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. By extensive experiments, 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.

1. Introduction

Recognition of all labels in an image, referred to as multi-label recognition, is a fundamental problem in computer vision with applications in self-driving cars, surveillance systems and assistive robots, among others. To successfully support real-world tasks, multi-label recognition systems must accurately learn tens of thousands of labels, handle unseen labels and localize them in images. Despite advances, in particular, using deep neural networks, [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], to date, there is no multi-label learning algorithm that can achieve all these goals. This paper takes steps towards addressing large-scale multi-label zero-shot learning and localization.

The majority of existing work on multi-label learning have focused on exploiting dependencies among labels to improve the recognition performance of methods that learn

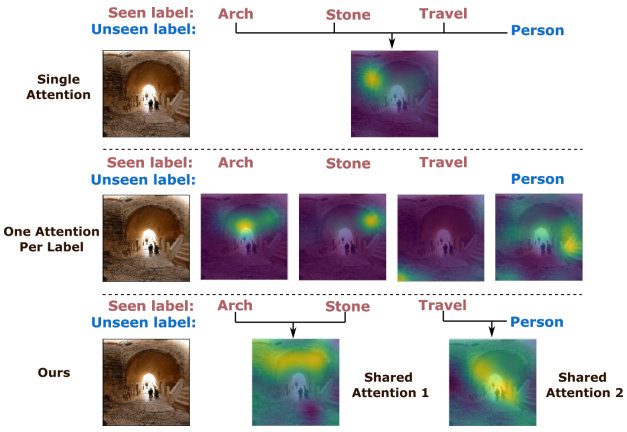

Figure 1: Visualization of attentions learned by a single attention for all labels, one attention per label and our shared multi-attention model. Our method successfully attends to relevant image regions for both seen and unseen labels while producing only a few number of attentions that significantly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

a separate classifier for each label [1, 11, 3, 12, 13, 14, 3, 8, 15]. However, they cannot handle the classification of (multiple) unseen labels in an image and cannot localize labels. A few recent work have incorporated attention mechanism into multi-label learning to focus on relevant image regions [16, 17, 18, 10], yet, they lack the ability to handle unseen labels. Moreover, the recurrent neural network employed in [16, 18], which has to sequentially compute the attention regions for the subsequent label to be predicted, imposes large training and inference time and limits the scalability to classify a large number of labels in an image.

On the other hand, a large body of work have focused on zero-shot learning with the goal of recognizing unseen labels [19, 20, 21, 7, 9, 22, 23, 24], with some of the recent work taking advantage of attention models [25, 26, 27] to improve the prediction accuracy. However, these methods address the multi-class zero-shot learning, where each image is assumed to have one label, hence cannot handle the multi-label setting, where an image contains several labels, some of which could be unseen. Moreover, as observed by [28, 29, 30], using a single feature vector to encode discrim-

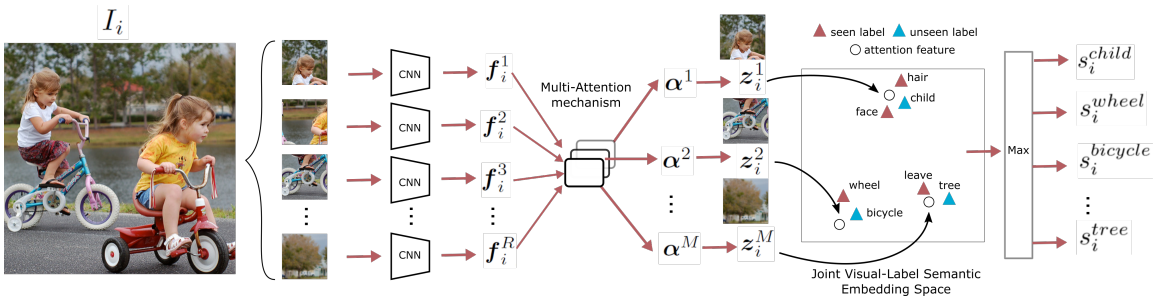

Figure 2: The overview of our shared multi-attention zero-shot learning. Image features of R regions are extracted and fed into our shared multi-attention mechanism to compute multiple attention features. The attention features are projected into the joint visual-label semantic embedding space to determine their labels.

inative information about all labels is restrictive, especially when dealing with a large number of labels.

