Semantic Diversity Learning for Zero-Shot Multi-label Classification

Avi Ben-Cohen   Nadav Zamir   Emanuel Ben Baruch   Itamar Friedman   Lihi Zelnik-Manor
DAMO Academy, Alibaba Group

{avi.bencohen, nadav.zamir, emanuel.benbaruch, itamar.friedman, lihi.zelnik}
@alibaba-inc.com

Abstract

Training a neural network model for recognizing multiple labels associated with an image, including identifying unseen labels, is challenging, especially for images that portray numerous semantically diverse labels. As challenging as this task is, it is an essential task to tackle since it represents many real-world cases, such as image retrieval of natural images. We argue that using a single embedding vector to represent an image, as commonly practiced, is not sufficient to rank both relevant seen and unseen labels accurately. This study introduces an end-to-end model training for multi-label zero-shot learning that supports the semantic diversity of the images and labels. We propose to use an embedding matrix having principal embedding vectors trained using a tailored loss function. In addition, during training, we suggest up-weighting in the loss function image samples presenting higher semantic diversity to encourage the diversity of the embedding matrix. Extensive experiments show that our proposed method improves the zero-shot model's quality in tag-based image retrieval achieving SoTA results on several common datasets (NUS-Wide, COCO, Open Images).

Figure 1: Our model extracts a set of principal embedding vectors used as a transformation matrix A_i where each row sets a ranking principal direction for labels in the word vector space based on their relevance. By using multiple directions, it can deal with multiple diverse image semantic concepts. In addition, we propose that images with larger semantic diversity (top images) should be up-weighted during training compared to ones with lower semantic diversity (bottom images).

1. Introduction

Identifying all the relevant labels that describe the objects or scene in an image is an essential task in computer vision real-world applications. With the ongoing increase of photos stored online comes a growing need for better image tagging and tag-based retrieval for various use cases such as search, organization, or data collection. Recent datasets in this field made progress in this direction by including a large number of classes with annotations of their presence in each image. Yet, annotating a large number of classes for many images, each with high semantic diversity, can be very time-consuming and practically infeasible for real-world applications. Although current conventional multi-label classification methods can deal with a large number of classes, they are still limited by the annotated (seen) set of
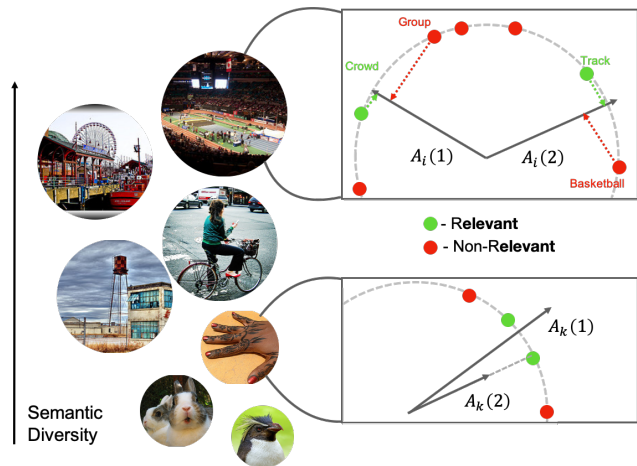
labels provided with the datasets.

On the other hand, Zero-shot (ZS) learning for multi-label classification adds the ability to recognize labels from additional categories that do not exist during training (unseen). This is usually done by transferring knowledge between the seen and unseen labels. In most cases, a text model [5, 32, 36] is used to transfer this knowledge using word vectors. Then, a visual model is trained, learning a transformation between the visual space and the word vector label space.

Most studies on ZS classification focus on the single label problem, i.e., recognizing a single unseen label in each image [1, 11, 14, 25, 29, 34, 50, 57, 53, 51, 52]. However, a single label per image does not provide a full description

of it and usually is not practical for real-world applications. Other studies tried to tackle the ZS multi-label classification problem. [35] trained a multi-label classifier on the seen classes and linearly combined word vectors of these classes using the prediction scores to represent an image. Based on that representative vector, the ranking of unseen labels was done by computing similarities to their word vectors. [58] trained a network to output per-image a single principal direction that ranks relevant labels higher than non-relevant ones. However, this method faces difficulties with images that include multiple labels with high semantic diversity. In these cases, the extracted principal direction needs to be robust to high variability in the relevant labels' word vectors. For example, the classes *"dog"* and *"rice"* are quite different, but might still be present in the same image. As we grow with the number of annotated classes in images, the probable semantic diversity in each image grows. This high semantic diversity problem requires special treatment, which is hard to achieve using a single principal direction.

Several works had approached the problem of high semantic diversity of the labels in an image using a pre-trained object detector and learning to select bounding boxes of seen or unseen labels [3, 37, 38, 39]. Yet, these approaches require annotated bounding boxes as ground truth, making it not scalable for a large number of labels. Alternatively, [21] used attention techniques to estimate the relevant regions based on a pre-trained model's features. However, this usually requires a large pre-trained model to get rich regional features (VGG-19) and a complex loss function to be tuned.

