CS772A COURSE PROJECT

A survey of Zero Shot Learning

Shreesh Ladha - 12679
Shreyash Pandey - 12683

Abstract

There are around 30,000 human-distinguishable basic object classes and many more fine grained ones. A major barrier to progress in visual recognition is thus collecting training data for many classes. To counter this problem, a technique known as Zero Shot Learning has recently been introduced through which one is able to detect classes which were not part of the training set. In this project, we have tried to survey a variety of techniques within this area, describing their algorithm, strengths and weaknesses.

1 Introduction

A lot of work has been done in the area of detecting objects within images and research continues to happen in this area. One of the major bottlenecks here is that the number of different classes that the image could comprise of are huge in number. One way of detecting them is to include all of them during the training stages. However, there are millions of fine grained classes that would need to be added, making the process too computationally inefficient. The other way is to design algorithms that simulate how humans overcome this issue. A human being can detect the object in question eventhough it may be the first time they may be seeing it. We are able to perform this inference by drawing information about that object from a different source (like text) and then using that to attempt to identify the object. This method is essentially what is used in practice to detect unseen classes and is referred to as Zero Shot Learning (ZSL).

A ZSL model typically utilizes information from text corpora, images and their labels and maps them to a common semantic space. Such a semantic space could either be a word space or an attribute space. Attribute space is defined using attributes(usually binary) such as 'hasFur', 'hasTail', 'isBrown' etc. and are usually not preferred since manually tagging images with such attributes is not scalable and is inefficient. Incase of a word space, where the labels are already mapped to that space, a mapping is learnt from the data to project images into that space. Now during test time, the input image is mapped in the semantic space and then a nearest neighbour search or some other similarity metric is used to select the closest unseen class.

As part of our survey, we have selected six papers. We start off with covering the basic aspects of all these techniques, followed by a comparison with respect to their performance on some benchmark datasets, and finally concluding by critiquing on their strengths and weaknesses.

2 Section I

2.1 ConSE

The first technique, known as ConSE or "Zero-Shot Learning by Convex Combination of Semantic Embeddings" (ConSE), employs a straightforward approach to dealing with ZSL tasks. It uses a classifier to obtain semantic embedding of images by a convex combination of class label embedding vectors from the training set. The intuition is that the semantic embedding of an unseen image would be close to a weighted combination of the most likely seen classes:

$$
\begin{aligned}
f(x) &= \frac{1}{Z}\sum_{t=1}^{T} p(\hat{y_0}(\mathbf{x}, t \mid \mathbf{x}).s(\hat{y_0}(\mathbf{x}, t)) \\
\text{where } Z &= \sum_{t=1}^{T} p(\hat{y_0}(\mathbf{x}, t \mid \mathbf{x})),
\end{aligned} \tag{1}
$$

T here is a hyperparameter, $\hat{y_0}(\mathbf{x}, t)$ gives the t^{th} most probable label and $s(y)$ gives the semantic embedding of an image y. Finally, the prediction is obtained by finding the class nearest to the obtained semantic embedding.

2.2 HierSE

The next paper we studied was an extension of ConSE : "Zero-shot Image Tagging by Hierarchical Semantic Embedding" (HierSE). It improves upon ConSE and obtains better semantic embedding by extracting hierarchical structure defined in the WordNet. This is to ensure that labels with low/no occurrence in the vocabulary, which are of particular interest in ZSL, get reliable embedding vectors. It also creates its semantic space from Flickr tags as opposed to Wikipedia in ConSE. This is based on the motivation that Flickr might be a better source since their tags better capture the label's visual context. The embedding vectors now are obtained by :

$$
\begin{aligned}
f(x) &= \frac{1}{Z}\sum_{t=1}^{T} p(\hat{y_0}(\mathbf{x}, t \mid \mathbf{x}) s_{hi}(\hat{y_0}(\mathbf{x}, t)) \\
\text{where } s_{hi}(y) &= \frac{1}{Z_{hi}} \sum_{y' \in y \bigcup super(y)} w(y' \mid y) s(y) \\
Z_{hi} &= \sum_{y' \in y \bigcup super(y)} w(y' \mid y),
\end{aligned} \tag{2}
$$

Here $super(y)$ refers to the ancestors of a label obtained using WordNet and $w(y' \mid y)$ is a weight subject to exponential delay with respect to the minimal path length from y to y'. Prediction is performed in a similar manner as described above.