A few work have addressed the problem of multi-label zero-shot learning [7, 9, 31] by taking advantage of the correlation between unseen and seen labels which is inferred from a global image representation. However, they only capture dominant labels and ignore the ones in smaller regions of images. To overcome this issue, [28, 29, 30] use pre-trained object detection modules and learn to select bounding boxes of seen and unseen labels. However, this approach is costly and not scalable to a large number of labels as it requires ground-truth bounding boxes for training. Moreover, it cannot handle abstract concepts, e.g., 'travel' or 'singing', which often do not have a clear bounding box. On the other hand, one can naively generalize attention techniques to the multi-label zero-shot setting by computing one attention per label [32]. However, this not only is computationally and memory expensive, but more importantly is prone to overfitting, due to a small number of training images for each label.

Paper Contributions. 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, see Figure 2. 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.

Our method automatically discovers related labels and assigns them to the same attention module. Moreover, it dynamically allocates an appropriate number of attention modules to each label depending on its complexity. By eliminating any recurrent attention structure and using a small number of attention modules compared to the large number of labels, our method significantly reduces the time and memory complexity of computing one attention per label or of recurrent attention mechanisms.

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. We conduct experiments on both multi-label zero-shot and generalize zero-shot learning on the NUS-Wide and the large-scale Open Images datasets, showing the effectiveness of our method, which improves the F1 score of the state of the art by 2.9% and 1.4%, respectively.

2. Related Work

Multi-label learning can be naively addressed by learning a binary classifier for each label [33, 34], which does not incorporate correlations among labels. Thus, the majority of multi-label learning methods have focused on incorporating label dependencies [11, 2, 15, 8, 35, 36]. However, some methods require training data with the full annotation of images [10, 18], some cannot generalize to unseen labels [11, 2, 15, 8, 35, 36, 33, 34], and some work with global feature representation of images, which is restrictive when dealing with a large number of labels, and cannot find regions of labels [35, 37, 38].

To localize labels, [39, 40, 28, 29, 30] find region proposals followed by applying CNN-based recognition on each proposal. This can recognize few labels for foreground objects (not concepts, e.g., 'travel' or 'singing') and requires costly bounding box annotations. On the other hand, attention modeling [10, 41, 42, 43, 32] has provided powerful tools to address the localization of labels by learning to fo-

cus on relevant parts of images. However, most existing methods cannot generalize to unseen labels [42, 10]. While image captioning [44, 45, 46] can be thought of as multi-label learning (labels are words in the generated caption), it requires training and predicted labels with a sequential order. While [3, 47, 48] have proposed methods to find semantic orders of labels, their sequential nature does not allow fast training or inference (e.g., via parallelization) and they cannot localize labels or generalize to unseen labels.

Zero-shot learning, on the other hand, addresses the problem of generalizing learning to unseen labels [49, 50, 5, 51, 52, 53, 54, 55]. This often requires using semantic information from seen and unseen labels in the form of attribute vectors [56, 57, 58] or word vector representations [51, 59, 57, 20, 60, 52]. The semantic label vectors are often combined with the image features via learning a compatibility score between the two, which then allows to classify unseen labels [20, 60, 50]. Having shown great success, the majority of zero-shot learning methods find only the dominant label in each image [5, 51, 20, 21, 25, 26, 27] and rely on using a global feature without localizing labels. Recently, both single attention [25] and double-attention [26] mechanisms have been employed for single class zero-shot learning. However, these works learn a single representation to predict all classes and cannot recognize diverse labels in an image.

The recent works in [7, 61, 62] address the problem of zero-shot multi-label learning by finding the joint embedding space of image and labels while optimizing the standard zero-shot ranking loss modified for multi-label learning. These works, however, do not localize labels and neglect the importance of discriminative features from local image regions by using global features. Moreover, [9] requires access to a knowledge graph between seen and unseen labels. [63, 28, 30] use multiple features generated by an object proposal algorithm for zero-shot prediction. However, the proposal is designed for objects and cannot generalize to abstract concepts. Multi-modal attention [32] can be used to generate specific attention for each label and generalize to unseen labels through label semantics. However, this has large time and memory complexity when computing thousands of attention for thousands of classes. Moreover, the extrapolation of seen to unseen attentions often focuses on irrelevant regions as there is no supervision on unseen attention (see the experiments).

Finally, our proposed method is different from [64, 27], which have proposed multi-attention models without attention sharing mechanism, thus can not effectively generalize to unseen labels. Moreover, they fuse all attention features into a single global feature which discards discriminative information obtained by each attention model.