In this paper, we propose a method that aims to properly cope and leverage the semantic diversity of the labels in each image, by allowing multiple principal directions, constructed as a transformation matrix in the loss function. Also, sample images with larger semantic diversity are up-weighted in the loss function as these images are considered hard examples. As a result, our model learns to extract a per-image transformation designed to handle the image label diversity challenges (Figure 1). We believe that by doing so, we learn a model that is better suited for understanding and recognizing multiple seen and unseen labels in an image.

We further show how we achieve results that are on par or better than SotA while keeping a relatively simple end-to-end training scheme using our suggested loss function.

The main contributions presented in this study include:

- A loss function tailored to the problem of ZS multi-label classification.

- We show that up-weighting samples with higher semantic diversity further improves the model generalization.

- An efficient end-to-end training scheme for ZS models

is proposed, reaching SotA results in tag-based image retrieval while keeping high-performance for image tagging on several datasets (NUS-Wide, Open-Images, and MS-COCO) with a smaller number of model parameters compared to other methods.

2. Related Work

Recent studies on multi-label classification reported notable success by exploiting dependencies among labels via graph neural networks to represent label relationships or word embeddings based on prior knowledge [6, 7, 13, 46]. Other approaches try to model the image parts using attentional regions [16, 47, 54, 56]. Although these approaches show promising results, they usually include a complex architecture, and other approaches reported similar and even better results using a more conventional training flow with advanced loss modifications [4]. While most of these approaches are effective for images, including the seen classes they were trained for, they don't generalize well to unseen classes.

The main objective of zero-shot learning is to overcome this challenge and extract both seen and unseen labels for each image. This is usually done using semantic label information like attributes [22] or word vector representation [14, 2]. The central concept is to combine the visual features with the semantic word vectors representing each label using a similarity metric. Based on the similarity, unseen labels could be classified [50, 42, 48]. Most of the methods for zero-shot learning concentrate on finding the most dominant label in an image [49, 50, 43]. Despite their great success, these solutions do not generalize well to the problem of zero-shot multi-label classification and do not tackle the multi-label diversity challenges included in it.

In contrast to the zero-shot single-label classification task, multiple seen/unseen labels are assigned to an image in the zero-shot multi-label classification task. There is a limited number of studies addressing this problem. An interesting concept was suggested by [35] where predictions of a classifier trained on seen tags were linearly combined in the word embedding space to form a semantic embedding for that image to tackle the zero-shot single-label classification problem. This semantic embedding was later used to rank unseen labels based on their word vector's similarity to that embedding vector. [28] extended their work by proposing a hierarchical semantic embedding to make the label embedding more representative for the multi-label task. [15] proposed a transductive learning strategy to promote the regression model learned from seen classes to generalize well to unseen classes. In the Fast0Tag approach [58], the authors proposed a fast zero-shot tagging method by estimating a principal direction for an image. They show that word vectors of relevant tags in a given image rank ahead of the irrelevant tags along this principal direction in the

word vector space. Another approach suggested by [27], is using structured knowledge graphs to describe the relationships between multiple labels from the semantic label space and show how it can be applied to multi-label and zero-shot multi-label classification tasks. Due to the difficulty in distinguishing between multiple instances in an image using only global features, some studies try to identify important sub-regions in the image that includes the relevant labels by utilizing region proposal methods [40, 37]. In recent work, [21] proposed a shared multi-attention model for multi-label zero-shot learning that can focus on the relevant regions, obviating the need for object detection or region proposal networks. VGG-19 backbone is used to extract rich regional features, and a 4-term loss function is formed to tackle multiple challenges encountered during training. The derived model is then used to extract multiple attentions projected into the joint visual-label semantic embedding space to determine their labels. While this method tackles the diversity challenge by using multiple attention features for comparison to seen and unseen tags, it includes a complicated loss function consisting of a ranking loss and 3 regularization terms that require careful parameter tuning during training, while the image label diversity isn't used implicitly in any of these loss functions.

Finally, in our proposed method, we use the image semantic label diversity directly during training to improve and generalize our model better to diverse images. As far as we know, this is the first work to analyze the zero-shot semantic diversity problem and offer a method to exploit this information in a novel loss function. In addition, our end-to-end training flow does not require a large backbone model or object proposals for training while still achieving state-of-the-art results.

3. Semantic Diversity Learning

In this section, we present our proposed method for training multi-label zero-shot models. The problem and network architecture will be presented first, following a detailed description of our semantic diversity-based loss function.

3.1. Problem Setting

Let us denote by $\mathcal{S}$ and $\mathcal{U}$ the seen and unseen sets of tags, respectively, where seen tags stand for tags that have been seen during training and 'unseen' means tags that were not included in the training annotations. The entire set of tags is defined by $\mathcal{C} = \mathcal{S} \cup \mathcal{U}$.

