Zero-Shot Learning for Image Classification

Deepesh V. Hada [1] **Jayant Priyadarshi** [2] **Kavita Wagh** [2]

Abstract

We consider the problem of zero-shot recognition: learning a visual classifier for a category with zero training examples, just using some embedding to represent the category and its relationship to the other categories, for which visual data are provided. The key to dealing with the unfamiliar or novel category is to transfer knowledge obtained from familiar classes to describe the unfamiliar class. Zero-shot learning is a promising learning method, in which the classes covered by the training instances and the classes we aim to classify are disjoint. It refers to a specific use case of machine learning (and therefore, deep learning) where we want the model to classify data based on very few or even no labeled examples, which means classifying data on the fly.

1. Introduction

In recent years, deep learning has achieved state-of-the-art performance across a wide range of computer vision tasks such as image classification. However, these deep learning methods rely on enormous amount of labeled data which is scarce for dynamically emerging objects. Practically, it is unrealistic to annotate everything around us, thus, making the conventional object classification methods infeasible.

We focus on the extreme case when there is no labeled data, *i.e.*, Zero-shot learning (ZSL), where the task is to recognize the novel categories' instances that have no labeled data available for training. ZSL assumes that the semantic label embeddings of both the seen and unseen classes are known apriori. ZSL, thus, learns to identify unseen classes by leveraging the semantic relationship between seen and unseen classes. It is important to note that the training and test classes are **disjoint** in ZSL.

ZSL algorithms generally project the seen and unseen class instances onto a latent space that is expected to be robust for learning unseen class labels. This latent space is shared by both seen and unseen categories and is aimed to be a mapping between images and their class embeddings. The learning is carried out with the help of semantic descriptors associated with each class. These approaches, thus, tend to learn a latent space that is aligned towards the semantic class embedding (Frome et al., 2013; Kodirov et al., 2017; Lampert et al., 2014; Romera-Paredes & Torr, 2015; Socher et al., 2013; Xian et al., 2016), on which the input data is transformed. This transformation of data onto the latent space, however, makes these ZSL approaches suffer from the **hubness problem**: the transformed data vectors become *hubs* for the nearby semantic class embeddings, leading to performance deterioration (Shigeto et al., 2015; Radovanović et al., 2010; Dinu & Baroni, 2014). To mitigate the hubness problem, other approaches (Zhang et al., 2016; Ba et al., 2015; Zhang & Saligrama, 2015; Shigeto et al., 2015; Paul et al., 2019) learn a latent space for recognizing the seen class labels by aligning the semantic class embeddings towards this latent space.

There exists an orthogonal approach, wherein artificial images (Reed et al., 2016) are generated to augment the training data. These synthetic images are usually generated from conditional generative models like conditional WGANs (Xian et al., 2017a). Xian et al. (2017a) improved image classification considerably, but having GAN-based loss functions, suffer from instability in training. Recent developments have then been made along this approach using Conditional VAEs (CVAEs). Schönfeld et al. (2019) proposed an aligned VAE model, CADA-VAE, which learns a latent embedding of image features and class embedding via aligned VAEs optimized with cross-alignment and distribution-alignment objectives, and subsequently trains a classifier on sampled latent features of seen and unseen classes. This model has achieved a significant performance boost in many datasets, including state-of-the-art results in the AWA and AWA2 datasets.

There has been another very different approach for zero-shot recognition using the semantic class embeddings and a **Knowledge Graph** (KG) that encodes the relationship of the novel category to some familiar categories. The first paradigm is to use implicit knowledge representations, *i.e.*, semantic embeddings, and hence, this is the same as the

[1]Department of CSA, Indian Institute of Science, Bangalore, India [2]Department of EECS, Indian Institute of Science, Bangalore, India. Correspondence to: Deepesh V. Hada <deepeshhada@iisc.ac.in>, Jayant Priyadarshi <jayantp@iisc.ac.in>, Kavita Wagh <kavitawagh@iisc.ac.in>.

first set of approaches. The alternative paradigm for zero-shot learning in these approaches is to use explicit knowledge bases or KGs, wherein one explicitly represents the knowledge as rules or relationships between objects. These relationships can then be used to learn zero-shot classifiers for new categories (Misra et al., 2017). There has been a combination of the two such that the model distills both the implicit knowledge representations (i.e. semantic embedding) and explicit relationships (i.e. knowledge graph) for learning **Graph Convolutional Network** (GCN) based visual classifiers of novel classes (Wang et al., 2018).