3 Section II

3.1 Novelty Detection through Cross Modal Transfer

The techniques that we discussed above are structured to work well only when the test set contains classes not seen during the training stage. However, this constraint would not hold true in real instances and assuming it would decrease the accuracies significantly. The next model in our report doesn't hold any such assumption and simultaneously operates on both seen and unseen classes during test time using a novelty detection approach.

Initially, as with most ZSL models, this paper tries to learn a mapping from the visual space to a semantic space. It achieves this by training a neural network to minimize the below objective function,

$$
J(\theta) = \sum_{y \in Y_s} \sum_{x^{(i)} \in X_y} \|w_y - \theta x^{(i)}\|^2. \tag{3}
$$

Here, w_y is the word vector corresponding to the class and x^i being the image. Y_s is the set of all seen classes and X_y is the set of images under a seen class. Its novelty detection algorithm is based on the principle that eventhough an unseen label might lie close to a seen label in the semantic space if the labels are semantically similar, it still doesn't lie on the manifold of seen labels which are tightly bounded together. Hence an outlier detection approach could be used at test time for identification of a seen/unseen label. In this regard, two approaches have been proposed.

The first outlier detection approach computes the marginal probability of an image : $P(x|X_y, w_y, F_y, \theta)$ as $P(f|F_y, w_y) = \mathcal{N}(f|w_y, \Sigma_y)$ where Σ_y is estimated from training data. The probability of an image belonging to the unseen class is chosen to be 1 if this marginal probability is below a certain threshold and 0 otherwise. The choice of the threshold determines the sensitivity of the model towards detection of seen/unseen class. If the threshold is low, then the majority of the classes would be detected as seen.

The second approach improves upon the first one by providing a probability of an image belonging to a certain class. For each point f in the semantic space, it defines a context set $C(f) \subseteq \mathcal{F}_s$ of k nearest neighbours in the training set of seen categories. Now using a probabilistic set distance (a measure of distance between two statistical objects) 'pdist' of each point from the points in $C(f)$, a local outlier factor is defined. If a point has a high relative pdist, then it would correspond to it being an outlier. Now during inference, if the label is detected as seen, a usual probabilistic classifier is used to predict the correct class. And similarly, in cases where the label is detected as unseen, an isometric gaussian distribution is assumed and classes are assigned based on their likelihood.

4 Section III

In all the above papers, a basic assumption is assumed of the source and target classes to be from the same distribution. This assumption does not hold true for many real world cases, especially in the present time, where data is being generated from many heterogeneous sources - in domains of texts, images and videos. The next class of papers specifically try to resolve this issue, also known as Projection Domain Shift(PDS), while performing Zero Shot Learning tasks.

4.1 Semantic Graph for Zero-Shot Learning

In this paper, a special absorbing Markov chain process is designed based on a semantic graph, and each unseen class is treated as an absorbing state. When a test image is incorporated into the semantic graph, the absorbing probabilities to each unseen class can be effectively computed; and zero-shot classification can be achieved by finding the class label with the highest absorbing probability.

This paper is of particular interest to us as it proposes a novel method of completely avoiding the PDS problem. The root cause of PDS is learning a mapping function that maps the low-level image features to a semantic space, and the effect is more pronounced when the seen and unseen classes vary significantly in their feature distributions. Instead of learning such a mapping function, the proposed method uses training data to learn a n-way probabilistic classifier in the visual feature space. The embedding space is only used for computing the semantic relatedness between seen and unseen classes.