3. Visual Attention Review

Visual attention generates a feature from the most relevant region of an image and has been shown to be effective for image classification, saliency detection and captioning, among others [42, 18, 10, 65, 44, 66]. More specifically, one divides an image I into R regions denoted by $I^1, \dots, I^R$, which can be arbitrary [41] or equal-size grid cells [44]. For simplicity and reproducibility, we use the latter approach. Let $\boldsymbol{f}^r = f_\Theta(I^r)$ denote the feature vector of the region r, extracted using a CNN parametrized by Θ. Given region features $\{\boldsymbol{f}^r\}_{r=1}^R$, the goal of the attention module, $g(\cdot)$, is to find the most relevant regions for the task. This is done by finding an attention feature, $\boldsymbol{z}$, defined as

$$\boldsymbol{z} = g\big(\boldsymbol{f}^1, \dots, \boldsymbol{f}^R\big) = \sum_{r=1}^{R} \alpha_r(\boldsymbol{f}^r)\boldsymbol{f}^r, \tag{1}$$

where $\alpha_r(\boldsymbol{f}^r)$ denotes the weight or preference of selecting the region r. These weights are unknown and the task of the attention module is to find them for an input image. In the soft-attention mechanism [44], which we use in the paper, one assumes that $\alpha_r \in [0, 1]$ and $\sum_{r=1}^{R} \alpha_r = 1$ to select different regions with different degrees of importance. The attention weights are often modeled by the output of a neural network, normalized using the softmax function.

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

In this section, we discuss our proposed framework for multi-label zero-shot learning. We first define the problem settings and then present our approach based on a shared multi-attention mechanism.

4.1. Problem Setting

Assume we have two sets of labels $\mathcal{C}_s$ and $\mathcal{C}_u$, where $\mathcal{C}_s$ denotes seen labels that have training images and $\mathcal{C}_u$ denotes unseen labels without training annotations. We denote the set of all labels by $\mathcal{C} \triangleq \mathcal{C}_s \cup \mathcal{C}_u$. Let $(I_1, \mathcal{Y}_1), \dots, (I_N, \mathcal{Y}_N)$ be N training samples, where I_i denotes the i-th training image and $\mathcal{Y}_i \subseteq \mathcal{C}_s$ denotes the set of labels present in the image. The goal of multi-label zero-shot learning is to find the labels in $\mathcal{C}$ that appear in a new test image. Given that there are no training images for the unseen labels, $\mathcal{C}_u$, similar to exiting work on zero-shot learning [67, 57, 20, 59], we assume access to semantic vectors $\{\boldsymbol{v}^c\}_{c\in\mathcal{C}}$ that provide descriptions of labels, e.g., using attributes or word embeddings [67, 57, 20, 59, 4, 50].

Naive Approach. To address the problem using the attention mechanism, we can consider two naive extreme cases. First, we can generate one attention feature $\boldsymbol{z}_i^c$ for each label $c \in \mathcal{C}$ in an image i. Thus, the prediction score for the label c in image i can be computed as

$$s_i^c = \langle \boldsymbol{\theta}^c, \boldsymbol{z}_i^c \rangle, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product and $\boldsymbol{\theta}^c$ denotes the parameters of the logistic classifier for the label c (the exact form of $\boldsymbol{\theta}^c$ will be defined later, see (12)). We can then determine the labels in the image i by ranking and picking the top prediction scores $\{s_i^c\}_{c \in \mathcal{C}}$ across all labels. This has two major drawbacks. First, it is not clear to how learn an attention model for an unseen label that has no training images. While we can extend and employ methods such as [32] by using the label semantic vector to generate an attention feature, learning would be prone to overfitting as a large number of attention models have to be learned with no or few training images, hence, will often focus on irrelevant regions (see Figure 1). Second, training and computing a separate attention feature for each label is computationally and memory expensive, especially when dealing with thousands of labels, e.g., in the Open Images dataset. However, for seen labels with sufficient number of training images, this approach allows to learn informative features that are able to focus on label-relevant regions of an image.

In the second approach, instead of learning label-specific attention features as above, we can compute a single attention feature for all labels. This approach has the benefit that it does not suffer from overfitting and is memory and computationally efficient. However, the model will not have enough capacity to represent and localize a large number of possibly diverse labels (see Figure 1).