Let $\{(I_n, Y_n); n = 1, 2, ..., N\}$ denote the training data where I_n is the n-th image, and Y_n is the corresponding set of seen tags. We assume that each tag will be represented by a semantic word vector $\{v^c\}_{c \in \mathcal{C}}$. Based on these notations, we define the task of multi-label zero-shot learning as assigning the relevant unseen tags $y_i \subset \mathcal{U}$ for a given image I_i, and generalized multi-label zero-shot learning as

assigning the relevant seen or unseen tags $y_i \subset \mathcal{C}$ for a given image I_i.

3.2. Network Architecture

The proposed network architecture is illustrated in Figure 2. We used TResNet-M [41] convolutional neural network (CNN) as a backbone for our visual model. TResNet-M is a GPU-optimized model that reports high accuracy and efficiency on several competitive computer vision datasets. Using an efficient model design allows us to train easily in an end-to-end manner. By modifying its last fully connected layer, the vision model is trained to output a $M \times d_w$ linear transformation matrix where d_w is the length of the word vectors, and M is a parameter of our architecture representing the number of principal direction vectors. This matrix provides an image-dependent word ranking. A similar concept was presented in [55] for multi-label classification, where the transformation was learned per image and optimized to reach a linear combination of word vectors that allows it to distinguish the relevant from the non-relevant tags. However, in our experiments, simply using the loss suggested in [55] for the zero-shot learning task, the resulting model failed to generalize well for the unseen tags. Hence, we propose a loss function more suitable for this task.

3.3. Loss

The per-image linear transformation A extracted from the vision model should provide a high rank for all relevant tags even when there exists a large difference in their word embedding representation. Hence, we design our loss function to deal with the semantic diversity problem in zero-shot multi-label learning. Namely, we would like to provide a higher rank for a positive word vector p_j and a lower rank for a negative word vector n_k, hence minimize the following:

$$u_{jk} = \max(An_k) - \max(Ap_j) \qquad (1)$$

The usage of a $\max$ function is crucial for this purpose as it allows each row in the matrix (principal direction) to be optimized in a different direction than other rows. In this formulation, one matrix row is sufficient to rank a label correctly, thus letting other rows focus on the additional relevant labels and output high scores for all of them. This is ideal if there is a high semantic diversity in the image that makes it difficult for a single row in the transformation matrix to deal with the multiple and diverse set of tags. Similar intuition is presented in multi-class support vector machines (SVM) [9], where the confidence value for the correct label is optimized to be larger by a certain margin than the confidences assigned to the rest of the labels. Otherwise, we get a loss that is linearly proportional to the difference between the confidence of the correct label and the maximum among
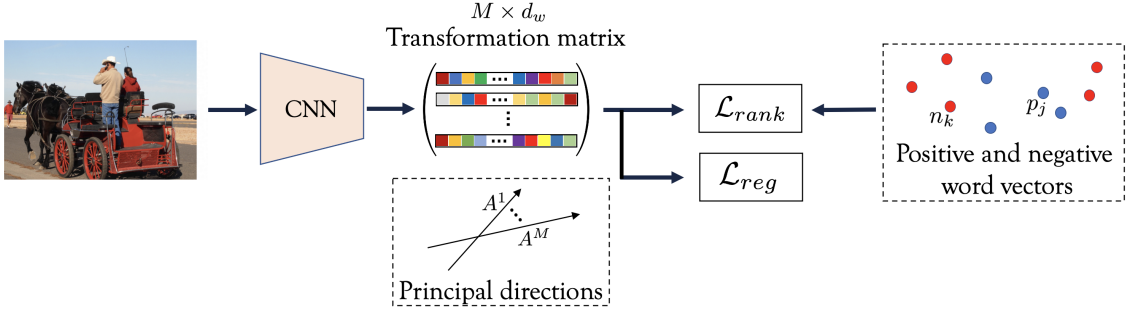
Figure 2: The overview of our end-to-end training scheme for multi-label zero-shot learning. A CNN model is used to extract a per-image matrix transformation A_i that includes several principal directions. $\mathcal{L}_{rank}$ loss pushes A_i towards ranking positive labels higher than negative ones, and $\mathcal{L}_{reg}$ regularizes its principal directions.

the other labels' confidences. Also, by using the max operation when comparing a specific pair of positive and negative labels, we allow gradients propagation only through the most dominant row in the matrix for each tag, hence, not modifying the entire matrix for each pair which allows each row to focus on different semantic concepts.

