Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly

Article written by Yongqin Xian, Christoph H. Lampert, Bernt Schiele and Zeynep Akata in September 2020

Presented by Vassily Baranov

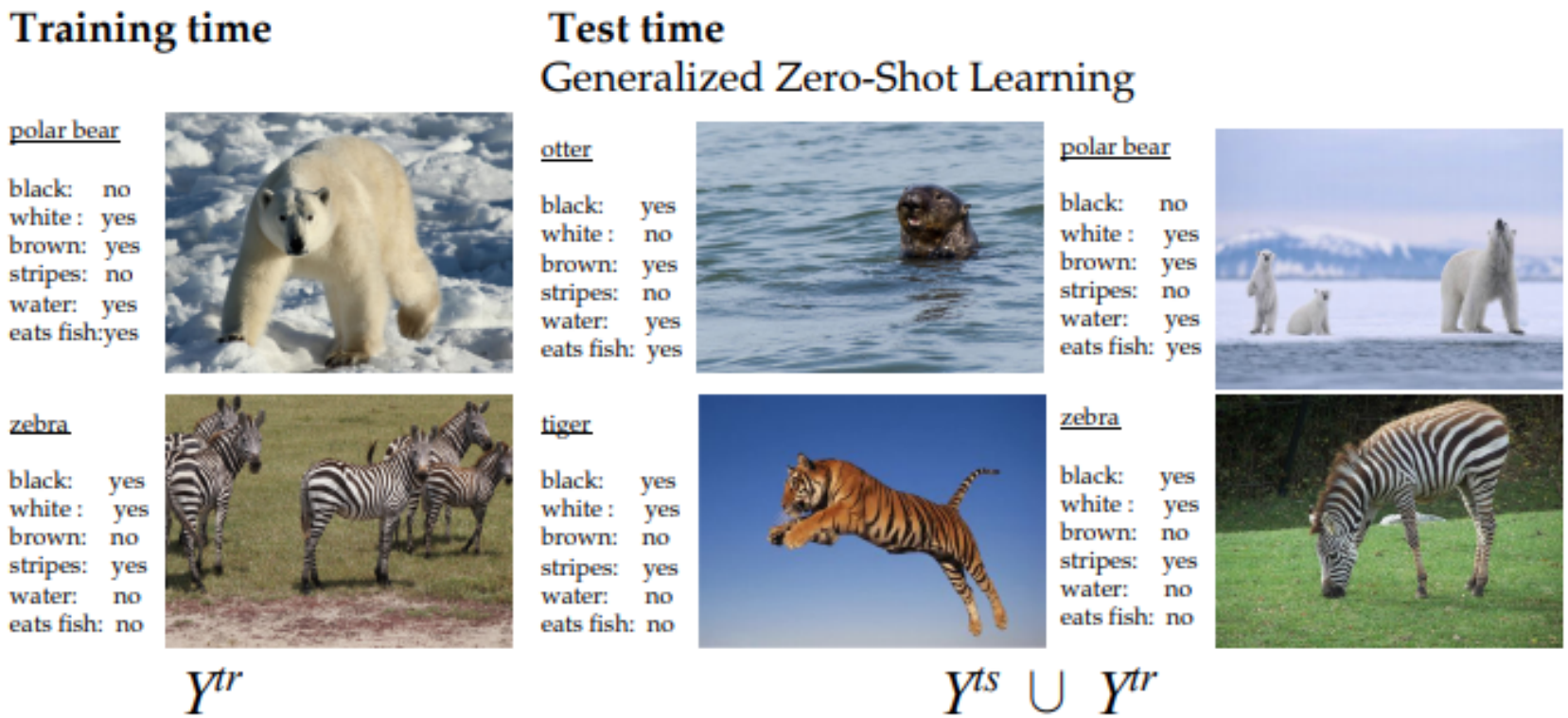
October 27, 2020

Content

- Introduction
- Evaluated methods
- Datasets
- Evaluation protocol
- Experiments
- Conclusion

Introduction

- Zero-shot learning aims to recognize objects whose instances may not have been seen during training.



