SDM-Net: A Simple and Effective Model for Generalized Zero-Shot Learning

Shabnam Daghaghi[1] **Tharun Medini**[1] **Anshumali Shrivastava**[1,2]

[1]Department of Electrical and Computer Engineering, Rice University
[2]Department of Computer Science, Rice University

Abstract

Zero-Shot Learning (ZSL) is a classification task where some classes referred to as *unseen classes* have no training images. Instead, we only have side information 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 widespread cross-entropy loss. We introduce a novel Similarity Distribution Matching Network (SDM-Net) which is a standard fully connected neural network architecture with a non-trainable penultimate layer consisting of class attributes. The output layer of SDM-Net consists of both seen and unseen classes. To enable zero-shot learning, during training, we regularize the model such that the predicted distribution of unseen class is close in KL divergence to the distribution of similarities between the correct seen class and all the unseen classes. We evaluate the proposed model on five benchmark datasets for zero-shot learning, AwA1, AwA2, aPY, SUN, and CUB datasets. We show that, despite the simplicity, our approach achieves competitive performance with state-of-the-art methods in Generalized-ZSL setting for all of these 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 regimes and detecting object categories with no labeled data [Xiao et al., 2015, Socher et al., 2013, Xian et al., 2017, Zhang and 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 and 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 the 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 along with the semantic information for a different set of classes, referred to as seen classes. The key to solving zero-shot learning problem is to train a 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.

Where samples are images, in order to create a bridge between visual space (training data) and semantic attribute space (semantic information), some methods utilize embedding techniques [Palatucci et al., 2009, Romera-Paredes and 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 and Zisserman, 2014, Frome et al., 2013, Xian et al., 2016, Zhang and Saligrama, 2016, Al-Halah et al., 2016, Zhang and Shi, 2019, Atzmon and Chechik, 2018] and the others use semantic similarity between seen and unseen classes [Zhang and Saligrama, 2015, Fu et al., 2015b, 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 and 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 and Baghshah, 2016, Ye and Guo, 2017], and mapping both visual and semantic space into a joint embedding space [Akata et al., 2016, 2015, Simonyan and Zisserman, 2014, Frome et al., 2013, Xian et al., 2016, Zhang and Saligrama, 2016, Al-Halah et al., 2016].

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, Felix et al., 2018, Kumar Verma et al., 2018]. Although generative models boost the results significantly, it was argued that they are more difficult to train [Sutskever et al., 2015, Salimans et al., 2016]. Furthermore, training requires generation of a large number of samples followed by training on much larger augmented data which hurts their scalability.

A recent notable model called DCN [Liu et al., 2018], is also based on mapping visual features and semantic attributes to a common embedding. DCN minimizes cross-entropy on seen classes to learn their visual features. For the model to not ignore the unseen class attributes, it employs a regularization. DCN chooses to minimize the entropy of unseen classes. In essence, it is forcing the network not to predict uniform distribution (maximum entropy). Instead, it forces all the probability mass on one of the unseen classes (least entropy). While this entropy regularization is simple and remarkably improves network accuracy on unseen classes, we argue that it is sub-optimal. Consider an example where the correct class is a squirrel, and we have rats, mice, and several similar rodents in the unseen class. These unseen classes likely have attributes similar to the squirrel's attribute. However, forcing the network to concentrate the probability only on one class is likely to lose information. Worse, nothing stops the entropy loss to converge on an utterly wrong class, which also minimized the entropy.

To resolve these issues, we propose a Similarity Distribution Matching (SDM) regularizer, which enforces a complete distribution on the unseen class to match the distribution obtained from the semantic similarity of class attributes with the correct class. Therefore, SDM uses attribute information in a more explicit way. The regularization imposes a larger structure on the network. We show that SDM outperforms DCN by a significant margin and achieves competitive performance with state-of-the-art algorithms on ZSL benchmark datasets.

1.1 OUR CONTRIBUTION

We propose a simple, fully connected neural network architecture with unified (both seen and unseen classes together) cross-entropy loss. Our proposal differs from a standard neural network for supervised classification in two ways. The first difference is that in the proposed network, the final layer is fixed and non-trainable. The weight vectors for neurons in the last non-trainable layer are precisely the available semantic attributes. We argue that this standard architecture is no less powerful than the popular embedding models. The second difference is a novel loss function based on semantic similarity-based regularization.