Having this in mind, we define the ranking loss, inspired by [58], as follows:

$$\mathcal{L}_{rank} = \omega_d \frac{1}{\omega_n} \sum_j \sum_k \log\left(1 + e^{u_{jk}}\right) \qquad (2)$$

Where $\omega_n = |P||\bar{P}|$, while $|P|$ is the size of the set of ground-truth tags for a single image and $|\bar{P}|$ is the size of the set of the non relevant tags. The division by ω_n is used as a normalization. Images with a large number of tags will be treated the same as these with a low number of tags. We consider images with high label diversity more difficult as they require our model to learn how to rank several different semantic concepts higher than others. Hence, we would like to put additional focus on these examples during training. We denote by ω_d the per-image semantic diversity weight (SDW). The SDW up-weights more diverse images (hard samples), thus increasing focus on them in our loss function. We define ω_d as the sum of variances across the relevant set of tags in an image:

$$\omega_d = 1 + \sum_{i=1}^{d_w} \text{var}(P_i) \qquad (3)$$

We note that our ranking loss function provides high flexibility when learning the transformation matrix which is good for learning diverse tags per image. However, this flexibility should be controlled when training on large and noisy datasets. The matrix rows could become too diverse and, by such, making it sensitive to outliers and drive it to not generalize well when training on large and noisy datasets. Hence, we add a regularization term to deal with

these cases defined by

$$\mathcal{L}_{reg} = \left\| \sum_{m=1}^{d_w} \text{var}(A^m) \right\|_1 \qquad (4)$$

This regularization term puts a constraint on the matrix rows from being too diverse. In practice, reducing the variance between rows encourages learning correlative information between tags known to be useful in multi-label setting [7]. This regularization term is similar in a way to multi-class SVM regularization [9]. However, by looking at the variance, we ignore the mean of each column in the matrix, making it invariant to translations. This is different from the standard l_2 norm regularization used in multi-class SVM. The main reasoning is that our loss function is optimized over a given word vector space that its features are not standardized. Hence, using the variance as a regularization instead of the standard l_2 norm is more suitable for our use case. Our final loss function is defined by

$$\mathcal{L}_{final} = \frac{1}{N} \sum_{i=1}^{N} \left[(1 - \tilde{\lambda})\mathcal{L}_{rank}(A_i, y_i) + \tilde{\lambda}\mathcal{L}_{reg}(A_i) \right] \qquad (5)$$

Where $\tilde{\lambda}$ sets the regularization weight. In practice we use $\lambda = \tilde{\lambda}|\bar{P}|$ as the regularization parameter which is invariant to the number of negative tags per image.

4. Experiments

Several experiments were conducted to analyze and evaluate our method for zero-shot multi-label classification. In section 4.2 we compare our approach to other state-of-the-art works in the field. To better understand each component's contribution in our framework, we conduct an ablation study as discussed in section 4.3, and the regularization parameter is further analyzed in section 4.4. Next, to visualize what the transformation matrix learns using our

Table 1: State-of-the-art comparison for ZSL and GZSL tasks on the NUS-WIDE and Open Images datasets. We report the results in terms of mAP, as well as precision (P), recall (R), and F1 score at $K\in\{3,5\}$ for NUS-WIDE and $K\in\{10,20\}$ for Open Images. Best results are in bold.

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 [35]	ZSL	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
	GZSL	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]	ZSL	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
	GZSL	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 [58]	ZSL	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
	GZSL	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 [23]	ZS	20.9	33.5	25.8	16.2	43.2	23.6	10.4	-	-	-	-	-	-	-
	GZSL	17.9	7.9	10.9	15.6	11.5	13.2	3.7	-	-	-	-	-	-	-
LESA (M=10) [21]	ZSL	**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
	GZSL	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
Ours (M=7)	ZSL	24.2	**41.3**	30.5	18.8	**53.4**	27.8	**25.9**	**6.1**	**47.0**	**10.7**	**4.4**	**68.1**	**8.3**	**62.9**
	GZSL	**27.7**	**13.9**	**18.5**	**23.0**	**19.3**	**21.0**	**12.1**	**35.3**	**40.8**	**37.8**	**23.6**	**54.5**	**32.9**	**75.3**

Table 2: State-of-the-art comparison on the MS COCO dataset split into 48 seen and 17 unseen classes. We report the results in terms of precision (P), recall (R), and F1 score at $K=3$ for ZSL and GZSL tasks.

Method	Task	P	R	F1
CONSE [35]	ZSL	11.4	28.3	16.2
	GZSL	23.8	28.8	26.1
Fast0tag [58]	ZSL	24.7	61.4	25.3
	GZSL	38.5	46.5	42.1
Deep0tag [37]	ZSL	26.5	65.9	37.8
	GZSL	43.2	52.2	47.3
Ours (M=2)	ZSL	26.3	65.3	37.5
	GZSL	59.0	60.8	59.9

method, a set of qualitative results are presented and discussed in section 4.5. As our method aims towards dealing with the semantic diversity challenge, we wish to analyze our results on the more diverse set of images as discussed in section 4.6. Finally, in section 4.7 we discuss and analyze our results using a different number of principal directions in the transformation matrix.