Evaluated methods

- Given a training set $\mathcal{S} = \{(x_n, y_n), n = 1 \dots N\}$, with $y_n \in \mathcal{Y}^{tr}$ belonging to training classes, the task is to learn $f : \mathcal{X} \to \mathcal{Y}$ by minimizing the regularized empirical risk:

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

- The mapping $f : \mathcal{X} \to \mathcal{Y}$ from input to output embeddings is defined as:

$$f(x; W) = \underset{y \in \mathcal{Y}}{\text{argmax}} F(x, y; W)$$

Learning Linear Compatibility

- Deep Visual Semantic Embedding (DEVISE) and Structured Joint Embedding (SJE) use bi-linear compatibility function to associate visual and auxiliary information:

$$F(x, y; W) = \theta(x)^T W \phi(y)$$

where $\theta(x)$ and $\phi(y)$, i.e. image and class embeddings, both of which are given.

- DEVISE uses pairwise ranking objective that is inspired from unregularized ranking SVM:

$$\sum_{y \in \mathcal{Y}^{tr}} \left[\Delta\left(y_n, y\right) + F\left(x_n, y; W\right) - F\left(x_n, y_n; W\right)\right]$$

where $\Delta\left(y_n, y\right)$ is equal to 1 if $y_n = y$, otherwise 0. The objective function is convex and is optimized by Stochastic Gradient Descent.

Learning Linear Compatibility

- SJE gives the full weight to the top of the ranked list and is inspired from the structured SVM:

$$\left[\max_{y \in \mathcal{Y}^{tr}} \left(\Delta\left(y_n, y\right) + F\left(x_n, y; W\right)\right) - F\left(x_n, y_n; W\right)\right]$$

  The prediction can only be made after computing the score against all the classifiers, i.e. so as to find the maximum violating class, which makes SJE less efficient than DEVISE

Learning Linear Compatibility

- ESZSL applies a square loss to the ranking formulation and adds the following implicit regularization term to the unregularized risk minimization formulation:

$$\gamma \|W\phi(y)\|^2 + \lambda \left\|\theta(x)^T W\right\|^2 + \beta \|W\|^2$$

- SAE also learns the linear projection from image embedding space to class embedding space, but it further constrains that the projection must be able to reconstruct the original image embedding. Similar to the linear auto-encoder, SAE optimizes the following objective:

$$\min_W \left\|\theta(x) - W^T\phi(y)\right\|^2 + \lambda \|W\theta(x) - \phi(y)\|^2$$

Learning Nonlinear Compatibility

- Latent Embeddings (LATEM) and Cross Modal Transfer (CMT) encode an additional non-linearity component to linear compatibility learning framework.
- LATEM constructs a piece-wise linear compatibility:

$$F(x, y; W_i) = \max_{1 \leq i \leq K} \theta(x)^T W_i \phi(y)$$

where every W_i models a different visual characteristic of the data and the selection of which matrix to use to do the mapping is a latent variable and K is a hyperparameter to be tuned. LATEM uses the ranking loss formulated in DEVISE and Stochastic Gradient Descent as the optimizer.

Learning Intermediate Attribute Classifiers

- Although Direct Attribute Prediction and Indirect Attribute Prediction have been shown to perform poorly compared to compatibility learning frameworks, authors included them to their evaluation for being historically the most widely used methods in the literature.
- DAP learns probabilistic attribute classifiers and makes a class prediction by combining scores of the learned attribute classifiers. A novel image is assigned to one of the unknown classes using:

$$f(x) = \underset{c}{\operatorname{argmax}} \prod_{m=1}^{M} \frac{p\left(a_m^c \mid x\right)}{p\left(a_m^c\right)}$$

Learning Intermediate Attribute Classifiers

- IAP indirectly estimates attributes probabilities of an image by first predicting the probabilities of each training class, then multiplying the class attribute matrix. Once the attributes probabilities are obtained by the following equation:

$$p\left(a_m \mid x\right) = \sum_{k=1}^{K} p\left(a_m \mid y_k\right) p\left(y_k \mid x\right)$$

the previous equation is used to predict the class label for which was trained a multi-class classifier on training classes with logistic regression.