The semantic graph is constructed with seen classes and unseen classes as its nodes. Each seen class node is connected to it's k-nearest neighbours in the semantic space. Each unseen class node is connected to it's k-nearest seen classes. A node is connected to its k nearest neighbours to keep the graph structure simpler and allow computationally efficient traversal. In the absorbing Markov chain formulation, the seen class nodes are transient, and the unseen class nodes are absorbing. The edge weights are calculated as the cosine similarity between two ends of the edge. The intuition is to arrive at the correct unseen class by traversing through seen classes that are the most similar to the correct unseen class. Zero Shot Learning task for any test image can be achieved by assigning to it the unseen label with maximum absorbing probability. The proposed model is linear in the number of test images and also has a closed form solution.

4.2 Transductive Multi-view Embedding for Zero-Shot Recognition and Annotation (TMV-BLP)

All the methods discussed so far use a single semantic space for ZSL. However, it makes sense to extract information from all semantic spaces, if multiple number of them exist. This paper suggests a method of learning a single latent embedding space using Canonical Correlation Analysis(CCA), i.e multiple semantic views are projected into this space to maximize their alignment. This procedure also reduces the bias that was earlier coming up as a result of learning a projection function directly from the low level image space. Since multiple spaces are now being aligned, the earlier bias is being reduced and this is in a way addressing the PDS issue. Note that this is more of a heuristic rather than a principled approach. The details of the algorithm are described below.

Consider three types of semantic views - the low level visual space , attribute space and word space denoted by $\mathcal{X}$, $\mathcal{A}$ and $\mathcal{V}$ respectively. Support vector regressors are used to learn the projections from $\mathcal{X}$ to other views for each dimension of the attribute and word vectors. CCA is a standard statistical technique for finding linear projections of two random vectors such that they are maximally correlated. A multiview CCA for E views(here $E = 3$) each denoted by Φ^i is performed as follows:

$$\begin{aligned} \min \quad & \textstyle\sum_{i,j=1}^{E} \quad Trace(W^i \Sigma_{ij} W^j) \\ = & \textstyle\sum_{i,j=1}^{E} \quad \parallel \Phi^i W^i - \Phi^j W^j \parallel_F^2 \\ \text{s.t.} \quad & [W^i]^T \Sigma_{ii} W^i = I \quad [\mathbf{w}_k^i]^T \; \Sigma_{ij} \mathbf{w}_l^j = 0 \\ i \neq j, k \neq l \quad & i,j = 1,\ldots,E \quad k,l = 1,\ldots,n_T, \end{aligned} \tag{4}$$

where W^i is the projection matrix which maps each view Φ^i ($\in \mathbb{R}^{n_T \times m_i}$) into the embedding space, $\mathbf{w}_k^i$ is the kth column of W^i and n_T the total number of images. Σ_{ij} is the covariance matrix between Φ^i and Φ^j.

During inference, a test image is projected into this space. Now since there are multiple representations of the same input in the shared space, a graph is considered for each view and using semi-supervised label propagation within and across the graphs, the correct label is chosen.

4.3 Unsupervised Domain Adaptation for Zero-Shot Learning (UDA)

ZSL in itself is not a domain adaption problem as the two domains have non-overlapping label spaces. But if we consider visual feature projection approach, that is projecting low-level image features to a semantic space, the learning of such projection function for unseen classes is a standard domain adaption problem, as both source domain (seen classes) and target domain (unseen classes) are projected into the same embedding space. This is an unsupervised domain adaptation as labels are not available in the target domain. The paper proposes a dictionary learning/sparse-coding approach to handle this problem - as sparse coding also involves learning a projection to a subspace. Embeddings for each visual feature vector are obtained as coefficients of the dictionary basis vectors.

