Zero-Shot Learning via Class-Conditioned Deep Generative Models

Wenlin Wang

Joint with Yunchen Pu, Vinay Kumar Verma, Kai Fan, Yizhe Zhang, Changyou Chen, Piyush Rai, Lawrence Carin

Duke University, IIT Kanpur, SUNY at Buffalo

February 6, 2018

Outline

Problem of Interest

- Many-Shot Learning

Train: $x_{\mathrm{tr}} \in \mathcal{X} \rightarrow y_{\mathrm{tr}} \in \mathcal{Y}$

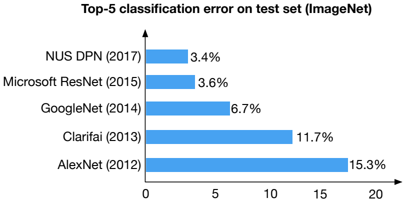
Test: $x_{\mathrm{te}} \in \mathcal{X} \rightarrow y_{\mathrm{te}} \in \mathcal{Y}$

Objective:

$$f : \mathcal{X} \rightarrow \mathcal{Y}$$

+

Annotations



Problem of Interest

- Zero-Shot Learning (ZSL): ZSL refers to the problem of recognizing objects from classes that are not seen at training time.

Train: $x_{\mathrm{tr}} \in \mathcal{X} \rightarrow y_{\mathrm{tr}} \in \mathcal{Y}$

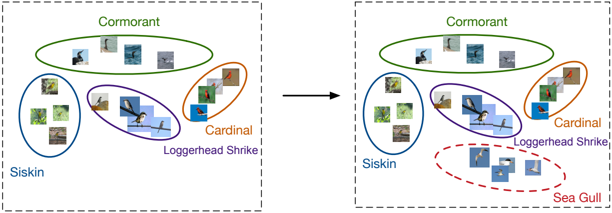
Test: $x_{\mathrm{te}} \in \mathcal{X} \rightarrow y_{\mathrm{te}} \in \mathcal{Y}'$

$y_{\mathrm{tr}} \in \mathcal{Y} \xleftrightarrow{\text{Knowledge transfer}} y_{\mathrm{te}} \in \mathcal{Y}'$

Objective:

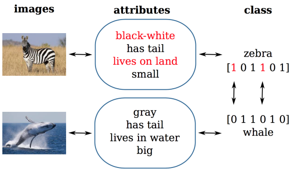
$$f : \mathcal{X} \rightarrow \mathcal{Y}'$$

$$\mathcal{Y} \cap \mathcal{Y}' = \varnothing$$

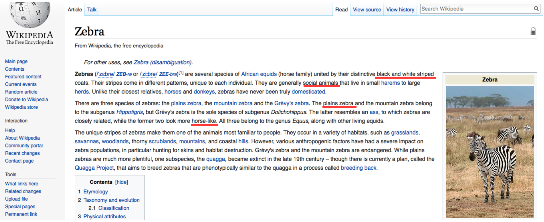


How to Transfer Knowledge ?[1]

- Attribute as side information [11]



- Wikipedia and WordNet [14, 17] as side information



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link

Article Talk Read View source View history Search Wikipedia

Zebra

From Wikipedia, the free encyclopedia

For other uses, see Zebra (disambiguation).

Zebras (/ˈzɛbrə/ ZEB-rə or /ˈziːbrə/ ZEE-brə)[1] are several species of African equids (horse family) united by their distinctive black and white striped coats. Their stripes come in different patterns, unique to each individual. They are generally social animals that live in small harems to large herds. Unlike their closest relatives, horses and donkeys, zebras have never been truly domesticated.

There are three species of zebras: the plains zebra, the mountain zebra and the Grévy's zebra. The plains zebra and the mountain zebra belong to the subgenus *Hippotigris*, but Grévy's zebra is the sole species of subgenus *Dolichohippus*. The latter resembles an ass, to which zebras are closely related, while the former two look more horse-like. All three belong to the genus *Equus*, along with other living equids.

