Cross-Modal Mapping for Generalized Zero-Shot Learning by Soft-Labeling

Shabnam Daghaghi
Electrical and Computer Engineering
Rice University
Houston, TX 77005
shabnam.daghaghi@rice.edu

Tharun Medini
Electrical and Computer Engineering
Rice University
Houston, TX 77005
tharun.medini@rice.edu

Anshumali Shrivastava
Department of Computer Science
Rice University
Houston, TX 77005
anshumali@rice.edu

Abstract

Zero-Shot Learning (ZSL) is a classification task where some classes referred as *unseen classes* have no labeled training images. Instead, we only have side information (or description) about seen and unseen classes, often in the form of semantic or descriptive attributes. Lack of training images from a set of classes restricts the use of standard classification techniques and losses, including the popular cross-entropy loss. Visual information from the training images and textual data as the semantic information offer a challenging multi-modal problem. State-of-the-art approaches aim to link visual and semantic spaces by learning a cross-modal transfer/embedding and then performing classification in the embedding space. In this paper, we propose a novel architecture of casting ZSL as a standard neural-network with cross-entropy loss to embed visual space to semantic space. During training in order to introduce unseen visual information to the network, we utilize soft-labeling based on semantic similarities between seen and unseen classes. To the best of our knowledge, such similarity based soft-labeling is not explored for cross-modal transfer and ZSL. We evaluate the proposed model on four benchmark datasets for zero-shot learning, AwA, aPY, SUN and CUB datasets, and show that our model achieves significant improvement over the state-of-the-art methods in Generalized-ZSL setting on all of these datasets consistently.

1 Introduction

Supervised classifiers, specifically Deep Neural Networks, need a large number of labeled samples to perform well. Deep learning frameworks are known to have limitations in fine-grained classification regime and detecting object categories with no labeled data [1, 2, 3, 4]. On the contrary, humans can recognize new classes using their previous knowledge. This power is due to the ability of humans to transfer their prior knowledge to recognize new objects [5, 6]. Zero-shot learning aims to achieve this human-like capability for learning algorithms, which naturally reduces the burden of labeling. In zero-shot learning problem, there are no training samples available for a set of classes, referred to as unseen classes. Instead, semantic information (in the form of visual attributes or textual features) is available for unseen classes [7, 8]. Besides, we have standard supervised training data for a different set of classes, referred to as seen classes along with the semantic information of seen classes. The

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

key to solving zero-shot learning problem is to leverage trained classifier on seen classes to predict unseen classes by transferring knowledge analogous to humans.

In order to create a bridge between visual space and semantic attribute space, some methods utilize embedding techniques [9, 10, 2, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21] and the others use semantic similarity between seen and unseen classes [22, 23, 24]. Semantic similarity based models represent each unseen class as a mixture of seen classes. While the embedding based models follow three various directions; mapping visual space to semantic space [9, 10, 2, 11, 12, 2], mapping semantic space to the visual space [13, 14, 25, 26], and finding a latent space then mapping both visual and semantic space into the joint embedding space [15, 16, 17, 18, 19, 20, 21].

Another recent methodology which follows a different perspective is deploying Generative Adversarial Network (GAN) to generate synthetic samples for unseen classes by utilizing their attribute information [27, 28, 29]. Although generative models boost the results significantly, it is difficult to train these models. Furthermore, the training requires generation of large number of samples followed by training on a much larger augmented data which hurts their scalability.

Our Contribution: We propose a simple fully connected neural network architecture with unified (both seen and unseen classes together) cross-entropy loss along with soft-labeling. Soft-labeling is the key novelty of our approach which enables the training data from the seen classes to also train the unseen class. We directly use attribute similarity information between the correct seen class and the unseen classes to create a soft unseen label for each training data. As a result of soft labeling, training instances for seen classes also serve as soft training instance for the unseen class without increasing the training corpus. This soft labeling leads to implicit supervision for the unseen classes that eliminates the need for any unsupervised regularization such as entropy loss in [30]. Soft-labeling along with crossentropy loss enables a simple MLP network to tackle GZSL problem. Our proposed model, which we call Soft-labeled ZSL (SZSL), is simple (unlike GANs) and efficient (unlike visual-semantic pairwise embedding models) which outperforms the current state-of-the-art methods in GZSL setting on four benchmark datasets with a significant margin.

