THE POWER OF SEMANTIC SIMILARITY BASED SOFT-LABELING FOR GENERALIZED ZERO-SHOT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-Shot Learning (ZSL) is a classification task where some classes referred as *unseen classes* have no labeled training images. Instead, we only have side information (or description) about seen and unseen classes, often in the form of semantic or descriptive attributes. Lack of training images from a set of classes restricts the use of standard classification techniques and losses, including the popular cross-entropy loss. The key step in tackling ZSL problem is bridging visual to semantic space via learning a nonlinear embedding. A well established approach is to obtain the semantic representation of the visual information and perform classification in the semantic space. In this paper, we propose a novel architecture of casting ZSL as a fully connected neural-network with cross-entropy loss to embed visual space to semantic space. During training in order to introduce unseen visual information to the network, we utilize soft-labeling based on semantic similarities between seen and unseen classes. To the best of our knowledge, such similarity based soft-labeling is not explored for cross-modal transfer and ZSL. We evaluate the proposed model on five benchmark datasets for zero-shot learning, AwA1, AwA2, aPY, SUN and CUB datasets, and show that, despite the simplicity, our approach achieves the state-of-the-art performance in Generalized-ZSL setting on all of these datasets and outperforms the state-of-the-art for some datasets.

1 INTRODUCTION

Supervised classifiers, specifically Deep Neural Networks, need a large number of labeled samples to perform well. Deep learning frameworks are known to have limitations in fine-grained classification regime and detecting object categories with no labeled data (Xiao et al., 2015; Socher et al., 2013; Xian et al., 2017; Zhang & Koniusz, 2018). On the contrary, humans can recognize new classes using their previous knowledge. This power is due to the ability of humans to transfer their prior knowledge to recognize new objects (Fu & Sigal, 2016; Lake et al., 2015). Zero-shot learning aims to achieve this human-like capability for learning algorithms, which naturally reduces the burden of labeling.

In zero-shot learning problem, there are no training samples available for a set of classes, referred to as unseen classes. Instead, semantic information (in the form of visual attributes or textual features) is available for unseen classes (Lampert et al., 2009; 2014). Besides, we have standard supervised training data for a different set of classes, referred to as seen classes along with the semantic information of seen classes. The key to solving zero-shot learning problem is to leverage trained classifier on seen classes to predict unseen classes by transferring knowledge analogous to humans.

Early variants of ZSL assume that during inference, samples are only from unseen classes. Recent observations (Chao et al., 2016; Scheirer et al., 2013; Xian et al., 2017) realize that such an assumption is not realistic. Generalized ZSL (GZSL) addresses this concern and considers a more practical variant. In GZSL there is no restriction on seen and unseen classes during inference. We are required to discriminate between all the classes. Clearly, GZSL is more challenging because the trained classifier is generally biased toward seen classes.

In order to create a bridge between visual space and semantic attribute space, some methods utilize embedding techniques (Palatucci et al., 2009; Romera-Paredes & Torr, 2015; Socher et al., 2013; Bucher et al., 2016; Xu et al., 2017; Zhang et al., 2017; Kodirov et al., 2015; Akata et al., 2016; 2015; Simonyan & Zisserman, 2014; Frome et al., 2013; Xian et al., 2016; Zhang & Saligrama, 2016; Al-Halah et al., 2016; Zhang & Shi, 2019; Atzmon & Chechik, 2019) and the others use semantic similarity between seen and unseen classes (Zhang & Saligrama, 2015; Fu et al., 2015; Mensink

et al., 2014). Semantic similarity based models represent each unseen class as a mixture of seen classes. While the embedding based models follow three various directions; mapping visual space to semantic space (Palatucci et al., 2009; Romera-Paredes & Torr, 2015; Socher et al., 2013; Bucher et al., 2016; Xu et al., 2017; Socher et al., 2013), mapping semantic space to the visual space (Zhang et al., 2017; Kodirov et al., 2015; Shojaee & Baghshah, 2016; Ye & Guo, 2017), and finding a latent space then mapping both visual and semantic space into the joint embedding space (Akata et al., 2016; 2015; Simonyan & Zisserman, 2014; Frome et al., 2013; Xian et al., 2016; Zhang & Saligrama, 2016; Al-Halah et al., 2016).

The loss functions in embedding based models have training samples only from the seen classes. For unseen classes, we do not have any samples. It is not difficult to see that this lack of training samples biases the learning process towards seen classes only. One of the recently proposed techniques to address this issue is augmenting the loss function with some unsupervised regularization such as entropy minimization over the unseen classes (Liu et al., 2018).

Another recent methodology which follows a different perspective is deploying Generative Adversarial Network (GAN) to generate synthetic samples for unseen classes by utilizing their attribute information (Mishra et al., 2018; Zhu et al., 2018; Xian et al., 2018). Although generative models boost the results significantly, it is difficult to train these models. Furthermore, the training requires generation of large number of samples followed by training on a much larger augmented data which hurts their scalability.