Also, irrespective of the latent space for transforming the data, there is an inherent **bias** in the model towards seen classes, which is referred to as the bias problem in literature. Due to this bias, the models generally perform poorly on unseen classes, and the predictions are primarily made in favour of the seen classes.

After considering these various approaches of GZSL for the task of image classification, we have implemented and modified **Semantically Aligned Bias Reducing** (SABR) *Generalized Inductive* ZSL (Paul et al., 2019), that gets rid of the hubness problem and the bias to a large extent and outperforms all the other models on **CUB** (Welinder et al., 2010), a fine-grained, medium scaled dataset. The modification in the model comes in the form of Spectral Normalization (Miyato et al., 2018).

2. Related Work

We discuss the state-of-the-art performing models in this section. The datasets which are predominantly used for training and evaluating ZSL models are **CUB** (Welinder et al., 2010), **SUN** (Xiao et al., 2010), **AwA** (Lampert et al., 2009), **AwA2** (Lampert et al., 2017) and **aPY** (Xian et al., 2017b).

Wang et al. (2018) take a different approach by constructing a new knowledge graph based on the Never-Ending Language Learning (NELL) for embeddings (Carlson et al., 2010) and Never-Ending Image Learning (NEIL) for images (Chen et al., 2013) datasets and WordNet for word embeddings and ImageNet for images.

Since they evaluate their performances only on these datasets, their work cannot be compared with the other other models which evaluate on the aforementioned datasets.

A significant progress in ZSL was achieved by (Schönfeld et al., 2019). Their approach takes feature generation one step further through a model (CADA-VAE) where a shared latent space of image features and class embeddings is learned through *modality-specific* aligned variational autoencoders, on which a SoftMax classifier is trained. CADA-VAE has achieved state-of-the-art results on *AwA* and *AwA2*.

Huang et al. (2018) proposed a model, GDAN, consisting of three components that perform GZSL, where the *Generator* represents the semantic $\rightarrow$ visual methods; the *Regressor* represents the visual $\rightarrow$ semantic methods; and the *Discriminator* represents metric learning. The *Regressor* is trained simultaneously with the CVAE using the cyclic consistency loss. GDAN has pulled off state-of-the-art results on the **SUN** (Xiao et al., 2010) dataset.

3. Task

On the basis of data available during the training phase, ZSL can also be divided into two categories: **Inductive** and **Transductive** ZSL. In inductive ZSL (Frome et al., 2013; Lampert et al., 2014; Kodirov et al., 2017; Romera-Paredes & Torr, 2015; Xian et al., 2016; Zhang et al., 2016; Annadani & Biswas, 2018; Xian et al., 2017a), we are provided with the labeled seen class instances and the semantic embedding of unseen class labels during training. In transductive ZSL (Verma & Rai, 2017; Ye & Guo, 2017), in addition to the labeled seen class data and the semantic embedding of all labels, we are also provided with the **unlabeled instances** of unseen classes data during the training time.

Here, we assume that we do NOT have any unlabeled inputs corresponding to the unseen classes, and hence, we'll be dealing with **Inductive** ZSL.

Further, we consider a more practical and realistic version of ZSL, the **Generalized Zero-shot learning** (GZSL) problem. In the **classical** ZSL, data emerges only from unseen classes at test time, *i.e.*, the performance of the algorithmic approach is solely judged on its classification accuracy on the novel unseen classes. In contrast, GZSL aims at maximizing performance on both seen and unseen classes, where the data during testing comes from both seen and unseen classes.

