Adaptive Confidence Smoothing for Generalized Zero-Shot Learning *

Yuval Atzmon
Bar-Ilan University, NVIDIA Research

Gal Chechik
Bar-Ilan University, NVIDIA Research

Abstract

Generalized zero-shot learning (GZSL) is the problem of learning a classifier where some classes have samples and others are learned from side information, like semantic attributes or text description, in a zero-shot learning fashion (ZSL). Training a single model that operates in these two regimes simultaneously is challenging. Here we describe a probabilistic approach that breaks the model into three modular components, and then combines them in a consistent way. Specifically, our model consists of three classifiers: A "gating" model that makes soft decisions if a sample is from a "seen" class, and two experts: a ZSL expert, and an expert model for seen classes. We address two main difficulties in this approach: How to provide an accurate estimate of the gating *probability without any training samples for unseen classes; and how to use expert predictions when it observes samples outside of its domain.*

The key insight to our approach is to pass information between the three models to improve each one's accuracy, while maintaining the modular structure. We test our approach, adaptive COnfidence SMOothing *(COSMO), on four standard GZSL benchmark datasets and find that it largely outperforms state-of-the-art GZSL models. COSMO is also the first model that closes the gap and surpasses the performance of generative models for GZSL, even-though it is a light-weight model that is much easier to train and tune.*

1. Introduction

People can easily learn to recognize visual entities based on few semantic attributes. For example, we can recognize a bird based on visual features (long beak, red crown), or find a location based on a language description (a 2-stories brick town house).

Taking into account the semantics of attributes becomes crucial when no training samples are available. This learning setup, called **Zero-Shot Learning (ZSL)** is the task of learning to recognize objects of classes without any training samples [6]. Instead, learning is based on semantic knowledge about the classes [7, 1], a "*class description*".

*An extended version of this work was published in the proceedings of *Conference on Computer Vision and Pattern Recognition, 2019* [2]

Generalized zero-shot learning (GZSL) [3] is a realistic task that extends ZSL to make predictions when test data has both seen and unseen classes. GZSL poses a unique combination of hard challenges: First, the model has to learn effectively for classes without samples (zero-shot), it also needs to learn well for classes with plenty of samples, and finally, the two very different regimes should be combined in a consistent way in a single model. GZSL can be viewed as an extreme case of classification with unbalanced classes, hence solving the last challenge can lead to better ways to address class imbalance, which is a key problem in learning with real-world data.

These learning problems operate in different setups, hence combining them into a single model is challenging. Here we describe an architecture that combines three modules, each focusing on one problem [7, 10]. It uses two (*Seen* / *Unseen*) *domain expert* classifiers and a soft gating mechanism that combines the expert predictions.

Unfortunately, softly combining expert predictions raises several difficulties. First, when training a gating module, it is hard to provide an accurate estimate of the probability that a sample is from the "unseen" classes, since by definition no samples have been observed from that class. Second, in soft combination, each model also contributes its beliefs to samples from the "other" domain, typically producing falsely confident spurious predictions that confuse the GZSL-mixture, because multi-class models tend to assign most of the softmax distribution mass to very few classes, even when their input is random noise [4].

Our main contributions address the main difficulties in this approach: (1) A novel confidence-based gating network that estimates the *gating* probability without any training samples for unseen classes. (2) A novel adaptive prior applied during inference that smooths an expert prediction, if it believes that an image is out of the expert domain.

2. The Zero-Shot Learning Setup

In **Zero-shot learning**, a training set has labeled samples from a set of *seen* classes $\mathcal{S}$. At test time, a new set of samples is given from a set of *unseen* classes $\mathcal{U}$. Our goal is to predict the class of each sample. As a supervision signal, each class $y \in \mathcal{S} \bigcup \mathcal{U}$ is accompanied with a semantic *class description* vector $\mathbf{a}_y$. ZSL Probabilistic ap-

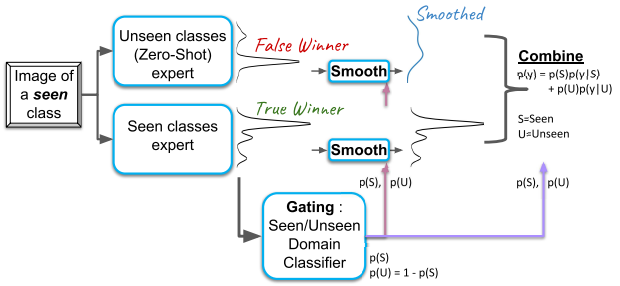

Figure 1: **Qualitative illustration our GZSL model.**

proaches learn a compatibility score for samples and class descriptions $F(\mathbf{a}_y, \mathbf{x})$, that assigns a probability for each class $p(Y{=}y|\mathbf{x}){=}F(\mathbf{a}_y, \mathbf{x})$. Y viewed as a random variable for the label y of a sample $\mathbf{x}$. In **Generalized ZSL**, samples are drawn from either *seen or unseen* domains: $Y \in \mathcal{S} \bigcup \mathcal{U}$.
Notation: For brevity, below we drop conditioning on $\mathbf{x}$.

3. Our approach

By the law of total probability, we can derive a probabilistic method that divides the GZSL problem to modules:

$$p(y) = p^S(y|\mathcal{S})p^{Gate}(\mathcal{S}) + p^{ZS}(y|\mathcal{U})p^{Gate}(\mathcal{U}). \quad (1)$$

Here, $p^S(y|\mathcal{S})$ is a model trained to classify seen $\mathcal{S}$ classes. Similarly, $p^{ZS}(y|\mathcal{U})$ is a model classifying *unseen* $\mathcal{U}$ classes, namely a ZSL model. Finally, $p^{Gate}(\mathcal{S})$ is a *gating* classifier, that distinguishes seen from unseen domains and combines the experts in a soft way.