The two most recent state-of-the-art GZSL methods, CRnet (Zhang & Shi, 2019) and COSMO (Atzmon & Chechik, 2019), both employ a complex mixture of experts approach. CRnet is based on k-means clustering with an expert module on each cluster (seen class) to map semantic space to visual space. The output of experts (cooperation modules) are integrated and finally sent to a complex loss (relation module) to make a decision. CRnet is a multi-module (multi-network) method that needs end-to-end training with many hyperparameters. Also COSMO is a complex gating model with three modules: a seen/unseen classifier and two expert classifiers over seen and unseen classes. Both of these methods have many modules, and hence, several hyperparameters; architectural, and learning decisions. A complex pipeline is susceptible to errors, for example, CRnet uses k-means clustering for training and determining the number of experts and a weak clustering will lead to bad results.

Our Contribution: We propose a simple fully connected neural network architecture with unified (both seen and unseen classes together) cross-entropy loss along with soft-labeling. Soft-labeling is the key novelty of our approach which enables the training data from the seen classes to also train the unseen class. We directly use attribute similarity information between the correct seen class and the unseen classes to create a soft unseen label for each training data. As a result of soft labeling, training instances for seen classes also serve as soft training instance for the unseen class without increasing the training corpus. This soft labeling leads to implicit supervision for the unseen classes that eliminates the need for any unsupervised regularization such as entropy loss in (Liu et al., 2018). Soft-labeling along with crossentropy loss enables a simple MLP network to tackle GZSL problem. Our proposed model, which we call Soft-labeled ZSL (SZSL), is simple (unlike GANs) and efficient (unlike visual-semantic pairwise embedding models) approach which achieves the state-of-the-art performance in Generalized-ZSL setting on all five ZSL benchmark datasets and outperforms the state-of-the-art for some of them.

2 Problem Definition

In zero-shot learning problem, a set of training data on seen classes and a set of semantic information (attributes) on both seen and unseen classes are given. The training dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$ includes n samples where $\mathbf{x}_i$ is the visual feature vector of the i-th image and $\mathbf{y}_i$ is the class label. All samples in $\mathcal{D}$ belong to seen classes $\mathcal{S}$ and during training there is no sample available from unseen classes $\mathcal{U}$. The total number of classes is $C = |\mathcal{S}| + |\mathcal{U}|$. Semantic information or attributes $\mathbf{a}_k \in \mathbb{R}^a$, are given for all C classes and the collection of all attributes are represented by attribute matrix $\mathbf{A} \in \mathbb{R}^{a \times C}$. In the inference phase, our objective is to predict the correct classes (either seen or unseen) of the test dataset $\mathcal{D}'$. The classic ZSL setting assumes that all test samples in $\mathcal{D}'$ belong to unseen classes $\mathcal{U}$ and tries to classify test samples only to unseen classes $\mathcal{U}$. While in a more realistic setting i.e. GZSL, there is no such an assumption and we aim at classifying samples in $\mathcal{D}'$ to either seen or unseen classes $\mathcal{S} \cup \mathcal{U}$.

3 PROPOSED METHODOLOGY

3.1 NETWORK ARCHITECTURE

As Figure 1 illustrates our architecture, We map visual space to semantic space, then compute the similarity score (dot-product) between true attributes and the attribute/semantic representation of the input ($\mathbf{x}$). Finally, the similarity score is fed into a Softmax, and the probability of all classes are computed. For the visual features as the input, in all five benchmark datasets, we use the extracted visual features by a pre-trained ResNet-101 on ImageNet provided by Xian et al. (2017). We do not fine-tune the CNN that generates the visual features unlike model in Liu et al. (2018). In this sense, our proposed model is also fast and straightforward to train.

3.2 SOFT LABELING

In ZSL problem, we do not have any training instance from unseen classes, so the output nodes corresponding to unseen classes are always inactive during learning. Standard supervised training with cross entropy loss biases the network towards seen classes only. The true labels (hard labels) used for training only represent seen classes so the cross entropy cannot penalize unseen classes. Moreover, the available similarity information between the seen and unseen attributed is never utilized.

We propose soft labeling based on the similarity between semantic attributes. For each seen sample, we represent its relationship to unseen categories by obtaining semantic similarity (dot-product) using the seen class attribute and all the unseen class attributes. In the simplest form, for every training data, we can find the nearest unseen class to the correct seen class label and assign a small probability q (partial membership or soft label) of this instance to be from the closest unseen class. Note, each training sample only contains a label which comes from the set of seen classes. With soft labeling, we enrich the label with partial assignments to unseen classes and as Hinton et al. (2015) shows, soft labels act as a regularizer which allows each training case to enforce much more constraint on weights.

In a more general soft labeling approach, we propose assigning a probability to all the unseen classes. A natural choice is to transform seen-to-unseen similarities to probabilities (soft labels) shown in Equation (1). The unseen distribution is obtained for each seen class by calculating dot-product of seen class attribute and all unseen classes attributes and squashing all these dot-product values by Softmax to acquire probabilities. In this case, we distribute the probability q among all unseen classes based on the obtained unseen distribution. This proposed strategy results in a soft label for each seen image during training, which as we show later helps the network to learn unseen categories.

