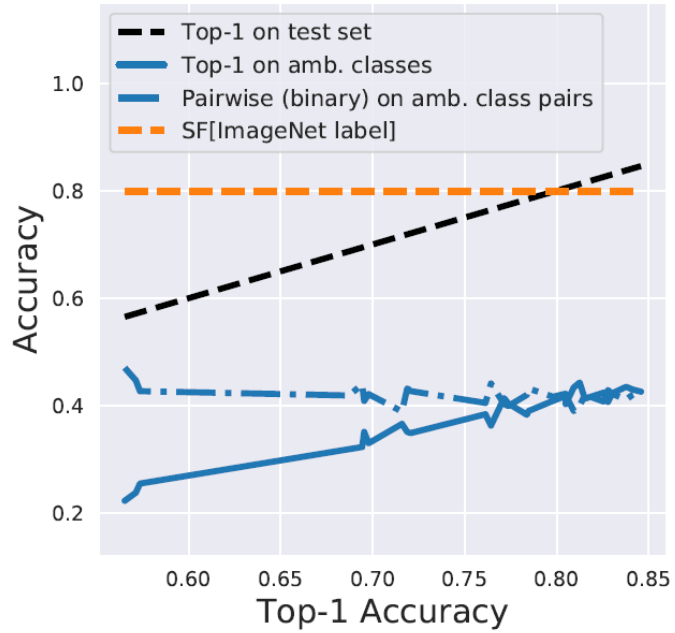
- Confusing class pairs
  - There are several pairs of ImageNet classes that annotators have trouble telling apart—they consistently select both labels as valid for images of either class.
  - On some of these pairs we see that models still perform well.
  - However, for other pairs, even state-of-the-art models have poor accuracy (below 40%).
  - If that is indeed the case, it is natural to wonder whether accuracy on ambiguous classes can be improved without overfitting to the ImageNet test set.

Bias in Label Validation

- Confusing class pairs
  - The annotators' inability to remedy overlaps in the case of ambiguous class pairs could be due to the presence of classes that are semantically quite similar, e.g., "rifle" and "assault rifle".
  - In some cases, there also were errors in the task setup. For instance, there were occasional overlaps in the class names (e.g., "maillot" and "maillot, tank suit") and Wikipedia links (e.g., "laptop computer" and "notebook computer") presented to the annotators.



(a)



(b)

Figure 9: (a) ImageNet class pairs for which annotators often deem both classes as valid. We visualize the top 10 pairs split based on the accuracy of EfficientNet B7 on these pairs being high (*top*) or low (*bottom*). (b) Model progress on ambiguous class pairs (from (a) *bottom*) has been largely stagnant—possibly due to substantial overlap in the class distributions. In fact, models are unable to distinguish between these pairs better than chance (cf. pairwise accuracy).

Beyond Test Accuracy: Human-In-The-Loop Model Evaluation

- Human assessment of model predictions
  - Selection frequency of the prediction
    - How often annotators select the predicted label as being present in the image (determined using the CONTAINS task).
  - Accuracy based on main label annotation
    - How frequently the prediction matches the main label for the image, as obtained using our CLASSIFY task.