4.1. Setup

Datasets: Three datasets were used to evaluate our proposed methodology. The **NUS-WIDE** [8] dataset includes 270K images with 81 human-annotated categories used as unseen classes in addition to a set of 925 labels obtained from Flickr users tags automatically that are used as seen classes. The **MS COCO** [30] dataset is divided into training and validation sets with 82,783 and 40,504 images, re-

spectively. This dataset is commonly used for multi-label zero-shot object detection [3, 19] and was also used in recent works of multi-label zero-shot classification [37]. We follow [3] with our split to seen and unseen tags, resulting in 48 seen and 17 unseen classes based on their cluster embedding in the semantic space and WordNet hierarchy [33]. We use the provided list of images, including 73,774 (44,154 after removing unseen classes) images with only seen objects for training, and 6,608 images containing both seen and unseen objects for testing.

The **Open Images** (v4) [26] dataset consists of 9 million training images, 41,620 validation images, and 125,456 test images. This dataset introduces several challenges: this large-scale dataset is larger by orders of magnitude when compared to NUS-WIDE or MS COCO, and its images are only partially annotated where not all labels were verified as true-positives or negatives in each image. Similar to [21], we use 7,186 labels, having at least 100 images in training set for each seen class. The most frequent 400 test labels not present in the training data are selected as unseen classes.

Evaluation Metrics: We follow [21] and use the mean Average Precision (mAP) and F1 score at *top-K* predictions in each image. The mAP evaluates the accuracy for tag-based retrieval, i.e., it answers the question of how good our model is at ranking images for each given label, while the *top-K* F1 score captures its accuracy for image tagging, measuring how good it is at ranking relevant labels for each image.

Implementation Details: Unless stated otherwise, all experiments were conducted with the following training configuration. We use as a backbone TResNet-M, pre-trained on the ImageNet dataset [10]. The model was fine-tuned using Adam optimizer [24] and 1-cycle cosine annealing policy [45] with maximal learning rate of 1e-4. We use cutout

	Fast0Tag [58]			a	b	c	d	f	Ours
	Orig	Impl.	Base.						
SDW				✓	✓	✓	✓		✓
M=2					✓	✓			
Reg. (0.1)						✓	✓		
M=7							✓	✓	✓
Reg. (0.3)								✓	✓
mAP GZS	3.7	9.7	9.5	10.2	10.6	11.0	12.2	11.8	12.1
ZS	15.1	21.6	20.6	22.7	22.2	23.8	25.1	25.8	25.9

Table 3: Ablation study showing the contribution of the different components in our training scheme compared to the loss presented in Fast0Tag showing the original implementation results, our implementation results using our training framework, and the baseline on NUS-Wide test set.

[12] with probability of 0.5, True-weight-decay [31] of 3e-4 and standard ImageNet augmentations. The regularization parameter λ was set to 0.3. We train the network for 7/7/20 epochs and a batch-size of 192/96/32 for NUS-Wide/Open Images/MS-COCO, respectively.

For our tag embedding representations we use a FastText pre-trained model [17] with a vector size of $d_w = 300$. The word vectors are ℓ_2 normalized. At inference, our trained model takes an image I_i as input and provides a corresponding transformation matrix A_i as output. Let $T = \{t_j\}$ denote the set of word vectors representing each tag in the label set. For image tagging, we compute $r_{ij} = \max(A_i t_j)$ for each seen/unseen tag and rank them such that higher values represent the more relevant tags. For tag-based image retrieval, for a query tag, we compute similarly r_{ij} for all given images and rank them from most relevant to least.

Baseline: We use as a baseline to our method implementation of Fast0Tag [58] loss function integrated within our training framework. In addition, we also compare our method to a baseline with multiple principal directions ($M = 7$), substituting our max function in equation (1) with l_2 norm similar to [55], and removing our regularization term and SDW.

4.2. Comparison to State-of-the-art

Table 1 shows a comparison of our proposed method to other state-of-the-art methods on NUS-WIDE and Open Images. Our method outperforms all other methods in terms of mAP for both datasets. We also present the *top-K* Precision (P) and Recall (R) in addition to the F1 score. Note that we used $K \in \{3, 5\}$ for NUS-WIDE and $K \in \{10, 20\}$ for Open Images due to a large number of available labels in it. Compared to the recently introduced shared multi-attention-based approach (LESA) [21] we achieve better performance on open images. We improved results in both zero-shot/generalized zero-shot learning tasks by 9.3%/20.4%, 7.3%/18.6%, and 21.2%/29.9% in $F1(K = 10)$, $F1(K = 20)$, and mAP respectively.

As for NUS-WIDE, although the LESA approach shows a moderate improvement in performance in terms of F1 for zero-shot learning, our proposed method shows a much higher gain in terms of mAP with an improvement of 6.5% for both zero-shot and generalized zero-shot. In addition, our method achieves improved results in terms of F1 for generalized zero-shot of 4.1% and 4.2% for $F1(K = 3)$, $F1(K = 5)$ respectively.