In order to control the flatness of the unseen distribution, we utilize temperature parameter τ. Higher temperature results in flatter distribution over unseen categories and lower temperature creates a more ragged distribution with peaks on nearest unseen classes. A small enough temperature basically results in the nearest unseen approach. The Impact of temperature τ on unseen distribution is depicted in Figure 3.a for a particular seen class. Soft labeling implicitly introduces unseen visual features into the network without generating fake unseen samples as in generative methods (Mishra et al., 2018; Zhu et al., 2018; Xian et al., 2018). Hence our proposed approach is able to reproduce same effect as in generative models without the need to create fake samples and train generative models that are known to be difficult to train. Below is the formal description of *temperature* Softmax:

$$y^u_{i,k} = q\frac{\exp\left(s_{i,k}/\tau\right)}{\sum_{j\in\mathcal{U}}\exp\left(s_{i,j}/\tau\right)} \quad \text{where} \quad s_{i,j} \triangleq \langle \mathbf{a}_i, \mathbf{a}_j \rangle \tag{1}$$

where $\mathbf{a}_i$ is the i-th column of attribute matrix $\mathbf{A} \in \mathbb{R}^{a\times C}$ which includes both seen and unseen class attributes: $\mathbf{A} = [\mathbf{a}_1 \mid \mathbf{a}_2 \mid \cdots \mid \mathbf{a}_C]$. And $s_{i,j}$ is the *true* similarity score between two classes i, j based on their attributes. τ and q are temperature parameter and total probability assigned to unseen distribution, respectively. Also $y^u_{i,k}$ is the soft label (probability) of unseen class k for seen class i. It should be noted that q is the sum of all unseen soft labels i.e. $\sum_{k\in\mathcal{U}} y^u_{i,k} = q$.

3.3 TRAINING STRATEGY

The proposed method is a multi-class probabilistic classifier that produces a C-dimensional vector of class probabilities $\mathbf{p}$ for each sample $\mathbf{x}_i$ as $\mathbf{p}(\mathbf{x}_i) = \textit{Softmax}\left(\mathbf{A}^T g_{\mathbf{w}}\left(\mathbf{x}_i\right)\right)$ where $\mathbf{A}^T g_{\mathbf{w}}\left(\mathbf{x}_i\right)$ is a

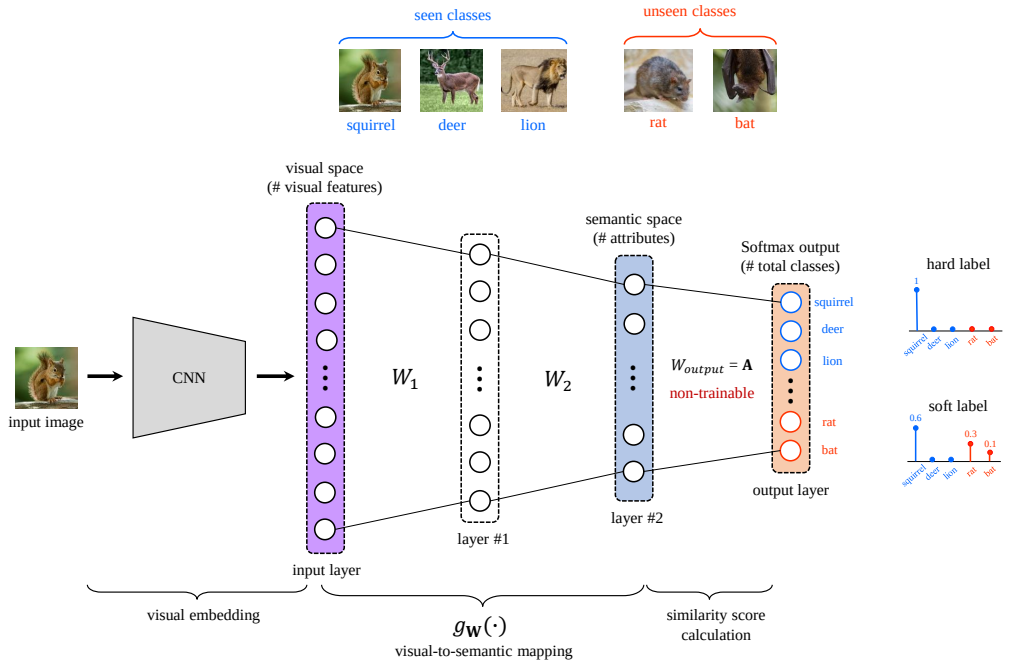


Figure 1: The overall workflow of the SZSL classifier and architecture of the proposed MLP. Layers #1 and #2 provide the nonlinear embedding $g_{\mathbf{W}}(.)$ to map visual features to attribute space and their weights W_1, W_2 are learned by SGD. The output layer with non-trainable weights $\mathbf{A}$, basically calculates dot-products of semantic representation of the input and all class attributes simultaneously. Soft-labels are also shown for a sample image from *squirrel* class.