Confidence-Based Gating Model: We train a network on top of the softmax output of the two experts, with the goal of discriminating $\mathcal{U}$ from $\mathcal{S}$ images. Since *no samples from* $\mathcal{U}$ *are available*, we create a hold-out set of classes from $\mathcal{S}$, which are not used for training the experts, and use them to estimate the output response of the expert classifiers over images of unseen classes.

Adaptive Confidence Smoothing: When a classifier is presented with a sample from an unfamiliar class, we would intuitively wish that it outputs uniform low probability to all classes, because they are all "equally wrong". However, classifiers tend to to assign most of the probability mass to few (incorrect) classes. To address this, for the unseen expert classifier $p(y|\mathcal{U})$, we apply the law of total probability and weigh two terms: (1) The classifier predictions $p(y|\mathcal{U})$; (2) A uniform smoothing prior $\pi^{\mathcal{U}}$. They are weighed by the belief that the input image is from a class that is familiar to the expert $p(\mathcal{U})$ or unfamiliar $(1 - p(\mathcal{U}))$.

$$p'(y|\mathcal{U}) \quad = \quad p(\mathcal{U})p(y|\mathcal{U}) + (1{-}p(\mathcal{U}))\pi^{\mathcal{U}}, \quad (2)$$

and similarly $p'(y|\mathcal{S}) = p(\mathcal{S})p(y|\mathcal{S}) + (1 - p(\mathcal{S}))\pi^{\mathcal{S}}$. The rightmost term means that *given that we know* an image is unfamiliar for an expert, we assign a uniform low probability, which is weighed by the belief that the input image is indeed unfamiliar.

To conclude, we clarify that $\mathcal{S}$ and $\mathcal{U}$ classes do not overlap. However, the unseen expert can also produce predictions about the seen classes which we use for estimating the gater probability. More details in the extended version [2].
Zero-Shot Expert: Here we use LAGO [1], which models new classes as compositions of soft AND-OR expressions.
Seen Classes Expert: For seen classes, we train a logistic regression classifier over pre-trained CNN [5] features.

4. Experiments

We tested our approach, *adaptive COnfidence SMOothing* (COSMO), on four GZSL benchmarks (AWA, SUN, CUB, FLOWER), following standard evaluation protocols [9, 8]. We report Acc_H, the harmonic mean of accuracy over seen classes, and accuracy over unseen classes. We also compute a seen-unseen accuracy curve by sweeping over decision threshold of the gating network. The curve trades the performance of seen and unseen domains.
Results: Table 1 describes test accuracy of COSMO and compared methods over the three benchmarks. Compared with non-generative models, COSMO improves the harmonic accuracy Acc_H by a large margin for all datasets. Notably, accuracy of COSMO is comparable with state-of-the-art generative models in spite of COSMO being much easier to train and tune than GAN-based methods.

METHOD / DATASET	AWA	SUN	CUB	FLOWER
NON-GENERATIVE				
DEM (CVPR 2017)	47.3	-	29.2	-
KERNEL (CVPR 2018)	29.8	23.6	28.9	-
TRIPLE (TIP 2019)	38.6	28.1	37.2	-
RN (CVPR 2018)	46.7	-	47	-
GENERATIVE				
F-CLSWGAN (CVPR 2018)	59.6	39.4	49.7	65.6
CYCLE-WGAN (ECCV 2018)	59.8	39.4	**53.0**	65.2
COSMO AND BASELINES				
DCN (NIPS 2018)	39.1	30.2	38.7	-
LAGO (UAI 2018)	33.7	23.9	35.1	-
CS+LAGO (ECCV 2016)	54.5	29.8	48.9	-
(OURS) COSMO (CVPR 2019)	**63.6**	**41**	**50.2**	**68.8**

Table 1: Comparing Acc_H of COSMO with state-of-the-art GZSL models.

The seen-unseen plane (Figure 2): We provide a full Seen-Unseen curve (blue dots) that shows how COSMO trades-off the metrics. COSMO can be tuned to select any operation point along the curve, and achieves better or equivalent performance at all regions.

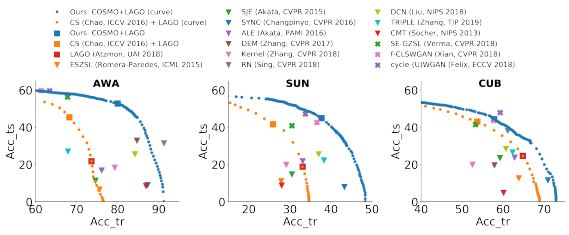

Figure 2: The Seen-Unseen curve for COSMO (blue dots). **Squares**: best COSMO model and its LAGO-based baselines, **Triangles**: non-generative approaches, **'X'**: approaches based on generative models.

References

[1] Y. Atzmon and G. Chechik. Probabilistic and-or attribute grouping for zero-shot learning. In *Proceedings of the Thirty-Forth Conference on Uncertainty in Artificial Intelligence*, 2018.

[2] Y. Atzmon and G. Chechik. Adaptive confidence smoothing for generalized zero-shot learning. In *CVPR*, 2019.

[3] R. Chao, S. Changpinyo, B. Gong, and S. F. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ICCV*, 2016.

[4] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

[5] K.He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[6] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*. IEEE, 2009.

[7] R. Socher, M. Ganjoo, C. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.

[8] Y. Xian, C. Lampert, B. Schiele, and Z. Akata. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *arXiv preprint arXiv:1707.00600*, 2017.

[9] Y. Xian, B. Schiele, and Z. Akata. Zero-shot learning - the good, the bad and the ugly. In *CVPR*, 2017.

[10] H. Zhang and P. Koniusz. Model selection for generalized zero-shot learning. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018.