Link the head to the "peak": Zero Shot Learning from Noisy Text description at Part Precision Supplementary Materials

Mohamed Elhoseiny[1,2*], Yizhe Zhu[1*], Han Zhang[1], and Ahmed Elgammal[1]
elhoseiny@fb.com, yizhe.zhu@rutgers.edu, {han.zhang, elgammal}@cs.rutgers.edu
[1]Rutgers University, Department of Computer Science, [2] Facebook Research

This supplementary document includes the following sections. For reproducing the results, our code, data, and models are publically available link [1].

1. λ_1 and λ_2 Setting in Experiments.
2. Gradient Derivations
3. More Qualitative Examples
4. More Figures and Detailed Results

1. λ_1 and λ_2 Setting

Similar to methods in the literature (e.g., [3, 2]), we learn the hyper-parameters by cross validation (CV) on the validation set, with the grid search in the range of $10^{[3:6]}$ for both λ_1 and λ_2 (i.e., 4×4 grid).

1. *CUB (Easy Split) :* $\lambda_1 = 10^5$ and $\lambda_2 = 10^4$.
2. *NABirds (Easy Split):* $\lambda_1 = 10^5$ and $\lambda_2 = 10^4$.
3. *CUB (Hard Split):* $\lambda_1 = 10^6$ and $\lambda_2 = 10^4$
4. *NABirds (Hard Split):* $\lambda_1 = 10^6$ and $\lambda_2 = 10^5$.

We find it intuitive to see higher values lambdas after cross-validation for the *Hard Split* since regularization becomes more important as shared information gets smaller. Morevoer, we did not find the method very sensitive to the hyper parameters. For instance, the performance on *CUB (Easy Split)* with $\lambda_1 = 10^5$, the performance of $\lambda_2 = 10^3$, $\lambda_2 = 10^4$, and $\lambda_2 = 10^5$ are 35.4%, 37.2%, and 35.9%, respectively.

2. Gradient Derivations

2.1. Gradients for Equation 5 : Fix $\mathbf{W_t}$, and optimize over $\mathbf{W_x}$

We name the loss in Equation 3 in the paper as L.
Let $\mathbf{X} = [\mathbf{X^{(1)}}, \mathbf{X^{(2)}} \cdots \mathbf{X^{(P)}}] \in \mathbb{R}^{d \times P \cdot d_X}$.
Let $\mathbf{W_x^\mathsf{T}} = [\mathbf{W_x^1}; \mathbf{W_x^2}; ...; \mathbf{W_x^7}]$, where $\mathbf{W_x^p} \in R^{d_X \times d}$; and $\mathbf{W_t} \in R^{d \times d_T}$.
(a) The first term:

$$||(\sum_{p=1}^{P} \mathbf{X}^{(p)^\mathsf{T}} \mathbf{W_x^p}^\mathsf{T}) \mathbf{W_t} \mathbf{T} - \mathbf{Y}||_F^2 = ||\mathbf{X}^\mathsf{T} \mathbf{W_x^\mathsf{T}} \mathbf{W_t} \mathbf{T} - \mathbf{Y}||_F^2 = \mathrm{Tr}((\mathbf{X}^\mathsf{T} \mathbf{W_x^\mathsf{T}} \mathbf{W_t} \mathbf{T} - \mathbf{Y})(\mathbf{X}^\mathsf{T} \mathbf{W_x^\mathsf{T}} \mathbf{W_t} \mathbf{T} - \mathbf{Y})^\mathsf{T}) \quad (1)$$

* Both authors contributed equally to this work

We can get the derivative of **the first term** in the objective function w.r.t. $\mathbf{W_x^p}$:

$$\frac{\partial ||\mathbf{X}^\intercal \mathbf{W_x^\intercal W_t T} - \mathbf{Y}||_F^2}{\partial \mathbf{W_x^p}} = 2\mathbf{W_t T T^\intercal W_t^\intercal W_x X}^{(p)} {\mathbf{X}^{(p)}}^\intercal - 2\mathbf{W_t T Y^\intercal} {\mathbf{X}^{(p)}}^\intercal \tag{2}$$