The unique stripes of zebras make them one of the animals most familiar to people. They occur in a variety of habitats, such as grasslands, savannas, woodlands, thorny scrublands, mountains, and coastal hills. However, various anthropogenic factors have had a severe impact on zebra populations, in particular hunting for skins and habitat destruction. Grévy's zebra and the mountain zebra are endangered. While plains zebras are much more plentiful, one subspecies, the quagga, became extinct in the late 19th century – though there is currently a plan, called the Quagga Project, that aims to breed zebras that are phenotypically similar to the quagga in a process called breeding back.

Contents [hide]

Zebra

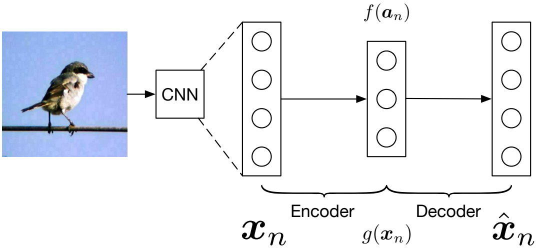
[1] http://isis-data.science.uva.nl/tmensink/docs/ZSL17.web.pdf

Existing Auto-encoder based ZSL

- Auto-encoder based non-linear models achieve state-of-the-art performance [22, 10]
- Objective function includes 3 general terms:

$$Loss(\boldsymbol{x}_n, y_n) = \underbrace{\beta \cdot D_1(\boldsymbol{x}_n, \hat{\boldsymbol{x}}_n)}_{\text{Reconstruction}} + \underbrace{D_2(g(\boldsymbol{x}_n), f(a_n))}_{\text{Supervision}} + \underbrace{\lambda \cdot R}_{\text{Regularizer}}$$

D_1 and D_2 are distance measurements (*e.g.* L2 Distance), λ and β are hyper-parameters, $a_n = \boldsymbol{A}_{y_n}$



Deep Generative Model for ZSL

- $y \in \{1, ..., S, S+1, ..., S+U\}$ is a class from the seen or the unseen classes
- Traditional auto-encoder based method represents each class as a *point* in the latent space: $f(\boldsymbol{A}_y)$
- Ours represent each class using a *class-specific latent-space distribution*: $\mathcal{N}(\mu_y, \Sigma_y)$, where

$$\boldsymbol{\mu}_y = f_\mu(\boldsymbol{A}_y) \quad \text{and} \quad \boldsymbol{\Sigma}_y = \text{diag}(\exp(f_{\sigma^2}(\boldsymbol{A}_y))) \tag{1}$$

- Let θ to be the parameter of the encoder, ϕ to be the parameter of the decoder, ψ to be the parameter of $f_*(\cdot)$, for $* = \mu, \sigma^2$.

Deep Generative Model for ZSL

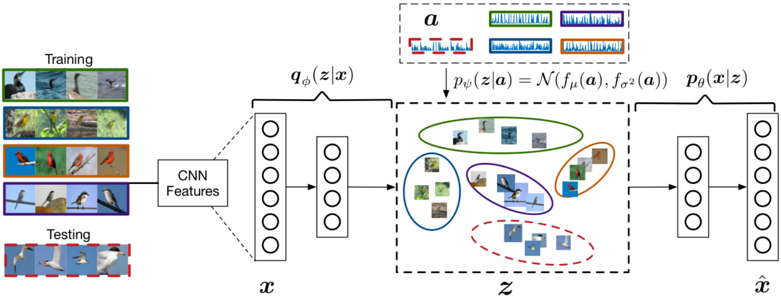


Figure: A diagram of our basic model; only the training stage is shown here. In the above figure, $\boldsymbol{a} \in \mathbb{R}^M$ denotes the class attribute vector (given). Red-dotted rectangle/ellipse correspond to the unseen classes. Note: The CNN module is not a part of our framework and is only used as an initial feature extractor, on top of which the rest of our model is built. The CNN can be replaced by any feature extractor depending on the data type