Human Assessment of Model Predictions

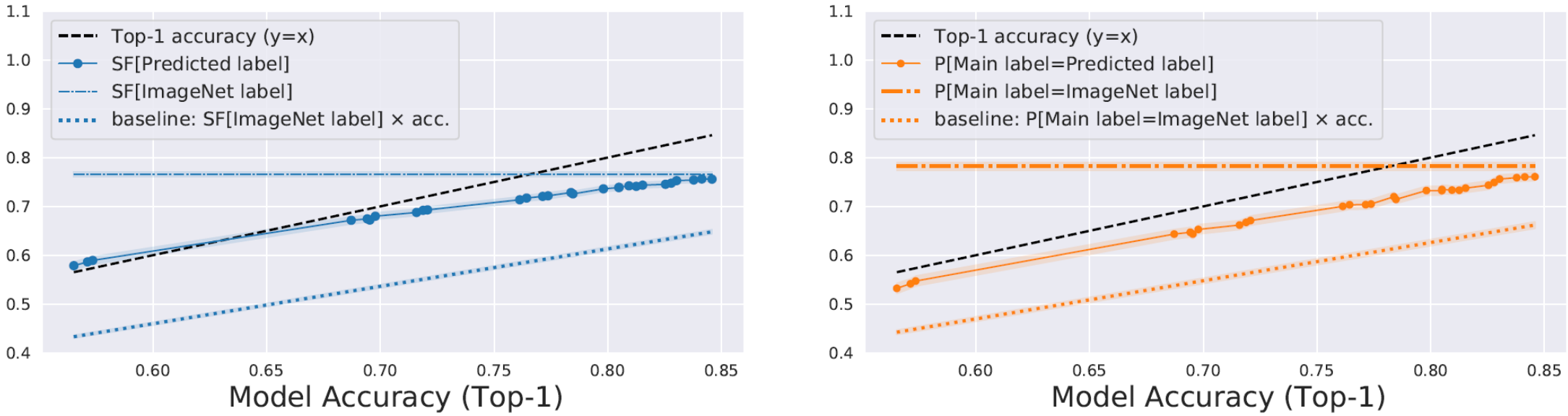
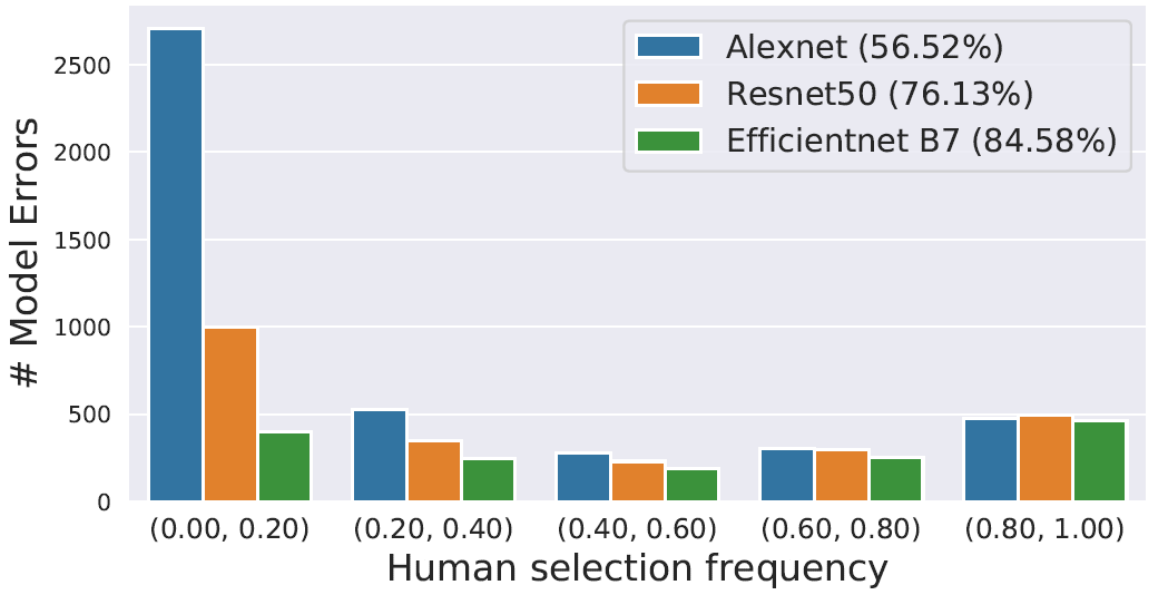


Figure 10: Using humans to assess model predictions—we measure how often annotators select the predicted/ImageNet label: (a) to be contained in the image (*selection frequency* [SF] from Section 3.1), and (b) to denote the *main* image object (cf. Section 3.2), along with 95% confidence intervals via bootstrap (shaded). We find that though state-of-the-art models have imperfect top-1 accuracy, their predictions are, on average, almost indistinguishable according to our annotators from the ImageNet labels themselves.