During this paper's writing, recent work in the field was published as a pre-print showing a Generative approach for zero-shot learning [18]. In this approach, the authors propose to train two separate classifiers, one focused on the zero-shot learning task and another on the generalized zero-shot learning task. This differs from our problem formulation, as we wish to have a single model that is trained for both zero-shot and generalized zero-shot tasks. The formulation used in our study seems to be more suitable for real-world applications, i.e., evaluating one single model for the two tasks under the same working point.

MS-COCO dataset is quite different from NUS-Wide and Open Images as it holds a relatively small number of seen and unseen labels. In Table 2 we compare to the method presented in [37] which is based on an object detection model. In comparison, our model achieves slightly lower results for zero-shot learning with significant improvement in the generalized zero-shot metrics. As presented in Figure 2, our framework does not include any additional modules such as object detectors, region proposals, or attention layers and can still achieve high-quality results. Note that for COCO, we used $M = 2$. We discuss the motivation for this in section 4.7.

In our experiments we use TResNet-M [41] as a backbone for our visual model, while most works use other backbone architectures such as VGG19 [44] or ResNet50 [20]. We also explore the usage of these popular backbones in the supplementary material.
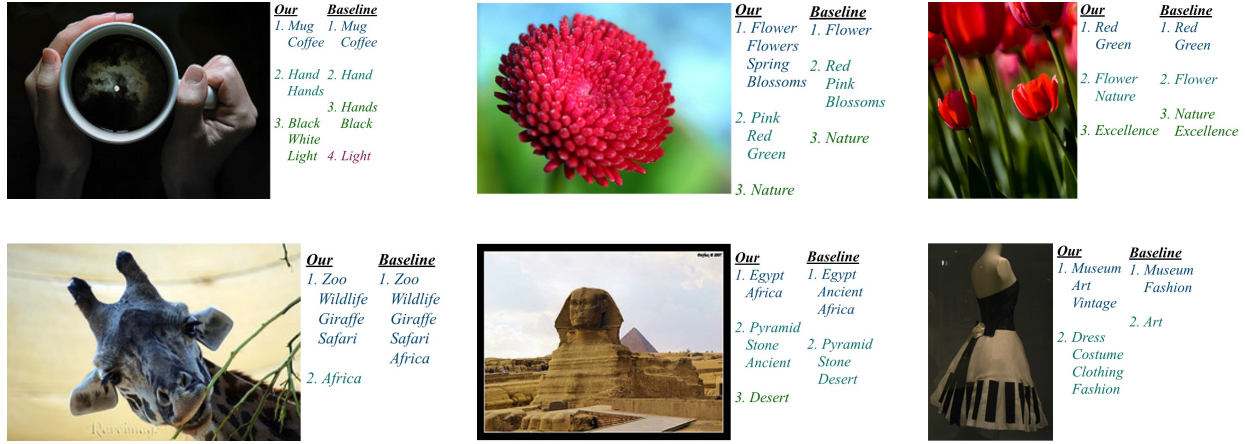
Figure 3: Qualitative results, showing the most relevant tags for each row in the transformation matrix using our proposed method and the baseline. Using our proposed method, we can see that different semantic concepts are being learned by different rows in the matrix.
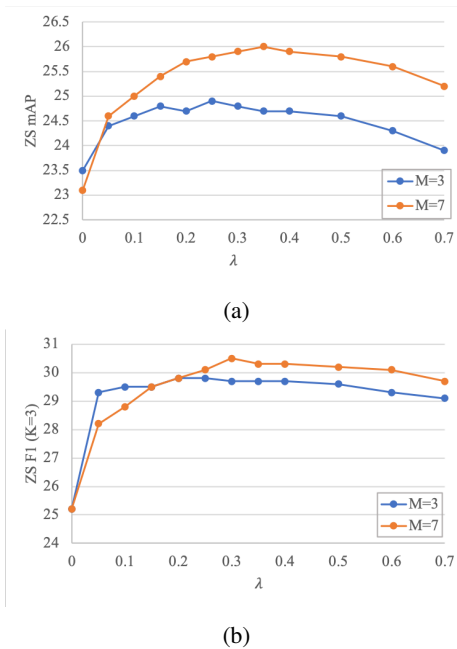

(a)

(b)

Figure 4: Zero-shot multi label classification quality as a function of λ and $M = \{3, 7\}$ in the transformation matrix. (a) The mAP for the multi-label ZS task on NUS-wide test set; (b) Similar to (a) showing $F1(K = 3)$.

4.3. Ablation Study

To better understand each component's contribution in our solution, we perform an ablation study as shown in Table 3. We compare to Fast0Tag [58] method that uses a single principal direction per image as a starting point.

We also implement Fast0Tag in our framework for a fair comparison, which already shows much better results compared to the original paper. However, the baseline with additional principal directions shows lower performance in terms of mAP, indicating that a naive addition of principal directions is insufficient to improve the model performance. The addition of SDW has been shown to improve the results when used together with the loss proposed in [58]. It also improves the results of our proposed method (column f to ours), indicating that SDW can be beneficial for different methods and loss functions that can support per-sample weighting. In addition, the regularization term has shown to improve the results when using $M = 2$ while for $M = 7$ higher regularization showed higher performance in terms of ZS mAP while keeping on par results in terms of GZS mAP.