Deep Generative Model for ZSL

- Our model assumes the data are generated from the class-specific normals, and we write down the marginal likelihood of the data as (we omit the subscript n)

$$
\begin{aligned}
\log p_\theta(\mathbf{x}) &= \log \int_z p_\psi(\mathbf{z}|y) p_\theta(\mathbf{x}|\mathbf{z}) dz \\
&= \log \int_z \frac{p_\psi(\mathbf{z}|y)}{q_\phi(\mathbf{z}|\mathbf{x})} q_\phi(\mathbf{z}|\mathbf{x}) p_\theta(\mathbf{x}|\mathbf{z}) dz \\
&= \log \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left(\frac{p_\psi(\mathbf{z}|y)}{q_\phi(\mathbf{z}|\mathbf{x})} p_\theta(\mathbf{x}|\mathbf{z}) \right) dz \\
&\geq \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left(p_\theta(\mathbf{x}|\mathbf{z}) \right)}_{reconstruction} - \underbrace{KL\left(q_\phi(\mathbf{z}|\mathbf{x}) || p_\psi(\mathbf{z}|y)\right)}_{\text{supversion}} \\
&= \mathcal{L}_{\theta,\phi,\psi}(\mathbf{x}, y) \qquad (2)
\end{aligned}
$$

Note we aim to maximizing the evidence lower bound (ELBO) $\mathcal{L}_{\theta,\phi,\psi}(\mathbf{x}, y)$.

- The variational auto-encoder (VAE) [9], as an unsupervsied model, assumes the data is generated from $p_o(\mathbf{z}) \sim \mathcal{N}(0, \mathbf{I})$, the marginal likelihood of the data can be similarly written as

$$
\begin{aligned}
\log p_\theta(\mathbf{x}) &= \log \int_z p_o(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z}) dz \\
&= \log \int_z \frac{p_o(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} q_\phi(\mathbf{z}|\mathbf{x}) p_\theta(\mathbf{x}|\mathbf{z}) dz \\
&= \log \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left(\frac{p_o(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} p_\theta(\mathbf{x}|\mathbf{z}) \right) dz \\
&\geq \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left(p_\theta(\mathbf{x}|\mathbf{z}) \right)}_{reconstruction} - \underbrace{KL\left(q_\phi(\mathbf{z}|\mathbf{x}) || p_o(\mathbf{z})\right)}_{\text{prior knowledge(unsupervised)}} \\
&= \mathcal{L}^V_{\theta,\phi}(\mathbf{x}) \qquad (3)
\end{aligned}
$$

Deep Generative Model for ZSL

- **Margin Regularizer**: promotes $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ to be far away from other class-specific distritions $p_\psi(\boldsymbol{z}|c), c \neq y$, defined as

$$\begin{aligned} R^* &= \min_{c:c\in\{1..,y-1,y+1,..,S\}} \{\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\psi(\boldsymbol{z}|c))\} \\ &= -\max_{c:c\in\{1..,y-1,y+1,..,S\}} \{-\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\psi(\boldsymbol{z}|c))\} \end{aligned} \tag{4}$$

since (4) is non-differentiable, we approximate R^* as

$$R = -\log \sum_{c=1}^{S} \exp(-\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\psi(\boldsymbol{z}|c))) \tag{5}$$