4.2. Proposed Shared Multi-Attention Approach

We develop a multi-label zero-shot learning method based on attention mechanism that overcomes the limitations of the two above approaches and enjoys the advantages of both, i.e., learns informative features that focus on label-relevant regions of an image, does not suffer from overfitting and generalizes well to unseen labels, and is computationally and memory efficient. To do so, we propose a shared multi-attention mechanism that consists of $M \ll |\mathcal{C}|$ 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 that contain seen labels without access to their ground-truth localization.

For an image i with 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

$$\begin{aligned} \boldsymbol{F}_i &\triangleq \left[\boldsymbol{f}_i^1 \quad \boldsymbol{f}_i^2 \quad \dots \quad \boldsymbol{f}_i^R\right], \\ \boldsymbol{\alpha}^m(\boldsymbol{F}_i) &\triangleq \left[\alpha_1^m(\boldsymbol{f}_i^1) \quad \dots \quad \alpha_R^m(\boldsymbol{f}_i^R)\right]^\top, \end{aligned} \tag{3}$$

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. Using the model (1), we can write the m-th attention feature of the image i, denoted by $\boldsymbol{z}_i^m$, as a linear combination of all region features as

$$\boldsymbol{z}_i^m = \boldsymbol{F}_i \, \boldsymbol{\alpha}^m(\boldsymbol{F}_i). \tag{4}$$

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

$$\boldsymbol{\alpha}^m(\boldsymbol{F}_i) = \frac{\exp(\boldsymbol{e}_i^m)}{\sum_{r=1}^R \exp(e_{i,r}^m)}, \; \boldsymbol{e}_i^m = \tanh\left(\boldsymbol{F}_i^\top \boldsymbol{W}_1^m\right)\boldsymbol{w}_2^m, \tag{5}$$

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 $c \in \mathcal{C}$ is obtained by the maximum response of the classifier c over the M attention features, i.e.,

$$s_i^c \triangleq \max_{m=1,\dots,M} \langle \boldsymbol{\theta}^c, \boldsymbol{z}_i^m \rangle. \tag{6}$$

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

$$\mathcal{L}_{div} \triangleq \sum_i \sum_{m \neq n} \frac{\langle \boldsymbol{e}_i^m, \boldsymbol{e}_i^n \rangle}{\|\boldsymbol{e}_i^m\|_2 \|\boldsymbol{e}_i^n\|_2}, \tag{7}$$

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 $\langle \boldsymbol{e}_i^m, \boldsymbol{e}_i^n \rangle$, 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. Notice that our diversity loss is less restrictive than [44] as we do not enforce the attention model to attend to *all* regions of an image, instead to attend to only regions that are *diverse* and relevant for prediction.

Relevant Multi-Attention Features: Given that the training data does not include information about locations of labels in images, unlike existing work [25, 29], 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 to produce higher scores than $\bar{s}_i^o$ for present labels and lower scores for absent labels. In other words, we define

$$\mathcal{L}_{rel} \triangleq \sum_i \sum_{o \in \mathcal{C}_s} \max\Big(\big(\bar{s}_i^o - s_i^o\big) y_i^o, 0 \Big), \tag{8}$$

where $y_i^o \triangleq 1$ for $o \in \mathcal{Y}_i$ and $y_i^o \triangleq -1$ otherwise.[1] 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 (6) 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,

$$\ell_m \triangleq \sum_i \sum_{o \in \mathcal{Y}_i} \mathrm{I}_m\big(\operatorname{argmax}_n \langle \boldsymbol{\theta}^o, \boldsymbol{z}_i^n \rangle \big), \tag{9}$$

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 (9) 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 module 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,

$$\mathcal{L}_{dist} \triangleq \sum_{m=1}^{M} \ell_m^2. \tag{10}$$

The difficulty of minimizing $\mathcal{L}_{dist}$ is that the ℓ_m defined in (9) is non-differentiable, due to the indicator function. We tackle this by using a softmax function instead, where

$$\ell_m \triangleq \sum_i \sum_{o \in \mathcal{Y}_i} \frac{\exp\big(\langle \boldsymbol{\theta}^o, \boldsymbol{z}_i^m \rangle\big)}{\sum_{n=1}^{M} \exp\big(\langle \boldsymbol{\theta}^o, \boldsymbol{z}_i^n \rangle\big)}. \tag{11}$$

[1] One can also use a margin in (8). However, in all our experiments, the above loss, which does not have hyperparameters, performed well.

