A Shared Multi-Attention Framework for Multi-Label Zero-Shot Learning

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

Ehsan Elhamifar
Northeastern University
eelhami@ccs.neu.edu

Abstract

In this work, we develop a shared multi-attention model for multi-label zero-shot learning. We argue that designing attention mechanism for recognizing multiple seen and unseen labels in an image is a non-trivial task as there is no training signal to localize unseen labels and an image only contains a few present labels that need attentions out of thousands of possible labels. Therefore, instead of generating attentions for unseen labels which have unknown behaviors and could focus on irrelevant regions due to the lack of any training sample, we let the unseen labels select among a set of shared attentions which are trained to be label-agnostic and to focus on only relevant/foreground regions through our novel loss. Finally, we learn a compatibility function to distinguish labels based on the selected attention. We further propose a novel loss function that consists of three components guiding the attention to focus on diverse and relevant image regions while utilizing all attention features. We show that our method improves the state of the art by 2.9% and 1.4% F1 score on the NUS-WIDE and the large scale Open Images datasets, respectively.

This is a short version of a full paper that is accepted for presentation in CVPR 2020.

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.

In this paper, we develop a framework for multi-label zero-shot learning based on a novel shared multi-attention mechanism that handles recognition of a large number of labels, can recognize multiple unseen labels in an image and finds relevant regions to each label. Our method consists of multiple label-agnostic attention modules that generate multiple attention features simultaneously and uses the semantic vector of each label to select the most suitable feature to compute the prediction score of the label. Thus, instead of generating one attention feature for all labels, which cannot encode discriminative information about labels, and instead of generating one attention feature per label, which cannot generalize well to unseen labels, we generate multiple shared attention features that capture both common and discriminative information about labels, hence not only do well for the prediction of seen labels, but also transfer attention to unseen labels as in Figure 1.

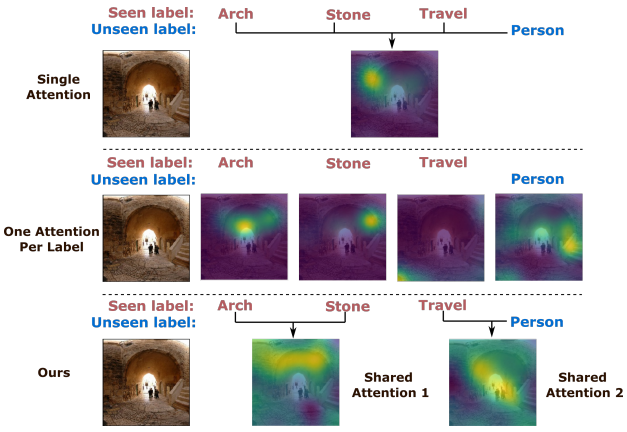

Figure 1: Visualization of attentions learned by single attention for all labels, one attention per label and our shared multi-attention model. Our shared multi-attention model successfully attends to relevant image regions for unseen labels while producing only a few number of attentions that significantly improves the memory and computational complexity for predicting thousands of labels.

Given that each training image only contains the list of present labels without ground-truth bounding box information, to effectively train our shared multi-attention method, we propose a novel loss function that enforces i) different attention modules to focus on diverse regions of an image,

covering different labels; ii) to find relevant regions that would lead to high prediction scores for present labels; iii) to effectively use all attention modules.

2. Multi-Label Zero-Shot Learning via Attention Sharing

Let $\mathcal{C}_s, \mathcal{C}_u$ be the set of seen and unseen classes respectively and $|\cdot|$ denote the cardinality of the set. We propose a shared multi-attention mechanism that consists of $M \ll |\mathcal{C}_s \cup \mathcal{C}_s|$ attention modules generating M attention features, where each feature will be used for the prediction of a subset of related labels, which are determined automatically. We also propose an efficient learning scheme that uses label semantic vectors and training images containing seen labels without access to their ground-truth localization. For an image i, where we divide the image into R equal regions and extract its region features $\{\boldsymbol{f}_i^r\}_{r=1}^R$, let $\{\boldsymbol{z}_i^m\}_{m=1}^M$ denote M attention features obtained via the attention modules $\{g_m(\cdot)\}_{m=1}^M$. We define

$$\boldsymbol{F}_i \triangleq \left[\boldsymbol{f}_i^1 \dots \boldsymbol{f}_i^R\right], \quad \boldsymbol{\alpha}^m(\boldsymbol{F}_i) \triangleq \left[\alpha_1^m(\boldsymbol{f}_i^1) \dots \alpha_R^m(\boldsymbol{f}_i^R)\right]^\top, \tag{1}$$

where $\boldsymbol{F}_i$ denotes a matrix whose columns are R region features and $\boldsymbol{\alpha}^m(\boldsymbol{F}_i)$ denotes the R-dimensional weight vector of the attention module m, for the image i. we formulate the m-th attention feature of the image i, denoted by $\boldsymbol{z}_i^m$, as a linear combination of all region features as

$$\boldsymbol{z}_i^m = \boldsymbol{F}_i\, \boldsymbol{\alpha}^m(\boldsymbol{F}_i). \tag{2}$$