It can be easily shown that

$$R^* \leq R \leq R^* + \log S \tag{6}$$

Deep Generative Model for ZSL

- Overall objective: the ELBO (2) together with the margin regularizer (5) as

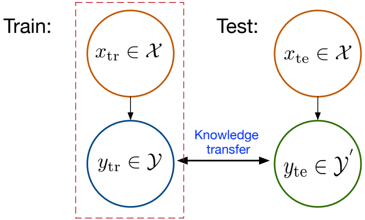
$$\hat{\mathcal{L}}_{\theta,\phi,\psi}(\boldsymbol{x}, y) = \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]}_{\text{reconstruction}} - \underbrace{\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\psi(\boldsymbol{z}|y))}_{\text{supervision}}$$
$$\underbrace{-\lambda \log \sum_{c=1}^{S} \exp(-\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\psi(\boldsymbol{z}|c)))}_{\textit{regularizer}} \qquad (7)$$

- Prediction: Given a test input $\hat{\boldsymbol{x}}$, we first predict its latent embeddings $\hat{\boldsymbol{z}}$ with the VAE recognition model, and find the "best" label by solving

$$\hat{y} = \arg\max_{y \in \mathcal{Y}_u} \hat{\mathcal{L}}_{\theta,\phi,\psi}(\hat{\boldsymbol{x}}, y)$$
$$= \arg\min_{y \in \mathcal{Y}_u} \mathrm{KL}(q_\phi(\hat{\boldsymbol{z}}|\hat{\boldsymbol{x}})||p_\psi(\hat{\boldsymbol{z}}|y)) \qquad (8)$$

Variations - Transductive ZSL

- Recall ZSL (Inductive ZSL)

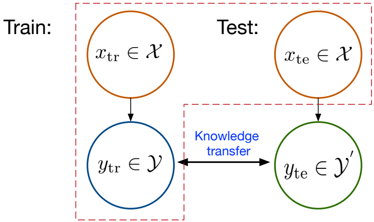


Objective:

$$f : \mathcal{X} \to \mathcal{Y}'$$

$$\mathcal{Y} \cap \mathcal{Y}' = \varnothing$$

- Transductive ZSL



Objective:

$$f : \mathcal{X} \to \mathcal{Y}'$$

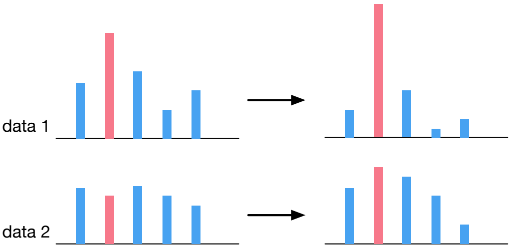
$$\mathcal{Y} \cap \mathcal{Y}' = \varnothing$$

Variations - Transductive ZSL

- A naïve approach for leveraging the unlabeled inputs would be to add the following reconstruction error

$$\tilde{\mathcal{L}}_{\theta,\phi,\psi}(\hat{\boldsymbol{x}}, y) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\hat{\boldsymbol{x}}|\boldsymbol{z})] \tag{9}$$

- A better method is to introduce our self-training regularizer.
- **Motivation**: inductive ZSL model is able to make confident predictions for unseen class test inputs, and these confident predicted class distributions can be emphsized in the regularizer to guide those ambiguous test inputs.



Variations - Transductive ZSL

- First, we define the *probability* of assigning an unseen class test input $\hat{\boldsymbol{x}}_i$ to class $c \in \{S+1, \ldots, S+U\}$ to be

$$q(\hat{\boldsymbol{x}}_i, c) = \frac{\exp(-\mathrm{KL}(q_\phi(\boldsymbol{z}|\hat{\boldsymbol{x}}_i)||p_\psi(\boldsymbol{z}|c)))}{\sum_c \exp(-\mathrm{KL}(q_\phi(\boldsymbol{z}|\hat{\boldsymbol{x}}_i)||p_\psi(\boldsymbol{z}|c)))} \tag{10}$$

- Second, we define a sharper version of the predicted class probabilities $q(\hat{\boldsymbol{x}}_i, c)$ as

$$p(\hat{\boldsymbol{x}}_i, c) = \frac{q(\hat{\boldsymbol{x}}_i, c)^2/g(c)}{\sum_{c'} q(\hat{\boldsymbol{x}}_i, c')^2/g(c')} \tag{11}$$

  where $g(c) = \sum_{i=1}^{N'} q(\hat{\boldsymbol{x}}_i, c)$ is the marginal probability of unseen class c. Note that normalizing the probabilities by $g(c)$ prevents large classes from distorting the latent space.