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, similar to previous work on zero-shot learning [57, 59, 51], we use the 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 (6), compute the compatibility score of each label $c \in \mathcal{C}$ in an image i as

$$s_i^c = \max_{m=1,\dots,M} \langle \boldsymbol{W}_3 \boldsymbol{v}^c, \boldsymbol{z}_i^m \rangle. \tag{12}$$

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

$$\mathcal{L}_{rank} \triangleq \sum_i \sum_{o \in \mathcal{Y}_i, o' \notin \mathcal{Y}_i} \max(1 + s_i^{o'} - s_i^o, 0), \tag{13}$$

in which the margin is set to one.

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

$$\min_{\Theta, \{\boldsymbol{W}_1^m, \boldsymbol{w}_2^m\}_m, \boldsymbol{W}_3} \mathcal{L}_{rank} + \lambda_{div}\mathcal{L}_{div} + \lambda_{rel}\mathcal{L}_{rel} + \lambda_{dist}\mathcal{L}_{dist}, \tag{14}$$

where $\lambda_{div}, \lambda_{rel}, \lambda_{dist} \geq 0$ are regularization parameters. We minimize this loss function using stochastic gradient descent (see experiments for details). In the experiments, we investigate the effectiveness of each loss function term and show the robustness of our methods with respect to the values of the regularization parameters.

5. Experiments

We evaluate our proposed shared multi-attention framework for multi-label zero-shot learning on NUS-WIDE [68] and the large-scale Open Images [69] datasets. Below, we discuss the datasets, evaluation metrics, baseline methods then present and analyze the results on both datasets. Given that our method handles multi-label learning, we also report the multi-label learning performance on both datasets in the supplementary material.

5.1. Experimental Setup

Datasets: We perform experiments on the NUS-WIDE [68] and the Open Images [69] datasets. In the NUS-WIDE, each

image has 81 labels, called 'ground-truth' labels, which are carefully labeled by human annotators, in addition to 925 labels extracted from Flicker user tags. Similar to [7], we use the 925 labels as seen and the other 81 labels as unseen. We run all methods on the full dataset that has 20% more training and testing samples than the data used in [7].

To demonstrate the effectiveness of our method on a larger number of labels and images and to investigate the localization performance of our method, we use the large-scale Open Images (v4) dataset, which consists of 9 millions training images in addition to $41,620$ and $125,436$ images for validation and testing, respectively. For the seen labels, we use $7,186$ labels in the training set, where each label has at least 100 training samples. We select 400 most frequent test set labels that are not observed in the training data as the unseen labels. Each unseen label has at least 75 test samples for evaluation. Due to the large number of classes, each image has unannotated labels.

Evaluation Metrics: Similar to other work on multi-label learning [2, 10], for evaluation, we use the mean Average Precision (mAP) [70] and F1 score at top K predictions [7] in each image. The details of computing the scores are provided in the supplementary materials. 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*.

Baselines: We compare with CONSE [59] (ensemble of classifiers), LabelEM [57] (joint image-label embedding) and Fast0Tag, which is a state-of-the-art multi-label zero-shot learning method. We also compare with [32], which uses one attention per label, hence learning a total of $|C|$ attention modules. This allows to investigate the effectiveness of our method that uses a small number of attention modules and share them across labels.

We refer to our method as LEarning by Sharing Attentions (LESA) and train the following variants of our method: i) LESA ($M = 1$), where we use a single attention module learned by the combination of the ranking and relevance losses (since there is one attention, there will be no diversity and distribution losses). This allows to demonstrate the effectiveness of sharing multiple attention modules; ii) LESA with M attention modules, learned via our proposed loss function in (14). In addition, we use the semantic vectors of labels to cluster them into M groups via kmeans and learn an attention module for the labels in each group using the combination of the ranking and relevance losses (referred to as One Attention per Cluster). This allows us to investigate the effectiveness of our multi-attention sharing framework that automatically allocates an appropriate number of attention modules for each label.

Implementation Details: Similar to other works [7], we use a pretrained VGG-19 for feature extraction in all methods. We extract the feature map at the last convolutional layer whose size is $14 \times 14 \times 512$ and treat it as a set of features from 14×14 regions. For our all variants of our method, we freeze the VGG network and learn an additional convolutional layer of size $2 \times 2 \times 512$ on top of the VGG's last convolutional layer. Thus, our convolutional layer has significantly smaller number of parameters than [7], which learns three fully connected layers. We extract the semantic vectors $\{\boldsymbol{v}^c\}_{c \in \mathcal{C}}$ using the GloVe model [71] trained on Wikipedia articles.