The task of GZSL is defined as follows:

$$D_s = \{(x, y, c(y)) \mid x \in X,\ y \in Y^s,\ c(y) \in C\} \quad (1)$$

is a set of training examples, consisting of image-features x, extracted by a CNN (ResNet-101), class labels y available during training and semantic class embeddings $c(y)$. The class-embeddings are 312-dimensional hand-annotated attribute vectors. Note that unlike the Transductive ZSL, no auxiliary training set is provided.

With conventional ZSL, the aim of the classifier during test time is to map,

$$X \rightarrow Y^u \quad (2)$$

However, in this work, we focus on the more realistic and challenging setup of GZSL where the aim is to learn a classifier that maps:

$$X \rightarrow Y^u \cup Y^s \quad (3)$$

4. Model

We first use a pre-trained deep network **Resnet-101** to extract features from the seen and unseen class images. For notational convenience, we'll denote x_i^s and x_i^u as 2048-dimensional feature vectors extracted from the pre-trained Resnet model. The model consists of four components with an adversarial setup. We'll define these components and their responsibilities now.

1. **Classifier (f_c)**: The classifier f_c, as the name suggests, learns to discriminate between the seen classes. A training instance, x_i^s is transformed by f_c to a one-hot encoding of its class label and then trained by minimizing the categorical cross-entropy loss:

$$L_C = -\frac{1}{N_s}\sum_{i=1}^{N_s} L(y_i^s, f_c(x_i^s)) \tag{4}$$

   where L is cross entropy loss between true and predicted labels of seen class instance x_s.

2. **Regressor (f_r)**: The regressor f_r preserves the semantic relationships among the seen class labels. It tries to ensure that the regressor output on the *embedding* of a seen instance, *i.e.*, $f_r(x_i^s)$, is closely related to the corresponding semantic embedding, $c(y_i^s)$. We use a similarity based cross-entropy loss between the predicted label embeddings of the regressor and the true semantic label embedding:

$$L_S = -\sum_{i=1}^{N_s} log\frac{exp(< f_r(x_i^s), c(y_i^s) >)}{\sum_{y^s \in S} exp(< f_r(x_i^s), c(y^s) >)} \tag{5}$$

   Here, $< f_r(x_i^s), c(y_i^s) >$ refers to the *similarity* between predicted label embedding, $f_r(x_i^s)$. The similarity function that we have used is a simple normalized cosine similarity. One could use some other similarity functions like the *Jaccard* similarity as well.

3. **Adversarial Pair:** In inductive GZSL setup, we do not have training instances of the unseen classes which could pose a problem in our objective. We follow the approach of Xian et al. (2017a), wherein we construct a *conditional generator* network that can generate artificial instances from the unseen classes.

   This conditional generator takes as input a random noise vector z and a class label embedding $c(y_s)$, and outputs an instance $\tilde{x}^s$ in the latent 2048-dimensional space.

   As we know the labels associated with each seen class training instance, we train the conditional generator using the *Wasserstein* adversarial loss defined by,

$$L_G^s = \mathbb{E}[D^s(x^s, c(y^s))] - \mathbb{E}[D^s(\tilde{x}^s, c(y^s))] - \lambda\,\mathbb{E}[(\,\left\|\nabla_{\hat{x}^s} D^s(\hat{x}^s, c(y^s))\right\| - 1)^2] \tag{6}$$

   where, D^s is the seen class conditional discriminator whose input is the seen class label embedding, $c(y^s)$, and the latent space instance, $\hat{x}^s = \alpha x^s + (1-\alpha)\tilde{x}^s$ with $\alpha \sim U(0,1)$ and λ is the gradient penalty coefficient (Gulrajani et al., 2017). Thus, the objective of the generator-discriminator pair is to,

$$\min_{G^s}\max_{D^s} L_G^s \tag{7}$$

Further, we want to encourage the generator network to synthesize feature vectors that are discriminative between seen classes and encode the semantic similarity between the label embeddings, which are precisely what the Classfier and Regressor do, respectively. Thus, the overall loss function for the generator-discriminator network is,

$$\min_{G^s}\max_{D^s} L_G^s + \beta(L_C + \gamma L_S) \tag{8}$$

The generator can now be used to synthesize 2048-dimensional feature vectors for the unseen classes. However, the generator can be overly biased towards the seen classes due to the training set that is presented to it, which is a stumbling block. This bias is mitigated using the principle of early stopping during training the generator.