- Third, we encourage $q(\hat{\boldsymbol{x}}_i, c)$ to be close to $p(\hat{\boldsymbol{x}}_i, c)$.

$$\mathrm{KL}(P(\hat{\mathbf{X}})||Q(\hat{\mathbf{X}})) \triangleq \sum_{i=1}^{N'} \sum_{c=S+1}^{S+U} p(\hat{\boldsymbol{x}}_i, c) \log \frac{p(\hat{\boldsymbol{x}}_i, c)}{q(\hat{\boldsymbol{x}}_i, c)} \tag{12}$$

Variations - Transductive ZSL

- We have the following objective defined exclusively over the unseen class unlabeled inputs

$$U(\hat{\mathbf{X}}) = \sum_{i=1}^{N'} \mathbb{E}_{q_\phi(\boldsymbol{z}|\hat{\boldsymbol{x}}_i)}[\log p_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_i|\boldsymbol{z})] - \mathrm{KL}(P(\hat{\mathbf{X}})||Q(\hat{\mathbf{X}})) \quad (13)$$

- Finally, we combine (7) and (13), which leads to the overall objective

$$\sum_{n=1}^{N} \hat{\mathcal{L}}_{\theta,\phi,\psi}(\boldsymbol{x}_n, y_n) + U(\hat{\mathbf{X}}) \quad (14)$$

defined over the seen class labeled training inputs $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ and the unseen class unlabeled test inputs $\{\hat{\boldsymbol{x}}_i\}_{i=1}^{N'}$.

Datasets

- We conduct experiments on the following datasets, (*i*) Animal with Attributes (AwA) [12]; (*ii*) Caltech-UCSD Birds-200-2011 (CUB-200) [24]; and (*iii*) SUN attribute (SUN) [16].
- For the large-scale dataset (ImageNet), we follow [6], for which 1000 classes from ILSVRC2012 [19] are used as seen classes, while 360 non-overlapped classes of ILSVRC2010 [4] are used as unseen classes.

Dataset	# Attribute	training(+validation)		testing	
		# of images	# of classes	# of images	# of classes
AwA	85	24,295	40	6,180	10
CUB-200	312	8,855	150	2,933	50
SUN	102	14,140	707	200	10
ImageNet	1,000	200,000	1,000	54,000	360

Table: Summary of datasets used in the evaluation

Setup

- VGG-19 fc7 features [20] is used as our raw input reprensentation (D=4096).
- Default class attribute features are used for AwA, CUB-200 and SUN.
- Word2vec [14] representation is used for ImageNet.
- λ=1 is set across all our experiments.
- Encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ are 2-layer multi-layer perceptron (MLP) with 500 nodes (1,000 for ImageNet).
- ReLU is used as the nonlinear activation function.
- Dropout with constant rate 0.8 is used to avoid overfitting.

Inductive ZSL

- We achieve state-of-the-art performance

Method	AwA	CUB-200	SUN	Average	Method	ImageNet
(Lampert et al., 2014)[12]	57.23	–	72.00	–	DeViSE [5]	12.8
ESZSL [18]	75.32 ± 2.28	–	82.10 ± 0.32	–	ConSE [15]	15.5
MLZSC [3]	77.32 ± 1.03	43.29 ± 0.38	84.41 ± 0.71	68.34	AMP [7]	13.1
SDL [31]	80.46 ± 0.53	42.11 ± 0.55	83.83 ± 0.29	68.80	SS-Voc [6]	16.8
BiDiLEL [25]	79.20	46.70	–	–		
SSE-ReLU [29]	76.33 ± 0.83	30.41 ± 0.20	82.50 ± 1.32	63.08		
JFA [30]	81.03 ± 0.88	46.48 ± 1.67	84.10 ± 1.51	70.53		
ReViSE [22]	78.00	56.60	-	-		
SAE [10]	83.40	56.60	84.50	74.83		
GFZSL [23]	80.83	56.53	86.50	74.59		
VZSL$^{\#}$	84.45 ± 0.74	55.37 ± 0.59	85.75 ± 1.93	74.52	-	22.88
VZSL	$\mathbf{85.28 \pm 0.76}$	$\mathbf{57.42 \pm 0.63}$	$\mathbf{86.75 \pm 2.02}$	**76.48**	-	**23.08**