4.4. Regularization Parameter

The regularization term presented in equation (4) provides control over the matrix transformation rows' diversity. Figure 4 presents the results in terms of mAP and F1 ($K = 3$) on NUS-Wide test set using different λ values for $M = \{3, 7\}$. The contribution of the regularization term is noticeable for a different number of rows. However, for a larger number of rows ($M = 7$) stronger regularization provides better performance. A possible reason for that is that a large number of rows in the matrix can lead to a decrease in utilization of all rows in practice. Using our proposed regularization, we better utilize the different rows in the matrix and thus better generalize on the test set. Especially for zero-shot learning, the generalizability of the model is crucial for retrieving images with unseen tags.

Table 4: ZS multi-label classification results with $M = 7$ rows in the transformation matrix, for samples with more than 6 unseen labels, in terms of precision (P), recall (R), and F1 for top 10 retrieved labels on NUS-Wide test set.

Method	P	R	F1
Baseline	31.7	44.8	37.1
Our w/o SDW	36.2	51.2	42.4
Our	**36.6**	**51.9**	**42.9**

4.5. Qualitative Assessment

Our proposed method focuses on semantic diversity learning using a matrix transformation. Each row in this matrix can be described as a principal direction responsible for a set of relevant labels. In Figure 3 we compare our method results to the baseline model. For each image, we show the most relevant results in the $top - 10$ retrieved labels. The numbers in the figure indicate the most dominant row that provided the highest score for the corresponding set of tags. In several of these sample images, we can see that the tags learned using our approach were separated based on their main concept, e.g., in the top left image, hand and hands belong to the same row, while for the baseline, they are separated. Moreover, using our method, we can see that more relevant tags were discovered in some cases, e.g., in the bottom right image, additional tags such as "clothing" and "dress" were discovered by the same row that learned to understand this concept in the image.

4.6. Performance on Diverse Images

As our method aims towards learning diverse concepts in an image, we wish to evaluate its performance on the more-diverse samples in the dataset. Since diverse images usually include more labels, we perform an experiment to evaluate our image tagging method on image samples that include more than 6 labels from both zero-shot and generalized zero-shot sets. Table 4 presents the results using the baseline and our proposed method ($M = 7$). Since SDW up-weights more diverse images in the loss function, we show results both with and without it compared to the baseline. Our method outperforms the baseline without SDW and achieves even higher results when adding it during training, demonstrating its effectiveness with managing diverse samples.

4.7. Matrix Principal directions

The size of the per-image transformation matrix A is set by the parameter M that defines the number of principal directions it has. In Figure 5 the mAP improvement in terms of zero-shot and generalized zero-shot is presented for a different number of rows in the matrix using a fixed set of parameters (e.g., regularization parameter $= 0.1$) on the NUS-
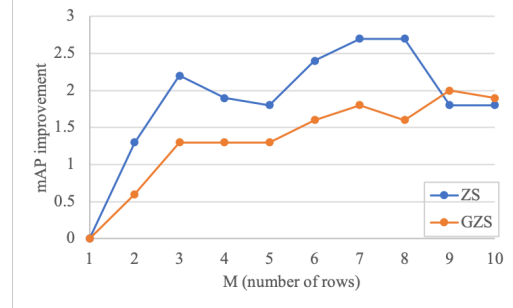

Figure 5: mAP improvement with different number of rows (M) in the transformation matrix compared to $M = 1$ for zero-shot and generalized zero-shot tasks on NUS-Wide test set.

Wide test set.

Noticeably, an increasing number of rows improves the generalized zero-shot results. While the set of principal directions used increases, it is easier for the model to learn the seen tags in the image and improve the generalized zero-shot performance. As for zero-shot learning, while there is an improvement when increasing M up to $7 - 8$ rows, a further increase in the number of rows causes a decrease in mAP. This indicates that the model does not generalize well to unseen tags in our method when using too many principal directions. A possible solution for this would be to increase the regularization parameter. In our experiments, we have found that $M = 7$ covers the semantic diversity in the image compared to other choices for NUS-Wide and Open Images. While for MS-COCO, which is much smaller in the number of labels, $M = 2$ was found experimentally to show superior results.

5. Conclusions

The zero-shot multi-label classification task introduces the challenge of recognizing multiple and diverse labels in an image, including categories not seen during the training process. This task is even more challenging for images with high semantic diversity.

In this study, we propose an end-to-end training scheme including a novel loss function tailored to semantic diversity in zero-shot learning. Our loss function consists of a semantic diversity weight for more diverse images, utilizing multiple principal directions to enable specialization of principal vectors in different semantic concepts and a matrix variance regularization term to improve model generalizability to unseen tags. Extensive experiments show that our proposed loss function improves the zero-shot model's quality in terms of tag-based image retrieval showing SoTA results while keeping high-performance for image tagging on several standard datasets (NUS-Wide, COCO, Open Images).