2 Proposed Methodology

Problem Definition: Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$ be training dataset includes n samples where $\mathbf{x}_i$ is the visual feature vector of the i-th image and $\mathbf{y}_i$ is the class label. All samples in $\mathcal{D}$ belong to seen classes $\mathcal{S}$ and during training there is no sample available from unseen classes $\mathcal{U}$. The total number of classes is $C = |\mathcal{S}| + |\mathcal{U}|$. Semantic information or attributes $\mathbf{a}_k \in \mathbb{R}^a$, are given for all C classes and the collection of all attributes are represented by attribute matrix $\mathbf{A} \in \mathbb{R}^{a \times C}$. In the inference phase, our objective is to predict the correct classes (either seen or unseen) of the test dataset $\mathcal{D}'$. The classic ZSL setting assumes that all test samples in $\mathcal{D}'$ belong to unseen classes $\mathcal{U}$ and tries to classify test samples only to unseen classes $\mathcal{U}$. While in a more realistic setting i.e. GZSL, there is no such an assumption and we aim at classifying samples in $\mathcal{D}'$ to either seen or unseen classes $\mathcal{S} \cup \mathcal{U}$.

Network Architecture: The proposed architecture is shown in Figure 1. For the visual features as the input, for all five benchmark datasets, we use the extracted visual features by a pre-trained ResNet-101 on ImageNet provided by [3]. We do not fine-tune CNN that generates the visual features unlike model in [30]. In this sense, our proposed model is also fast and straightforward to train.

Soft Labeling: In GZSL problem, we do not have any training instance from unseen classes, so the output nodes corresponding to unseen classes are always inactive during learning. The true labels (hard labels) used for training only represent seen classes so the cross entropy cannot penalize unseen classes. Moreover, the available similarity information between the seen and unseen attributed is never utilized.

We propose soft labeling based on the similarity between semantic attributes. With soft labeling, during training we enrich each label with partial assignments to unseen classes and as [31] shows, soft labels act as a regularizer which allows each training case to enforce much more constraint on weights. To assign a distribution to all unseen classes, a natural choice is to transform seen-to-unseen similarities to probabilities (soft labels) shown in Equation (1). In order to control the flatness of the unseen distribution, we utilize temperature parameter τ. The Impact of temperature τ on unseen distribution is depicted in Figure 2.a for a particular seen class. Soft labeling implicitly introduces

