**SLIDE 2 and 3**

This project is discussing about the Large scale vision dataset creation pipeline, shortfalls of the pipeline, the level of misalignment between the benchmark of the models that are built using this datasets and the real world task to which these models are serving as a proxy for. This project proposes methodology to overcome these issues.

And this project focuses on emphasizing the need to augment current model training and evaluation toolkit to take such misalignments into account.

Why I chose this ?

What are the

**SLIDE 4**

Large scale datasets are usually created through scalable methods like crowd-source annotation and automated data retrieval. The annotators will be asked to validate a given label for an image without the knowledge of other available classes

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

**SLIDE 5 - Image**

While the labels shown above appear valid for the corresponding images, none of them actually match the ImageNet labels (which are “projectile”, “acoustic guitar”, “church”, and “American Staffordshire terrier”, respectively

**SLIDE 6**

In this work, human studies are used to investigate the consequences of employing such a pipeline, focusing on the popular ImageNet dataset.

Consider the most popular vision dataset for this study. ImageNet data has 1000 class with over million images.

This dataset is created through automated data collection and crowd-sourced filtering.

This creation process comprises of 2 stages

Image and label collection

Image validation via CONTAINS task

**SLIDE 7**

How the image and label are collected is, the ImageNet creators first selected a set of classes using wordnet hierarchy. Then for each class, they sourced images by querying several search engines. The queries are made of Wordnet synonyms of the class and augments with synonyms of its wordnet parent nodes

For example, the query for class “whippet” (parent: “dog”) would also include “whippet dog” [Rus+15].

To further expand the image pool, the dataset creators also performed queries in several languages.

**SLIDE 8**

Image creators have employed MTurk crowd sourcing platform.

Specifically, for every class, annotators were presented with its description (along with links to the relevant Wikipedia pages) and a grid of candidate images.

They will have to select all images in that grid that contained an object of that class.

The grids were shown to multiple annotators and only images that received a “convincing majority” of votes were included in ImageNet. This filtering procedure is called the CONTAINS task.

**SLIDE 9**

So how are the ImageNet labels going to be. As a result of above discussed data creation pipeline, resulting dataset might not capture the ground truth.

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 shown in isolation, is valid for a given image. Crucially, annotators are never asked to choose among different possible labels for the image and, in fact, they have no knowledge of what the other classes even are.

So, This can introduce discrepancies in the dataset in two ways:

**SLIDE 10 – here 4th slide**

If there are Image with multiple objects

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

As you can see from the image, There are other objects available and they are not considered as valid labels like for example The imagenet label is cabinet but the other mutually exclusive objects are goblet, vase, tray etc

**SLIDE 11**

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 (the labels in figure 1 appear reasonable until one becomes aware of the other possible classes in ImageNet).

So yeah, one cannot reasonably expect non-experts to distinguish, e.g., between all the 24 terrier breeds that are present in ImageNet.

As a result, if annotators are shown images containing objects of a different, yet similar class, they are likely to select them as valid. This implies that potential errors in the collection process are unlikely to be corrected during validation (via the CONTAINS task) and thus can propagate to the final dataset.

**SLIDE 12**

So in order to address all the mentioned issues, have developed a methodology for obtaining fine-grained data annotations via large-scale human studies. They have presented a framework for improving this refinement of existing labels and then use that framework to investigate the discrepancies highlighted above and their impact on ImageNet trained models

**SLIDE 13**

The framework concentrates on understanding of the ground truth for ImageNet data. To achieve this, rather than asking annotators to validate a single proposed label for an image (as in the original pipeline), we 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.

So, For the analysis, the we have 10,000 images from the ImageNet validation set—i.e., 10 randomly selected images per class.

since both ImageNet training and ImageNet validation sets were created using the same procedure, analyzing the latter is sufficient to understand systematic issues in that dataset.

To overcome this difficulty, our pipeline consists of two phases

**SLIDE 14**

First, they obtain a small set of (potentially) relevant candidate labels for each image

How they do this is, by combining top 5 predictions of 10 models from different parts of accuracy spectrum with the existing ImageNet label. (yields approximately 14 labels per image).

Then, to prune this set further, Reuse the CONTAINS task – 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

We find that relatively few of these labels had high selection frequency (typically less than five per image).

Thus, restricting potential labels to this smaller set of candidate labels allows us to hone in on the most likely ones, while ensuring that the resulting annotation task is still controllable

**SLIDE 15**

Show and move on

**SLIDE 16**

Once identified a small set of candidate labels for each image, it is presented to annotators to obtain fine-grained label information

As we can see from the image, annotators are asked 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.
* From each annotator’s response, we 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

**SLIDE 17**

Different annotators may choose different labels for the same object and thus we need to map their selections to a single set of distinct objects

So the we would like to partition selections in a way that avoids grouping labels together if annotators identified them as distinct objects

So we have employed exhaustive search to find a partition that optimizes for this criterion.