Human Assessment of Model Predictions – Incorrect Predictions



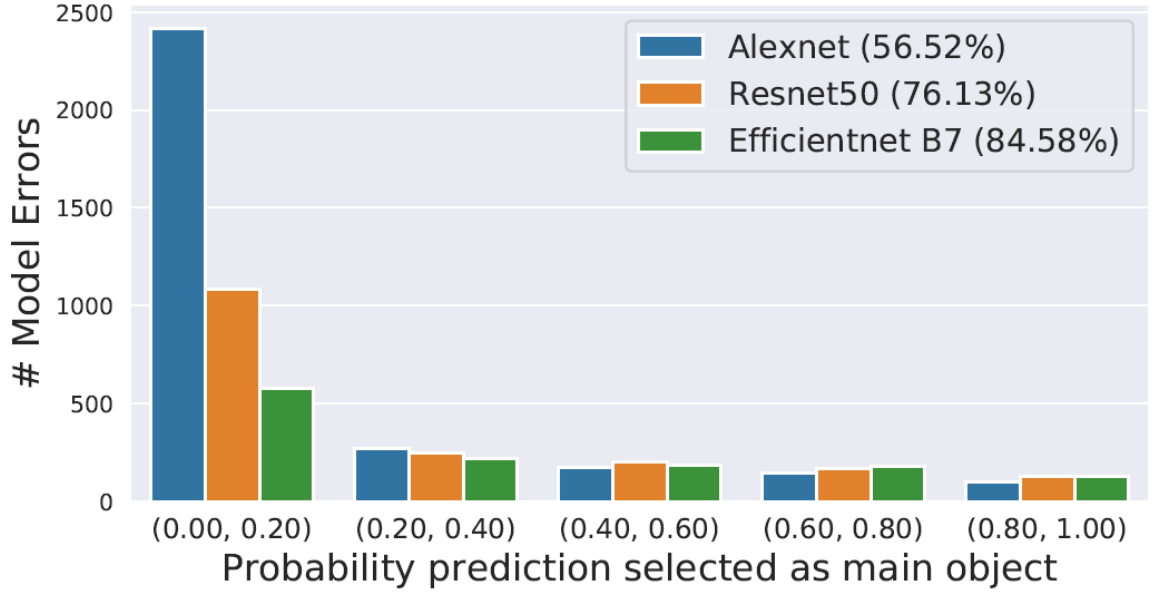
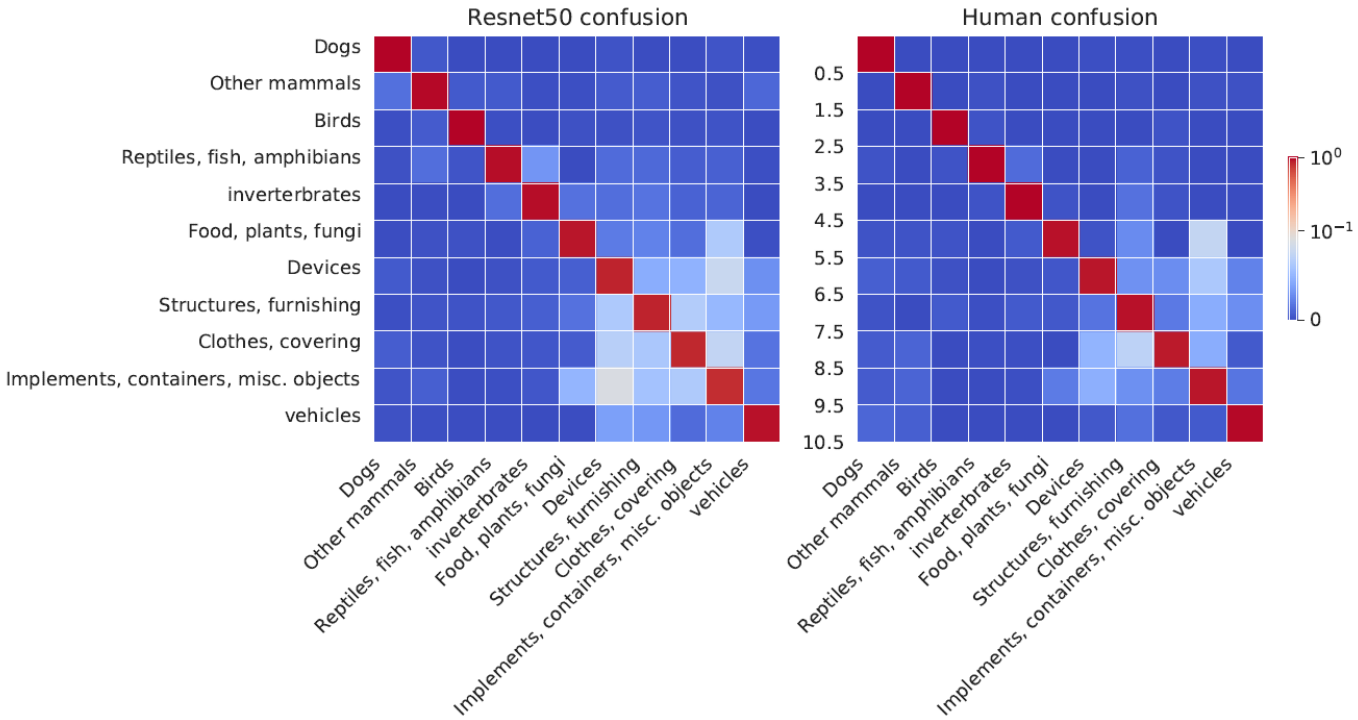


