Supplementary Material
Adaptive Confidence Smoothing for Generalized Zero-Shot Learning

Yuval Atzmon
Bar-Ilan University, NVIDIA Research
yuval.atzmon@biu.ac.il

Gal Chechik
Bar-Ilan University, NVIDIA Research
gal.chechik@biu.ac.il

A. A walk-through example

Figure S.1 demonstrates the inference process of COSMO with an without smoothing. An image (*panel a*) is processed by two experts: (1) An expert of unseen classes produces a distribution of confidence scores $p^{ZS}(y, \mathcal{U})$ (2) An expert of seen classes produces a distribution of confidence scores $p^{S}(y, \mathcal{S})$. Next, the CBG gating network (Section 4.1) combines these confidence scores into a belief $p^{Gate}(\mathcal{U})$.

Without smoothing (*panel b*): Here, $p^{ZS}(y, \mathcal{U})$ and $p^{S}(y, \mathcal{S})$ are normalized to $p^{ZS}(y|\mathcal{U})$, $p^{S}(y|\mathcal{S})$ and then a joint prediction is estimated by soft combining the modules with Eq. (2). In the example, the unseen expert produces overly confident prediction for a wrong (*distractor*) class (red bar). When soft combining the expert decisions, this overwhelms the correct decision of the seen expert (blue bar), producing a false positive detection of distractor class.

With smoothing (*panel c*): Here, $p^{ZS}(y, \mathcal{U})$ and $p^{S}(y, \mathcal{S})$ are smoothed to $p'(y|\mathcal{U})$, $p'(y|\mathcal{S})$ with Eq. (4) and then a joint prediction is estimated by soft combining the modules with Eq. (2). In the example, the over confident prediction of the unseen expert is smoothed (red bar). When soft combining the expert decisions, it allows the model to reach a correct decision (blue bar).

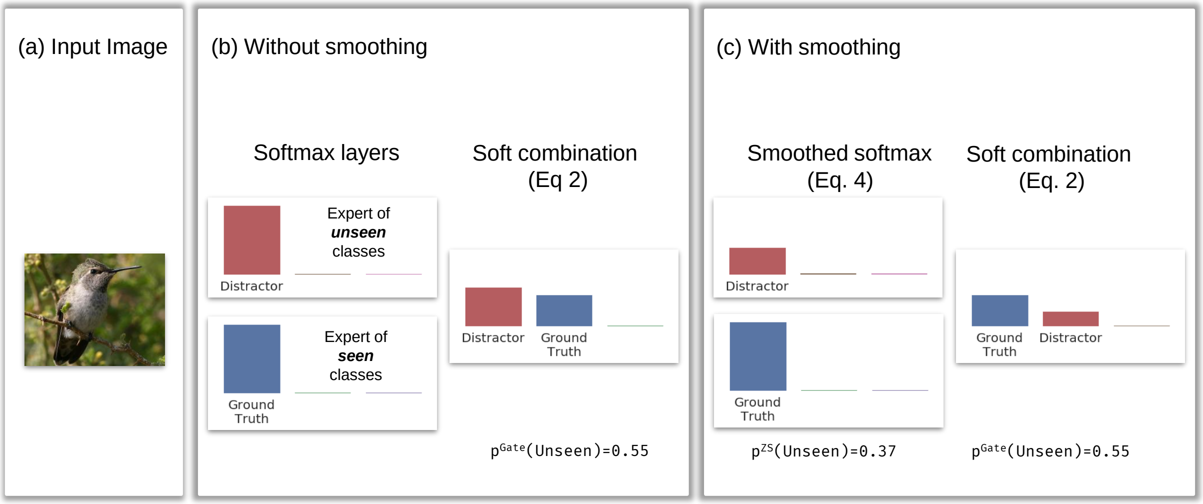

Figure S.1. A walk-through example

B. Negative results for OOD methods

We tested two state-of-the-art methods for out-of-distribution detection: *ODIN* [2] and *Ensemble* (Ensemble, [3]). We observed that taking a perturbation hurts OOD metrics with both these methods. In addition, in *Ensemble*, although quality metrics improved for the left-out training subsets during training time, the ensemble models learned to overfit the left-out subsets and failed to generalize to Unseen-Val set, better than using the baseline Max-Softmax-1 .

We believe this result may be due to two factors: **(1) Fine-grained datasets are harder:** CUB, SUN and AWA are fine grained datasets. For an un-trained eye, all their unseen samples may appear as in-distribution. For example, only a few fine-grained details discriminate "Black Throated Blue Warbler" ($\in \mathcal{S}$) of "Cerulean Warbler" ($\in \mathcal{U}$). Therefore we believe that a perturbation would have a similar effect on images from $\mathcal{S}$ or $\mathcal{U}$. **(2) Shallow vs Deep:** In the standard GZSL protocol we use, each sample is represented as a feature vector extracted from a deep CNN pre-trained on ImageNet. We found that the best classifier for this data is a shallow logistic regression classifier. This is different than ODIN and Ensemble that make the perturbation along a deep network.