C-dimensional vector of all similarity scores of an input sample. The *predicted* similarity score between semantic representation of sample $\mathbf{x}_i$ and attribute $\mathbf{a}_k$ is $\hat{s}_{i,k} \triangleq \langle g_{\mathbf{W}}(\mathbf{x}_i), \mathbf{a}_k \rangle$. Each element of vector $\mathbf{p}$, represents an individual class probability that can be shown below:

$$p_k(\mathbf{x}_i) = \frac{\exp(\hat{s}_{i,k})}{\sum_{j=1}^{C} \exp(\hat{s}_{i,j})} \tag{2}$$

This Softmax as the activation function of the last layer of the network is calculated on all classes. An established choice to train a multi-class probabilistic classifier is the cross-entropy loss which we later show naturally integrates our idea of soft labeling. Inspired by Hinton et al. (2015), in addition to the cross-entropy loss over soft targets, we also consider cross entropy-loss over true labels (hard labels) to improve the performance. During training, we aim at learning the nonlinear mapping $g_{\mathbf{W}}(.)$ i.e. obtaining network weights $\mathbf{W}$ through:

$$\min_{\mathbf{W}} \sum_{i=1}^{n} L(\mathbf{x}_i) + \lambda \|\mathbf{W}\|_F^2 + \gamma \|\mathbf{W}\| \tag{3}$$

where λ and γ are regularization factors which are obtained through hyperparameter tuning, and $L(\mathbf{x}_i)$ is the weighted sum of cross-entropy loss over soft labels (L^{soft}) and cross-entropy loss over hard labels (L^{hard}) for each sample as shown below:

$$L(\mathbf{x}_i) = \alpha L^{soft}(\mathbf{x}_i) + (1-\alpha) L^{hard}(\mathbf{x}_i) \tag{4}$$

where $\alpha \in [0, 1]$ is a hyperparameter. For better understanding, the hard-loss and soft-loss terms for each sample $\mathbf{x}_i$ (or $\mathbf{x}$ for simplicity) are expanded and elaborated. The hard-loss term is a conventional cross-entropy loss $L^{hard}(\mathbf{x}) = -\sum_{k=1}^{C} z_k \log(p_k)$, where z_k is the hard label. Clearly, hard-loss term alone does not work in ZSL regime since it does not penalize unseen classes. The soft-loss term is expanded to seen and unseen terms as follows:

$$L^{soft}(\mathbf{x}) = -\sum_{k \in \mathcal{S}} y_k^s \log(p_k^s) - \sum_{k \in \mathcal{U}} y_k^u \log(p_k^u) \tag{5}$$

Let $\bar{p}_k^s$ and $\bar{p}_k^u$ be the normalized versions of p_k^s and p_k^u , respectively. Also the total predicted unseen probability is $\sum_{k\in\mathcal{U}} p_k^u \triangleq \hat{q}$, consequently for seen classes $\sum_{k\in\mathcal{S}} p_k^s \triangleq 1-\hat{q}$. Plugging normalized probabilities in Equation (5), we have:

$$L^{soft}(\mathbf{x}) = -\sum_{k\in\mathcal{S}} y_k^s \log(\bar{p}_k^s) - \sum_{k\in\mathcal{U}} y_k^u \log(\bar{p}_k^u) - \sum_{k\in\mathcal{S}} y_k^s \log(1-\hat{q}) - \sum_{k\in\mathcal{U}} y_k^u \log\hat{q} \tag{6}$$

Utilizing Equation (1), we have $y_k^u = q\bar{y}_k^u$, where y_k^u are soft labels of unseen classes and $\bar{y}_k^u$ is the temperature softmax where $\sum_{k\in\mathcal{U}} \bar{y}_k^u = 1$. Similarly, the normalized seen labels $\bar{y}_k^s$ can be obtained by $y_k^s = (1-q)\bar{y}_k^s$. Replacing normalized labels in Equation (6) leads to:

$$L^{soft}(\mathbf{x}) = -(1-q)\sum_{k\in\mathcal{S}} \bar{y}_k^s \log(\bar{p}_k^s) - q\sum_{k\in\mathcal{U}} \bar{y}_k^u \log(\bar{p}_k^u) - (1-q)\log(1-\hat{q}) - q\log\hat{q} \tag{7}$$

Hence the first two terms of $L^{soft}(\mathbf{x})$ is the weighted sum of cross-entropy of seen classes and cross-entropy of unseen classes. In particular, first term penalizes and controls the relative (normalized) probabilities within all seen classes and the second term acts similarly within unseen classes. We also require to penalize the total probability of all seen classes $(1-\hat{q})$ and total probability of all unseen classes $(\hat{q})$. This is accomplished through the last two terms of Equation (7) which is basically a binary cross entropy loss. Intuitively soft-loss in Equation (7) works by controlling the balance *within* seen/unseen classes (first two terms) as well as the balance *between* seen and unseen classes (last two terms).

As we have shown in Equation (7), soft-loss enables the classifier to learn unseen classes by only being exposed to samples from seen classes. Hyperparameter q acts as a trade-off coefficient between seen and unseen cross-entropy losses (Figure 2). We can see that the regularizer is a weighted cross entropy on unseen class, which leverages similarity structure between attributes compared to uniform entropy function of DCN (Liu et al., 2018). DCN and all prior works use uniform entropy as regularizer, which does not capitalize on the known semantic similarity information between seen and unseen class attributes.