We implement all methods in Tensorflow and optimize with the default setting of RMSprop [72] with the learning rate 0.001 and batch size of 32, and use exponential learning rate decay of 0.8 whenever the training model degrades performance on the validation set. We also use early stopping [73] as a form of regularization in all models. We train all models on an NVIDIA V100 GPU for 40 epochs for the NUS-WIDE and 2 epochs for Open Images. In our method, 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. For simplicity, we share the parameters $\boldsymbol{W}_1^m$ across the attention modules.

5.2. 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% 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[2], due to its class myopic nature and lack of ability to capture shared characteristics of different labels to transfer to unsee ones.

– Clustering labels based on semantic vectors and learning

[2]Notice that on NUS-WIDE, Fast0Tag for GZS achieves 3% lower score than in [7]. This is due to the fact that [7] used a subset of images for testing, instead of all images. We contacted authors of [7] who replied they did not record the set of images used for their testing.

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 [59]	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 [57]	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 [7]	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 [32]	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	-	-	-	-	-	-	-
One Attention per Cluster ($M = 10$)	ZS	20.0	31.9	24.6	15.7	41.9	22.9	12.9	0.6	22.9	1.2	0.4	32.4	0.9	40.7
	GZS	10.4	4.6	6.4	9.1	6.7	7.7	2.6	15.7	18.3	16.9	9.6	22.4	13.5	44.9
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.

an attention for each cluster as well as only learning one attention module for all labels do not do as well as our LESA ($M = 10$), which shows the importance of allowing each label to use more than one attention module, and generally a number of attentions depending on the complexity of the label (i.e., visual variations across images).

– 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 4 (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 '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.

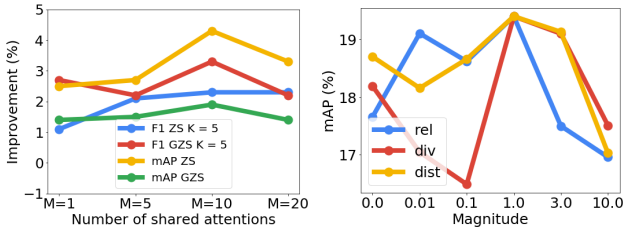

Figure 3: F1/mAP improvement (%) over Fast0Tag for different numbers of attention features (left) and effect of λ_{rel},λ_{div},λ_{dist} on multi-label zero-shot mAP (%) (right) on NUS-WIDE.

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.

Effect of Hyperparameters: Figure 3 (left) shows our model's improvement over Fast0Tag for (generalized) zero-shot learning with different number of attention features on test images in NUS-WIDE. We observe improvement by using shared attention regardless of the number of attention modules (for one attention, we use our new loss $\mathcal{L}_{rank}+\mathcal{L}_{rel}$). Notice in all cases, the performance saturates or peaks at 10 attention features and drops if more attention features are used. This again verifies that large number of attention features could harm by overfitting to seen labels.

Figure 3 (right) shows the effect of hyperparameters for multi-label zero-shot learning on NUS-WIDE test images. We first set $(\lambda_{div}, \lambda_{rel}, \lambda_{dist})$ to $(1e^{-2}, 1e^{-3}, 1e^{-1})$, and fixing two regularizations, change the other one by a magnitude shown on the horizontal axis. Notice that, generally, the score improves as the regularization parameters increase and is stable around the nominal values.

