Creativity Inspired Zero-Shot Learning Supplementary Materials

Mohamed Elhoseiny[1,2] Mohamed Elfeki[3]
[1]Facebook AI Research (FAIR), [2]King Abdullah University of Science and Technology (KAUST), [3]University of Central Florida
mohamed.elhoseiny@kaust.edu.sa elfeki@cs.ucf.edu

1 Parakeet Auklet vs Crested Auklet AUC on CUB dataset (SCS split)

We hypothesized that our method is better in generalization than standard generative ZSL approaches at L51-151 in the main paper. We conduct an additional experiment to verify this claim by plotting the Seen-Unseen curves for only Parakeet Auklet among the seen classes and Crested Auklet among the unseen classes. We note that the prediction space (T) still includes the 200 CUB species (see Fig 1), but with a focus on analyzing these two categories. The AUC for the baseline GAZSL is 0.139 and for our CIZSL (GAZSL + our loss) is 0.271 $\approx$ 100% relative improvement for discriminating these two classes. This demonstrates how the confusion between those two classes is drastically reduced by using our loss, especially for the unseen Crested Auklet (x-axis).

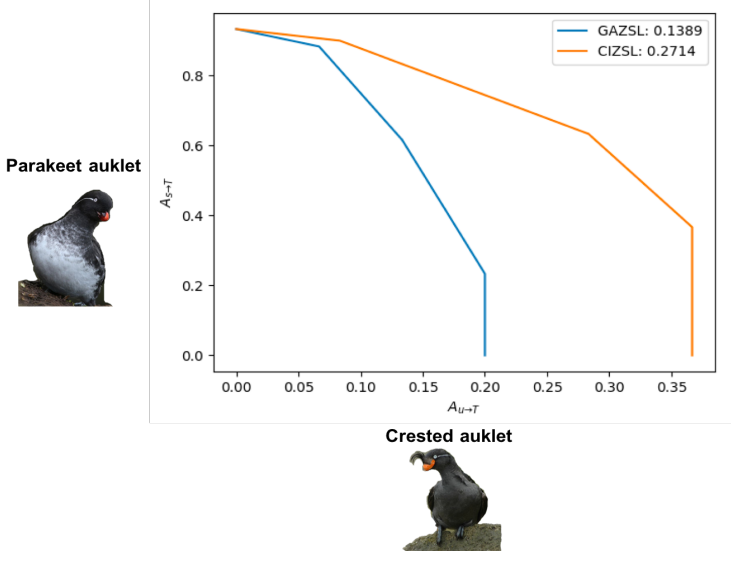


Fig. 1: Seen Unseen Curve for Parakeet Auklet (Seen) on the y-axis versus Crested Auklet (unseen) on the x-axis for GAZSL and CIZSL (GAZSL+our loss)

2 Divergence Measures

We generalize the expression of the creativity term to a broader family of divergences, unlocking new way of enforcing deviation from seen classes.

In [1], Sharma-Mittal divergence was studied, originally introduced [2]. Given two parameters (α and β), the Sharma-Mittal (SM) divergence $SM_{\alpha,\beta}(p\|q)$, between two distributions p and q is defined $\forall \alpha > 0,\ \alpha \neq 1,\ \beta \neq 1$ as

$$SM(\alpha,\beta)(p||q) = \frac{1}{\beta-1}\left[\sum_i (p_i^{1-\alpha} {q_i}^{\alpha})^{\frac{1-\beta}{1-\alpha}} - 1\right] \tag{1}$$

It was shown in [1] that most of the widely used divergence measures are special cases of SM divergence. For instance, each of the Rényi, Tsallis and Kullback-Leibler (KL) divergences can be defined as limiting cases of SM divergence as follows:

$$\begin{aligned} R_\alpha(p\|q) &= \lim_{\beta\to 1} SM_{\alpha,\beta}(p\|q) = \frac{1}{\alpha-1}\ln(\sum_i p_i^{\alpha} q_i^{1-\alpha})), \\ T_\alpha(p\|q) &= \lim_{\beta\to \alpha} SM_{\alpha,\beta}(p\|q) = \frac{1}{\alpha-1}(\sum_i p_i^{\alpha} q_i^{1-\alpha}) - 1), \\ KL(p\|q) &= \lim_{\beta\to 1,\alpha\to 1} SM_{\alpha,\beta}(p\|q) = \sum_i p_i \ln(\frac{p_i}{q_i}). \end{aligned} \tag{2}$$

