A Stduy about Zero Shot Learning

Abstract

This study contains **1).** The state of the art of Zero-Shot Learning(*ZSL*). **2).** Benchmark and evaluation metric for ZSL. **3).** Insights and proposals for our stduy.

Contents

1. Task Formulation

- *Tranfer Learning/Domain Adaption* (Goodfellow et al., 2016): Use learned feature in one setting (i.e., distribution P_1) to improve generalization in another setting (say distribution P_2).
- *One-shot Learning*: A extreme form of transfer learning.
- *Zero-shot Larning*: Compare to tradition learning scenario that needs inputs $\mathbf{x}$ and targets $\mathbf{y}$, zero-shot learning must need **side information** exploited during training, that is the task T. The model is trained to estimate the conditional distribution $p(\mathbf{y}|\mathbf{x}, T)$.

- Side information includes: Attributes, WordNet, detailed visual descriptions and its deep representations, human gaze and its embeddings.

1.1. Zero Shot Learning(ZSL)

Gvien $\mathcal{S} = \{(x_n, y_n), n = 1...N\}$, learn $f : \mathcal{X} \rightarrow \mathcal{Y}$ by minimizing the regularized empirical risk

$$\frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n, \mathbf{W})) + \Omega(\mathbf{W})$$

For Zero Shot Learning, $\mathcal{Y}^{tr} \cup \mathcal{Y}^{ts} = \Phi$; for Generalized Zero Shot Learning, $\mathcal{Y}^{ts} \subseteq \mathcal{Y}^{tr}$, *i.e.*, test image can be labeled with both seen and unseen classes.

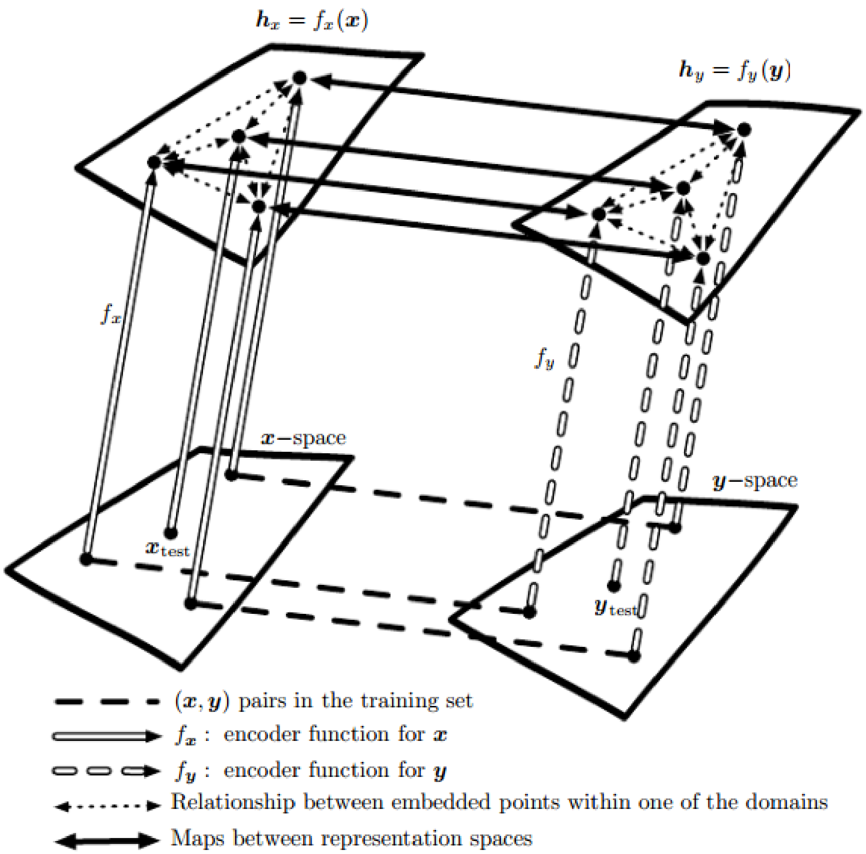


Figure 1: The Method for ZSL. Class in **y**-space is embeded to another latent space $\mathbf{h}_y$. There is internal structure between classes! (May be hierarchy or overlapping!) ZSL aim to learn word and image representation and the relations between them. The relation may be learned by end-to-end training/expolore similarity between the structure og two space.

2. Methodolity

- Compatibility Learning:
  - Linear: ALE, DEVISE, SJE, ESZSL, SAE
  - Non-linear: LATEM, CMT
- Two-stage Inference/Learn Intermediate Attribute Classifier: DAP
- Unseen is a Mixture of Seen: SYNC, CONSE

2.1. Sythetic classifier(SYNC)

Synthesized Classifiers for Zero-Shot Learning (Changpinyo et al., 2016). This paper makes lots of simplication.