At the inference time, our proposed SZSL method works the same as a conventional classifier, we only need to provide the test image and the network will produce class probabilities for all seen and unseen classes.

4 EXPERIMENT

We conduct comprehensive comparison of our proposed SZSL model with the state-of-the-art methods for GZSL settings on five benchmark datasets (Table 1). We present the detailed description of datasets in Appendix A. Our model outperforms the state-of-the-art methods on GZSL setting for all benchmark datasets.

Table 1: Statistics of five ZSL benchmark datasets

Dataset	#Attributes	#Seen Classes	#Unseen Classes	#Images
AwA1	85	40	10	30475
AwA2	85	40	10	37322
CUB	312	150	50	11788
aPY	64	20	12	18627
SUN	102	645	72	14340

4.1 EVALUATION METRIC

For the purpose of validation, we employ the validation splits provided along with the Proposed Split (PS) (Xian et al., 2017) to perform cross-validation for hyper-parameter tuning. The main objective of GZSL is to simultaneously improve seen samples accuracy and unseen samples accuracy i.e. imposing a trade-off between these two metrics. As the result, the standard GZSL evaluation metric is harmonic average of seen and unseen accuracy. This metric is chosen to encourage the network not

be biased toward seen classes. Harmonic average of accuracies is defined as $A_H = \frac{2A_S A_U}{A_S + A_U}$ where A_S and A_U are seen and unseen accuracies, respectively.

Table 2: Results of GZSL methods on ZSL benchmark datasets under Proposed Split (PS) Xian et al. (2017). U, S and H respectively stand for Unseen, Seen and Harmonic average accuracies.

	AwA1			AwA2			aPY			CUB			SUN		
Method	U	S	H	U	S	H	U	S	H	U	S	H	U	S	H
Non-Generative Models															
ALE (Akata et al., 2013)	16.8	76.1	27.5	14.0	81.8	23.9	4.6	73.7	8.7	23.7	62.8	34.4	21.8	33.1	26.3
SJE (Akata et al., 2015)	11.3	74.6	19.6	8.0	73.9	14.4	3.7	55.7	6.9	23.5	59.2	33.6	14.7	30.5	19.8
ConSE (Norouzi et al., 2013)	0.4	88.6	0.8	0.5	90.6	1.0	0.0	**91.2**	0.0	1.6	**72.2**	3.1	6.8	39.9	11.6
Sync (Changpinyo et al., 2016)	8.9	87.3	16.2	10.0	90.5	18.0	7.4	66.3	13.3	11.5	70.9	19.8	7.9	**43.3**	13.4
DeViSE (Frome et al., 2013)	13.4	68.7	22.4	17.1	74.7	27.8	4.9	76.9	9.2	23.8	53.0	32.8	16.9	27.4	20.9
CMT (Socher et al., 2013)	0.9	87.6	1.8	8.7	89.0	15.9	1.4	85.2	2.8	7.2	49.8	12.6	8.1	21.8	11.8
Generative Models															
f-CLSWGAN (Xian et al., 2018)	57.9	61.4	59.6	-	-	-	-	-	-	43.7	57.7	49.7	42.6	36.6	39.4
SP-AEN (Chen et al., 2018)	23.3	90.9	37.1	-	-	-	13.7	63.4	13.7	34.7	70.6	46.6	24.9	38.6	30.3
cycle-UWGAN (Felix et al., 2018)	59.6	63.4	59.8	-	-	-	-	-	-	47.9	59.3	53.0	47.2	33.8	39.4
SE-GZSL (Kumar Verma et al., 2018)	56.3	67.8	61.5	-	-	-	-	-	-	46.7	53.3	41.5	40.9	30.5	34.9
ZSKL (Zhang & Koniusz, 2018)	18.9	82.7	30.8	-	-	-	10.5	76.2	18.5	21.6	52.8	30.6	20.1	31.4	24.5
DCN (Liu et al., 2018)	25.5	84.2	39.1	-	-	-	14.2	75.0	23.9	28.4	60.7	38.7	25.5	37.0	30.2
COSMO (Atzmon & Chechik, 2019)	52.8	80.0	63.6	-	-	-	-	-	-	44.4	57.8	50.2	44.9	37.7	**41.0**
CRnet (Zhang & Shi, 2019)	58.1	74.7	**65.4**	52.6	52.6	63.1	32.4	68.4	44.0	45.5	56.8	**50.5**	34.1	36.5	35.3
SZSL (Ours)	**57.2**	75.8	65.2	55.1	78.6	**64.7**	**42.7**	57.2	**48.7**	**47.1**	52.5	49.6	**47.2**	32.6	38.6

4.2 Implementation Details

To evaluate SZSL, we follow the popular experimental framework and the Proposed Split (PS) in (Xian et al., 2017) for splitting classes into seen and unseen classes to compare GZSL/ZSL methods. Utilizing PS ensures that none of the unseen classes have been used in the training of ResNet-101 on ImageNet. The input to the model is the visual features of each image sample extracted by a pre-trained ResNet-101 (He et al., 2016) on ImageNet provided by (Xian et al., 2017). The dimension of visual features is 2048.