In ZSL, due to the lack of training data for the unseen class, after minimizing any loss function over the training data, the classifier will always prefer seen classes over unseen classes. This is the main challenge of ZSL problem where for any given input, the predicted class will likely only come from the seen classes. We propose Similarity Distribution Matching (SDM) to regularize this minimization problem which enables training data from seen classes to also learn and even predict unseen classes.

In particular, we directly use attribute similarity information between the correct seen class and the unseen classes to create a regularizer. Among all classifiers with small training loss on the seen data, we prefer classifiers whose predicted probability distribution on unseen classes, matches the normalized similarity distribution. The similarity distribution is defined by the attribute similarity between the correct seen class and all the unseen classes. As a result of SDM, training instances for seen classes also serve as proxy training instances for the unseen class without increasing the training corpus. Thus SDM after simplification leads to a straightforward regularizer, which we argue is more informative than the recently proposed entropy regularizer [Liu et al., 2018]. SDM regularization, along with cross-entropy loss, enables a simple MLP network to tackle GZSL problem. Our proposed model achieves competitive performance with state-of-the-art methods in Generalized-ZSL setting on all five ZSL benchmark datasets.[1]

2 RELATED WORKS

The main goal of ZSL problem is to bridge the gap between visual features and semantic representations of unseen classes. Semantic representations are usually available in the form of word embeddings learned on text corpus or human annotation attributes. Some early ZSL methods utilized a two-step approach, [Lampert et al., 2009, Al-Halah et al., 2016] learn a probabilistic attribute classifier and then estimate class posteriors, [Norouzi et al., 2013] predict seen

[1] The code is available at https://github.com/shabnamD70/SDM-GZSL

class posteriors then take the convex combination of class label embeddings to project images into the semantic space. These two-step methods suffer from projection domain shift [Fu et al., 2015a]. On the other hand, recent works in ZSL directly learn an embedding between visual and semantic representations. [Akata et al., 2015, 2013] learn a bilinear compatibility function through structural SVM loss and ranking loss, respectively. DeViSE [Frome et al., 2013] utilizes a pairwise ranking loss to learn a mapping between visual space and semantic space. ESZSL [Romera-Paredes and Torr, 2015] introduces a simple analytical approach and utilizes square loss with L_2 norm regularization to learn the compatibility function between visual and semantic space.

Loss functions in embedding based models have training samples only from seen classes and there is no sample from unseen classes. 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].

The two most recent state-of-the-art discriminative GZSL methods, CRnet [Zhang and Shi, 2019] and COSMO [Atzmon and Chechik, 2018], 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.

Utilizing generative models is a totally different approach in GZSL setting. Given the semantic representation of classes, GANs aim to synthesize visual features and turn GZSl problem into a conventional classification problem. Xian et al. [2018] generates discriminative visual features through paring Wasserstein GAN [Ishaan et al., 2017] with classification loss. Sariyildiz and Cinbis [2019] extends this notion with gradient matching loss and learning an unconditional discriminator. Mishra et al. [2018] train a conditional Variational Auto-Encoder (cVAE) to generates samples from given semantic representations. Kumar Verma et al. [2018] follows a similar approach and adds a multivariate regressor to map the generated samples to the relevant semantic representation. Xian et al. [2019] employs both GAN and VAE with an additional discriminator to generate more discriminative features.

Our proposed model follows a discriminative framework to solve GZSL setting. We map visual features to semantic space, calculate similarity measures in semantic space and finally apply a softmax classifier. By utilizing SDM regularization, We efficiently implement all three components by a simple MLP network.

3 PROBLEM DEFINITION

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 correct classes for the test dataset $\mathcal{D}'$. 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 assumption and we aim at classifying samples in $\mathcal{D}'$ to either seen or unseen classes $\mathcal{S} \cup \mathcal{U}$.

4 PROPOSED METHODOLOGY

In the next few sections, we outline the specific components of our method. The proposed network architecture, the SDM regularization details, and the employed training strategy are presented.