This dictionary learning problem is formulated separately for the source and target domains. This is because for the source domain, the sparse coding coefficient vector is already known (in the form of word/attribute vectors). For each x_i in the source domain, of class label z_i's, the corresponding embedding in the semantic space is already known, and denoted by y_i. Thus the problem is different from conventional sparse coding in which both y_i and the dictionary D's have to be estimated. Considering that there are n_s seen classes, $X_s = [x_1, ..., x_{n_s}]$, and $Y_s = [y_1, ..., y_{n_s}]$ are defined accordingly. If we look to minimize the reconstruction error in the semantic space, and add a regularization term to favour a solution of smaller norm, the problem reduces to the following ridge regression problem:

$$\{D_t,\ Y_t\} = \min_{D_t, Y_t} \|X_t - D_t Y_t\|_F^2 + \lambda \|Y_t\|_1 s.t.\ \|d_i\|_2^2 \leq 1, \tag{5}$$

The sparse coding formulation for the target domain is the conventional one, as for a given $X_t = [x_1, ..., x_{n_t}]$, both $Y_t = [y_1, ..., y_{n_t}]$ and D_t are unknown. But this doesn't constrain D_t to be meaningful in terms of it capturing a learned representation of a suitable embedding space. To ensure the validity of this model, two important regularization terms are added in the cost function.

- Adaptation regularisation constraint: The D_t learned from the unseen target classes should be similar to D_s learned from the seen source classes. This constraint ensures that D_t is also a semantic dictionary that projects target data into the same semantic space as D_s.

- Visual-semantic similarity constraint: This is to ensure the closeness of learned coefficient vector y_i to its true class label z_t^i, embedded in the semantic embedding space as p_t^i. This is done by training a classifier on all unseen classes, and the probability of x_i being the j-th unseen class is used as weight w_{ij}. These weights are found using the IAP approach[8]. So for high w_{ij}, the distance between y_i and p_t^j would have to be minimized even more, resulting in enforced semantic similarity for visually similar classes. This constraint ensures that their approach is not prone to PDS.

The final optimization problem is stated below:

$$\{D_t, Y_t\} = \min_{D_t, Y_t} \|X_t - D_t Y_t\|_F^2 + \lambda_1 \|D_t - D_s\|_F^2 + \lambda_2 \sum_{i,j} w_{ij} \|\mathbf{y}_i - \mathbf{p}_j^t\|_2^2 + \lambda_3 \|Y_t\|_1 \ s.t. \ \|d_i\|_2^2 \leq 1. \quad (6)$$

The objective function in this problem is convex for both D_t and Y_t separately, but not simultaneously. The paper proposes an alternating optimization method to solve it. Once the sparse coding coefficients Y_t is estimated, the Zero-Shot classification task can now be performed by either a nearest neighbour (NN) approach in the semantic space, or a semi-supervised label propagation (LP) framework.

5 Experiments

We have performed two comparisons. The first comparison has been performed between between ConSE and HierSE. Note that we implemented ConSE ourselves since its code was unavailable, while we executed the codes for other methods to confirm the results. We have reported hit@1 and hit@10 accuracies for both the methods. A more detailed table is available in the mid-term report. The accuracies have been reported on ImageNet 2011 1k dataset. Note that these test classes are completely disjoint from the training ones and the test images are visually and semantically similar to the images from the training set. The semantic space was created using GloVe word vectors. As is evident from the table, HierSE significanlty outperformed ConSE. This was because of its particular choice of semantic space, tuned for the given experiment and the use hierarchical semantic embeddings. However, since this method relied on such a hierarchy of the dataset to create embeddings, it is not generalizable and hence was not considered for other comparisons.

