An Embarrassingly Simple Approach To Zero-Shot Learning

Bernardino Romera Paredes & Philip H. S. Torr
Proceedings of the 32nd International Conference on Machine Learning, 2015

Presented By: Shriyash Upadhyay

Zero-Shot Learning With Attributes

- Describe classes in terms of attributes
  - The list of attributes is called the signature
- Given a new class with no training examples, classify new examples

Image: Y. Xian, C. H. Lampert, B. Schiele, Z. Akata. "Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly", IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI) 40(8), 2018

Zero-Shot Learning In Two Parts

- Attribute Learning (Training)
- “Attribute Based Prediction” (Inference)

Zero-Shot Learning In Two Parts

- **Attribute Learning (Training)**
  - Given training instances + attribute signatures for training classes
  - Learn to identify the attributes
- "Attribute Based Prediction" (Inference)

polar bear

black: no
white: yes
brown: no
stripes: no
water: yes
eats fish: yes

Image: Y. Xian, C. H. Lampert, B. Schiele, Z. Akata. "Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly", IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI) 40(8), 2018

Zero-Shot Learning In Two Parts

- Attribute Learning (Training)
- **“Attribute Based Prediction” (Inference)**
    - Given the attribute signatures for new classes, classify new instances

zebra

black:	yes
white:	yes
brown:	no
stripes:	yes
water:	no
eats fish:	no

Image:Y. Xian, C. H. Lampert, B. Schiele, Z. Akata. "Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly", IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI) 40(8), 2018

A Very Simple Zero-Shot Learning Model

- Train a linear model (e.g. logistic regression) for each of the attributes
- Create a second model to predict classes based on the attributes
- This is known as Directed Attribute Prediction (DAP)

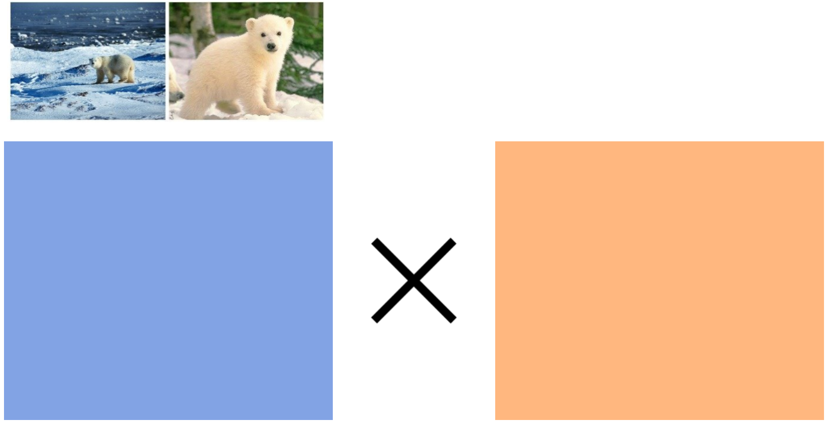
Outline Of The Presentation

Paper's Goal: Improve on DAP by combining the two parts

1. How does ESZSL work?
    a. Trains quickly
    b. Simple to implement
    c. Outperforms DAP
2. What are the weaknesses of the model?
    a. Underperforms other models
    b. More performant variants lose useful properties

