A survey of Zero Shot Learning

Shreesh Ladha, Shreyash Pandey

IIT Kanpur

Introduction

1. Zero Shot Learning is the ability to detect classes not part of the training procedure
2. 30,000 human-distinguishable basic object classes - major barrier is thus collecting training data for many classes
3. Model human's ability to identify unseen objects

General Methodology

1. An unseen class related to a seen class by representing both in a semantic embedding space
2. Test image is projected to the semantic space (using regression or classification), and a similarity measure with each unseen class is used for prediction
3. This method is prone to the Projection Domain Shift (PDS) problem which arises because the projection function learnt from source domain is applied to target domain without any adaptation

ConSE

1. Semantic embedding of an unseen image is being obtained using a weighted combination of the most likely seen classes
2. $f(x) = \frac{1}{Z}\Sigma p(\hat{y}_0(\mathbf{x}, t \mid \mathbf{x}).s(\hat{y}_0(\mathbf{x}, t))$
3. Prediction :
   $\bar{y}_1(x, 1) = \arg\max_{y' \in \mathcal{Y}} cos(f(x), s(y'))$

HierSE

1. Builds upon ConSE to obtain better semantic embeddings by extracting hierarchical structure defined in the WordNet
2. Ensures that labels with low/no occurrence in the vocabulary get reliable embedding vectors
3. $f(x) = \frac{1}{Z}\Sigma p(\hat{y}_0(\mathbf{x}, t \mid \mathbf{x}) s_{hi}(\hat{y}_0(\mathbf{x}, t))$
4. Here $s_{hi}(y) = \frac{1}{Z_{hi}}\Sigma w(y' \mid y)s(y)$

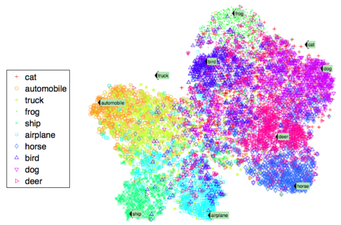

Fig. 1: Visualization of the semantic word space

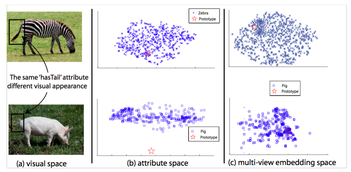

Fig. 2: Projection Domain shift problem

Novelty Detection using Cross Modal Transfer

1. All above models fail if the test set contains both seen and unseen classes
2. Simultaneously operates on both seen and unseen classes during test time using a novelty detection approach
3. Outlier detection at test time is based on the property that an image from an unseen class won't be very close to the existing training images but will still be roughly in the same semantic region

Transductive Multi-View Embedding

1. Counters PDS issue by aligning different semantic views using CCA to a shared embedding space
2. Unlike other methods, exploits information from multiple semantic representations
3. A graph is constructed from the projection of each view in the embedding space
4. Label prediction is performed using semi-supervised label propagation from the prototypes to the target data points within and across the graphs

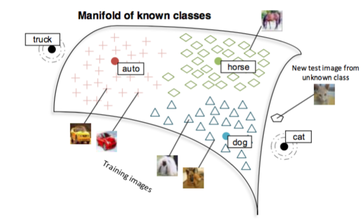

Fig. 3: Overview of the cross-modal zero-shot model

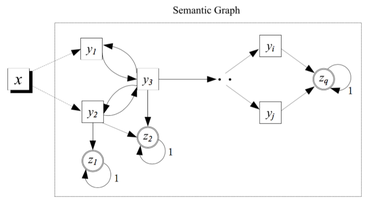

Fig. 4: ZSL via semantic graph

ZSL using Semantic Graph

1. Models semantic relationships between both seen and unseen classes in the form of a graph
2. Avoids PDS by learning an n-way classifier in the visual feature space, and using the embedding space only to compute semantic relatedness between seen and unseen classes
3. An absorbing Markov chain process is designed in which unseen classes are treated as absorbing states and for a test image, ZSL classification is achieved by finding the class label with highest absorbing probability

Unsupervised Domain Adaptation

1. Identifies visual feature projection as an unsupervised domain adaptation problem and proposes a dictionary learning approach to solve this problem
2. Semantic embeddings are obtained as sparse coding coefficients and Dictionary learning is performed separately for source and target domains
3. $D_s = \min_{D_s} \|X_s - D_s Y_s\|_F^2 + \lambda \|D_s\|_F^2$
   s.t.$\|d_i\|_2^2 \leq 1$
4. $\{D_t, Y_t\} = \min_{D_t, Y_t} \|X_t - D_t Y_t\|_F^2 + \lambda_1 \|D_t - D_s\|_F^2 + \lambda_2 \Sigma w_{ij} \|\mathbf{y}_i - \mathbf{p}_j^t\|_2^2 + \lambda_3 \|Y_t\|_1 s.t. \|d_i\|_2^2 \leq 1$
5. After estimating the sparse coding coefficients (embeddings) Y_t, ZSL classification is performed by a NN search in the semantic space

Experiments

Datasets used is Animals With Attributes(AwA) with the semantic word space created using wikipedia articles

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 1: Summary of Results

Dataset used is Imagenet 2011 with semantic space for ConSE created using GloVe vectors while for HierSE using flickr tags

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

Table 2: Comparison between ConSE and HierSE

References

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

[2] Li, Xirong, et al. "Zero-shot image tagging by hierarchical semantic embedding." (2015)

[3] Socher, Richard, et al. "Zero-shot learning through cross-modal transfer." (2013)

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

[5] Fu, Yanwei, et al. "Transductive multi-view embedding for zero-shot recognition and annotation." (2014)

[6] Kodirov, Elyor, et al. "Unsupervised Domain Adaptation for Zero-Shot Learning." (2015)