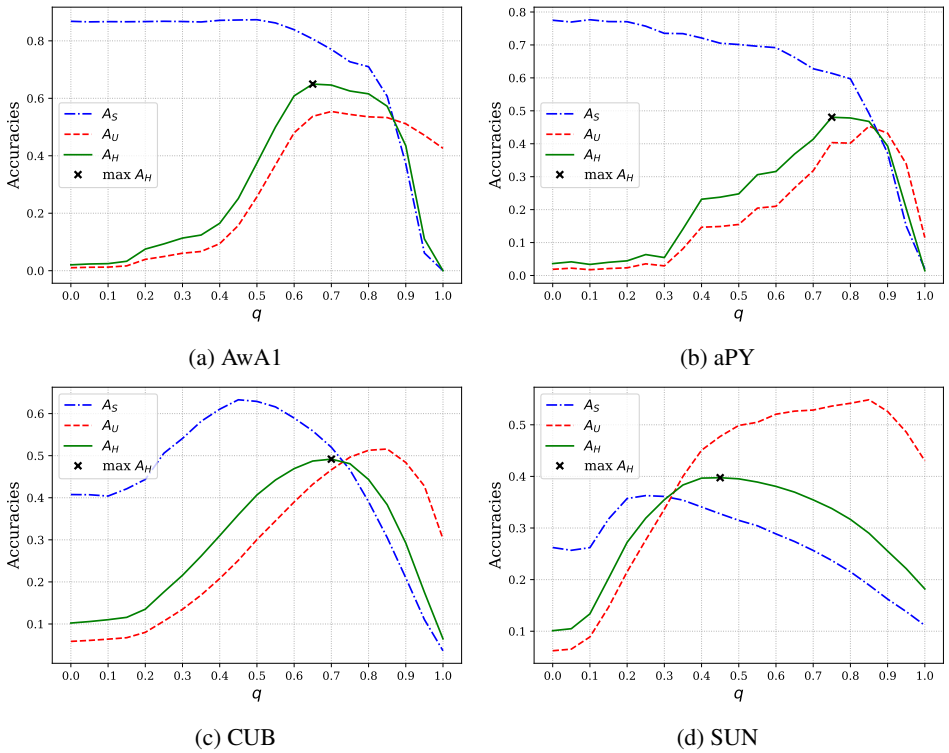
We utilized Keras (Chollet, 2015) with TensorFlow back-end (Abadi et al., 2016) to implement our model. We used Xian et al. (2017) proposed unseen classes for validation (3-fold CV) and added 20% of train samples (seen classes) as seen validation samples to obtain GZSL validation sets. We cross-validate $\tau \in [10^{-2}, 10]$, *mini-batch size* $\in \{64, 128, 256, 512, 1024\}$, $q \in [0, 1]$, $\alpha \in [0, 1]$, *hidden layer size* $\in \{128, 256, 512, 1024, 1500\}$ and *activation function* $\in$\{tanh, sigmoid, hard-sigmoid, relu\} to tune our model. To obtain statistically consistent results, the reported accuracies are averaged over 5 trials (using different initialization) after tuning hyper-parameters with cross-validation. Also we ran our experiments on a machine with 56 vCPU cores, Intel(R) Xeon(R) CPU E5-2660 v4 @ 2.00GHZ and 2 NVIDIA-Tesla P100 GPUs each with 16GB memory. The code is provided in the supplementary material.

4.3 Generalized Zero-Shot Learning Results

To demonstrate the effectiveness of SZSL model in GZSL setting, we comprehensively compare our proposed method with state-of-the-art GZSL models in Table 2. Since we use the standard proposed split, the published results of other GZSL models are directly comparable.

As reported in Table 2, accuracies of our model achieves the state-of-the-art GZSL performance on all five benchmark datasets and outperforms the state-of-the-art on AwA2 and aPY datasets. It is exciting and motivating while our architecture is much simpler compared to recently proposed CRnet and COSMO, yet, we achieve similar or better accuracies compared to them. We have only one simple fully connected neural network with 2 trainable layers, compared to CRnet with K mixture of experts followed by relation module with complex loss functions (pairwsie).

Soft labeling employed in SZSL gives the model new flexibility to trade-off between seen and unseen accuracies during training and attain a higher value of harmonic accuracy A_H, which is the standard



(a) AwA1 (b) aPY

(c) CUB (d) SUN

Figure 2: Plots of seen (A_S), unseen (A_U) and harmonic average (A_H) accuracies versus total probability (q) assigned to unseen classes are shown for all five ZSL datasets. The maximum obtained harmonic accuracy is also marked by ($\times$).