References

[1] Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele. Multi-cue zero-shot learning with strong supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 59–68, 2016. 1

[2] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2015. 2, 5

[3] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 2, 5

[4] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. *arXiv preprint arXiv:2009.14119*, 2020. 2

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016. 1

[6] Zhao-Min Chen, Xiu-Shen Wei, Xin Jin, and Yanwen Guo. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 622–627. IEEE, 2019. 2

[7] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019. 2, 4

[8] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *CIVR*, 2009. 5

[9] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001. 3, 4

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 5

[11] Shay Deutsch, Soheil Kolouri, Kyungnam Kim, Yuri Owechko, and Stefano Soatto. Zero shot learning via multi-scale manifold regularization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7112–7119, 2017. 1

[12] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout, 2017. 6

[13] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 647–657, 2019. 2

[14] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 1, 2

[15] Yanwei Fu, Yongxin Yang, Tim Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-label zero-shot learning. *arXiv preprint arXiv:1503.07790*, 2015. 2

[16] Bin-Bin Gao and Hong-Yu Zhou. Multi-label image recognition with multi-class attentional regions. *arXiv preprint arXiv:2007.01755*, 2020. 2

[17] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018. 6

[18] Akshita Gupta, Sanath Narayan, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Joost van de Weijer. Generative multi-label zero-shot learning. *arXiv preprint arXiv:2101.11606*, 2021. 6

[19] Nasir Hayat, Munawar Hayat, Shafin Rahman, Salman Khan, Syed Waqas Zamir, and Fahad Shahbaz Khan. Synthesizing the unseen for zero-shot object detection. In *ACCV*, 2020. 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[21] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8776–8786, 2020. 2, 3, 5, 6

[22] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. *Advances in neural information processing systems*, 27:3464–3472, 2014. 2

[23] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, 2018. 5

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 5

[25] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3174–3183, 2017. 1

[26] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018. 5

[27] Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Multi-label zero-shot learning with structured knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1576–1585, 2018. 3

[28] Xirong Li, Shuai Liao, Weiyu Lan, Xiaoyong Du, and Gang Yang. Zero-shot image tagging by hierarchical semantic embedding. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 879–882, 2015. 2

[29] Yanan Li, Donghui Wang, Huanhang Hu, Yuetan Lin, and Yueting Zhuang. Zero-shot recognition using dual visual-semantic mapping paths. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3279–3287, 2017. 1

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 6

[32] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 1

[33] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 5

[34] Pedro Morgado and Nuno Vasconcelos. Semantically consistent regularization for zero-shot recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6060–6069, 2017. 1

[35] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 2, 5

[36] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 1

[37] Shafin Rahman, Salman Khan, and Nick Barnes. Deep0tag: Deep multiple instance learning for zero-shot image tagging. *IEEE Transactions on Multimedia*, 22(1):242–255, 2019. 2, 3, 5, 6

[38] Shafin Rahman, Salman Khan, and Nick Barnes. Transductive learning for zero-shot object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6082–6091, 2019. 2

[39] Shafin Rahman, Salman Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *Asian Conference on Computer Vision*, pages 547–563. Springer, 2018. 2

[40] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan L Yuille. Multiple instance visual-semantic embedding. In *BMVC*, 2017. 3

[41] Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1400–1409, 2021. 3, 6

[42] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015. 2

[43] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8247–8255, 2019. 2

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[45] Leslie N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay, 2018. 5

[46] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12265–12272, 2020. 2

[47] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *Proceedings of the IEEE international conference on computer vision*, pages 464–472, 2017. 2

[48] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 69–77, 2016. 2

[49] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018. 2

[50] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591, 2017. 1, 2

[51] Guo-Sen Xie, Li Liu, Xiaobo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, and Ling Shao. Attentive region embedding network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9384–9393, 2019. 1

[52] Guo-Sen Xie, Li Liu, Fan Zhu, Fang Zhao, Zheng Zhang, Yazhou Yao, Jie Qin, and Ling Shao. Region graph embedding network for zero-shot learning. In *European Conference on Computer Vision*, pages 562–580. Springer, 2020. 1

[53] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeurIPS*, 2020. 1

[54] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *European Conference on Computer Vision*, pages 649–665. Springer, 2020. 2

[55] Mei-Chen Yeh and Yi-Nan Li. Multilabel deep visual-semantic embedding. *IEEE transactions on pattern analysis and machine intelligence*, 42(6):1530–1536, 2019. 3, 6

[56] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *AAAI*, pages 12709–12716, 2020. 2

[57] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2021–2030, 2017. 1

[58] Yang Zhang, Boqing Gong, and Mubarak Shah. Fast zero-shot image tagging. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5985–5994. IEEE, 2016. 2, 4, 5, 6, 7