Method	**hit@1**	**hit@10**
ConSE	14.5	31.5
HierSE	17.8	50.9

Table 1: Comparison between ConSE and HierSE

Method	**A**	**W**	**Accuracy**
ConSE	-	✓	35.1
Semantic Gr.	-	✓	43.1
Semantic Gr.	✓	-	49.5
TMV-BLP	✓	✓	47.1
UDA	✓	-	47.5
UDA	✓	✓	49.7

Table 2: Summary of Results (A : Attribute space, W : Word space)

The second comparison has been performed across methods that try to resolve the PDS issue. The dataset used is the Animals with Attributes dataset, which is a popular dataset used in many domain shift tasks. The training was performed for 40 classes and testing over 10 classes completely left out during the training procedure. The utilization of the attribute and the word space was based on how the particular method inherently had structured its model. ✓ represents the spaces that were used for learning the semantic space. As one can see, since ConSE did not address the PDS issue, the accuracies it obtained were the worst of the lot. Unsupervised domain adaptation, which handled the PDS approach in the most principled fashion outperformed all other methods with an accuracy of 49.7%. Note that this accuracy(fir UDA) shoots up to 75% simply by using CNN features instead of the low level features.

6 Conclusion

We discussed six different techniques for performing Zero-Shot Learning. The first two in the list were ConSE and HierSE. These two were very basic techniques for ZSL and their strength lied in their simplicity and efficiency of implementation. HierSE relied on a presence of a hierarchical structure of dataset and hence wasn't generalizable. The next technique used a novelty detection approach to filter seen/unseen classes and classify using respective classifiers. The final three were specifically aimed at resolving the PDS issue which may occur more frequently than one may expect in today's scenario. Semantic graph completely avoided it and was very efficient in its implementation. Transductive multi-view embedding allowed information extraction from multiple semantic views, which in turn addressed the PDS issue. Unsupervised domain adaptation, provided the most principled approach to handling PDS through specific optimizing contraints, effects of which were visible in the accuracies. To build a complete model for ZSL, one may draw upon the specific novelties in all the above approaches and combine them to build a more robust model. Building a model that extracts hierarchical information if present to learn embeddings, extracts information through multiple views, uses the approach of UDA for learning the projections and finally implements an outlier detection at the end, seems like a complete model, which we wanted to try out but couldn't, owing to a shortage of time. Nevertheless, we believe we did a comprehensive survey, covering the breadth of various methods and issues involving ZSL. We learnt about how ZSL is performed in general and how this same problem can be formulated in different ways to handle certain specific issues. Also, since feature projection was identified as a domain adaptation problem, standard techniques that involve kernel learning, can be tried for countering it. One more issue, however, is gaining traction in recent times of that of hubness. Hubness arises in high dimensions where a same object is detected as the nearest neighbour for any query object. Reader may refer to [7] for details about this issue.

References

[1] Norouzi, Mohammad, et al. "Zero-shot learning by convex combination of semantic embeddings." *arXiv preprint arXiv:1312.5650* (2013).

[2] Li, Xirong, et al. "Zero-shot image tagging by hierarchical semantic embedding." *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 2015.

[3] Socher, Richard, et al. "Zero-shot learning through cross-modal transfer." *Advances in neural information processing systems.* 2013.

[4] Fu, Zhen-Yong, Tao Xiang, and Shaogang Gong. "Semantic Graph for Zero-Shot Learning." *arXiv preprint arXiv:1406.4112* (2014).

[5] Fu, Yanwei, et al. "Transductive multi-view embedding for zero-shot recognition and annotation." *Computer Vision–ECCV 2014. Springer International Publishing*, 2014. 584-599.

[6] Kodirov, Elyor, et al. "Unsupervised Domain Adaptation for Zero-Shot Learning." *Proceedings of the IEEE International Conference on Computer Vision.* 2015.

[7] Shigeto, Yutaro, et al. "Ridge Regression, Hubness, and Zero-Shot Learning." *Machine Learning and Knowledge Discovery in Databases. Springer International Publishing*, 2015. 135-151.

[8] Lampert, Christoph H., Hannes Nickisch, and Stefan Harmeling. "Attribute-based classification for zero-shot visual object categorization." *Pattern Analysis and Machine Intelligence, IEEE Transactions on 36.3*, (2014): 453-465.