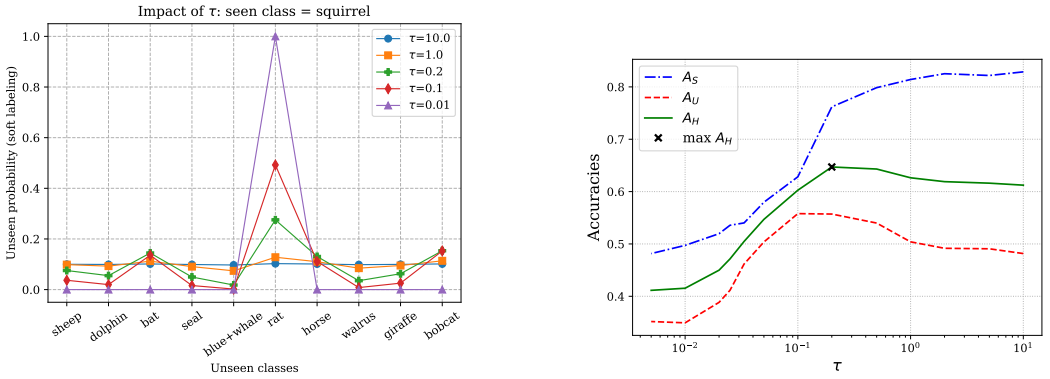
metric for GZSL. Assigned unseen soft labels (unseen probability q) enables the classifier to gain more confidence in recognizing unseen classes, which in turn results in considerably higher unseen accuracy A_U. As the classifier is now discriminating between more classes we get marginally lower seen accuracy A_S. However, balancing A_S and A_U with the cost of deteriorating A_S leads to much higher A_H. This trade-off phenomenon is depicted in Figure 2 for all datasets. The flexibility provided by soft labeling is examined by obtaining accuracies for different values of q. In Figure 2.a and 2.b, by increasing total unseen probability q, A_U increases and A_S decreases as expected. From the trade-off curves, there is an optimal q where A_H takes its maximum value as shown in Figure 2. Maximizing A_H is the primary objective in a GZSL problem that can be achieved by semantic similarity based soft labeling and the trade-off knob, q.

It should be noted that both AwA and aPY datasets (Figure 2.a and 2.b) are coarse-grained class datasets. In contrast, CUB and SUN datasets are fine-grained with hundreds of classes and highly unbalanced seen-unseen split, and hence their accuracies have different behavior concerning q, as shown in Figure 2.c and 2.d. However, harmonic average curve still has the same behavior and possesses a maximum value at an optimal q.

4.3.1 INTUITION

We illustrate the intuition with AwA dataset (Lampert et al., 2009), a ZSL benchmark dataset. Consider a seen class *squirrel*. We compute closest unseen classes to the class *squirrel* in terms of attributes. We naturally find that the closest class is *rat* and the second closest is *bat*, while other classes such as *horse*, *dolphin*, *sheep*, etc. are not close (Figure 3.a). This is not surprising as *squirrel* and *rat* share several attribute. It is naturally desirable to have a classifier that gives *rat* higher probability than other classes. If we force this softly, we can ensure that classifier is not blind towards unseen classes due to lack of any training example.

From a learning perspective, without any regularization, we cannot hope classifier to classify unseen classes accurately. This problem was identified in Liu et al. (2018), where they proposed entropy-



(a) (b)

Figure 3: The impact of temperature parameter τ for AwA dataset. (a) unseen soft labels (before multiplying q) produced by temperature Softmax Equation (1) for various τ, (b) accuracies versus τ for proposed Z-Softmax DU classifier.

based regularization in the form of Deep Calibration Network (DCN). DCN uses cross-entropy loss for seen classes, and regularize the model with entropy loss on unseen classes to train the network. Authors in DCN postulate that minimizing the uncertainty (entropy) of predicted unseen distribution of training samples, enables the network to become aware of unseen visual features. While minimizing uncertainty is a good choice of regularization, it does not eliminate the possibility of being confident about the wrong unseen class. Clearly in DCN's approach, for the above *squirrel* example, the uncertainty can be minimized even when the classifier gives high confidence to a wrong unseen class *dolphin* on an image of seen class *squirrel*. Utilizing similarity based soft-labeling implicitly regularizes the model in a supervised fashion. The similarity values naturally has information of how much certainty we want for specific unseen class. We believe that this supervised regularization is the critical difference why our model outperforms DCN with a significant margin.

4.4 Illustration of Soft Labeling

Figure 3 shows the effect of τ and the consequent assigned unseen distribution on accuracies for AwA dataset. Small τ enforces q to be concentrated on nearest unseen class, while large τ spread q over all the unseen classes and basically does not introduce helpful unseen class information to the classifier. The optimal value for τ is 0.2 for AwA dataset as depicted in Figure 3.b. The impact of τ on the assigned distribution for unseen classes is shown in Figure 3.a when seen class is *squirrel* in AwA dataset. Unseen distribution with $\tau = 0.2$, well represents the similarities between seen class (*squirrel*) and similar unseen classes (*rat*, *bat*, *bobcat*) and basically verifies the result of Figure 3.b where $\tau = 0.2$ is the optimal temperature. While in the extreme cases, when $\tau = 0.01$, distribution on unseen classes in mostly focused on the nearest unseen class, *rat*, and consequently the other unseen classes' similarities are ignored. Also $\tau = 10$ flattens the unseen distribution which results in high uncertainty and does not contribute helpful unseen class information to the learning.

