ZERO-SHOT LEARNING (ZSL)

Yannick LE CACHEUX

Manager: Hervé LE BORGNE – CEA LIST
Thesis director: Michel CRUCIANU – CNAM

JOURNAL CLUB CNAM, 8TH JUNE 2018

list
ceatech

- **Paper presented: *Zero-Shot Learning: a Comprehensive Evaluation of the Good, the Bad and the Ugly***

- **Initially published at CVPR 2017 as *Zero-Shot Learning: the Good, the Bad and the Ugly***

  by Y. Xian, B. Schiele and Z. Akata

- **Journal version: *Zero-Shot Learning: a Comprehensive Evaluation of the Good, the Bad and the Ugly***

  with C. Lampert as new 2nd author

- **Not yet published (but available on arXiv)**

Zero-Shot Learning - The Good, the Bad and the Ugly

Yongqin Xian[1] Bernt Schiele[1] Zeynep Akata[1,2]

[1]Max Planck Institute for Informatics, Saarland Informatics Campus [2]Amsterdam Machine Learning Lab, University of Amsterdam

Abstract

Due to the importance of zero-shot learning, the number of proposed approaches has increased steadily recently. We argue that it is time to take a step back and to analyze the status quo of the area. The purpose of this paper is three-fold. First, given the fact that there is no agreed upon zero-shot learning benchmark, we first define a new benchmark by unifying both the evaluation protocols and data splits. This is an important contribution as published results are often not comparable and sometimes even flawed due to, e.g. pre-training on zero-shot test classes. Second, we compare and analyze a significant number of the state-of-the-art methods in depth, both in the classic zero-shot setting but also in the more realistic generalized zero-shot setting. Finally, we discuss limitations of the current status of the area which can be taken as a basis for advancing it.

1. Introduction

Zero-shot learning aims to recognize objects whose instances may not have been seen during training [17, 22, 23, 30, 40]. The number of new zero-shot learning methods proposed every year has been increasing rapidly, i.e. the good aspects as our title suggests. Although each new method has been shown to make progress over the previous one, it is difficult to quantify this progress without an established evaluation protocol, i.e. the bad aspects. In fact, the quest for improving numbers has lead to even flawed evaluation protocols, i.e. the ugly aspects. Therefore, in this work, we propose to extensively evaluate a significant number of recent zero-shot learning methods in depth on several small to large-scale datasets using the same evaluation protocol both in zero-shot, i.e. training and test classes are disjoint, and the more realistic generalized zero-shot learning settings, i.e. training classes are present at test time.

We benchmark and systematically evaluate zero-shot learning w.r.t. three aspects, i.e. methods, datasets and evaluation protocol. The crux of the matter for all zero-shot learning methods is to associate observed and non observed classes through some form of auxiliary information which encodes visually distinguishing properties of objects. Different flavors of zero-shot learning methods that we evaluate in this work are linear [11, 2, 4, 32] and nonlinear [39, 34] compatibility learning frameworks whereas an orthogonal direction is learning independent attribute [22] classifiers and finally others [42, 7, 26] propose a hybrid model between independent classifier learning and compatibility learning frameworks.

We thoroughly evaluate the second aspect of zero-shot learning, by using multiple splits of several small to large-scale datasets [28, 38, 22, 10, 9]. We emphasize that it is hard to obtain labeled training data for fine-grained classes of rare objects recognizing which requires expert opinion. Therefore, we argue that zero-shot learning methods should be evaluated mainly on least populated or rare classes.

We propose a unified evaluation protocol to address the third aspect of zero-shot learning which is arguably the most important one. We emphasize the necessity of tuning hyper-parameters of the methods on a validation class split that is disjoint from training classes as improving zero-shot learning performance

4582

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

Yongqin Xian, Student Member, IEEE, Christoph H. Lampert, Bernt Schiele, Fellow, IEEE, and Zeynep Akata, Member, IEEE

arXiv:1707.00600v1 [cs.CV] 3 Jul 2017

INSTITUT CARNOT TN@UPSaclay
université PARIS-SACLAY

• Main contributions

- Review of state-of-the-art methods under a unified formalism for ZSL
- Highlighting of possible (big) bias in standard ZSL evaluation protocols
- Introduction of a new ZSL dataset
- Objective evaluation of state-of-the art methods with a common evaluation protocol, both in a ZSL and Generalized ZSL setting

Contents

What is Zero-Shot Learning ?