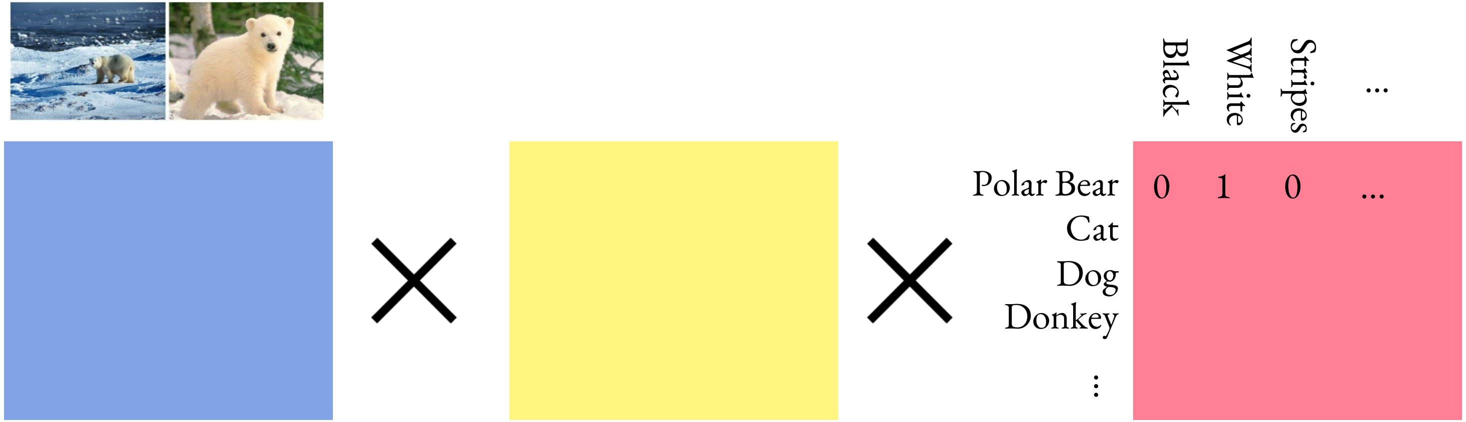
The General Form Of A Linear Layer

$$\underset{W \in \mathbb{R}^{d \times z}}{\text{minimise}} \; L\left(X^{\top} W, Y\right) + \Omega\left(W\right)$$



The ESZSL Linear Layer

$$\underset{V \in \mathbb{R}^{d \times a}}{\text{minimise}}\, L\left(X^\top V S, Y\right) + \Omega\left(V\right)$$



How To Use The Model

- We classify an instance x based on a new set of classes S' using $\operatorname{argmax}\, x^\top V S'_i$

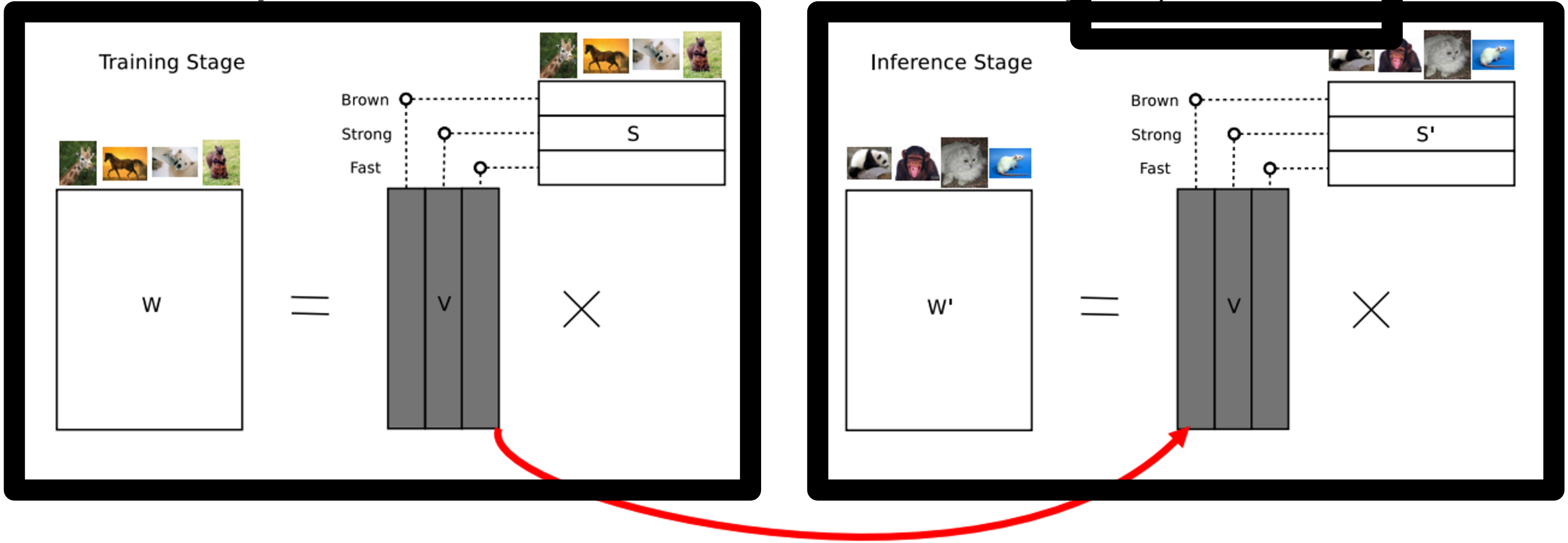


Figure: B. Romera-Paredes, P. Torr. "An Embarrassingly Simple Approach To Zero-Shot Learning", Proceedings of the 32nd International Conference on Machine Learning, 2015