Finally, labelling each distinct object with its most frequently selected label.

The resulting annotations characterize the content of an image in a more fine-grained manner compared to the original ImageNet labels.

Note that these annotations may still not perfectly match the ground truth.

**SLIDE 18**

Quantifying benchmark-task alignment of ImageNet

In this section of the framework, the goal is to examine potential sources of discrepancy between ImageNet data and the object recognition task, using the obtained refined image annotations

And then assess the impact, these deviations have on models developed using this benchmark.

Two important topics to be considered while quantifying the misalignment is “Multi object images and bias in label validation”

will measure how the accuracy of a diverse set of models (see Appendix A.2 for a list), is affected when they are evaluated on different sets of images.

Using any image dataset,

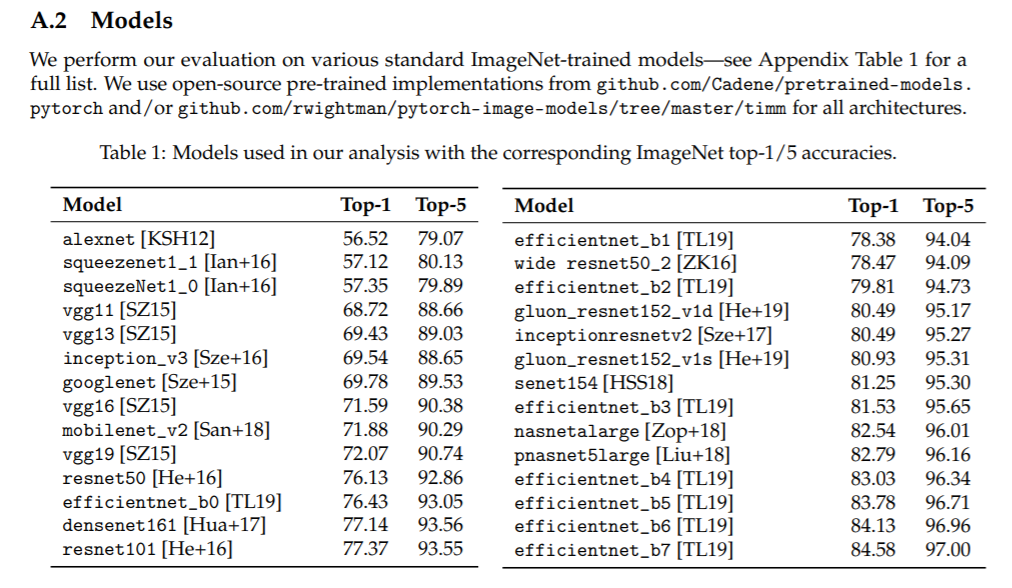
Model performance is typically measured using (top-1 or top-5) accuracy with respect to a single ImageNet label, treating it as the ground truth

It has been observed a significant fraction of images—more than a fifth—contains at least two objects

Models have performed worse on multi-label images based on top-1 accuracy

accuracy drops by more than 10% across all models.

**SLIDE 19**



* 1. There are pairs of classes which consistently co-occur.
  2. Model accuracy is especially low on certain classes that systematically co- occur.
  3. 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.
  4. If this metric is considered, the aforementioned performance drop disappears and the model performs similarly on single and multi-object images
  5. This indicates that the way typically accuracy is measured with respect to a single label can be overly pessimistic, as it penalizes the model for predicting the label of another valid object in the image

**SLIDE 20**

So on top of this how human-label disagreement comes up is

Slide pathu padichiru

**SLIDE 21**

these disagreements often arise when the ImageNet label corresponds to a very distinctive object (e.g., “street sign”), but the image also contain another more prominent object (caldron)

**SLIDE 22**

Bias in label validation

* Recalling 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.

SLIDE 23

Image

Explanation from slide itself

**SLIDE 24**

* Confusing class pairs
  1. 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.
  2. On some of these pairs we see that models still perform well.
  3. However, for other pairs, even state-of-the-art models have poor accuracy (below 40%).
  4. 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.

**SLIDE 25**

Image 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).

* 1. The annotators’ inability to remedy these overlaps in the case of ambiguous class pairs could be due to the presence of classes that are quite similar, e.g., “rifle” and “assault rifle”.
  2. In some cases, there also were errors in the task setup.
  3. 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.

**SLIDE 26**

So far, we have discussed what can be done to avoid misalignments, to improve the accurate model and its performance and how else the data can be curated.

Coming to the final part – model evaluation

Beyond test accuracy

There is this human assessment of model predictions

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

Figure 10: how often annotators select the predicted/ IN 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.

**SLIDE 27**

Paathu graph padichiru

**SLIDE 28**

Pathu padichiru

**SLIDE 30**

Pathu padichiru