Commissariat à l'énergie atomique et aux énergies alternatives
Institut List | CEA SACLAY NANO-INNOV | BAT. 861 – PC142
91191 Gif-sur-Yvette Cedex - FRANCE
www-list.cea.fr

What is Zero-Shot Learning?

- "Zero-Shot Learning aims to recognize objects whose instances may not have been seen during training."
- First proposed in *Learning to detect unseen objects by between-class attribute transfer* by Lampert et al. (CVPR 2009) and *Zero-shot learning with semantic output codes by Palatucci* et al. (NIPS 2009).
- In practice, this is achieved with the use of additional semantic knowledge for each class.
  - Most of the time, semantic knowledge consists in vectors of attributes.

- **Training phase**

Zebra | *Tiger* — Seen classes

Visual samples

Zebra		Tiger	
Stripes:	✔	Stripes:	✔
Orange:	✘	Orange:	✔
Hooves:	✔	Hooves:	✘

Attributes

- **Testing phase**



Test images

Stripes: 0.1
Orange: 0.8
Hooves: 0.2

Stripes: 0.0
Orange: 0.4
Hooves: 0.8

Stripes: 0.0
Orange: 0.0
Hooves: 0.9

Predicted attributes

Stripes: ✘
Orange: ✔
Hooves: ✘

Horse

Stripes: ✘
Orange: ✘
Hooves: ✔

Fox

Candidate unseen classes

- **Formally**

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

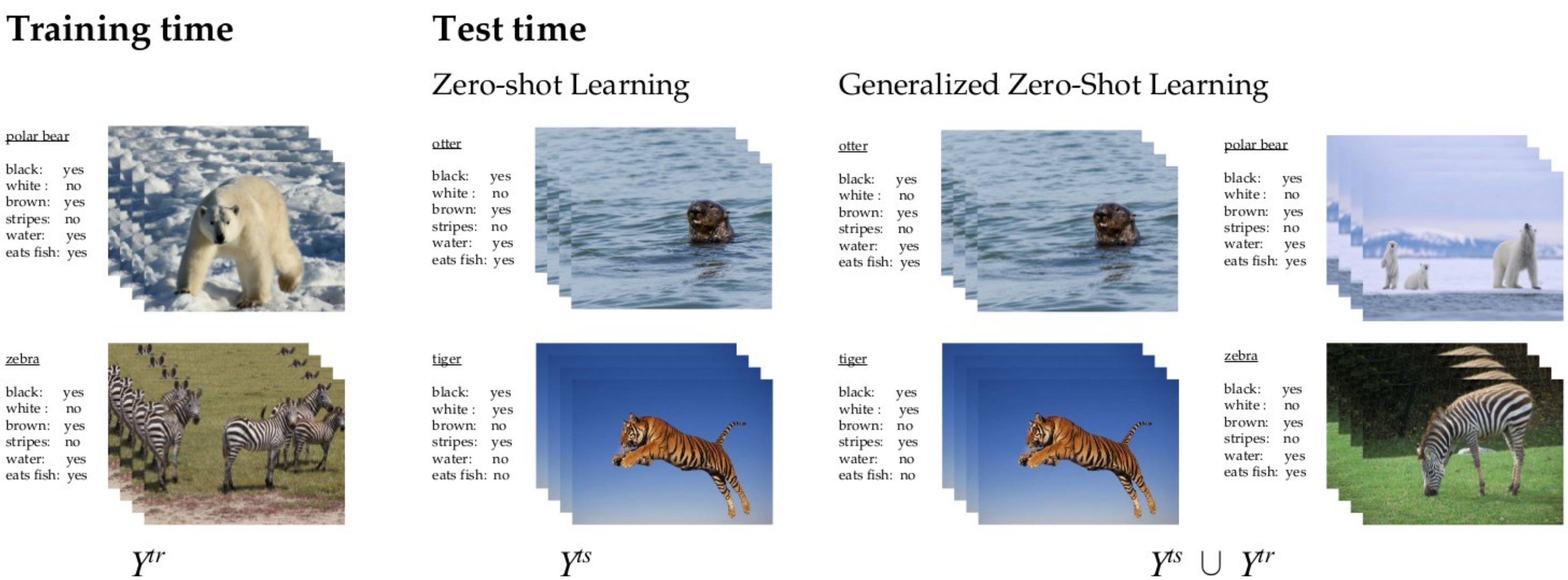
$$\frac{1}{N} \sum_{n=1}^{N} L(y_n, f(x_n; W)) + \Omega(W) \tag{1}$$