5 Conclusion

We proposed a discriminative GZSL classifier with visual-to-semantic mapping and cross-entropy loss. During training, while SZSL is trained on a seen class, it simultaneously learns similar unseen classes through soft labels based on semantic class attributes. We deploy similarity based soft labeling on unseen classes that allows us to learn both seen and unseen signatures simultaneously via a simple architecture. Our proposed soft-labeling strategy along with cross-entropy loss leads to a novel regularization via generalized similarity-based weighted cross-entropy loss that can successfully tackle GZSL problem. Soft-labeling offers a trade-off between seen and unseen accuracies and provides the capability to adjust these accuracies based on the particular application. We achieve state-of-the-art performance, in GZSL setting, on all five ZSL benchmark datasets while keeping the model simple, efficient and easy to train.

REFERENCES

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 819–826, 2013.

Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2927–2936, 2015.

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, 2016.

Ziad Al-Halah, Makarand Tapaswi, and Rainer Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5975–5984, 2016.

Yuval Atzmon and Gal Chechik. Adaptive confidence smoothing for generalized zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11671–11680, 2019.

Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *European Conference on Computer Vision*, pp. 730–746. Springer, 2016.

Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5327–5336, 2016.

Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European Conference on Computer Vision*, pp. 52–68. Springer, 2016.

Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1043–1052, 2018.

Franois Chollet. keras. `https://github.com/fchollet/keras`, 2015.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1785. IEEE, 2009.

Rafael Felix, Vijay BG Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 21–37, 2018.

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 C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 2121–2129. Curran Associates, Inc., 2013. URL `http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model.pdf`.

Yanwei Fu and Leonid Sigal. Semi-supervised vocabulary-informed learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5337–5346, 2016.

Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot object recognition by semantic manifold distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2635–2644, 2015.

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*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2452–2460, 2015.

Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision and pattern*

recognition, pp. 4281–4289, 2018.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 951–958. IEEE, 2009.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Generalized zero-shot learning with deep calibration network. In *Advances in Neural Information Processing Systems*, pp. 2005–2015, 2018.

Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2441–2448, 2014.

Ashish Mishra, Shiva Krishna Reddy, Anurag Mittal, and Hema A Murthy. A generative model for zero shot learning using conditional variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2188–2196, 2018.

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.

Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (eds.), *Advances in Neural Information Processing Systems 22*, pp. 1410–1418. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3650-zero-shot-learning-with-semantic-output-codes.pdf.

Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2751–2758. IEEE, 2012.

Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2152–2161, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/romera-paredes15.html.

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7): 1757–1772, 2013.

Seyed Mohsen Shojaee and Mahdieh Soleymani Baghshah. Semi-supervised zero-shot learning by a clustering-based approach. *arXiv preprint arXiv:1605.09016*, 2016.

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

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. CUB Dataset. Technical report, 2011.

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*, pp. 69–77, 2016.

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*, pp. 4582–4591, 2017.

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*, pp. 5542–5551, 2018.

Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern*

Recognition, pp. 842–850, 2015.
Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive zero-shot action recognition by word-vector embedding. *International Journal of Computer Vision*, 123(3):309–333, 2017.
Meng Ye and Yuhong Guo. Zero-shot classification with discriminative semantic representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7140–7148, 2017.
Fei Zhang and Guangming Shi. Co-representation network for generalized zero-shot learning. In *International Conference on Machine Learning*, pp. 7434–7443, 2019.
Hongguang Zhang and Piotr Koniusz. Zero-shot kernel learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7670–7679, 2018.
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*, pp. 2021–2030, 2017.
Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE international conference on computer vision*, pp. 4166–4174, 2015.
Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6034–6042, 2016.
Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1004–1013, 2018.

A Appendix

A.1 Dataset

The proposed method is evaluated on five benchmark ZSL datasets. The statistics for the datasets are shown in table 1. Animal with Attributes (AwA1) (Lampert et al., 2014) dataset is a coarse-grained benchmark dataset for ZSL/GSZl. It has 30475 image samples from 50 classes of different animals and each class comes with side information in the form of attributes (e.g. animal size, color, specific feature, place of habitat). Attribute space dimension is 85 and this dataset has a standard split of 40 seen and 10 unseen classes introduced in (Lampert et al., 2014). AWA2 Xian et al. (2017) is the public licensed version of AWA1 with roughly the same amount of samples and the same number of attributes and seen/unseen classes as AWA1.

Caltech-UCSD-Birds-200-2011 (CUB) (Wah et al., 2011) is a fine-grained ZSL benchmark dataset. It has 11,788 images from 200 different types of birds and each class comes with 312 attributes. The standard ZSL split for this dataset has 150 seen and 50 unseen classes (Akata et al., 2016).

SUN Attribute (SUN) (Patterson & Hays, 2012) is a fine-grained ZSL benchmark dataset consists of 14340 images of different scenes and each scene class is annotated with 102 attributes. This dataset has a standard ZSL split of 645 seen and 72 unseen classes.

attribute Pascal and Yahoo (aPY) (Farhadi et al., 2009) is a small and coarse-grained ZSL benchmark dataset which has 14340 images and 32 classes of different objects (e.g. aeroplane, bottle, person, sofa, ...) and each class is provided with 64 attributes. This dataset has a standard split of 20 seen classes and 12 unseen classes.