4.1 NETWORK ARCHITECTURE

As Figure 1 illustrates our architecture, the architecture is a fully connected neural network that takes visual features and predicts the class using standard softmax and Similarity Distribution Matching (described later). The final softmax layer has one node for every class, both seen and unseen. To incorporate the semantic attributes of every class in the learning process, we replace the weight corresponding to each class node, in the softmax layer, with the given attributes and make this layer non-trainable or immutable during training. The penultimate layer of the network has the same size as the dimension of semantic embedding of the class.

4.1.1 Why is this Architecture Sufficient?

We contrast our approach with popular embedding based approaches for zero-shot learning and argue that the two approaches are equally powerful in terms of architecture.

The whole learning process is about finding the right compatibility score between the input visual features, call it v, and the class attributes, call it a. In embedding models, this score is computed via inner product between embedding of visual features and the embedding of attributes. Lets $f(v)$ denotes the embedding of v and $g(a)$ denotes embedding of a. Here, f and g are non-linear functions which are generally a neural networks. Thus, the compatibility between v and a, call it $C(v,a)$ is given by

$$C(v,a) = \langle f(v), g(a) \rangle$$

In our architecture, the score is softmax which is monotonic in the inner product of the penultimate embedding, which is a function of v (call it $f'(v)$), and the semantic attributed a. Thus, with our model we can write

$$C(v,a) = \langle f'(v), a \rangle$$

Since f' is non-linear neural network and can be as complex as we want, given f and g, we can always choose f' complex enough such that

$$\langle f'(v), a \rangle = \langle f(v), g(a) \rangle$$

As a result, the power of our architecture is no less than the power of standard embedding models. Instead, softmax is more natural in terms of modeling the class dependencies and particularly incorporating similarity distribution, which is one of our contributions. Unlike embedding models which are limited to modeling pairwise compatibility, softmax models the complete conditional distribution. We will be precisely needing the information of the complete joint distribution to propose a superior regularizer. Besides, the simple softmax and its standard probability interpretation will eliminate the need for fancy normalization which otherwise is a concern for embedding models.

4.2 SIMILARITY DISTRIBUTION MATCHING

In the ZSL problem, the network output nodes corresponding to unseen classes are always inactive during learning since cross-entropy loss cannot penalize unseen classes. Moreover, the available similarity information between seen and unseen attributes is never utilized explicitly.

We overcome this inherent bias by regularizing the network to reproduce a predetermined probability distribution on unseen classes where this probability is dictated by semantic similarity. We propose creating unseen probability distribution based on the similarity between semantic attributes. For each seen sample, we represent its relationship to unseen categories by obtaining the semantic similarity (dot-product) of its attribute with the attributes of unseen classes.

We then squash all these dot-product values by softmax to acquire probabilities (Equation 1). While training, we use this similarity distribution to regularize the classifier. In particular, we enforce the predicted probability distribution on the unseen class close to the prescribed similarity distribution.

In order to control the flatness of the unseen distribution, we utilize temperature parameter τ in softmax [Hinton et al., 2015]. Higher temperature results in flatter distribution over unseen classes and lower temperature creates a more ragged distribution with peaks on nearest unseen classes. The impact of temperature τ on unseen distribution is depicted in Figure 4.*left* for a particular seen class. SDM regularizer 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]. Below is the formal description of *temperature* softmax to produce similarity distribution of unseen class k for seen class i (unseen probability distribution):

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

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: $\mathbf{A} = \left[\mathbf{a}_1 \mid \mathbf{a}_2 \mid \cdots \mid \mathbf{a}_C\right]$. 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 unseen similarity distribution regularization factor, respectively.

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 $\mathbf{p}(\mathbf{x}_i) = \mathit{softmax}\left(\mathbf{A}^T g_{\mathbf{w}}(\mathbf{x}_i)\right)$ where $\mathbf{A}^T g_{\mathbf{w}}(\mathbf{x}_i)$ is a C-dimensional vector of all similarity scores for an input sample (Figure 1).