Regularization

$$\Omega\left(V; S, X\right) = \gamma \left\|VS\right\|_{\text{Fro}}^{2} + \lambda \left\|X^{\top}V\right\|_{\text{Fro}}^{2} + \boxed{\beta \left\|V\right\|_{\text{Fro}}^{2}}$$

- Weights shouldn't be too large

Regularization

$$\Omega\left(V; S, X\right) = \gamma \left\|VS\right\|_{\text{Fro}}^2 + \boxed{\lambda \left\|X^\top V\right\|_{\text{Fro}}^2} + \boxed{\beta \left\|V\right\|_{\text{Fro}}^2}$$

- Weights shouldn't be too large
- All training instances should have a comparable impact on the weight

Regularization

$$\Omega(V; S, X) = \boxed{\gamma \|VS\|_{\mathrm{Fro}}^2} + \boxed{\lambda \|X^\top V\|_{\mathrm{Fro}}^2} + \boxed{\beta \|V\|_{\mathrm{Fro}}^2}$$

- Weights shouldn't be too large
- All training instances should have a comparable impact on the weight
- All signatures should have a comparable impact on the weights

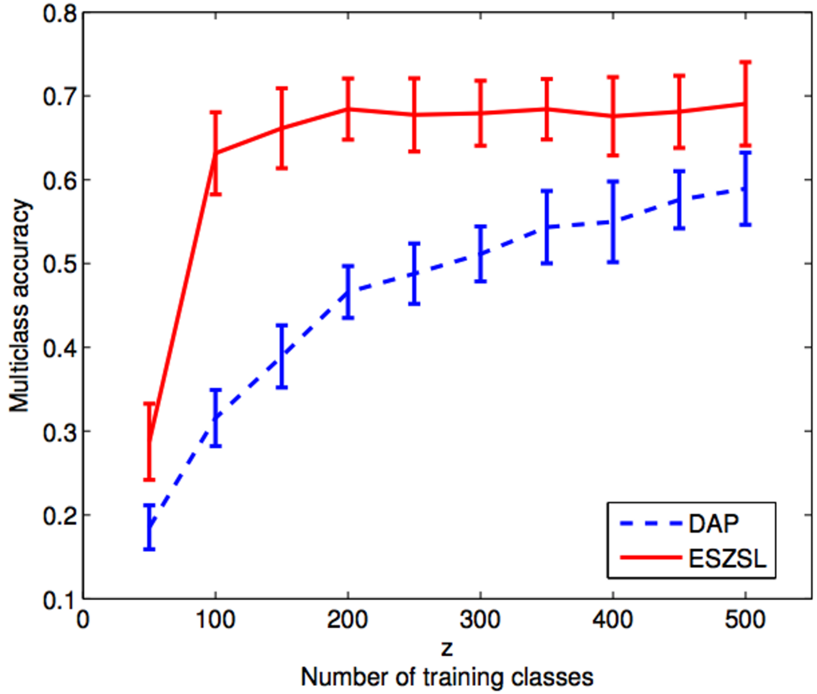
Closed Form Solution

If we let the following be true, the solution is closed form (easy to implement)

- $L(P, Y) = \|P - Y\|_{\text{Fro}}^2$
- $\beta = \gamma\lambda$

$$\boxed{V = \left(XX^\top + \gamma I\right)^{-1} XYS^\top \left(SS^\top + \lambda I\right)^{-1}}$$

Validation With Synthetic Data



- Create random classes
  - Each class has 100 attributes
  - Attributes are randomly selected to be 0 or 1
- Generate examples for classes
  - 50 examples for each class
  - Each example has dimension of 10 with added gaussian noise
- Goal: see how the number of training classes impacts model performance

Figure: B. Romera-Paredes, P. Torr. “An Embarrassingly Simple Approach To Zero-Shot Learning”, Proceedings of the 32nd International Conference on Machine Learning, 2015

Validation With Real Data

- Multiclass classification accuracy on 3 standard datasets
- aPY has a large number of attributes relative to classes
  - ESZSL-AS is a variation of ESZSL used to accommodate this by creating new classes

Method/Dataset	AwA	aPY	SUN
DAP	40.50	18.12	52.50
ESZSL	**49.30 $\pm$0.21**	15.11 $\pm$2.24	**65.75 $\pm$0.51**
ESZSL-AS	–	**27.27 $\pm$1.62**	61.53 $\pm$1.03

Table: B. Romera-Paredes, P. Torr. "An Embarrassingly Simple Approach To Zero-Shot Learning", Proceedings of the 32nd International Conference on Machine Learning, 2015