C. Seen-Unseen curves for COSMO+fCLSWGAN [4]

Figure S.2 provides a full Seen-Unseen curve (pink dots) that shows how COSMO+fCLSWGAN trades-off the metrics. We compare it with a curve that we computed for the CS+fCLSWGAN baseline (gray dots) and also show the results (operation-points) reported for the compared methods (pink-square), selected with cross-validation by choosing the best Acc_H on GZSL-Val set.

The pink curve shows that on all datasets, COSMO produces equivalent or better performance compared to fCLSWGAN baseline (pink-**X**). However in most cases the operation-point selected with cross validation (pink-square) is inferior to fCLSWGAN baseline (olive-square and Table 1).

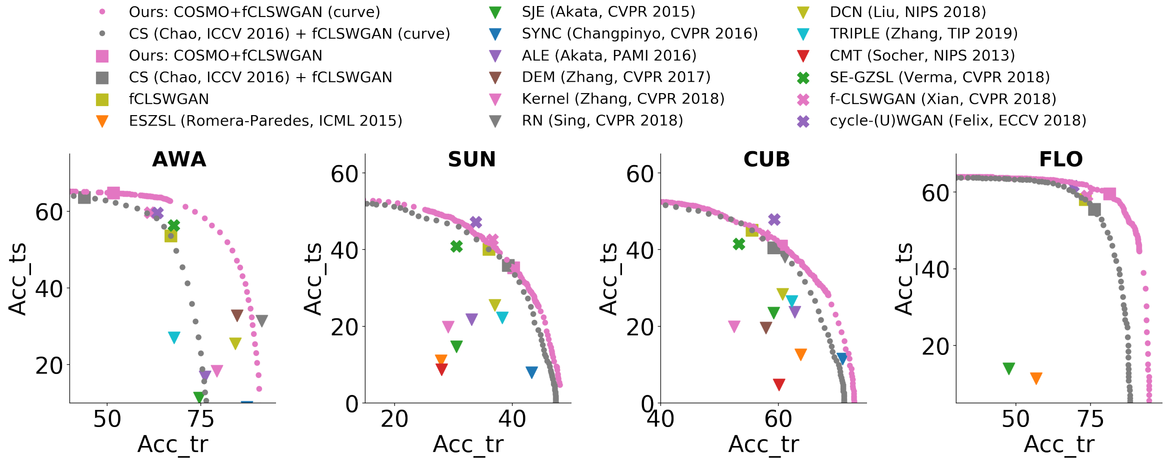

Figure S.2. The Seen-Unseen curve for COSMO+fCLSWGAN , compared to: (1) The curve of *CS [1] +fCLSWGAN* baseline, (2) 15 baseline GZSL models. Dot markers denote samples of each curve. **Squares**: COSMO cross-validated model and its fCLSWGAN*-based baselines. **Triangles**: non-generative approaches, **'X'**: approaches based on generative-models.

D. Joint training of all modules

We now explain why the GZSL setup prevents from training the gater jointly with the $\mathcal{S}$ and $\mathcal{U}$ experts. Basiclaly, in GZSL, one cannot mix seen and unseen samples during the same learning phase. More specifically, to adhere to the standard GZSL protocol by [5] in which some test samples come from unseen validation classes, one has two options. (1) Do not use these classes when training the seen expert S. This decimates S's accuracy on them. (2) Do use them for training S. In that case, all labeled samples are seen and the *gater* cannot learn to discriminate seen from unseen.

We ran two experiments on CUB to evaluate these two options, training the components jointly with a unified loss.

In the first case, accuracy on seen classes degrades from 72.8% to 53%, and on the GZSL task Acc_H degrades from 50.2% (COSMO) to 26.3%. In the second case, there were no samples of unseen classes when training the model. This greatly hurts the accuracy, leading to: Acc_{ts} = 0.1%, Acc_{tr} = 72.8%, Acc_H =1.9% far worse than original COSMO.

References

[1] R. Chao, S. Changpinyo, B. Gong, and Sha F. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ICCV*, 2016. 2

[2] S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 1

[3] A. Vyas, N. Jammalamadaka, X. Zhu, S. Das, B. Kaul, and T. L. Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *ECCV*, 2018. 1

[4] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018. 2

[5] Y. Xian, B. Schiele, and Z. Akata. Zero-shot learning - the good, the bad and the ugly. In *CVPR*, 2017. 2