A natural choice to train such classifiers is the cross-entropy loss which we later show naturally integrates our idea of similarity distribution matching. Therefore, the optimization problem of our framework is as:

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

where λ is the weight decay regularization factor, and $\mathcal{L}(\mathbf{x}_i)$ is the cross-entropy loss over true probability distribution (L), regularized by cross-entropy over probability distribution based on semantic similarity (R) for each sample as shown below:

$$\mathcal{L}(\mathbf{x}_i) = (1-\alpha)L(\mathbf{x}_i) + \alpha R(\mathbf{x}_i) \tag{3}$$

where $\alpha \in [0,1]$ is SDM regularization parameter. Through R, the SDM regularizer, we want to regularize the overconfidence of the classifier toward seen classes and enrich the network with the ability to also identify unseen samples.

The regularizer term can be expanded to seen and unseen terms as follows (omitting i subscript for simplicity):

$$R(\mathbf{x}) = -\sum_{k\in\mathcal{S}} y^s_k \log(p^s_k) - \sum_{k\in\mathcal{U}} y^u_k \log(p^u_k) \tag{4}$$

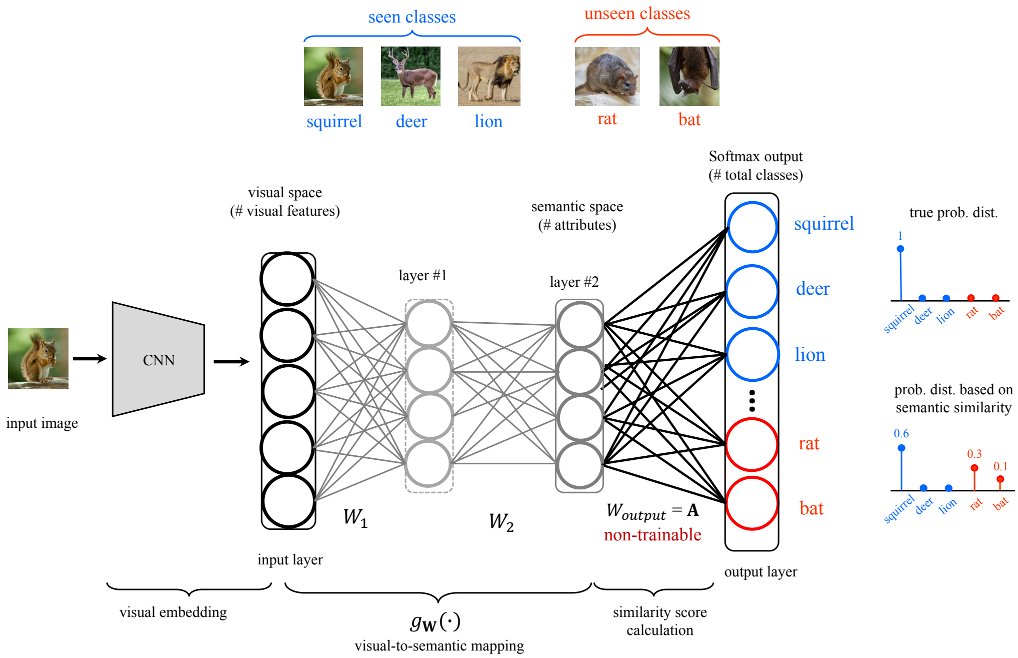


Figure 1: Overall workflow of the SDM classifier and architecture of the proposed MLP. Layers #1 and #2 provide the nonlinear embedding $g_{\mathbf{W}}(.)$ to map visual features to attribute space and their weights W_1, W_2 are learned by SGD. The output layer with non-trainable weights $\mathbf{A}$, basically calculates dot-products of semantic representation of the input and all class attributes simultaneously. Probability distribution based on semantic similarity is shown for a sample image from *squirrel* class, where the probability distribution assigned to unseen classes, *rat* and *bat*, is based on their semantic similarity with true class *squirrel*.

where y_k^s and y_k^u are probability distributions (based on semantic similarity) for seen and unseen class k, respectively. y_k^u is in fact the unseen similarity distribution (Equation 1) that the model attempts to match via SDM regularizer. The second term of Equation 4 is the KL divergence (constant entropy term on y_k^u is omitted) that matches the similarity distribution of unseen classes and their corresponding predicted probability distribution.

We observe that the SDM regularizer is a weighted cross-entropy on unseen class, which leverages similarity structure between attributes as opposed to uniform entropy function of DCN [Liu et al., 2018]. DCN and all prior works use uniform entropy as a regularizer, which does not capitalize on the known semantic similarity information between seen and unseen class attributes.