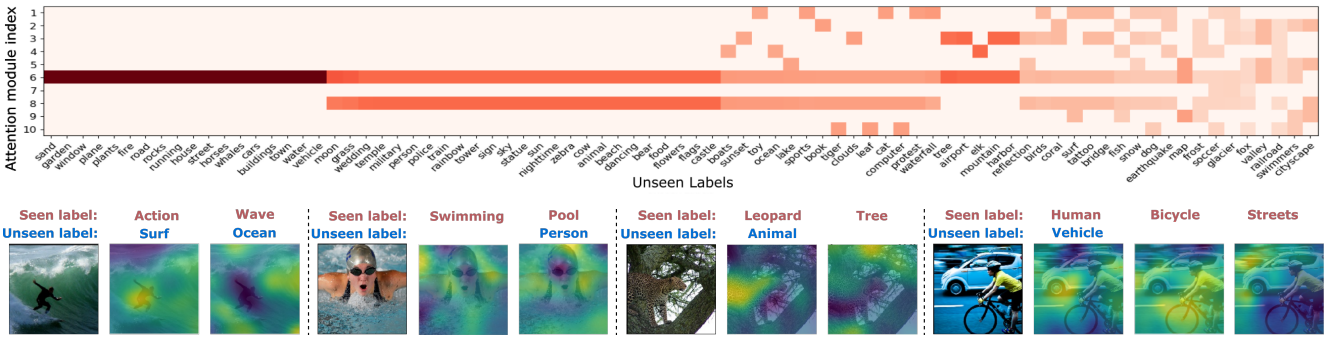

Figure 4: Top: Visualization of the frequency of using attention modules. For each label, we count over all training images the number of times that a prediction is made using each attention module. Each column shows the frequency over each label. Bottom: Visualization of learned attentions for few images from NUS-WIDE.

Method	mAP (seen)	mAP (unseen)	Harmonic mean (seen+unseen)
One attention per label	**10.8**	2.0	3.4
One attention for all labels	8.7	2.4	3.8
Ours	9.4	**2.7**	**4.2**

Table 3: mAP for label localization on the Open Images test set.

Zero-Shot Label Localization: To demonstrate the effectiveness of shared multi-attention method on localization of labels, we measure the mean average precision score of localization. We follow [74], which uses this score to measure localization on Pascal VOC and MSCOCO. Roughly speaking the score captures whether the attention(s) puts maximal weights on the ground-truth bounding box(es) of the label(s) and whether the model is confident about the label(s) (see the supplementary materials for the precise definition). We report the mAP on seen and unseen labels as well as the harmonic mean of seen and unseen predictions to measure the seen/unseen trade off.

In Open Images, out of all trainable labels with least 100 training samples for each, there are 558 labels that have bounding box annotations in the test set. Thus, we divide these labels into 420 seen labels and 138 unseen labels. We train our method, one attention for all labels and one attention per label on 420 seen labels in the training set and evaluate their localization score on the test set.

Table 3 shows the localization score of our method compared with other baselines. Notice that, as expected and discussed earlier, one attention per label does well on seen labels and performs worst on unseen labels, while one attention for all labels does better on generalization to unseen labels, yet performs poorly on seen labels. On the other hand, our shared multi-attention model, which combines the advantages of both, does well on both seen and unseen and achieves the largest overall performance, measured by the harmonic mean. Finally, Figure 5 further shows the localization mAP improvement of our method with respect to one attention per label for 20 unseen labels with the largest improvement and 20 unseen labels with the largest drop. Notice that our method significantly improves (more than

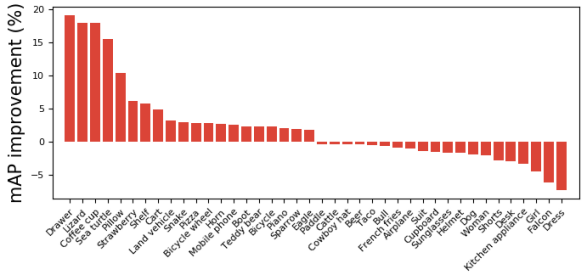

Figure 5: Localization mAP improvement over one attention per label on unseen labels in Open Images test set.

15%) on some unseen labels such as 'Drawer', 'Lizard' and 'Coffee cup' while having negative impact, however much smaller (less than 6%), on labels such as 'Dress' or 'Kitchen appliance', which have wide appearance change, hence better captured by a specialized attention module.

Qualitative results: Figure 4 (bottom) visualizes learned attentions by our method for a few images from NUS-WIDE. Notice that our method learns to successfully focus on both seen and unseen labels, including abstract concepts. For instance, in the first image, the model focuses on the person and the surrounding wave to recognize the seen label 'action', while uses the same attention feature to predict the unseen label 'surf'.

6. Conclusion

We proposed a novel shared multi-attention mechanism which predicts all labels in an image, including multiple unseen ones. We proposed 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. By extensive experiments on NUS-WIDE dataset and the large-scale Open Images dataset, we showed that our framework improves the state of the art.

Acknowledgements

This work is partially supported by DARPA Young Faculty Award (D18AP00050), NSF (IIS-1657197), ONR (N000141812132) and ARO (W911NF1810300).