Table: Top-1 classification accuracy (%) on AwA, CUB-200, SUN and Top-5 accuracy(%) on ImageNet under inductive ZSL. VZSL$^{\#}$ denotes our model trained with the reconstruction term from (7) ignored.

Transductive ZSL

- We also achieve state-of-the-art performance

Method	AwA	CUB-200	SUN	Average
SMS [8]	78.47	–	82.00	–
ESZSL [18]	84.30	–	37.50	–
JFA+SP-ZSR [30]	88.04 ± 0.69	55.81 ± 1.37	85.35 ± 1.56	77.85
SDL [31]	92.08 ± 0.14	55.34 ± 0.77	86.12 ± 0.99	76.40
DMaP [13]	85.66	61.79	–	–
TASTE [27]	89.74	54.25	–	–
TSTD [28]	90.30	58.20	–	–
GFZSL [23]	94.25	63.66	87.00	80.63
VZSL$^{\#}$	93.49 ± 0.54	59.69 ± 1.22	86.37 ± 1.88	79.85
VZSL*	87.59 ± 0.21	61.44 ± 0.98	86.66 ± 1.67	77.56
VZSL	$\mathbf{94.80 \pm 0.17}$	$\mathbf{66.45 \pm 0.88}$	$\mathbf{87.75 \pm 1.43}$	**83.00**

Table: Top-1 classification accuracy (%) obtained on AwA, CUB-200 and SUN under transductive setting. VZSL$^{\#}$ denotes our model with VAE reconstruction term ignored. VZSL* denotes our model with only (9) for unlabeled data. The '-' indicates the results was not reported

Visualization

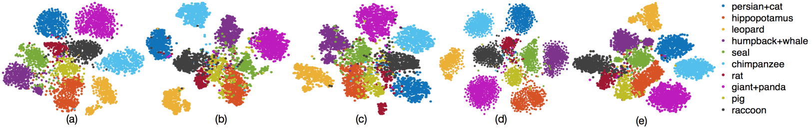


Figure: t-SNE visualization for AwA dataset (a) Original CNN features (b) Latent code for our VZSL under inductive zero-shot setting (c) Reconstructed features under inductive zero-shot setting (d) Latent code for our VZSL under transductive zero-shot setting (e) Reconstructed features under transductive setting. Different colors indicate different classes.

Summary

- Summary
  1. We present a deep generative framework for learning to predict unseen classes, focusing on inductive and trandsuctive ZSL.
  2. Our framework models each seen/unseen class using a class-specific latent-space distribution.
  3. Distribution method provides more robustness as compared to othe existing ZSL method use point based distance metric.
  4. We achieve state-of-the-art results.

Thank you !

References I

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid.
Label-embedding for attribute-based classification.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 819–826, 2013.

Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele.
Evaluation of output embeddings for fine-grained image classification.
In *CVPR*, pages 2927–2936, 2015.

Maxime Bucher, Stéphane Herbin, and Frédéric Jurie.
Improving semantic embedding consistency by metric learning for zero-shot classiffication.
In *European Conference on Computer Vision*, pages 730–746. Springer, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei.
Imagenet: A large-scale hierarchical image database.
In *CVPR*, pages 248–255. IEEE, 2009.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al.
Devise: A deep visual-semantic embedding model.
In *NIPS*, pages 2121–2129, 2013.