At the inference time, our proposed SDM 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.

5 EXPERIMENT

We conduct a comprehensive comparison of our proposed SDM model with the state-of-the-art discriminative methods for GZSL settings on five benchmark datasets (Table 1). Our model achieves competitive performance with the state-of-the-art methods on GZSL setting for all benchmark datasets.

5.1 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, place of habitat). The 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].

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

Table 1: Statistics of five ZSL benchmark datasets

SUN Attribute (SUN) [Patterson and Hays, 2012] is a fine-grained ZSL benchmark dataset that 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 that has 14340 images and 32 classes of different objects (e.g. airplane, 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.

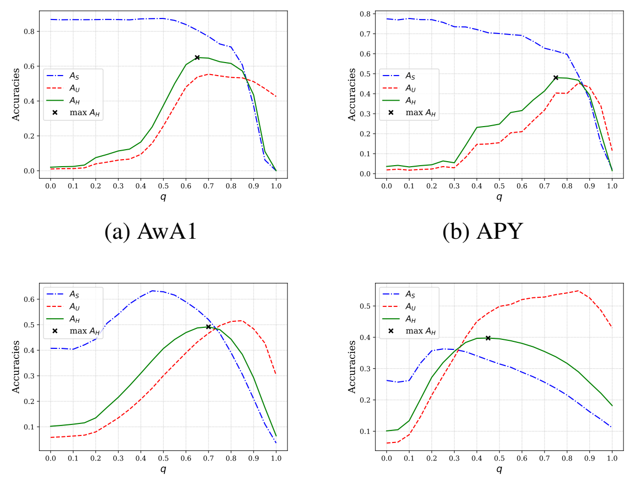
5.2 EVALUATION METRIC

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 the harmonic average of seen and unseen accuracy. This metric is chosen to encourage the network not to 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.

5.3 IMPLEMENTATION DETAILS

To evaluate SDM, 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 fairly 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 do not fine-tune the CNN that generates visual features unlike the model in Liu et al. [2018]. In this sense, our proposed model is also fast and straightforward to train.

We utilized Keras [Chollet, 2015] with TensorFlow backend [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* $\in \{64, 128, 256, 512, 1024\}$, $q \in [0, 1]$, $\alpha \in [0, 1]$, $\lambda \in \{0, 10^{-6}, 10^{-5}, 10^{-4}\}$, *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.



(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 four datasets. The maximum obtained harmonic accuracy is also marked by ($\times$). q provides a flexibility to trade-off between A_S, A_U, and A_H.

5.4 GENERALIZED ZERO-SHOT LEARNING RESULTS

To demonstrate the effectiveness of the SDM 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, our model achieves competitive performance with state-of-the-art GZSL methods on all five benchmark datasets and outperforms the state-of-the-art models 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 train-