To learn and infer $\boldsymbol{\alpha}^m(\boldsymbol{F}_i)$, we use a simple two-layer neural network model

$$\boldsymbol{\alpha}^m(\boldsymbol{F}_i) = \frac{\exp(\boldsymbol{e}_i^m)}{\sum_{r=1}^R \exp(e_{i,r}^m)},\ \boldsymbol{e}_i^m = \tanh\left(\boldsymbol{F}_i^\top \boldsymbol{W}_1^m\right)\boldsymbol{w}_2^m, \tag{3}$$

where $\{\boldsymbol{W}_1^m, \boldsymbol{w}_2^m\}_{m=1}^M$ are the model parameters, $\tanh(\cdot)$ is the element-wise hyperbolic tangent function, $\boldsymbol{\alpha}^m(\boldsymbol{F}_i)$ is the softmax normalization on each elements $e_{i,r}^m$ of the R-dimensional unnormalized attention weights, $\boldsymbol{e}_i^m$, (i.e., before applying softmax) from the attention module m. Given M attention features $\{\boldsymbol{z}_i^m\}_{m=1}^M$, we propose a model in which the score of each label $c \in \mathcal{C}_s \cup \mathcal{C}_u$ is obtained by the maximum response of the classifier $\boldsymbol{\theta}^c$ (which is defined latter in Eq 10) over the M attention features, i.e.,

$$s_i^c \triangleq \max_{m=1,\dots,M} \langle \boldsymbol{\theta}^c, \boldsymbol{z}_i^m \rangle. \tag{4}$$

Thus, different attention features can be used for the prediction of different labels. To learn the parameters of the M attention modules, we propose an efficient learning scheme with a novel loss function, which we discuss next.

Diverse Multi-Attention Features: We ideally want different attention modules to attend to different regions of an image. Thus, we define a diversity loss that promotes obtaining diverse attention features for an image. More specifically, using the cosine similarity between distinct pairs of unnormalized attention weight vectors, we define

$$\mathcal{L}_{div} \triangleq \sum_i \sum_{m \neq n} \frac{\langle \boldsymbol{e}_i^m, \boldsymbol{e}_i^n \rangle}{\|\boldsymbol{e}_i^m\|_2 \|\boldsymbol{e}_i^n\|_2}, \tag{5}$$

whose minimization promotes small or no overlap in the focus regions of different attention modules. For efficient learning, we use unnormalized attention weights $\boldsymbol{e}$ instead of normalized weights $\boldsymbol{\alpha}$, since the gradient of $\boldsymbol{\alpha}$ vanishes when softmax function saturates. Also, we do not minimize the inner product $\langle \boldsymbol{e}_i^m, \boldsymbol{e}_i^n \rangle$, since it reduces not only the cosine similarity but also the ℓ_2-norm of each weights vector, which prevents the weights of an attention module to concentrate on a single region.

Relevant Multi-Attention Features: Given that the training data does not include information about locations of labels in images, we cannot learn attention models by enforcing that attention weights on ground-truth regions be larger than weights on irrelevant regions. Here, we are only given the set of existing labels in each image. To tackle the problem, we use the prediction scores as surrogates for relevant regions to attend.

Our key observation is that when a seen label $o \in \mathcal{C}_s$ is present in an image, there must be a region containing o on which we have a high score for the label o. Thus, when successfully focused on the region of a label, the score of our multi-attention mechanism must be larger than simply weighting all regions equally. More specifically, let $\bar{s}_i^o \triangleq \frac{1}{R}\sum_r \langle \boldsymbol{\theta}^o, \boldsymbol{f}_i^r \rangle$ be the average score of the label o across all regions, i.e., the score when all regions contribute equally. We define a region relevance loss function that promotes our multi-attention mechanism produce higher scores than $\bar{s}_i^o$ for present labels and lower scores for absent labels. In other words, we define

$$\mathcal{L}_{rel} \triangleq \sum_i \sum_{o \in \mathcal{C}_s} \max\Big(\big(\bar{s}_i^o - s_i^o\big) y_i^o, 0\Big), \tag{6}$$

where $y_i^o \triangleq 1$ for $o \in \mathcal{Y}_i$ and $y_i^o \triangleq -1$ otherwise. Notice that with the above loss, attention modules find not only regions of present labels, but also indicative regions of absent labels, e.g., to predict the absence of the label 'desert', the attention may focus on a region with the label 'ocean'.