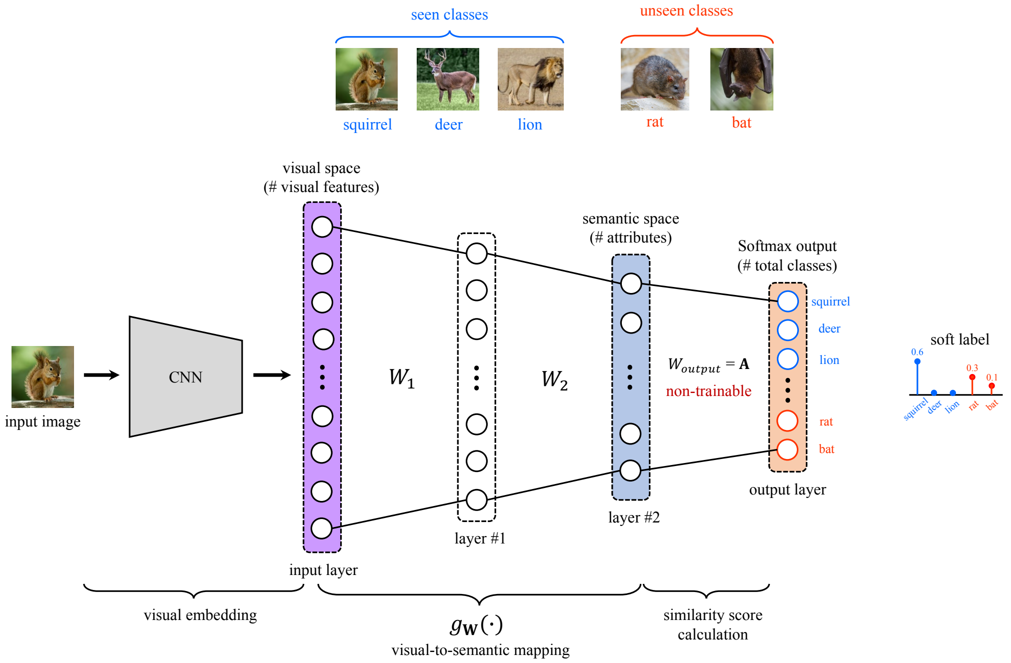


Figure 1: The overall workflow of the SZSL classifier and architecture of the proposed MLP. Layers #1 and #2 provide the nonlinear embedding $g_{\mathbf{W}}(.)$ to map visual features to attribute space and their weights W_1, W_2 are learned by SGD. The output layer with non-trainable weights $\mathbf{A}$, basically calculates dot-products of semantic representation of the input and all class attributes simultaneously. Soft-labels are also shown for a sample image from *squirrel* class.

unseen visual features into the network without generating fake unseen samples as in generative methods [27, 28, 29]. Hence our proposed approach is able to reproduce same effect as in generative models without the need to create fake samples and train generative models that are known to be difficult to train. Below is the formal description of *temperature* Softmax:

$$y^u_{i,k} = q\frac{\exp\left(s_{i,k}/\tau\right)}{\sum_{j\in\mathcal{U}}\exp\left(s_{i,j}/\tau\right)} \qquad \text{where} \qquad s_{i,j} \triangleq \langle \mathbf{a}_i, \mathbf{a}_j \rangle \tag{1}$$

where $\mathbf{a}_i$ is the i-th column of attribute matrix $\mathbf{A} \in \mathbb{R}^{a\times C}$ which includes both seen and unseen class attributes: $\mathbf{A} = [\mathbf{a}_1 \mid \mathbf{a}_2 \mid \cdots \mid \mathbf{a}_C]$. And $s_{i,j}$ is the *true* similarity score between two classes i, j based on their attributes. τ and q are temperature parameter and total probability assigned to unseen distribution, respectively. Also $y^u_{i,k}$ is the soft label (probability) of unseen class k for seen class i. It should be noted that q is the sum of all unseen soft labels i.e. $\sum_{k\in\mathcal{U}} y^u_{i,k} = q$.

Training Strategy: The proposed classifier produces a C-dimensional vector of class probabilities $\mathbf{p}$ for each sample $\mathbf{x}_i$ as $\mathbf{p}(\mathbf{x}_i) = \mathit{Softmax}\left(\mathbf{A}^T g_{\mathbf{w}}\left(\mathbf{x}_i\right)\right)$ where $\mathbf{A}^T g_{\mathbf{w}}\left(\mathbf{x}_i\right)$ is a C-dimensional vector of all similarity scores of an input sample. Therefore, the *predicted* similarity score between semantic representation of sample $\mathbf{x}_i$ and attribute $\mathbf{a}_k$ is $\hat{s}_{i,k} \triangleq \langle g_{\mathbf{w}}(\mathbf{x}_i), \mathbf{a}_k \rangle$.

During training, we aim at learning the nonlinear mapping $g_{\mathbf{w}}(.)$ i.e. obtaining network weights $\mathbf{W}$ through:

$$\min_{\mathbf{W}} \sum_{i=1}^{n} L(\mathbf{x}_i) + \lambda \left\|\mathbf{W}\right\|_F^2 + \gamma \left\|\mathbf{W}\right\| \tag{2}$$

where λ and γ are regularization factors which are obtained through hyperparameter tuning, and $L(\mathbf{x}_i)$ is the cross-entropy loss over soft labels (L^{soft}) for each sample $\mathbf{x}_i$ (or $\mathbf{x}$ for simplicity).