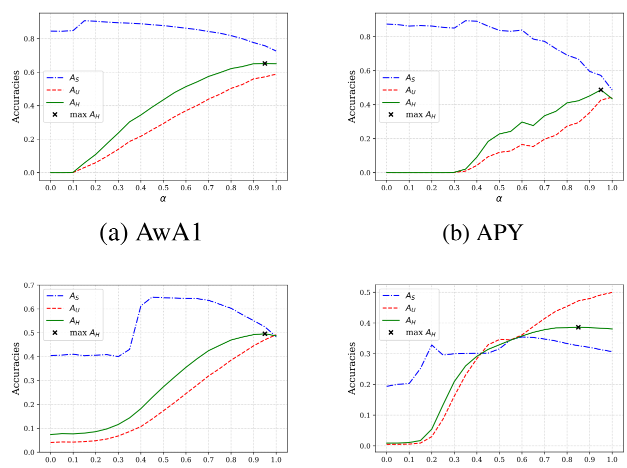
Method	AwA1 U	AwA1 S	AwA1 H	AwA2 U	AwA2 S	AwA2 H	aPY U	aPY S	aPY H	CUB U	CUB S	CUB H	SUN U	SUN S	SUN H
DAP [Lampert et al., 2009]	0.0	88.7	0.0	-	-	-	4.8	78.3	9.0	1.7	67.9	3.3	4.2	25.1	7.2
ALE [Akata et al., 2013]	16.8	76.1	27.5	14.0	81.8	23.9	4.6	73.7	8.7	23.7	62.8	34.4	21.8	33.1	26.3
SJE [Akata et al., 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
LATEM [Xian et al., 2016]	7.3	71.7	13.3	11.5	77.3	20.0	0.1	73.0	0.2	15.2	57.3	24.0	14.7	28.8	19.5
SSE [Zhang and Saligrama, 2015]	7.0	80.5	12.9	8.1	82.5	14.8	0.2	78.9	0.4	8.5	46.9	14.4	2.1	36.4	4.0
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
ESZSL [Romera-Paredes and Torr, 2015]	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6	12.6	63.8	21.0	11.0	27.9	15.8
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
f-CLSWGAN [Xian et al., 2018]	57.9	61.4	59.6	-	-	-	-	-	-	43.7	57.7	49.7	42.6	36.6	39.4
RN [Sung et al., 2018]	31.4	**91.3**	46.7	30.0	**93.4**	45.3	-	-	-	38.1	61.1	47.0	-	-	-
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
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 and 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 and Chechik, 2018]	52.8	80.0	63.6	-	-	-	-	-	-	44.4	57.8	50.2	44.9	37.7	**41.0**
CRnet [Zhang and 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
LFGAA+SA [Liu et al., 2019]	-	-	-	50.0	90.3	64.4	-	-	-	43.4	79.6	**56.2**	20.8	34.9	26.1
SDM-Net (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

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. Our model achieves highly competitive performance with state-of-the-art baselines despite its simplicity.

able layers, compared to CRnet with K mixture of experts followed by relation module with complex loss functions (pairwise).

Semantic similarity distribution employed in SDM 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 metric for GZSL. Assigned 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 four datasets. The flexibility provided by SDM 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 distribution matching and the trade-off knob, q.

Moreover, SDM alleviates overconfidence in seen classes and introduces information about unseen classes during the training phase. Figure 3 shows the impact of α on seen (A_S), unseen (A_U), and harmonic average (A_H) accuracies. The plots represent that conventional cross-entropy loss ($\alpha = 0$) results in almost zero unseen and harmonic average accuracies. This underscores the importance of similarity distribution regularization. As shown in Figure 3 for four datasets, not only unseen accuracy but also seen accuracy benefits from SDM, as the maximum of seen accuracy occurs at some α greater than zero which confirms the significance of probability distribution created by similarity values for cross-entropy loss in GZSL setting.



(a) AwA1 (b) APY

(c) CUB (d) SUN

Figure 3: Plots of seen (A_S), unseen (A_U) and harmonic average (A_H) accuracies versus α. The maximum obtained harmonic accuracy is also marked by ($\times$).

It should be noted that both AwA1 and aPY datasets (Figures 2.a, 2.b, 3.a ,3.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 and α, as shown in Figures (2.c 2.d 3.c, 3.d). However, the harmonic average curve still has the same behavior and possesses a maximum value at an optimal q and α.

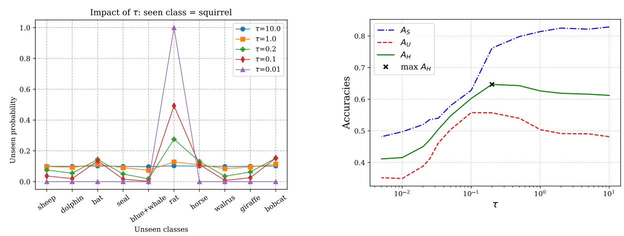


Figure 4: The impact of temperature parameter τ for AwA1 dataset. *left figure*: unseen probabilities (before multiplying q) produced by temperature softmax Equation (1) for various τ. *right figure*: accuracies versus τ for proposed SDM classifier. The optimal τ is 0.2 for AWA1 dataset where the unseen probability distribution is mostly focused on three unseen classes, *rat*, *bat*, and *bobcat*.

