Seeing Many Unseen Labels via Shared Multi-Attention Models

Dat Huynh
Northeastern University
huynh.dat@husky.neu.edu

Ehsan Elhamifar
Northeastern University
eelhami@ccs.neu.edu

Abstract

In this work, we address the challenging problem of learning attention mechanisms for unseen classes without any training samples. We consider multi-label zero-shot learning, where multiple unseen classes could appear in the same image. To tackle the problem, we introduce a novel shared attention framework that shifts the attention from only learning specific characteristics of each seen class to both class-specific and common visual features shared by all seen classes. These common features are the basis for transferring attention to new unseen classes. We also propose two novel losses that guide the attention to focus on both diverse and relevant image regions for prediction. Our method gains 2.2% *and* 0.7% *F1 score at* $K = 5$ *with respect to Fast0Tag on both NUS-WIDE and Open Images.*

1. Introduction

Multi-label learning is an important problem in image understanding since it captures a wide variety of objects appearing in the same image along with their correlation. It has been shown that using attention mechanism, which learns to localize labels under weakly supervision, significantly improves multi-label performance. In this work, we explore the possibility of generalizing attention mechanism to multiple unseen labels in an image. This setting is relevant because of the difficulty of collecting all possible samples during training. Therefore, the ability to extrapolate to novel classes not only significantly reduces annotation but also improve the system robustness when deploying in practice where novel classes could be encountered.

To successfully focus on image regions of unseen classes, we argue that simply fitting attention on them will not be able to effectively attend to unseen classes as there is no intersection between seen and unseen classes. Moreover, the model could overfit to specific visual appearances of seen classes which cannot be transferred to unseen classes. Our key observation is that instead of learning to attend to each seen class, we can train a multi-attention module to look for common characteristics among all seen classes. Although attentions are being shared, they are trained by a classification ranking loss to learn discriminative class-specific features. This results in each attention mainly becomes an expert in a subset of related classes as shown in Figure 2. These common visual features are more likely to be shared among unseen classes, see Figure 1.

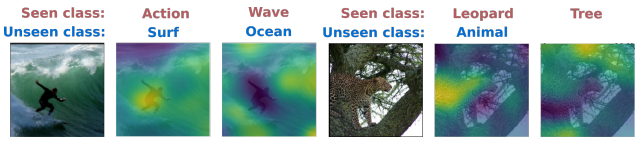

Figure 1: Visualization of our attention modules where seen and unseen classes share same attention features.

To achieve these goals, we propose a novel attention model which produces fewer attention features than the number of classes, hence, forcing classes to share attentions. In return, the attention mechanism would learn generic, transferable as well as discriminative visual appearance across classes. We further propose two novel losses to facilitate the learning of relevant and diverse image regions for recognition.

2. Proposed Framework

Multiple Shared-Attention We design a multi-attention model that generates a few regions of interests compared to the number of classes for prediction. An input image I_i is divided into equal regions in which visual features $\{\boldsymbol{f}_i^r\}_r$ are extracted through a convolutional neural network. Conditioned on these region features, the attention model learns to produce a set of important vectors $\{\boldsymbol{\alpha}_i^k\}_k$ where each vector α_i^k dictates which regions should be focused on. Then the representation for each attention $\boldsymbol{z}_i^k$ is formed by taking the weighted sum of image regions according to $\boldsymbol{\alpha}_i^k$. Given the set of visual feature candidates $\{\boldsymbol{z}_i^k\}_k$, each class selects the most compatible feature which produces the maximal prediction according to the class word2vec semantic representation. Notice that the number of attention features is much smaller than the number of classes. This not only significantly speeds up inference time but also forces the attention to pick up visual features that are common in many classes.

Attention features from diverse regions If kept unchecked, the attention model would produce redundant attention features that all focus on the same image region(s), hence degrading the performance. The model could then confuse existing classes as the same class since all attentions are paid towards a single region containing one class. Therefore, we impose a penalty loss $\mathcal{L}_{div}$ on important weights to distribute the weights across image regions.