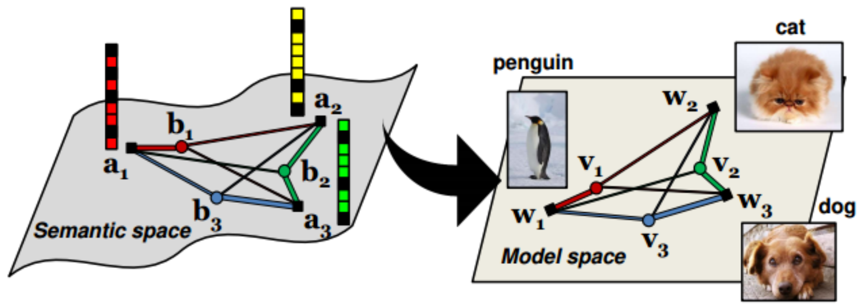


Figure 2: How to align two representation space? Embedding semantic space into image feature space.

The final unified Structure Loss with best performance of (Changpinyo et al., 2016) is:

$$\min_{\{\boldsymbol{v}_r\}_{r=1}^{\mathsf{R}}, \{\beta_{rc}\}_{r,c=1}^{\mathsf{R,S}}} \sum_{c=1}^{\mathsf{S}} \sum_{n=1}^{\mathsf{N}} \ell(\boldsymbol{x}_n, \mathbb{I}_{y_n,c}; \boldsymbol{w}_c)$$

$$+ \frac{\lambda}{2} \sum_{c=1}^{\mathsf{S}} \|\boldsymbol{w}_c\|_2^2 + \eta \sum_{r,c=1}^{\mathsf{R,S}} |\beta_{rc}| + \frac{\gamma}{2} \sum_{r=1}^{\mathsf{R}} (\|\boldsymbol{b}_r\|_2^2 - h^2)^2,$$

$$\text{s.t.} \quad \boldsymbol{w}_c = \sum_{r=1}^{\mathsf{R}} s_{cr} \boldsymbol{v}_r, \quad \forall c \in \mathcal{T} = \{1, \cdots, \mathsf{S}\},$$

- The objective varible is $\mathbf{v}_r$, that is, phantom base code for classifier.
- x_n is hand-crafted shallow feature or deep feature extracted from GoogleNet.
- After read the code, I find β_{rc} is constant and plays no role. As explained by the author, he assume the phantom base code $\mathbf{b}_r$ can be made from existed human labeled attributes or word2vec feature matrix $\mathbf{a}_c$, that is $\begin{cases} \mathbf{b}_r & = \sum_{c=1}^{\mathsf{S}} \beta_{rc} \mathbf{a}_c \\ \mathbf{b}_r & = \mathbf{a}_r \end{cases}$, which means number of phantom

code bases is exactly equal to number of classes.[1]

- For the constrains contains a similarity s_{cr} is designed as $s_{cr} = \frac{exp(-d(\mathbf{a}_c, \mathbf{b}_r))}{\Sigma_{r=1}^{\mathsf{R}} exp(-d(\mathbf{a}_c, \mathbf{b}_r))}$. The author tuned a lot, tried many hyperparameter. There are more hyperparameters and (Xian et al., 2017) suspect its performance partially comes from tuning attribute embedding from word2vec.
- The constrains can be explained as enforcing similarity structure in semantic space to feature space. The author explains that it is analytical solution of Laplacian eigenmaps that minimize distortion error $\min_{\mathbf{w}_c, \mathbf{v}_r} \|\mathbf{w}_c - \Sigma_{r=1}^{\mathsf{R}} s_{cr} \mathbf{v}_r\|_2^2$.
- In fact, this model is non-convex but not complex. Use SGD, heuristic initialization and simple toolbox(The backward gradient is written by hand) to solve this simple model.

2.2. *Atrribute Learning Embedding(ALE)*

Final lose function is (Akata et al., 2016):

$$R(\mathcal{S}; W, \Phi) = \frac{1}{N} \sum_{n=1}^{N} \frac{\beta_{r_\Delta(x_n, y_n)}}{r_\Delta(x_n, y_n)} \sum_{y \in \mathcal{Y}} \max\{0, \ell(x_n, y_n, y)\} \tag{12}$$

where

$$r_\Delta(x_n, y_n) = \sum_{y \in \mathcal{Y}} \mathbb{1}(\ell(x_n, y_n, y) > 0) \tag{13}$$

It will enforce correct labels to rank higher specially desined for classification task.