5.5 INTUITION

We illustrate the intuition with AwA1 dataset [Lampert et al., 2009]. Consider a seen class *squirrel*. We compute the 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 4.*left*). 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 for the classifier to classify unseen classes accurately. This problem was identified in Liu et al. [2018], where they proposed entropy-based regularization in the form of Deep Calibration Network (DCN). DCN uses cross-entropy loss for seen classes and regularizes 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 distribution matching implicitly regularizes the model in a supervised fashion. The similarity values naturally have information of how much certainty we want for a specific unseen class. We believe that this supervised regularization is the critical difference in why our model outperforms DCN with a significant margin.

5.6 ILLUSTRATION OF SIMILARITY DISTRIBUTION

Figure 4 shows the effect of τ and the assigned unseen distribution on seen, unseen and harmonic accuracies for AwA1 dataset. Small τ enforces q to be concentrated on the 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 AwA1 dataset as depicted in Figure 4.*right*. The impact of τ on the assigned distribution for unseen classes is shown in Figure 4.*left* when seen class is *squirrel* in AwA1 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 4.*right* where $\tau = 0.2$ is the optimal temperature. While in the extreme cases, when $\tau = 0.01$, distribution on unseen classes is 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.

6 CONCLUSION

We proposed a discriminative GZSL classifier with visual-to-semantic mapping and cross-entropy loss. During training, while SDM is trained on a seen class, it simultaneously learns similar unseen classes through probability distribution based on semantic similarity. We construct similarity distribution on unseen classes which allows us to learn both seen and unseen signatures simultaneously via a simple architecture. Our proposed similarity distribution matching 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. SDM offers a trade-off between seen and unseen accuracies and provides the capability to adjust these accuracies based on the particular application. We achieve competitive performance with state-of-the-art methods in GZSL setting, on all five ZSL benchmark datasets while keeping the model simple, efficient, and easy to train.

Acknowledgements

This work was supported by National Science FoundationIIS-1652131, BIGDATA-1838177, AFOSR-YIP FA9550-18-1-0152, ONR DURIP Grant, and the ONR BRC grant on Randomized Numerical Linear Algebra.

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)*, pages 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*, pages 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*, pages 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*, pages 5975–5984, 2016.

Atzmon and Chechik. Domain-aware generalized zero-shot learning. *CoRR*, abs/1812.09903, 2018. URL http://arxiv.org/abs/1812.09903.

Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *European Conference on Computer Vision*, pages 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*, pages 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*, pages 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*, pages 1043–1052, 2018.

François 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*, pages 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)*, pages 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, editors, *Advances in Neural Information Processing Systems 26*, pages 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*, pages 5337–5346, 2016.

Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-view zero-shot learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(11): 2332–2345, November 2015a. ISSN 0162-8828. doi: 10.1109/TPAMI.2015.2408354. URL http://dx.doi.org/10.1109/TPAMI.2015.2408354.

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*, pages 2635–2644, 2015b.

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.

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

Gulrajani Ishaan, A Faruk, A Martin, D Vincent, and C Aaron. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*. 2017.

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*, pages 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*, pages 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*, pages 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*, pages 2005–2015, 2018.

Yang Liu, Jishun Guo, Deng Cai, and Xiaofei He. Attribute attention for semantic disambiguation in zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

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*, pages 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*, pages 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, editors, *Advances in Neural Information Processing Systems 22*, pages 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*, pages 2751–2758. IEEE, 2012.

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

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

Mert Bulent Sariyildiz and Ramazan Gokberk Cinbis. Gradient matching generative networks for zero-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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, editors, *Advances in Neural Information Processing Systems 26*, pages 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

Ilya Sutskever, Rafal Jozefowicz, Karol Gregor, Danilo Rezende, Tim Lillicrap, and Oriol Vinyals. Towards principled unsupervised learning. *arXiv preprint arXiv:1511.06440*, 2015.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. CUB Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 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*, pages 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*, pages 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*, pages 5542–5551, 2018.

Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10275–10284, 2019.

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*, pages 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*, pages 7140–7148, 2017.

Fei Zhang and Guangming Shi. Co-representation network for generalized zero-shot learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7434–7443, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/zhang19l.html.

Hongguang Zhang and Piotr Koniusz. Zero-shot kernel learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 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*, pages 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*, pages 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*, pages 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*, pages 1004–1013, 2018.