Hierarchical Novelty Detection for Visual Object Recognition

Cheng Guan

August 7, 2018

1. Approach

From the last reading, I know that the primary focus of ZSL and GZSL tasks is on transfer learning for a new domain, and they assume that semantic information of all test classes is given, e.g., attributes or text description [1, 3, 2, 4, 5] of the objects. Therefore, GZSL cannot recognize a novel class if prior knowledge about the specific novel class is not provided. In this section, the authors define terminologies to describe hier- archical taxonomy and then propose models for hierarchical classification combined with novelty detection.

1.1. Taxonomy

A taxonomy represents a hierarchical relationship among classes, where each node in the taxonomy corresponds to a class or a set of indistinguishable classes. They define three types of classes as follows: 1) *known leaf classes* are nodes with no child, which are known and seen during training, 2) *super classes* are ancestors of the leaf classes, which are also known, and 3) *novel classes* are unseen during training, so they do not explicitly appear in the taxonomy. They note that all known leaf and novel classes have no child and are disjoint, i.e., they are neither ancestor nor descendant of each other. In the example in Figure 1, four species of cats and dogs are leaf classes, "cat," "dog," and animal are super classes, and any other classes un- seen during training, e.g.,"Angola cat," "Dachshund," and "Pika" are novel classes.

In the proposed hierarchical novelty detection framework, they first build a taxonomy with known leaf classes and their super classes, and at test time, aim at predict- ing in the most fine-grained way using the taxonomy. In other words, if an image is predicted as novweel, then they try to assign one of the super classes, implying that the input is in a novel class whose closest known class in the taxonomy is that super class.

To represent the hierarchical relationship, let $\mathcal{T}$ be the taxonomy of known classes, and for a class y, $\mathcal{P}(y)$ be the set of parents, C(y) be the set of children, $\mathcal{A}(y)$ be the set of ancestors including itself, and $\mathcal{N}(y)$ be the set of novel classes whose closest known class is y. And let $\mathcal{L}(y)$ be the set of all descendant leaves under a taxonomy $\mathcal{P}$.

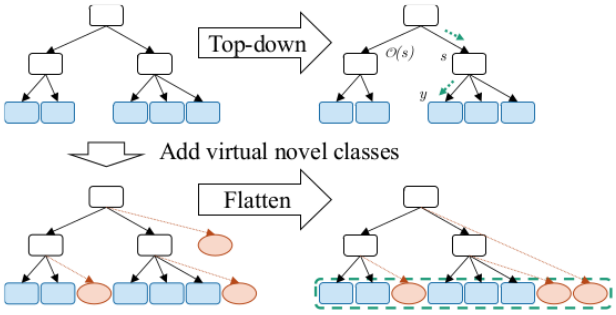

Figure 1. Illustration of two proposed approaches. In the top-down method, classification starts from the root class, and propagates to one of its children until the prediction arrives at a known leaf class (blue) or stops if the prediction is not confident, which means that the prediction is a novel class whose closest super class is the predicted class. In the flatten method, they add a virtual novel class (red) under each super class as a representative of all novel classes, and then flatten the structure for classification.

1.2. Top-down method

A natural way to perform classification using a hierarchical taxonomy is following *top-down* classification decisions starting from the root class, as shown in the top of Fig. 1 Let $(x, y) \sim P_r(x, y|s)$ be a pair of an image and its label sampled from data distribution at a super class s, where $y \in \mathcal{C}(s) \cup \mathcal{N}(s)$. Then, the classification rule is defined as Eq.1:

$$\hat{y} = \begin{cases} \arg\max_{y'} P_r(y'|x, s; \theta_s) \\ \mathcal{N}(s) \end{cases} \tag{1}$$

where θ_s and $P_r(\cdot|x, s; \theta_s)$ are the model parameters of $\mathcal{C}(s) \cup \mathcal{N}(s)$ and the posterior categorical distribution for an image x, respectively. They measure the prediction confidence using the KL divergence with respect to the uniform distribution: intuitively, a confidence-calibrated classifier generates near- uniform posterior probability vector if the classifier is not confident about its prediction. Hence, they interpret that the prediction is confident at a super class

s if

$$D_{KL}\left(U\left(\cdot|s\right)||P_r\left(\cdot|x,s;\theta_s\right)\right) \geq \lambda_s \tag{2}$$

where λ_s is a threshold, D KL denotes the **KL** divergence, and $U\left(\cdot|s\right)$ is the uniform distribution when the classification is made under a super class s.

References

[1] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016. 1

[2] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 1

[3] Y. Fu and L. Sigal. Semi-supervised vocabulary-informed learning. In *CVPR*, 2016. 1

[4] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 1

[5] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016. 1