The soft-loss term is expanded to seen and unseen terms as follows:

$$L^{soft}(\mathbf{x}) = -\sum_{k\in\mathcal{S}} y_k^s \log(p_k^s) - \sum_{k\in\mathcal{U}} y_k^u \log(p_k^u) \tag{3}$$

Let $\bar{p}_k^s$ and $\bar{p}_k^u$ be the normalized versions of p_k^s and p_k^u , respectively. Also the total predicted unseen probability is $\sum_{k\in\mathcal{U}} p_k^u \triangleq \hat{q}$, consequently for seen classes $\sum_{k\in\mathcal{S}} p_k^s \triangleq 1-\hat{q}$. Plugging normalized probabilities in Equation (3), we have:

$$L^{soft}(\mathbf{x}) = -\sum_{k\in\mathcal{S}} y_k^s \log(\bar{p}_k^s) - \sum_{k\in\mathcal{U}} y_k^u \log(\bar{p}_k^u) - \sum_{k\in\mathcal{S}} y_k^s \log(1-\hat{q}) - \sum_{k\in\mathcal{U}} y_k^u \log\hat{q} \tag{4}$$

Utilizing Equation (1), we have $y_k^u = q\bar{y}_k^u$, where y_k^u are soft labels of unseen classes and $\bar{y}_k^u$ is the temperature softmax where $\sum_{k\in\mathcal{U}} \bar{y}_k^u = 1$. Similarly, the normalized seen labels $\bar{y}_k^s$ can be obtained by $y_k^s = (1-q)\bar{y}_k^s$. Replacing normalized labels in Equation (4) leads to:

$$L^{soft}(\mathbf{x}) = -(1-q)\sum_{k\in\mathcal{S}} \bar{y}_k^s \log(\bar{p}_k^s) - q\sum_{k\in\mathcal{U}} \bar{y}_k^u \log(\bar{p}_k^u) - (1-q)\log(1-\hat{q}) - q\log\hat{q} \tag{5}$$

Hence the first two terms of $L^{soft}(\mathbf{x})$ is the weighted sum of cross-entropy of seen classes and cross-entropy of unseen classes. In particular, first term penalizes and controls the relative (normalized) probabilities within all seen classes and the second term acts similarly within unseen classes. We also require to penalize the total probability of all seen classes $(1-\hat{q})$ and total probability of all unseen classes $(\hat{q})$. This is accomplished through the last two terms of Equation (5) which is basically a binary cross entropy loss. Intuitively soft-loss in Equation (5) works by controlling the balance *within* seen/unseen classes (first two terms) as well as the balance *between* seen and unseen classes (last two terms). As we have shown in Equation (5), soft-loss enables the classifier to learn unseen classes by only being exposed to samples from seen classes. Hyperparameter q acts as a trade-off coefficient between seen and unseen cross-entropy losses. We can see that the regularizer is a weighted cross entropy on unseen class, which leverages similarity structure between attributes.

At the inference time, our proposed SZSL method works the same as a conventional classifier, we only need to provide the test image and the network will produce class probabilities for all seen and unseen classes.

3 Experiment

We conduct comprehensive comparison of our proposed SZSL with the state-of-the-art methods for GZSL setting on four benchmark datasets (Table 1): AwA [7], SUN attribute [32], CUB-200-2011 [33] and aPY [34]. We present the detailed description of datasets and implementation details in the Appendix A. The evaluation metric is harmonic average of seen and unseen accuracies. Since we use the standard split, the published results of other GZSL models are directly comparable. Our model outperforms the state-of-the-art methods in GZSL setting (Table 2) for all benchmark datasets.

Table 1: Statistics of four ZSL benchmark datasets

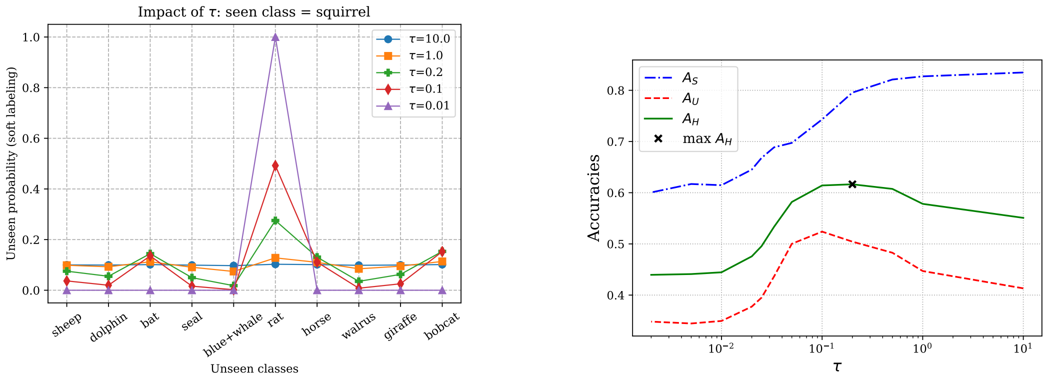
Dataset	#Attributes	#Seen Classes	#Unseen Classes	#Images
AwA	85	40	10	30475
CUB	312	150	50	11788
aPY	64	20	12	18627
SUN	102	645	72	14340