where $L(.)$ is the loss function and $\Omega(.)$ is the regularization term. Here, the mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ from input to output embeddings is defined as:

$$f(x; W) = \operatorname*{argmax}_{y \in \mathcal{Y}} F(x, y; W) \tag{2}$$

• Generalized Zero-Shot Learning

- Generalized ZSL: candidate classes can be either seen or unseen



Review of state-of-the-art methods

Commissariat à l'énergie atomique et aux énergies alternatives
Institut List | CEA SACLAY NANO-INNOV | BAT. 861 – PC142
91191 Gif-sur-Yvette Cedex - FRANCE
www-list.cea.fr

- **Four broad types of methods**

Intermediate Attribute Classifiers

Simple probabilistic models.
Evaluated for "historical" reasons:
Used in *Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer* by Lampert et al. (CVPR 2009)

- Direct Attribute Prediction (**DAP**)
- Indirect Attribute Prediction (**IAP**)

Linear Compatibility

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

- Attribute Label Embedding (**ALE**)
- Deep Visual Semantic Embedding (**DeViSE**)
- Structured Joint Embedding (**SJE**)

Nonlinear Compatibility

Compatibility function F can be non linear

- Latent Embeddings (**LATEM**)
- Cross Modal Transfer (**CMT**)

Hybrid Models

- Semantic Similarity Embedding (**SSE**)
- Convex Combination of Semantic Embeddings (**ConSE**)
- Synthesized Classifiers (**SynC**)

Intermediate attribute classifiers

- **Direct Attribute Prediction (DAP): maximum a posteriori estimation**

$$f(x) = \operatorname*{argmax}_c \prod_{m=1}^{M} \frac{p(a_m^c|x)}{p(a_m^c)}$$

- **Indirect Attribute Prediction (IAP)**

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

Linear Compatibility – Part 1

Linear compatibility function F: $F(x, y; W) = \theta(x)^T W \phi(y)$

Image embedding (visual features)

Class embedding (attributes)

- **Deep Visual Semantic Embedding (DeViSE): triplet loss**

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

where $\Delta(y_n, y)$ is equal to 1 if $y_n \neq y$, otherwise 0.

- **Structured Joint Embedding (SJE)**

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

- **Attribute Label Embedding (ALE)**

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

where $l_k = \sum_{i=1}^{k} \alpha_i$ and $r_{\Delta(x_n, y_n)}$ is defined as:

$\alpha_i = 1/i$

$$\sum_{y \in \mathcal{Y}^{tr}} \mathbb{1}(F(x_n, y; W) + \Delta(y_n, y) \geq F(x_n, y_n; W))$$

Linear Compatibility – Part 2

- **Embarrassingly Simple Approach to ZSL (ESZSL): linear mapping with triple regularization**

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

- **Semantic Autoencoder (SAE):**

$$\min_{W} ||\theta(x) - W^T\phi(y)||^2 + \lambda||W\theta(x) - \phi(y)||^2$$

Nonlinear Compatibility

- **Latent Embeddings (LATEM): triplet loss and latent embedding**

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

- **Cross Modal Transfer (CMT): simple neural network**

$$\sum_{y \in \mathcal{Y}^{tr}} \sum_{x \in \mathcal{X}_y} \|\phi(y) - W_1 \tanh(W_2.\theta(x)\|^2$$

Hybrid models

- Semantic Similarity Embedding (SSE): common embedding space

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

- Convex Combination of Semantic Embeddings (CONSE):

$$\frac{1}{Z}\sum_{i=1}^{T} p_{tr}(f(x,t)|x).s(f(x,t)) \qquad f(x,t) = \underset{y \in \mathcal{Y}^{tr}}{\operatorname{argmax}}\, p_{tr}(y|x)$$