Comparison With Existing Models

Xian et al., 2018 compares a large number of zero-shot learning models

Method	SUN ts	SUN tr	SUN H	CUB ts	CUB tr	CUB H	AWA1 ts	AWA1 tr	AWA1 H	AWA2 ts	AWA2 tr	AWA2 H	aPY ts	aPY tr	aPY H
DAP [1]	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 [1]	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 [15]	6.8	35.9	11.4	2.0	70.6	3.9	0.4	**89.6**	0.8	0.5	**90.6**	1.0	0.0	**91.2**	0.0
CMT [12]	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* [12]	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 [13]	2.1	36.4	4.0	8.5	46.9	14.4	7.0	80.5	12.9	8.1	82.5	14.8	0.3	78.4	0.6
LATEM [11]	14.7	28.8	19.5	15.2	57.3	24.0	7.3	71.7	13.3	11.5	77.3	20.0	1.3	71.4	2.6
ALE [30]	**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 [7]	16.9	27.4	20.9	**23.8**	53.0	32.8	13.4	68.7	22.4	**17.1**	74.7	**27.8**	3.5	78.4	6.7
SJE [9]	14.4	29.7	19.4	23.5	59.2	33.6	11.3	74.6	19.6	8.0	73.9	14.4	1.3	71.4	2.6
ESZSL [10]	11.0	27.9	15.8	14.7	56.5	23.3	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6
SYNC [14]	7.9	**43.3**	13.4	11.5	**70.9**	19.8	9.0	88.9	16.3	9.7	89.7	17.5	7.4	66.3	13.3
SAE [33]	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
GFZSL [41]	0.0	39.6	0.0	0.0	45.7	0.0	1.8	80.3	3.5	2.5	80.1	4.8	0.0	83.3	0.0

SJE vs ESZSL

Structured Joint Embeddings (Akata et al., 2015) maps attributes and images to word embeddings

	SUN			CUB			AWA1			AWA2			aPY		
Method	**ts**	**tr**	**H**	**ts**	**tr**	**H**	**ts**	**tr**	**H**	**ts**	**tr**	**H**	**ts**	**tr**	**H**
SJE [9]	14.4	29.7	19.4	23.5	59.2	33.6	11.3	74.6	19.6	8.0	73.9	14.4	1.3	71.4	2.6
ESZSL [10]	11.0	27.9	15.8	14.7	56.5	23.3	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6

Model Introduced: Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele. “Evaluation Of Output Embeddings for Fine-Grained Image Classification”, 2015

Potential Variations On This Model

- Introduce non-linearities
- Introduce more layers
- This might improve performance
  - Trains more slowly
  - More complicated to implement

Summary

- ESZSL has the following benefits
  - Convex (trains efficiently)
  - Closed form (simple to implement)
  - Better performance than DAP
- The specifics of the method are too simple to remain useful
  - Lower performance than contemporary and subsequent methods
  - However, the core idea of making weights a function of the signature is compelling

References

[1] B. Romera-Paredes, P. Torr. “An Embarrassingly Simple Approach To Zero-Shot Learning”, Proceedings of the 32nd International Conference on Machine Learning, 2015

[2] Y. Xian, C. H. Lampert, B. Schiele, Z. Akata. "Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly", IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI) 40(8), 2018. (arXiv:1707.00600 [cs.CV])

[3] Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele. “Evaluation Of Output Embeddings for Fine-Grained Image Classification”, 2015

Appendix On Transfer Learning
&
Domain Adaptation

Transfer Learning & Domain Adaptation

- Transfer Learning
  - Similar to zero-shot learning, but the information about the new task is given in the form of labeled instances
- Domain Adaptation
  - Trying to learn the same thing from two different types of data
  - E.g. training a autonomous vehicle in San Francisco and then getting it to work in Hong Kong

ESZSL & Domain Adaptation

- If we look at the outer product space of the instances and the signatures, this becomes a domain adaptation problem
  - There is no good visualization of this
  - There is no good intuitive reason to do this
- We can then apply tools from the study of domain adaptation to this method
- I omit discussion of these results because they are not practically applicable
  - The bounds do not apply to variants of this model
  - The specific cases on the bounds which are analyzed are very simple (i.e. if the distributions are unrelated, our model cannot do better than random; if the distributions are identical, we can have a perfect classifier)
  - The bounds require knowledge of the underlying distribution from which we are drawing our data
  - No other papers I found appear to use this method in order to compute bounds, even in other compatibility learning methods