Illustration of Soft Labeling: Figure 2 shows the effect of τ and the consequent assigned unseen distribution on accuracies for AwA1 dataset. Small τ enforces q to be concentrated on nearest unseen class while large τ, spread q over all the unseen classes and basically does not introduce helpful

Table 2: Results of GZSL methods on ZSL benchmark datasets under Proposed Split (PS) [3]. U, S and H respectively stand for Unseen, Seen and Harmonic average accuracies.

	AwA			aPY			CUB			SUN		
Method	U	S	H	U	S	H	U	S	H	U	S	H
Generative Models												
f-CLSWGAN [29]	57.9	61.4	59.6	-	-	-	43.7	57.7	49.7	42.6	36.6	39.4
SP-AEN [35]	23.3	**90.9**	37.1	13.7	63.4	13.7	34.7	70.6	46.6	24.9	38.6	30.3
cycle-UWGAN [36]	59.6	63.4	59.8	-	-	-	47.9	59.3	53.0	47.2	33.8	39.4
SE-GZSL [37]	56.3	67.8	61.5	-	-	-	46.7	53.3	41.5	40.9	30.5	34.9
Non-Generative Models												
ALE [38]	16.8	76.1	27.5	4.6	73.7	8.7	23.7	62.8	34.4	21.8	33.1	26.3
SJE [16]	11.3	74.6	19.6	3.7	55.7	6.9	23.5	59.2	33.6	14.7	30.5	19.8
ConSE [39]	0.4	88.6	0.8	0.0	**91.2**	0.0	1.6	**72.2**	3.1	6.8	39.9	11.6
Sync [40]	8.9	87.3	16.2	7.4	66.3	13.3	11.5	70.9	19.8	7.9	**43.3**	13.4
DeViSE [18]	13.4	68.7	22.4	4.9	76.9	9.2	23.8	53.0	32.8	16.9	27.4	20.9
CMT [2]	0.9	87.6	1.8	1.4	85.2	2.8	7.2	49.8	12.6	8.1	21.8	11.8
ZSKL [4]	18.9	82.7	30.8	10.5	76.2	18.5	21.6	52.8	30.6	20.1	31.4	24.5
DCN [30]	25.5	84.2	39.1	14.2	75.0	23.9	28.4	60.7	38.7	25.5	37.0	30.2
SZSL (Ours)	**58.8**	72.5	**64.9**	**36.6**	57.3	**44.5**	**49.1**	48.0	**48.5**	**42.2**	32.8	**36.9**

unseen class information to the classifier. The optimal value for τ is 0.2 for AwA dataset as depicted in Figure 2.b. The impact of τ on the assigned distribution for unseen classes is shown in Figure 2.a when seen class is *squirrel* in AwA dataset. Unseen distribution with $\tau = 0.2$, well represents the similarities between seen class (*squirrel*) and similar unseen classes (*rat*, *bat*, *bobcat*) and basically verifies the result of Figure 2.b where $\tau = 0.2$ is the optimal temperature. While in the extreme cases, when $\tau = 0.01$, distribution on unseen classes in mostly focused on the nearest unseen class, *rat*, and consequently the other unseen classes' similarities are ignored. Also $\tau = 10$ flattens the unseen distribution which results in high uncertainty and does not contribute helpful unseen class information to the learning.



(a)

(b)

Figure 2: The impact of temperature parameter τ for AwA dataset. (a) unseen soft labels (before multiplying q) produced by temperature Softmax for various τ, (b) accuracies versus τ for proposed SZSL classifier.