(b) The derivative of **the second term** in the objective function w.r.t. every part $\mathbf{W_x^p}$.

$$\frac{\partial \lambda_1 ||\mathbf{W_x}^\intercal \mathbf{W_t T}||_F^2}{\partial \mathbf{W_x^p}} = \quad 2\lambda_1 \mathbf{W_t T T^\intercal} \mathbf{W_t}^\intercal \mathbf{W_x^p} \tag{3}$$

(c) For **the third term** in the objective function, we do the partial derivative for each part:

$$\lambda_2 Tr(\mathbf{W_x^p W_t D}_l^p \mathbf{W_t^\intercal} {\mathbf{W_x^p}}^\intercal) \tag{4}$$

The derivative of **the third term** in the objective function w.r.t. every part $\mathbf{W_x^p}$:

$$\frac{\partial \lambda_2 Tr(\mathbf{W_x^p W_t D}_l^p \mathbf{W_t^\intercal} {\mathbf{W_x^p}}^\intercal)}{\partial \mathbf{W_x^p}} = 2\lambda_2 \mathbf{W_x^p W_t D}_l^p \mathbf{W_t^\intercal} \tag{5}$$

Therefore, the partial derivative of the loss function w.r.t. $\mathbf{W_x^p}$ is:

$$\frac{\partial L}{\partial \mathbf{W_x^p}} = 2\mathbf{W_t T T^\intercal W_t^\intercal W_x X}^{(p)} {\mathbf{X}^{(p)}}^\intercal - 2\mathbf{W_t T Y^\intercal} {\mathbf{X}^{(p)}}^\intercal + 2\lambda_1 \mathbf{W_t T T^\intercal} \mathbf{W_t}^\intercal \mathbf{W_x^p} + 2\lambda_2 \mathbf{W_x^p W_t D}_l^p \mathbf{W_t^\intercal} \tag{6}$$

2.2. Gradients for Equation 4: Fix $\mathbf{W_x^p}$, and optimize over $\mathbf{W_t}$

The loss function is rewritten as:

$$L = ||\mathbf{X}^\intercal \mathbf{W_x^\intercal W_t T} - \mathbf{Y}||_F^2 + \lambda_1 ||\mathbf{W_x^\intercal W_t T}||_F^2 + \lambda_2 \sum_{p=1}^{P} \mathrm{Tr}(\mathbf{W_x^p W_t D}_l^p \mathbf{W_t^\intercal} {\mathbf{W_x^p}}^\intercal) \tag{7}$$

The partial derivative over $\mathbf{W_t}$:

$$\frac{\partial L}{\partial \mathbf{W_t}} = 2\mathbf{W_x X X^\intercal} \mathbf{W_x}^\intercal \mathbf{W_t T T^\intercal} - 2\mathbf{W_x X Y T^\intercal} + 2\lambda_1 \mathbf{W_x W_x^\intercal W_t T T^\intercal} + 2\lambda_2 \sum_{i=1}^{P} {\mathbf{W_x^p}}^\intercal \mathbf{W_x^p W_t D}_l^p \tag{8}$$

3. More Qualitative Results

We show more qualitative examples in this section.