Using All Multi-Attention Modules: Given the ability to select among M different attention features in (4) and the non-convexity of learning, the model could potentially learn to use only some attention modules for prediction of all labels and not use the rest. Thus, we propose a loss function to encourage that each of the M attention modules will be used for the prediction of some of the seen labels. We start by defining a score ℓ_m that measures the utility of the m-th

attention module by computing the number of labels across training images that use the attention module m,

$$\ell_m \triangleq \sum_i \sum_{o \in \mathcal{Y}_i} \mathrm{I}_m\big(\operatorname{argmax}_n \langle \boldsymbol{\theta}^o, \boldsymbol{z}_i^n \rangle\big), \tag{7}$$

where $\mathrm{I}_m(x)$ is the indicator function, which outputs 1 when $x = m$ and 0 otherwise. Notice that the term inside the first sum in (7) corresponds to the number of labels of the image i that use the attention model m, hence, ℓ_m measures the utility of the attention module m across all training images. Ideally, we want every attention modules to be used for predictions, hence, we want to avoid having a few large ℓ_m's while most being zero. Thus, we propose to minimize the attention distribution loss,

$$\mathcal{L}_{dist} \triangleq \sum_{m=1}^{M} \ell_m^2. \tag{8}$$

The difficulty of minimizing $\mathcal{L}_{dist}$ is that the ℓ_m defined in (7) is non-differentiable, due to the indicator function. We tackle this by using a softmax function instead, where

$$\ell_m \triangleq \sum_i \sum_{o \in \mathcal{Y}_i} \frac{\exp\big(\langle \boldsymbol{\theta}^o, \boldsymbol{z}_i^m \rangle\big)}{\sum_{n=1}^{M} \exp\big(\langle \boldsymbol{\theta}^o, \boldsymbol{z}_i^n \rangle\big)}. \tag{9}$$

Notice that softmax function approximates the indicator of argmax, with the two coinciding when the magnitude of $\langle \boldsymbol{\theta}^o, \boldsymbol{z}_i^m \rangle$ is significantly larger than other $\langle \boldsymbol{\theta}^o, \boldsymbol{z}_i^n \rangle$.

Bilinear Compatibility Function: Given that we do not have training images for $\mathcal{C}_u$, we cannot directly optimize over and learn $\boldsymbol{\theta}^u$ for $u \in \mathcal{C}_u$. Thus, we use the pre-trained GloVe representation of label names as semantic vectors $\{\boldsymbol{v}^c\}_{c \in \mathcal{C}}$ of labels, allowing to transfer knowledge from seen to unseen labels. More specifically, we express the parameters of each classifier as a function of its semantic vector $\boldsymbol{\theta}^c = \boldsymbol{W}_3 \boldsymbol{v}^c$ and substituting in (4), compute the compatibility score of each label $c \in \mathcal{C}$ in an image i as

$$s_i^c = \max_{m=1,\dots,M} \langle \boldsymbol{W}_3 \boldsymbol{v}^c, \boldsymbol{z}_i^m \rangle. \tag{10}$$

Once we learn $\boldsymbol{W}_3$, as discussed below, we can determine the labels in an image i by ranking and picking the top prediction scores $\{s_i^c\}_{c \in \mathcal{C}}$ across all labels.

To learn the parameters of the compatibility function, $\boldsymbol{W}_3$, and the attention models, we use the ranking loss that imposes the scores of present labels in each image be larger by a margin than the scores of absent labels. More specifically, we define the ranking loss as

$$\mathcal{L}_{rank} \triangleq \sum_i \sum_{o \in \mathcal{Y}_i, o' \notin \mathcal{Y}_i} \max(1 + s_i^{o'} - s_i^o, 0), \tag{11}$$

in which the margin is set to one.