4 Conclusion

We proposed a discriminative GZSL classifier with visual-to-semantic mapping and cross-entropy loss. During training, while SZSL is trained on a seen class, it simultaneously learns similar unseen classes through soft labels based on semantic class attributes. We deploy similarity based soft labeling on unseen classes that allows us to learn both seen and unseen signatures simultaneously via a simple architecture. Our proposed soft-labeling strategy along with cross-entropy loss leads to a novel regularization via generalized similarity-based weighted cross-entropy loss that can successfully tackle GZSL problem. Soft-labeling offers a trade-off between seen and unseen accuracies and provides the capability to adjust these accuracies based on the particular application. We achieve state-of-the-art performance, in GZSL setting, on all four ZSL benchmark datasets while keeping the model simple, efficient and easy to train.

References

[1] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015.

[2] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 935–943. Curran Associates, Inc., 2013.

[3] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591, 2017.

[4] Hongguang Zhang and Piotr Koniusz. Zero-shot kernel learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7670–7679, 2018.

[5] Yanwei Fu and Leonid Sigal. Semi-supervised vocabulary-informed learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5337–5346, 2016.

[6] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

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

[8] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[9] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1410–1418. Curran Associates, Inc., 2009.

[10] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2152–2161, Lille, France, 07–09 Jul 2015. PMLR.

[11] 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.

[12] Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive zero-shot action recognition by word-vector embedding. *International Journal of Computer Vision*, 123(3):309–333, 2017.

[13] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2021–2030, 2017.

[14] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2452–2460, 2015.

[15] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2016.

[16] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015.

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

[18] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2121–2129. Curran Associates, Inc., 2013.

[19] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 69–77, 2016.

[20] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6034–6042, 2016.

[21] Ziad Al-Halah, Makarand Tapaswi, and Rainer Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5975–5984, 2016.

[22] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 4166–4174, 2015.

[23] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot object recognition by semantic manifold distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2635–2644, 2015.

[24] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2441–2448, 2014.

[25] Seyed Mohsen Shojaee and Mahdieh Soleymani Baghshah. Semi-supervised zero-shot learning by a clustering-based approach. *arXiv preprint arXiv:1605.09016*, 2016.

[26] Meng Ye and Yuhong Guo. Zero-shot classification with discriminative semantic representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7140–7148, 2017.

[27] Ashish Mishra, Shiva Krishna Reddy, Anurag Mittal, and Hema A Murthy. A generative model for zero shot learning using conditional variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2188–2196, 2018.

[28] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1004–1013, 2018.

[29] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018.

[30] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Generalized zero-shot learning with deep calibration network. In *Advances in Neural Information Processing Systems*, pages 2005–2015, 2018.

[31] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[32] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758. IEEE, 2012.

[33] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. CUB Dataset. Technical report, 2011.

[34] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785. IEEE, 2009.

[35] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1043–1052, 2018.

[36] Rafael Felix, Vijay BG Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37, 2018.

[37] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4281–4289, 2018.

[38] 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.

[39] 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.

[40] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5327–5336, 2016.

[41] François Chollet. keras. `https://github.com/fchollet/keras`, 2015.

[42] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

A Appendix

A.1 Evaluation Metric

For the purpose of validation, we employ the validation splits provided along with PS [3] to perform cross-validation for hyper-parameter tuning. The main objective of GZSL is to simultaneously improve seen samples accuracy and unseen samples accuracy i.e. imposing a trade-off between these two metrics. As the result, the standard GZSL evaluation metric is harmonic average of seen and unseen accuracy. This metric is chosen to encourage the network not be biased toward seen classes. Harmonic average of accuracies is defined in Equation 6 where A_S and A_U are seen and unseen accuracies, respectively.