In particular, the Bhattacharyya divergence [3], denoted by $B(p\|q)$ is a limit case of SM and Rényi divergences as follows as $\beta \to 1, \alpha \to 0.5$

$$B(p\|q) = 2 \lim_{\beta\to 1,\alpha\to 0.5} SM_{\alpha,\beta}(p\|q) = -\ln\Big(\sum_i p_i^{0.5} q_i^{0.5}\Big). \tag{3}$$

Since the notion of creativity in our work is grounded to maximizing the deviation from existing shapes and textures through KL divergence, we can generalize our MCE creativity loss by minimizing Sharma Mittal (SM) divergence between a uniform distribution and the softmax output $\hat{D}$ as follows

$$\mathcal{L}_{SM} = SM(\alpha,\beta)(\hat{D}||u) = SM(\alpha,\beta)(\hat{D}||u) = \frac{1}{\beta-1}\sum_i (\frac{1}{K}^{1-\alpha} \hat{D_i}^{\alpha})^{\frac{1-\beta}{1-\alpha}} - 1 \tag{4}$$

3 Training Algorithm

To train our model, we consider visual-semantic feature pairs, images and text, as a joint observation. Visual features are produced either from real data or synthesized by our generator. We illustrate in algorithm 1 how G and D are alternatively optimized with an Adam optimizer. The algorithm summarizes the training procedure. In each iteration, the discriminator is optimized for n_d steps (lines $6-11$), and the generator is optimized for 1 step (lines $12-14$). It is important to mention that when L_e has

parameters parameters like γ and β for Sharma-Mittal(SM) divergence, in Eq. 7, that we update these parameters as well by an Adam optimizer and we perform min-max normalization for L_e within each batch to keep the scale of the loss function the same. We denote the parameters of the entropy function as θ_E (lines 15). Also, we perform min-max normalization at the batch level for the entropy loss in equation 5

Algorithm 1 Training procedure of our approach. We use default values of $n_d = 5$, $\alpha = 0.001$, $\beta_1 = 0.5$, $\beta_2 = 0.9$

1: **Input:** the maximal loops N_{step}, the batch size m, the iteration number of discriminator in a loop n_d, the balancing parameter λ_p, Adam hyperparameters α_1, β_1, β_2.
2: **for** iter $= 1, ..., N_{step}$ **do**
3: Sample random text minibatches t_a, t_b, noise z^h
4: Construct t^h using Eq.6 with different α for each row in the minibatch
5: $\tilde{x}^h \leftarrow G(t^h, z^h)$
6: **for** $t = 1, ..., n_d$ **do**
7: Sample a minibatch of images x, matching texts t, random noise z
8: $\tilde{x} \leftarrow G(t, z)$
9: Compute the discriminator loss L_D using Eq. 4
10: $\theta_D \leftarrow \text{Adam}(\bigtriangledown_{\theta_D} L_D, \theta_D, \alpha_1, \beta_1, \beta_2)$
11: **end for**
12: Sample a minibatch of class labels c, matching texts T_c, random noise z
13: Compute the generator loss L_G using Eq. 5
14: $\theta_G \leftarrow \text{Adam}(\bigtriangledown_{\theta_G} L_G, \theta, \alpha_1, \beta_1, \beta_2)$
15: $\theta_E \leftarrow \text{Adam}(\bigtriangledown_{\theta_E} L_G, \theta, \alpha_1, \beta_1, \beta_2)$
16: **end for**

4 Zero-Shot Retrieval Qualitative Samples

Figure 2 shows qualitative examples of successful and unsuccessful retrieval in CUB-SCS(easy). Even when the model fails to retrieve the exact unseen class, it tends to retrieve visually similar images.

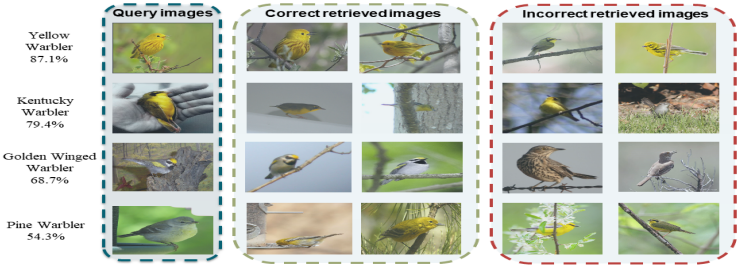


Fig. 2: Qualitative results of zero-shot retrieval on CUB dataset using SCS setting.