Figure 11: Distribution of annotator selection frequencies (cf. Section 3.1) for model predictions deemed incorrect w.r.t. the ImageNet label. Models that are more accurate also seem to make fewer mistakes that have low human selection frequency (for the corresponding image-prediction pair).

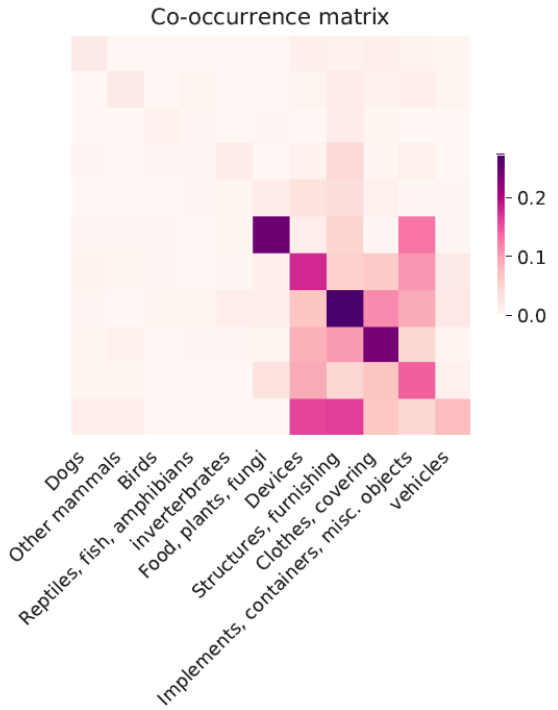
Human Assessment of Model Predictions

- The predictions of state-of-the-art models have gotten, on average, quite close to ImageNet labels.
- That is, annotators are almost equally likely to select the predicted label as being present in the image (or being the main object) as the ImageNet label.
- This indicates that model predictions might be closer to what non-expert annotators can recognize as the ground truth.
- This also means, we are at a point where we may no longer by able to easily identify (e.g., using crowd-sourcing) the extent to which further gains in accuracy correspond to such improvements.

Confusion Matrix



(a) Confusion matrices



(b) Co-occurrence matrix

Figure 12: Characterizing similarity between model and human predictions (indicated by main object selection in the CONTAINS task): (a) Model (ResNet-50) and human confusion matrices on 11 ImageNet superclasses (cf. Section A.1.1). (b) Superclass co-occurrence matrix: how likely a specific pair of superclasses is to occur together (based on annotation from Section 3.2).

Conclusion

- There are many images with multiple valid labels, and ambiguous classes in ImageNet.
- Top-1 accuracy often underestimates the performance of models by unduly penalizing them for predicting the label of a different, but also valid, image object.
- Taking a step towards evaluation metrics that circumvent these issues, authors utilize human annotators to directly judge the correctness of model predictions.
- On the positive side, they find that models that are more accurate on ImageNet also tend to be more human-aligned in their errors.
- While this might be reassuring, it also indicates that we are at a point where we cannot easily gauge (e.g., via simple crowd-sourcing) whether further progress on the ImageNet benchmark is meaningful, or is simply a result of overfitting to this benchmarks' idiosyncrasies.