- Synthesized Classifiers (SYNC) – it's complicated...

$$\min_{w_c} \|w_c - \sum_{r=1}^{R} s_{cr} v_r\|_2^2. \qquad \ldots$$

Experiments

Commissariat à l’énergie atomique et aux énergies alternatives
Institut List | CEA SACLAY NANO-INNOV | BAT. 861 – PC142
91191 Gif-sur-Yvette Cedex - FRANCE
www-list.cea.fr

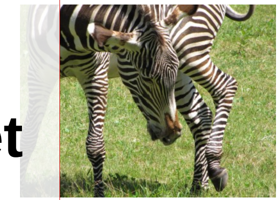
Usual datasets

- **SUN: scenes and places (*abbey*, … , *zoo*)**
  14340 images, 717 classes, 102 attributes


- **CUB: bird species (*black footed albatros*, … , *common yellowthroat*)**
  11788 images, 200 classes, 312 attributes


- **AwA a[illegible]AwA2: animal species (*antelope*, … , *zebra*[illegible]**
  304[illegible]ma[illegible] 50 classes, 85 attributes (AwA)
  37[illegible]ma[illegible]0 classes, 85 attributes (AwA2)

Huge possible bias as many classes are present in ImageNet

- **aPY: a[illegible]im[illegible]ls, objects and vehicules (*aer*[illegible]*an*[illegible]… , *zebra*)**
  15[illegible]ma[illegible]32 classes, 64 attributes


Protocol setting and metrics

- Visual features extracted with last pooling layer of ResNet-101, no finetuning
- Attributes normalized with l2-norm

- Metrics:

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

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

Results with ZSL setting on usual datasets

- Average per-class accuracy, evaluated on Standard Split (SS) and Proposed Split (PS)

	SUN		CUB		AWA1		AWA2		aPY	
Method	**SS**	**PS**	**SS**	**PS**	**SS**	**PS**	**SS**	**PS**	**SS**	**PS**
DAP [2]	38.9	39.9	37.5	40.0	57.1	44.1	58.7	46.1	35.2	33.8
IAP [2]	17.4	19.4	27.1	24.0	48.1	35.9	46.9	35.9	22.4	36.6
CONSE [14]	44.2	38.8	36.7	34.3	63.6	45.6	67.9	44.5	25.9	26.9
CMT [11]	41.9	39.9	37.3	34.6	58.9	39.5	66.3	37.9	26.9	28.0
SSE [12]	54.5	51.5	43.7	43.9	68.8	60.1	67.5	61.0	31.1	34.0
LATEM [10]	56.9	55.3	49.4	49.3	74.8	55.1	68.7	55.8	34.5	35.2
ALE [29]	59.1	**58.1**	53.2	54.9	78.6	59.9	80.3	**62.5**	30.9	39.7
DEVISE [6]	57.5	56.5	53.2	52.0	72.9	54.2	68.6	59.7	35.4	**39.8**
SJE [8]	57.1	53.7	**55.3**	53.9	76.7	**65.6**	69.5	61.9	32.0	32.9
ESZSL [9]	57.3	54.5	55.1	53.9	74.7	58.2	75.6	58.6	34.4	38.3
SYNC [13]	**59.1**	56.3	54.1	**55.6**	72.2	54.0	71.2	46.6	**39.7**	23.9
SAE [32]	42.4	40.3	33.4	33.3	**80.6**	53.0	**80.7**	54.1	8.3	8.3

- Significant difference between SS and PS for AwA1 and AwA2

Results with ZSL setting on ImageNet

- Average per-class accuracy, evaluated on classes 2 / 3 hops (H) away from training classes, on the 500 /1K /5K most and least populated classes, and on all 20K testing classes.