5. Experiments and Results

TODO

Software and Data

The repository can be found here and has been implemented in Python 3.7.4 using PyTorch.

References

Annadani, Y. and Biswas, S. Preserving semantic relations for zero-shot learning, 2018.

Ba, J., Swersky, K., Fidler, S., and Salakhutdinov, R. Predicting deep zero-shot convolutional neural networks using textual descriptions, 2015.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., and Mitchell, T. M. Toward an architecture for never-ending language learning, 2010.

Chen, X., Shrivastava, A., and Gupta, A. Neil: Extracting visual knowledge from web data, 2013.

Dinu, G. and Baroni, M. Improving zero-shot learning by mitigating the hubness problem, 12 2014.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. Devise: A deep visual-semantic embedding model, 2013.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans, 2017.

Huang, H., Wang, C., Yu, P. S., and Wang, C. Generative dual adversarial network for generalized zero-shot learning, 2018.

Kodirov, E., Xiang, T., and Gong, S. Semantic autoencoder for zero-shot learning, 2017.

Lampert, C. H., Nickisch, H., Harmeling, S., and Weidmann, J. Animals with Attributes, 2009.

Lampert, C. H., Nickisch, H., and Harmeling, S. Attribute-based classification for zero-shot visual object categorization, 2014.

Lampert, C. H., Pucher, D., and Dostal, J. Animals with Attributes 2, 2017.

Misra, I., Gupta, A., and Hebert, M. From red wine to red tomato: Composition with context, 2017.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks, 2018.

Paul, A., Krishnan, N. C., and Munjal, P. Semantically aligned bias reducing zero shot learning, 2019.

Radovanović, M., Nanopoulos, A., and Ivanović, M. Hubs in space: Popular nearest neighbors in high-dimensional data, December 2010. ISSN 1532-4435.

Reed, S. E., Akata, Z., Schiele, B., and Lee, H. Learning deep representations of fine-grained visual descriptions, 2016.

Romera-Paredes, B. and Torr, P. H. S. An embarrassingly simple approach to zero-shot learning, 2015.

Schönfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., and Akata, Z. Generalized zero- and few-shot learning via aligned variational autoencoders, 2019.

Shigeto, Y., Suzuki, I., Hara, K., Shimbo, M., and Matsumoto, Y. Ridge regression, hubness, and zero-shot learning, 2015.

Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. Zero-shot learning through cross-modal transfer, 2013.

Verma, V. K. and Rai, P. A simple exponential family framework for zero-shot learning, 2017.

Wang, X., Ye, Y., and Gupta, A. Zero-shot recognition via semantic embeddings and knowledge graphs, 2018.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-UCSD Birds 200, 2010.

Xian, Y., Akata, Z., Sharma, G., Nguyen, Q. N., Hein, M., and Schiele, B. Latent embeddings for zero-shot classification, 2016.

Xian, Y., Lorenz, T., Schiele, B., and Akata, Z. Feature generating networks for zero-shot learning, 2017a.

Xian, Y., Schiele, B., and Akata, Z. Zero-shot learning - the good, the bad and the ugly, 2017b.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo, June 2010. ISSN 1063-6919.

Ye, M. and Guo, Y. Zero-shot classification with discriminative semantic representation learning, 2017.

Zhang, L., Xiang, T., and Gong, S. Learning a deep embedding model for zero-shot learning, 2016.

Zhang, Z. and Saligrama, V. Classifying unseen instances by learning class-independent similarity functions, 2015.