Figure 1: Part-to-Term connectivity demonstrated on falsely labeled samples

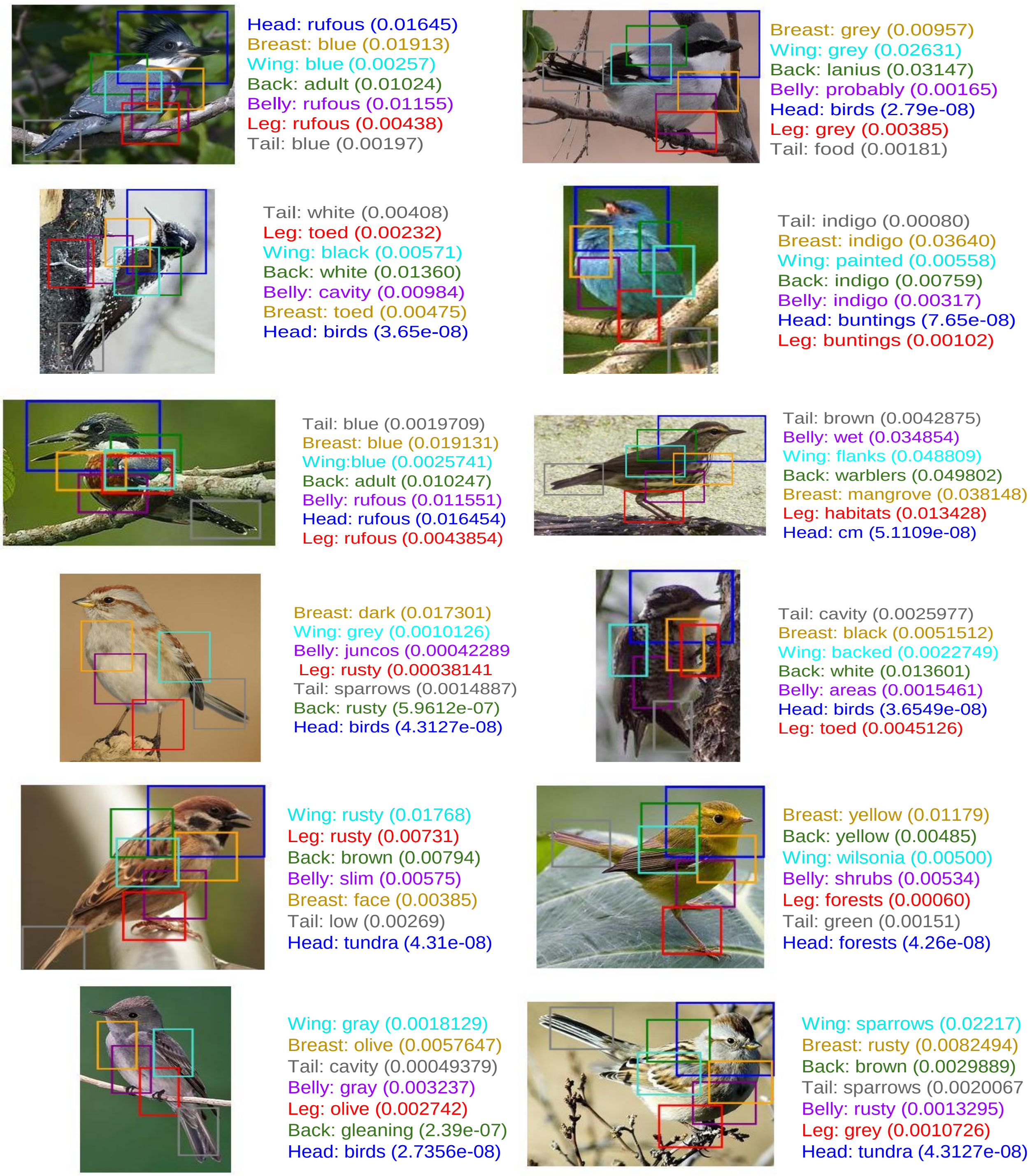
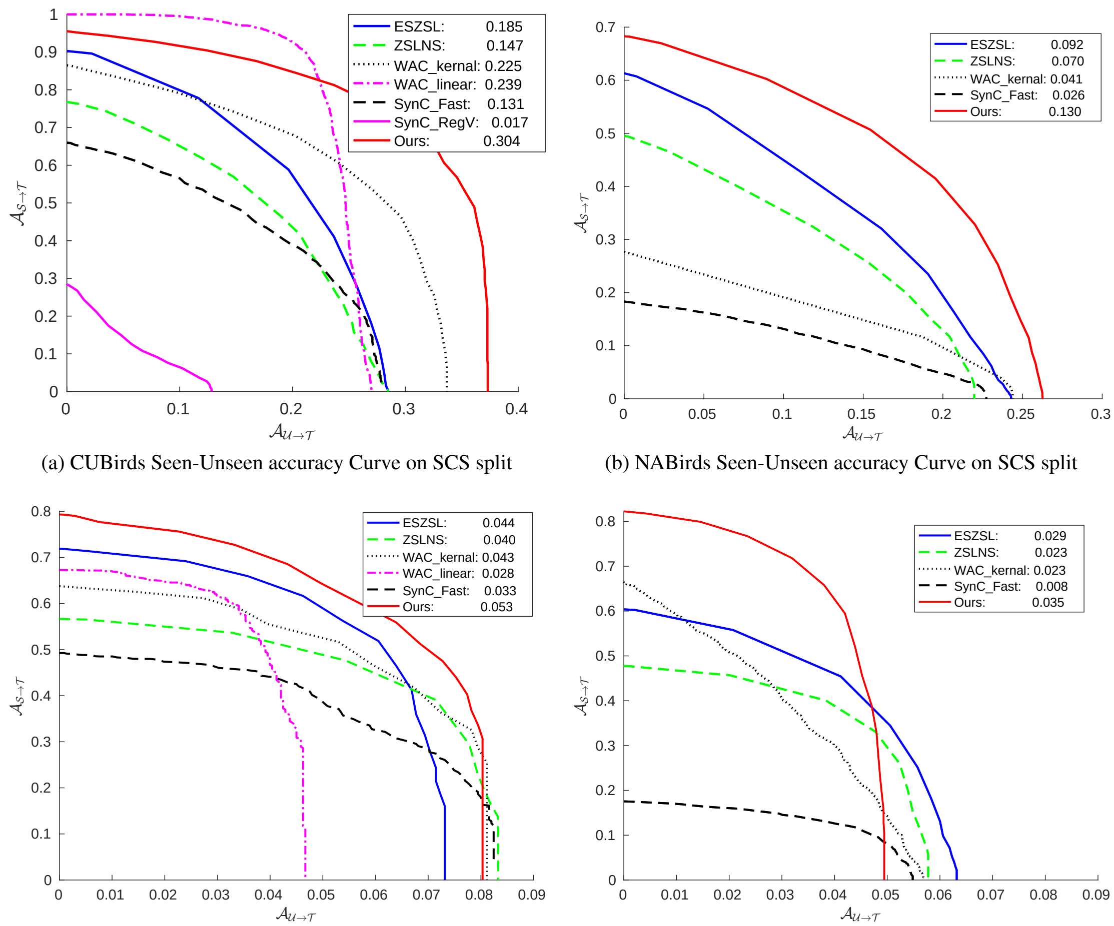


Figure 2: Part-to-Term connectivity demonstrated on correctly labeled samples

4. More Figures and Detailed Results

More Generalized Zero-Shot Learning Curves; see the captions for the corresponding benchmark.



(a) CUBirds Seen-Unseen accuracy Curve on SCS split

(b) NABirds Seen-Unseen accuracy Curve on SCS split

(c) CUBirds Seen-Unseen accuracy Curve on SCE split

(d) NABirds Seen-Unseen accuracy Curve on SCE split

Figure 3: Result comparison with Seen-Unseen accuracy Curves on different split settings.

References

[1] Our implementation: ZSL PP. https://github.com/EthanZhu90/ZSL_PP, 2017. 1

[2] R. Qiao, L. Liu, C. Shen, and A. v. d. Hengel. Less is more: zero-shot learning from online textual documents with noise suppression. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on, year=2016*. 1

[3] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 2152–2161, 2015. 1