[1]The dimension of arttribute is often more than number of classes. According to low-rank hypothysis/regularization, we can explore dimension/structure of phantom code base to be less than number of classes.

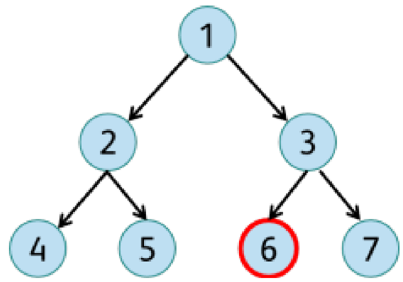


Figure 3: Class Hierarchical Embedding: *e.g.* $\phi^{\mathcal{H}}(y)|_{y=6} = [1, 0, 1, 0, 0, 1, 0]$

2.3. Future Work

- imagenet hierachy 2-hop 3-hop
- why ZLS generalize poorly: unseen image embedding close to existed seen embedding.
- class unbalance
- active learning

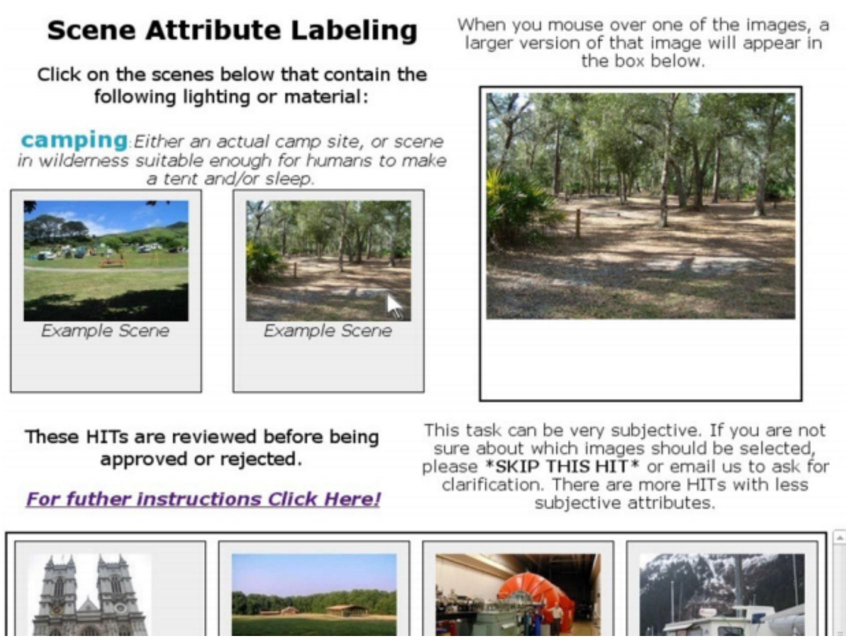
3. Benchmark

Zero-Shot Learning - The Good, the Bad and the Ugly (Xian et al., 2017).

3.1. Dataset

Figure 4: Example images from a-Pascal (top row) and a-Yahoo (bottom row). Images in a-Pascal and a-Yahoo are from disjoint categories.

Animals with Attributes (AWA) – Most widely used non-fine-grained dataset in literature



SUN Attributes – Most widely uses fine-grained datasets

3.2. Metric

- *Proposed Split*: 101-layerd ResNet pretrained on ImageNet 1K, and this 1K classes should not exist in $\mathcal{Y}^{ts}$.

Dataset	Classes $\mathcal{Y}^{tr}$	Classes $\mathcal{Y}^{ts}$
SUN	580+65	72
CUB	100+50	50
AWA	27+13	10
aPY	15+5	12
ImageNet	800+200	500/1K/5K[2]

- *Evaluatioin Critetia*: First calculate top-1 accuracy for each class, then take average on **classes**.

Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C., 2016. Label-embedding for image classification. IEEE transactions on pattern analysis and machine intelligence 38 (7), 1425–1438.

Changpinyo, S., Chao, W.-L., Gong, B., Sha, F., 2016. Synthesized classifiers for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5327–5336.
URL https://arxiv.org/abs/1603.00550

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, http://www.deeplearningbook.org.

Xian, Y., Schiele, B., Akata, Z., 2017. Zero-shot learning - the good, the bad and the ugly. CoRR abs/1703.04394.
URL http://arxiv.org/abs/1703.04394

[2]Make sure to be far away from training classes, *i.e.*, 2-hops/3-hops far away. Test classes can be most-populated classes/least populated classes(containing few images in this class).