Yanwei Fu and Leonid Sigal.
Semi-supervised vocabulary-informed learning.
In *CVPR*, pages 5337–5346, 2016.

Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong.
Zero-shot object recognition by semantic manifold distance.
In *CVPR*, pages 2635–2644, 2015.

References II

Yuchen Guo, Guiguang Ding, Xiaoming Jin, and Jianmin Wang.
Transductive zero-shot recognition via shared model space learning.
In *AAAI*, volume 3, page 8, 2016.

Diederik P Kingma and Max Welling.
Auto-encoding variational bayes.
In *ICLR*, 2014.

Elyor Kodirov, Tao Xiang, and Shaogang Gong.
Semantic autoencoder for zero-shot learning.
In *CVPR*, 2017.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling.
Learning to detect unseen object classes by between-class attribute transfer.
In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling.
Attribute-based classification for zero-shot visual object categorization.
TPAMI, 36(3):453–465, 2014.

Yanan Li and Donghui Wang.
Zero-shot learning with generative latent prototype model.
arXiv preprint arXiv:1705.09474, 2017.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean.
Distributed representations of words and phrases and their compositionality.
In *NIPS*, pages 3111–3119, 2013.

References III

Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean.
Zero-shot learning by convex combination of semantic embeddings.
arXiv preprint arXiv:1312.5650, 2013.

Genevieve Patterson and James Hays.
Sun attribute database: Discovering, annotating, and recognizing scene attributes.
In *CVPR*, pages 2751–2758. IEEE, 2012.

Jeffrey Pennington, Richard Socher, and Christopher Manning.
Glove: Global vectors for word representation.
In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Bernardino Romera-Paredes and Philip HS Torr.
An embarrassingly simple approach to zero-shot learning.
In *ICML*, pages 2152–2161, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al.
Imagenet large scale visual recognition challenge.
IJCV, 115(3):211–252, 2015.

Karen Simonyan and Andrew Zisserman.
Very deep convolutional networks for large-scale image recognition.
arXiv preprint arXiv:1409.1556, 2014.

References IV

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng.
Zero-shot learning through cross-modal transfer.
In *NIPS*, pages 935–943, 2013.

Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov.
Learning robust visual-semantic embeddings.
arXiv preprint arXiv:1703.05908, 2017.

Vinay Kumar Verma and Piyush Rai.
A simple exponential family framework for zero-shot learning.
arXiv preprint arXiv:1707.08040, 2017.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie.
The caltech-ucsd birds-200-2011 dataset.
2011.

Qian Wang and Ke Chen.
Zero-shot visual recognition via bidirectional latent embedding.
arXiv preprint arXiv:1607.02104, 2016.

Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele.
Latent embeddings for zero-shot classification.
In *CVPR*, pages 69–77, 2016.

Yunlong Yu, Zhong Ji, Jichang Guo, and Yanwei Pang.
Transductive zero-shot learning with adaptive structural embedding.
arXiv preprint arXiv:1703.08897, 2017.

References V

Yunlong Yu, Zhong Ji, Xi Li, Jichang Guo, Zhongfei Zhang, Haibin Ling, and Fei Wu.
Transductive zero-shot learning with a self-training dictionary approach.
arXiv preprint arXiv:1703.08893, 2017.

Ziming Zhang and Venkatesh Saligrama.
Zero-shot learning via semantic similarity embedding.
In *ICCV*, pages 4166–4174, 2015.

Ziming Zhang and Venkatesh Saligrama.
Learning joint feature adaptation for zero-shot recognition.
arXiv preprint arXiv:1611.07593, 2016.

Ziming Zhang and Venkatesh Saligrama.
Zero-shot learning via joint latent similarity embedding.
In *CVPR*, pages 6034–6042, 2016.