Final Loss Function: Putting all loss functions, discussed above, together we propose to minimize

$$\min_{\Theta, \{\boldsymbol{W}_1^m, \boldsymbol{w}_2^m\}_m, \boldsymbol{W}_3} \mathcal{L}_{rank} + \lambda_{div}\mathcal{L}_{div} + \lambda_{rel}\mathcal{L}_{rel} + \lambda_{dist}\mathcal{L}_{dist}, \tag{12}$$

where $\lambda_{div}, \lambda_{rel}, \lambda_{dist} \geq 0$ are regularization parameters. We minimize this loss function using stochastic gradient descent. In the experiments, we investigate the effectiveness of each loss function term.

3. Experiments

3.1. Experimental Setup

Datasets: We perform experiments on the NUS-WIDE [5] and the Open Images [6] datasets. In the NUS-WIDE, we use the 925 labels as seen and the other 81 labels as unseen as in [3] and follow the original training and testing splits in [5]. To demonstrate the effectiveness of our method on a larger number of labels and images, we use the large-scale Open Images (v4) dataset, which consists of 9 millions training images and $125,436$ testing images. We use the 7186 training classes as the seen classes and select 400 disjoint classes from seen classes in test set as unseen classes.

Evaluation Metrics: We use the mean Average Precision (mAP) and F1 score at top K predictions in each image. Notice that the mAP score captures how accurate the model ranks images for each *label*, while the $F1$ score measures how accurate the model ranks present labels in each *image*.

Implementation Details: Similar to other works [3], we use VGG-19 to extract the feature map of the last convolutional layer ($14 \times 14 \times 512$) as the region features. We implement all methods in Tensorflow and optimize with the default setting of RMSprop with the learning rate 0.001 and batch size of 32. We do not perform heavy hyperparameter tuning and set $(\lambda_{div}, \lambda_{rel}, \lambda_{dist})$ to $(1e^{-2}, 1e^{-3}, 1e^{-1})$ for both datasets. We set the number of attention modules to $M = 10$, unless stated otherwise. We refer to our method as LEarning by Sharing Attentions (LESA).

3.2. Experimental Results

Multi-Label Zero-Shot Learning: We consider both multi-label zero-shot learning, where models are trained on seen labels and tested only on unseen labels, and multi-label generalized zero-shot learning, where models are tested on both seen and unseen labels. Table 1 shows the mAP score and F1 score at $K \in \{3, 5\}$ for NUS-WIDE and at $K \in \{10, 20\}$ for Open Images. We use a larger K for Open Images, since models need to make a larger number of predictions due to a much larger number of labels. From the results, we make the following observations:

– Our method outperforms the state of the art on both datasets, improving the mAP score on NUS-WIDE by 4.3%

Method	Task	NUS-WIDE (#seen / #unseen = 925 / 81)							Open Images (#seen / #unseen = 7186 / 400)						
		$K = 3$			$K = 5$			mAP	$K = 10$			$K = 20$			mAP
		P	R	F1	P	R	F1		P	R	F1	P	R	F1	
CONSE [1]	ZS	17.5	28.0	21.6	13.9	37.0	20.2	9.4	0.2	7.3	0.4	0.2	11.3	0.3	40.4
	GZS	11.5	5.1	7.0	9.6	7.1	8.1	2.1	2.4	2.8	2.6	1.7	3.9	2.4	43.5
LabelEM [2]	ZS	15.6	25.0	19.2	13.4	35.7	19.5	7.1	0.2	8.7	0.5	0.2	15.8	0.4	40.5
	GZS	15.5	6.8	9.5	13.4	9.8	11.3	2.2	4.8	5.6	5.2	3.7	8.5	5.1	45.2
Fast0Tag [3]	ZS	22.6	36.2	27.8	18.2	48.4	26.4	15.1	0.3	12.6	0.7	0.3	21.3	0.6	41.2
	GZS	18.8	8.3	11.5	15.9	11.7	13.5	3.7	14.8	17.3	16.0	9.3	21.5	12.9	45.2
One Attention per Label [4]	ZS	20.9	33.5	25.8	16.2	43.2	23.6	10.4	-	-	-	-	-	-	-
	GZS	17.9	7.9	10.9	15.6	11.5	13.2	3.7	-	-	-	-	-	-	-
LESA ($M = 1$)	ZS	24.3	38.8	29.8	18.9	50.3	27.5	17.6	0.6	23.2	1.2	**0.5**	35.3	1.0	41.2
	GZS	22.6	10.0	13.8	19.1	14.0	16.2	5.1	15.2	17.7	16.4	9.6	22.3	13.4	45.3
LESA ($M = 10$)	ZS	**25.7**	**41.1**	**31.6**	**19.7**	**52.5**	**28.7**	**19.4**	**0.7**	**25.6**	**1.4**	**0.5**	**37.4**	**1.0**	**41.7**
	GZS	**23.6**	**10.4**	**14.4**	**19.8**	**14.6**	**16.8**	**5.6**	**16.2**	**18.9**	**17.4**	**10.2**	**23.9**	**14.3**	**45.4**

Table 1: Multi-Label Zero-Shot (ZS) and Multi-Label Generalized Zero-Shot (GZS) performance on NUS-WIDE and Open Images.