Hybrid Models

- Semantic Similarity Embedding (SSE), Convex Combination of Semantic Embeddings (CONSE) and Synthesized Classifiers (SYNC) express images and semantic class embeddings as a mixture of seen class proportions, hence we group them as hybrid models.
- SSE leverages similar class relationships both in image and semantic embedding space. An image is labeled with:

$$\underset{u \in \mathcal{U}}{\operatorname{argmax}} \pi(\theta(x))^T \psi(\phi(y_u))$$

- CONSE learns the probability of a training image belonging to a training class:

$$f(x,t) = \underset{y \in \mathcal{Y}^{tr}}{\operatorname{argmax}} p_{tr}(y \mid x)$$

where y denotes the most likely training label (t=1) for image x.

Hybrid Models

- SYNC learns a mapping between the semantic class embedding space and a model space. In the model space, training classes and a set of phantom classes form a weighted bipartite graph. The objective is to minimize distortion error:

$$\min_{w_c} \left\| w_c - \sum_{r=1}^{R} s_{cr} v_r \right\|_2^2$$

Transductive Zero-Shot Learning Setting

- In zero-shot learning, transductive setting implies that unlabeled images from unseen classes are available during training.
- GFZSL-tran uses an Expectation-Maximization based procedure that alternates between inferring the labels of unlabeled examples of unseen classes and using the inferred labels to update the parameter estimates of unseen class distributions.
- DSRL proposes to simultaneously learn image features with non-negative matrix factorization and align them with their corresponding class attributes. To improve the prediction score matrix by transductive learning, a graph-based label propagation algorithm is applied.

Datasets

- Attribute Pascal and Yahoo (aPY)
- Animals with Attributes2 (AWA2) Dataset
- Caltech-UCSDBirds 200-2011
- SUN
- Large-Scale ImageNet

Evaluation Protocol

- Authors extractes image features, namely image embeddings, from the entire image. Image embeddings are 2048-dim top-layer pooling units of the 101-layered ResNet.
- Average per-class top-1 accuracy:

$$acc_{\mathcal{Y}} = \frac{1}{\|\mathcal{Y}\|} \sum_{c=1}^{\|\mathcal{Y}\|} \frac{\#\text{ correct predictions in c}}{\#\text{ samples in c}}$$

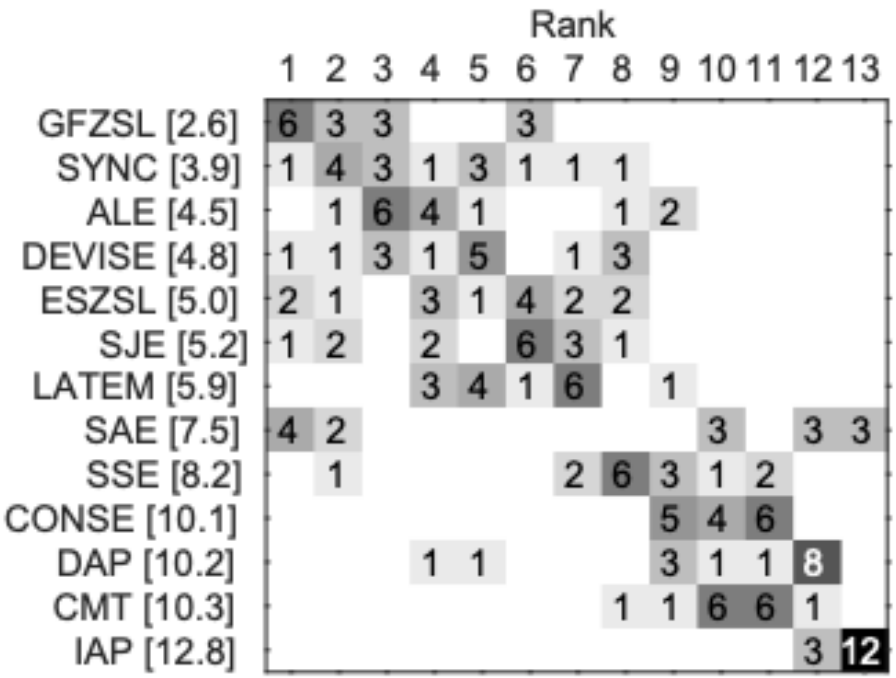
- Generalized zero-shot learning setting:

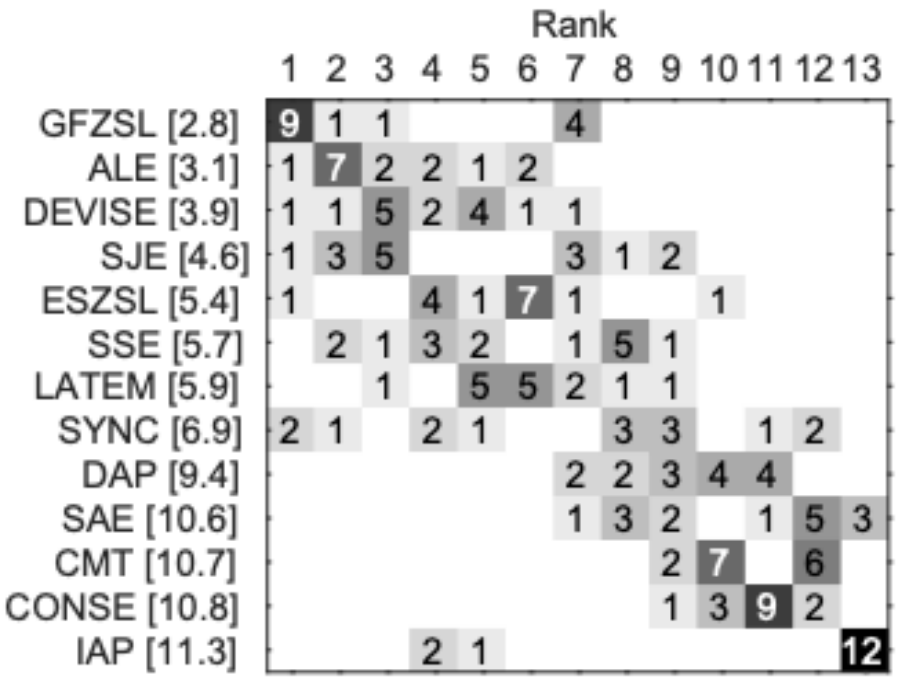
$$H = \frac{2 * acc_{\mathcal{Y}_{tr}} * acc_{\mathcal{Y}_{ts}}}{acc_{\mathcal{Y}_{tr}} + acc_{\mathcal{Y}^{ts}}}$$

Experiments

	SUN		CUB		AWA1		aPY	
Model	R	O	R	O	R	O	R	O
DAP [1]	22.1	22.2	–	–	41.4	41.4	19.1	19.1
SSE [13]	83.0	82.5	44.2	30.4	64.9	76.3	45.7	46.2
LATEM [11]	–	–	45.1	45.5	71.2	71.9	–	–
SJE [9]	–	–	50.1	50.1	67.2	66.7	–	–
ESZSL [10]	64.3	65.8	–	–	48.0	49.3	14.3	15.1
SYNC [14]	62.8	62.8	53.4	53.4	69.7	69.7	–	–
SAE [33]	–	–	–	–	84.7	84.7	–	–
GFZSL [41]	86.5	86.5	56.6	56.5	80.4	80.8	–	–
GFZSL-tran [41]	87.0	87.0	63.8	63.7	94.9	94.3	–	–
DSRL [71]	86.0	85.4	57.6	57.1	87.7	87.2	47.8	51.3

Experiments





The End