Attention feature from relevant regions As the attention model is trained in a weakly supervised setting without any class localization, it is difficult for the model to find regions of each classes. In addition, our model needs to further find common feature across all regions of all classes. To reduce the complexity of the task, we introduce a proxy loss $\mathcal{L}_{rel}$ in addition to the classification loss that prefers image regions where predictions is better than predictions without the attention model.

3. Experimental Results

We evaluate our shared multi-attention for multi-label zero-shot learning on NUS-WIDE [4] and the large-scaled Open Images [5]. Since NUS-WIDE has many corrupted images, we only manage to retrieve 66,207 and 44,340 images for training and testing respectively. Following [3], we use 925 classes collected from Flicker for seen classes and 81 classes carefully annotated for unseen classes. On the other hand, Open Images consists of 9 million training images and 126,436 testing images. We use the 5000 training classes as the seen classes and select 400 disjoint classes from seen classes in test set as unseen classes. For evaluation, we select top K unseen classes prediction in each image as predicted classes. We measure precision, recall and F1 score of the prediction against ground-truth average across unseen classes. For all experiments, we set the effects of diverse loss and relevant loss to $1e-2$ and $1e-3$ respectively compared to classification loss. The number of attention features is fixed to $M = 10$.

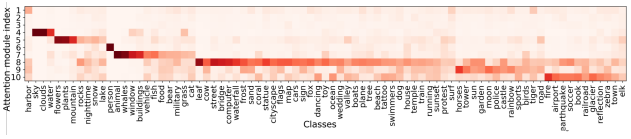

Figure 2: Visualization of the frequency of using attention modules. For each label, we count over all training images the number of times that a prediction is made using each attention module. Each column shows the frequency over each label.

Compared to other methods in Table 1, our method gains 2.2% and 0.7% F1 score at $K = 5$ with respect to Fast0Tag on both NUS-WIDE and Open Images. We also validate the proposed losses by performing ablation study on NUS-WIDE dataset in Table 2 which confirms that having both relevant losses $\mathcal{L}_{rel}$ and diverse loss $\mathcal{L}_{div}$ significantly boost performance.

Method	Dataset	**Zero-shot** multi-label learning $K = 3$			$K = 5$		
		P	R	F1	P	R	F1
CONSE [1]	NW	5.8	16.9	8.6	5.0	21.4	8.1
	OI	0.4	3.7	0.8	0.4	5.7	0.7
LabelEM [2]	NW	12.7	22.8	16.3	11.1	30.3	16.2
	OI	0.3	3.0	0.6	0.3	4.4	0.5
Fast0Tag [3]	NW	15.1	**32.6**	20.7	13.0	**42.6**	19.9
	OI	0.7	5.0	1.2	0.5	7.8	1.0
Our	NW	**19.2**	30.3	**23.5**	**15.4**	39.2	**22.1**
	OI	**0.9**	**6.9**	**1.6**	**0.9**	**9.6**	**1.7**

Table 1: Performance on NUS-WIDE (NW) and Open Images (OI) datasets.

Method	F1 ($K = 3$)	F1 ($K = 5$)
Multi-attention only	16.1	16.4
Multi-attention + $\mathcal{L}_{rel}$	20.0	20.0
Multi-attention + $\mathcal{L}_{rel}$ + $\mathcal{L}_{div}$	**23.5**	**22.1**

Table 2: Ablation study on NUS-WIDE.

Qualitative result: Finally, we quantitatively show that each attention module learns to pick up common visual appearance among related classes in Figure 2. We observe that semantically similar classes such as cloud, sky, water or flower, plant, mountain frequently choose the same attention module for prediction. This indicates these attention modules have learned to detect common visual features among these classes.

References

[1] M. Norouzi et al., "Zero-shot learning by convex combination of semantic embeddings," *ICLR*, 2014. 2

[2] Z. Akata et al., "Label-embedding for image classification," *IEEE TPAMI*, 2016. 2

[3] Y. Zhang et al., "Fast zero-shot image tagging," *2016 CVPR*, pp. 5985–5994, 2016. 2

[4] T. S. Chua et al., "Nus-wide: A real-world web image database from national university of singapore," *ACM CIVR*, 2009. 2

[5] I. Krasin et al., "Open images: A public dataset for large-scale multi-label and multi-class image classification," *available from https://github.com/openimages*, 2016. 2