	Hierarchy		Most Populated			Least Populated			All
Method	**2 H**	**3 H**	**500**	**1K**	**5K**	**500**	**1K**	**5K**	**20K**
CONSE [14]	7.63	2.18	12.33	8.31	3.22	3.53	2.69	1.05	0.95
CMT [11]	2.88	0.67	5.10	3.04	1.04	1.87	1.08	0.33	0.29
LATEM [10]	5.45	1.32	10.81	6.63	1.90	4.53	2.74	0.76	0.50
ALE [29]	5.38	1.32	10.40	6.77	2.00	4.27	2.85	0.79	0.50
DEVISE [6]	5.25	1.29	10.36	6.68	1.94	4.23	2.86	0.78	0.49
SJE [8]	5.31	1.33	9.88	6.53	1.99	4.93	2.93	0.78	0.52
ESZSL [9]	6.35	1.51	11.91	7.69	2.34	4.50	3.23	0.94	0.62
SYNC [13]	**9.26**	**2.29**	**15.83**	**10.75**	**3.42**	**5.83**	**3.52**	**1.26**	**0.96**
SAE [32]	4.89	1.26	9.96	6.57	2.09	2.50	2.17	0.72	0.56

- **Results with GZSL setting**

- Harmonic mean of per-class accuracy between test samples from seen classes and test samples from unseen classes, evaluated on Proposed Split.

	SUN			CUB			AWA1			AWA2			aPY		
Method	**ts**	**tr**	**H**	**ts**	**tr**	**H**	**ts**	**tr**	**H**	**ts**	**tr**	**H**	**ts**	**tr**	**H**
DAP [2]	4.2	25.1	7.2	1.7	67.9	3.3	0.0	**88.7**	0.0	0.0	84.7	0.0	4.8	78.3	9.0
IAP [2]	1.0	37.8	1.8	0.2	72.8	0.4	2.1	78.2	4.1	0.9	87.6	1.8	5.7	65.6	10.4
CONSE [14]	6.8	**39.9**	11.6	1.6	**72.2**	3.1	0.4	88.6	0.8	0.5	**90.6**	1.0	0.0	**91.2**	0.0
CMT [11]	8.1	21.8	11.8	7.2	49.8	12.6	0.9	87.6	1.8	0.5	90.0	1.0	1.4	85.2	2.8
CMT* [11]	8.7	28.0	13.3	4.7	60.1	8.7	8.4	86.9	15.3	8.7	89.0	15.9	**10.9**	74.2	**19.0**
SSE [12]	2.1	36.4	4.0	8.5	46.9	14.4	7.0	80.5	12.9	8.1	82.5	14.8	0.2	78.9	0.4
LATEM [10]	14.7	28.8	19.5	15.2	57.3	24.0	7.3	71.7	13.3	11.5	77.3	20.0	0.1	73.0	0.2
ALE [29]	**21.8**	33.1	**26.3**	23.7	62.8	**34.4**	**16.8**	76.1	**27.5**	14.0	81.8	23.9	4.6	73.7	8.7
DEVISE [6]	16.9	27.4	20.9	**23.8**	53.0	32.8	13.4	68.7	22.4	**17.1**	74.7	**27.8**	4.9	76.9	9.2
SJE [8]	14.7	30.5	19.8	23.5	59.2	33.6	11.3	74.6	19.6	8.0	73.9	14.4	3.7	55.7	6.9
ESZSL [9]	11.0	27.9	15.8	12.6	63.8	21.0	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6
SYNC [13]	7.9	43.3	13.4	11.5	70.9	19.8	8.9	87.3	16.2	10.0	90.5	18.0	7.4	66.3	13.3
SAE [32]	8.8	18.0	11.8	7.8	54.0	13.6	1.8	77.1	3.5	1.1	82.2	2.2	0.4	80.9	0.9

- Accuracy is much higher on samples from seen classes

Conclusion and perspectives

Commissariat à l'énergie atomique et aux énergies alternatives
Institut List | CEA SACLAY NANO-INNOV | BAT. 861 – PC142
91191 Gif-sur-Yvette Cedex - FRANCE
www-list.cea.fr

Conclusion

- **Major keypoint: importance of a proper training / testing split for ZSL to avoid possible bias**

- **Many methods were reviewed and evaluated**

- **A new dataset was introduced**

- ***Evaluation protocol of iterative methods could be more robust***

- ***The GZSL process is just one example of such a process***