Method	Task	F1 $K = 3$	F1 $K = 5$	mAP
$\mathcal{L}_{rank}$	ZS	28.3	26.1	13.5
	GZS	12.4	14.8	3.8
$\mathcal{L}_{rank} + \mathcal{L}_{rel}$	ZS	31.0	28.1	16.9
	GZS	**14.5**	**16.8**	5.3
$\mathcal{L}_{rank} + \mathcal{L}_{rel} + \mathcal{L}_{div}$	ZS	31.3	28.6	18.0
	GZS	14.4	**16.8**	5.0
$\mathcal{L}_{rank} + \mathcal{L}_{rel} + \mathcal{L}_{div} + \mathcal{L}_{dist}$	ZS	**31.6**	**28.7**	**19.4**
	GZS	14.4	**16.8**	**5.6**

Table 2: Ablation study for multi-label zero-shot (ZS) and multi-label generalized zero-shot (GZS) performance on NUS-WIDE.

for zero-shot and by 2.9% on generalized zero-shot learning. On Open Images, our method improves F1@10 by 0.7% for zero-shot learning and by 1.4% for generalized zero-shot learning, similarly for F1@20, we obtain 0.4% and 1.4% improvement, respectively.

– Learning one attention module per label cannot scale to thousands of labels as in Open Images, due to significantly large memory requirement, hence, we do not report it. Moreover, on NUS-WIDE, it does not do as well as our method or Fast0Tag , due to its class myopic nature and lack of ability to capture shared characteristics of different labels to transfer to unsee ones.

– The F1 score of all methods on Open Images is much smaller than on NUS-WIDE. This comes from the fact that Open Images has significantly larger number of labels, hence, ranking the right labels within an image becomes more challenging, which results in the F1 score drop. On the other hand, the mAP scores of all methods is larger on Open Images than NUS-WIDE. This is because Open Images has more number of positive samples per label, hence, the model has higher change of retrieving relevant images.

Figure 2 (top) shows the frequency of using each attention module for each of the 81 unseen labels in the NUS-WIDE. Notice that our method learns one main attention module (attention module 6) to predict most unseen labels and depending on the complexity of each label, it would use more attention modules, if needed. In particular, simple labels such as 'window' and 'road' use only one attention module, 'flowers' and 'tree' use two attention modules, while more visually varying labels such as 'cityscape' and

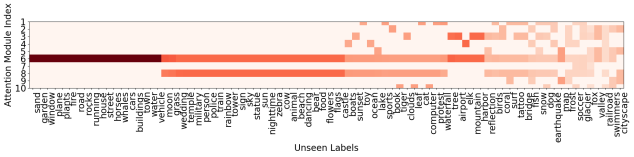

Figure 2: Top: Visualization of the frequency of using attention modules. For each label, we count over all training images the number of times that a prediction is made using each attention module. Each column shows the frequency over each label.

'coral' use multiple attentions. This is a unique property of our framework that dynamically allocates the right number of attention modules to labels and allows different labels to be predicted by different modules, if needed, and the quantitative results in Table 1 verify its importance.

Ablation Studies: Table 2 shows the F1 and mAP scores for our method on the NUS-WIDE for multi-label zero-shot and multi-label generalized zero-shot learning by using different components of our proposed loss function. Notice that only using $\mathcal{L}_{rank}$ as in standard zero-shot learning does perform worst. We obtain 2.7% (3.4%) improvement in F1@3 (mAP) scores when using $\mathcal{L}_{rel}$, which promotes to select relevant regions to labels in images. Enforcing attention diversity further improves F1@3 (mAP) by 0.3% (1.1%). Finally, adding the distribution loss $\mathcal{L}_{dist}$ obtains the best result with 31.6% F1@3 and 19.4% mAP score.

References

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

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

[3] Y. Zhang et al., "Fast zero-shot image tagging," *CVPR*, 2016.

[4] J. w. Kim et al., "Bilinear attention networks," *neurIPS*, 2018.

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

[6] 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.