$$A_H = \frac{2A_S A_U}{A_S + A_U} \tag{6}$$

A.2 Dataset Description

The proposed method is evaluated on four benchmark ZSL datasets. The statistics for the datasets are shown in table 3. Animal with Attributes (AwA) [7, 8] dataset is a coarse-grained benchmark dataset for ZSL/GSZl. It has 30475 image samples from 50 classes of different animals and each class comes with side information in the form of attributes (e.g. animal size, color, specific feature, place of habitat). Attribute space dimension is 85 and this dataset has a standard split of 40 seen and 10 unseen classes introduced in [8]. Caltech-UCSD-Birds-200-2011 (CUB) [33] is a fine-grained ZSL benchmark dataset. It has 11,788 images from 200 different types of birds and each class comes with 312 attributes. The standard ZSL split for this dataset has 150 seen and 50 unseen classes [15]. SUN Attribute (SUN) [32] is a fine-grained ZSL benchmark dataset consists of 14340 images of different scenes and each scene class is annotated with 102 attributes. This dataset has a standard ZSL split of 645 seen and 72 unseen classes. attribute Pascal and Yahoo (aPY) [34] is a small and coarse-grained ZSL benchmark dataset which has 14340 images and 32 classes of different objects (e.g. aeroplane, bottle, person, sofa, ...) and each class is provided with 64 attributes. This dataset has a standard split of 20 seen classes and 12 unseen classes.

Table 3: Statistics of four ZSL benchmark datasets

Dataset	#Attributes	#Seen Classes	#Unseen Classes	#Images
AwA	85	40	10	30475
CUB	312	150	50	11788
aPY	64	20	12	18627
SUN	102	645	72	14340

A.3 Implementation Details

We utilized Keras [41] with TensorFlow back-end [42] to implement our model

The input to the model is the visual features of each image sample extracted by a pre-trained ResNet-101 [43] on ImageNet provided by [3]. The dimension of visual features is 2048.

To evaluate SZSL, we follow the popular experimental framework and the Proposed Split (PS) in [3] for splitting classes into seen and unseen classes to compare GZSL/ZSL methods. Utilizing PS ensures that none of the unseen classes have been used in the training of ResNet-101 on ImageNet. To obtain statistically consistent results, the reported accuracies are averaged over 30 trials (using different initialization) after tuning hyper-parameters with cross-validation.

We cross-validate $\tau \in [10^{-2}, 10]$, *mini-batch size* $\in \{64, 128, 256, 512, 1024\}$, $q \in [0, 1]$, *hidden layer size* $\in \{128, 256, 512, 1024, 1500\}$ and *activation function* $\in$\{tanh, sigmoid, hard-sigmoid, relu\} to tune our model. Also we ran our experiments on a machine with 56 vCPU cores, Intel(R) Xeon(R) CPU E5-2660 v4 @ 2.00GHZ and 2 NVIDIA-Tesla P100 GPUs each with 16GB memory.

A.4 Intuition

Figure 2.a illustrates the intuition of our methodology with AwA dataset. Consider a seen class *squirrel*. We compute unseen classes closest to the class *squirrel* in terms of attributes. We naturally find that the closest class is *rat* and the second closest is *bat*, while other classes such as *horse*, *dolphin*, *sheep*, etc. are not close. This is not surprising as *squirrel* and *rat* share several attribute. It is naturally desirable to have a classifier that gives *rat* higher probability than other classes. If we force this softly, we can ensure that classifier is not blind towards unseen classes due to lack of any training example.

From a learning perspective, without any regularization, we cannot hope classifier to classify unseen classes accurately. This problem was identified in [30], where they proposed entropy-based regularization in the form of Deep Calibration Network (DCN). DCN uses cross-entropy loss for seen classes, and regularize the model with entropy loss on unseen classes to train the network. Authors in DCN postulate that minimizing the uncertainty (entropy) of predicted unseen distribution of training samples, enables the network to become aware of unseen visual features. While minimizing uncertainty is a good choice of regularization, it does not eliminate the possibility of being confident about the wrong unseen class. Clearly, in our example above, the uncertainty can be minimized even when the classifier gives high confidence to an unseen class *dolphin* on an image of seen class *squirrel*. Furthermore, in many cases if several unseen classes are close to the correct class, we may not actually want low uncertainty. Utilizing similarity based soft-labeling implicitly regularizes the model in a supervised fashion. The similarity values naturally has information of how much certainty we want for specific unseen class. We believe that this supervised regularization is the critical difference why our model outperforms DCN with a significant margin.