Large-Scale Long-Tailed Recognition in an Open World

Chen Zhao

CVPR 2019 https://arxiv.org/pdf/1904.05160.pdf

1 Motivation

In existing visual recognition setting, the training data and testing data are both balanced under a closed-world setting, e.g. the ImageNet dataset. However, this setting is not a good proxy of the real-world scenario. For example, it is never possible for ecologists to gather balanced wildlife datasets because animal distribution is imbalanced. Similarly, people are bothered by the imbalanced and open-ended distribution from all sorts of datasets: street signs, fashion brands, faces, weather conditions, street conditions, etc. To faithfully reflect these aspects, we formally study "Open Long-Tailed Recognition" (OLTR) arising in natural data settings. A practical system shall be able to classify among a few common and many rare categories, to generalize the concept of a single category from only a few known instances, and to acknowledge novelty upon an instance of a never seen category. We define OLTR as learning from long-tail and open-end distributed data and evaluating the classification accuracy over a balanced test set which includes head, tail, and open classes in a continuous spectrum.

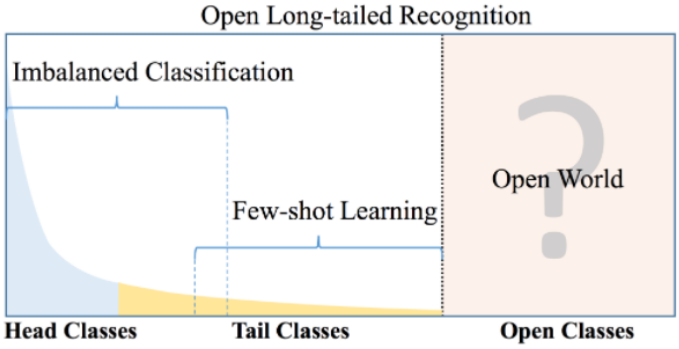


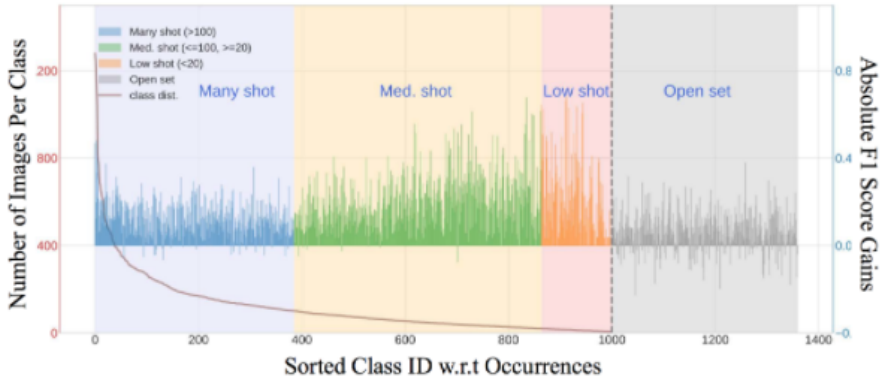
Figure 1: Our task of open long-tailed recognition must learn from long-tail distributed training data in an open world and deal with imbalanced classification, few-shot learning, and open-set recognition over the entire spectrum

This paper formulated the task of Open Long-Tailed Recognition (OLTR) that combined many-shot, medium-shot, few-shot and open-set classification into one. This would ask for recognition robustness and open-set sensitivity at the same time which are to some extent competing. This paper proposed a unified solution which leverages the self-attention,memorized meta feature, cosine classifier and large margin loss.

2 Method

2.1 OLTR

There has been several image classification tasks spanning over the spectrum: Imbalanced classification, few-shot learning, and open-set recognition – either of them focusing on solving one aspect of classification problem, i.e. one particular range of frequency of train/test data samples (per class). The authors propose the open Long-Tailed Recognition (OLTR) task where one integrated algorithm is asked to solve them all at once, dealing with classes that have no training data (zero-shot/open-set), 20ish (few-shot), all the way to more than hundreds (many shots).



2.2 Dynamic Meta-Embedding

The first important module propose is Dynamic meta-embedding. When an income image are from a new category, or category with less samples, the naive-trained encoder usually does not extract good enough features v^{direct}. One way to fix this is to augmented the feature vector with the encoding of other concepts. Here centroids of classes $M = c_i^K$ is used as the memory embedding. Concept Selector and Hallucinator will re-weight the memory embedding across classes as well as dimensions of v^{direct}:

$$v^{reweighted} = v^{direct} + tanh(T_{sel}(v^{direct})) \otimes T_{hal}(v^{direct})^T M$$

We can further leverage the memory we have by penalize the embedding with the distance of nearest centroid:

$$v^{meta} = \frac{1}{\gamma} v^{reweight}, \quad \gamma := \min_i ||v^{direct} - c_i||_2$$

Therefore if the direct embedding is far away from the any classes embedding, v^{meta} will be very small and in the extreme case nothing will be passed to the classifier – this gives the downstream module a chance to switch between few-shot and zero-shot.

2.3 Modulated Attention

To improve v^{meta}, we can plug in self attention on top of the features generated:

$$f^{att} = f + MA(f) \otimes SA(f)$$

where $MA(\cdot)$ and $SA(\cdot)$ are not explained in details but inherited from the *Attention Is All You Need* paper.

This paper choose to use cosine similarity as the classier to mitigate 1 the huge difference of sample numbers per classes. It will normalize the v^{meta} to slightly shorter than 1 an the weigh ts to be 1 before calculating the dot product. In most of the experiments, if the softmax score is less than 0.1 the image is flagged as novel.

The author constructed three long tail dataset by sampling according to a Pareto distribution upon existing dataset. The model achieve consistent better performance for all region of OLTR. In the ablation study, the dynamic meta-embedding seems to be biggest driver.