We show several examples of the retrieval on CUB dataset using SCS split setting. Given a query semantic representation of an unseen class, the task is to retrieve images from this class. Each row is an unseen class. We show three correct retrievals as well as one incorrect retrieval, randomly picked. We note that, even when the method fails to retrieve the correct class, it tends to retrieve visually similar images. For instance, in the Red bellied Woodpecker example (last row in the first subfigure). Our algorithm mistakenly retrieves an image of the red headed woodpecker. It is easy to notice the level of similarity between the two classes, given that both of them are woodpeckers and contain significant red colors on their bodies.

Fig. 3: Qualitative results of zero-shot retrieval on CUB dataset using SCS setting.

5 Ablation Study

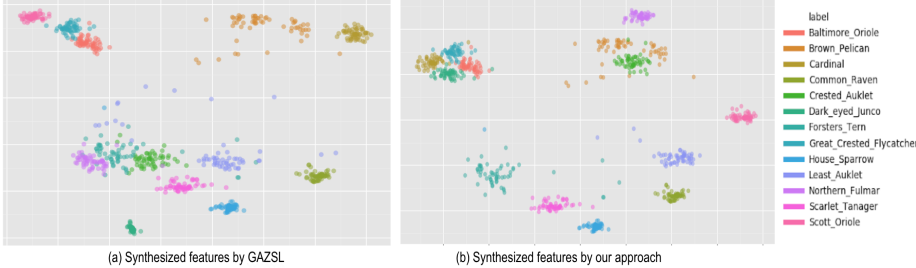
In this section we perform an ablation study to investigate best distribution for α in Eq. 6. Unlike our experiments in section 5 of original text where λ is cross validated, in this ablation we fix λ to examine the effect of changing α distribution on α, we achieve better performance. We observe that when we introduce more variation. Note that generalized Seen-Unseen AUC accuracy is very similar to the results reported in Table 4 of the main paper.

Metric	Top-1 Accuracy (%)				Seen-Unseen AUC (%)			
Dataset	CUB		NAB		CUB		NAB	
Split-Mode	SCS	SCE	SCS	SCE	SCS	SCE	SCS	SCE
GAZSL [4]- No creative loss	43.7	10.3	35.6	8.6	35.4	8.7	20.4	5.8
$\alpha = 0.5$	**45.7**	**13.9**	38.6	9.1	39.6	11.2	24.2	6.0
$\alpha \sim \mathcal{U}(0, 1)$	45.3	13.2	38.4	9.7	39.7	11.4	24.1	**7.3**
$\alpha \sim \mathcal{U}(0.2, 0.8)$	45.3	13.7	**38.8**	**9.7**	**39.7**	**11.8**	**24.6**	6.7

Table 1: Ablation Study using Zero-Shot recognition on **CUB** & **NAB** datasets with two split settings. We experiment the best α distribution in Eq. 6 of original text.

6 Visual Representation

Zhang *et al.* [5] showed that fine-grained recognition of bird species can be improved by detecting objects parts and learning a part-based learning representations on top. More specifically, ROI pooling is performed on the detected bird parts (e.g., wing, head) then semantic features are extracted for each part as a representation. They named their network Visual Part Detector/Encoder network (VPDE-net) which has VGG [6] as backbone architecture. We use the VPDE-net as our feature extractor of images for all our experiments on fine-grained bird recognition data sets, so are all the baselines.



(a) Synthesized features by GAZSL (b) Synthesized features by our approach

Fig. 4: t-SNE visualization of features of randomly selected unseen classes. Compared to GAZSL[4], our method preserves more inter-class discrimination.

References

1. Akturk, E., Bagci, G., Sever, R.: Is sharma-mittal entropy really a step beyond tsallis and rényi entropies? arXiv preprint cond-mat/0703277 (2007)
2. Sharma, B.D., Mittal, D.P.: New nonadditive measures of entropy for discrete probability distributions. J. Math. Sci **10** (1975) 28–40
3. Kailath, T.: The divergence and bhattacharyya distance measures in signal selection. IEEE transactions on communication technology **15**(1) (1967) 52–60
4. Zhu, Y., Elhoseiny, M., Liu, B., Peng, X., Elgammal, A.: A generative adversarial approach for zero-shot learning from noisy texts. In: CVPR. (2018)
5. Zhang, H., Xu, T., Elhoseiny, M., Huang, X., Zhang, S., Elgammal, A., Metaxas, D.: Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In: CVPR. (2016) 1143–1152
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. (2015)