Few-Shot Zero-Shot Learning: Knowledge Transfer with Less Supervision – Supplementary Material –

Nanyi Fei[1], Jiechao Guan[1], Zhiwu Lu[2] (✉), and Yizhao Gao[2]

[1] School of Information, Renmin University of China, Beijing, China
[2] Beijing Key Laboratory of Big Data Management and Analysis Methods, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
luzhiwu@ruc.edu.cn

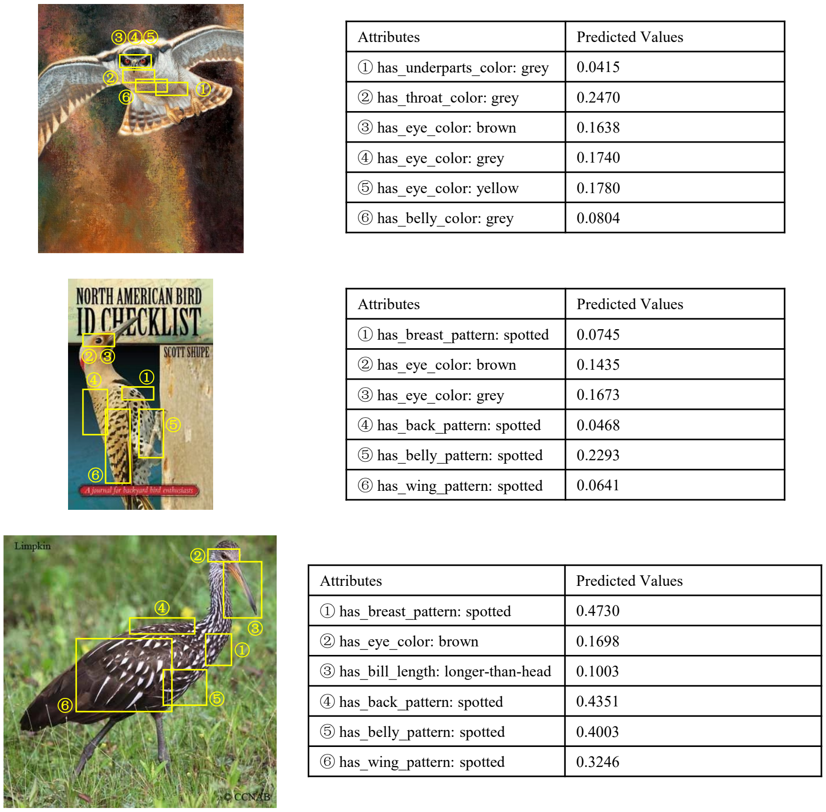
In this document, we provide more support results to show the effectiveness of our algorithm. Firstly, we present more comparative results with MixMatch [1] being adopted as the propagation method. Secondly, we show several examples of propagated attributes of web images. Thirdly, we show the effect of different parameter settings on our algorithm. Next, we give a convergence analysis of our algorithm. Finally, we demonstrate the general applicability of our model on solving the social image annotation (SIA) problem.

1 More Results for Standard FSZSL

We show more comparative results under our standard FSZSL setting on the CUB [2] dataset (with 5 annotated images per seen class) in Table 1. Four baselines (i.e., RPL [3], ESZSL [4], SAE [5], and ZSKL [6]) are selected here because they can utilize the attributes (with continuous, rather than binary values) as inputs for ZSL. For each model, the nearest neighbor based label propagation (NN-LP) and the semi-supervised learning (SSL) method MixMatch [1] (one of the strongest, but without denoising) are respectively adopted to propagate attributes from a few labelled seen class samples to unlabelled seen class ones. We can see from Table 1 that NN-LP is generally comparable to MixMatch under our FSZSL setting. One explanation is that stronger SSL methods tend to induce too much noise due to the insufficient initial supervision. In contrast, with only one-step propagation, NN-LP induces much less noise. Therefore, it is reasonable to use NN-LP for all compared ZSL models in our main paper.

Table 1. Comparative results (%) under the standard FSZSL setting on the CUB dataset. Average top-1 accuracy is reported (with standard deviation in bracket).

Propagation Method	K	RPL [3]	ESZSL [4]	SAE [5]	ZSKL [6]
NN-LP	5	29.4(1.4)	24.1(1.4)	29.7(1.5)	32.2(1.6)
MixMatch [1]	5	27.7(1.4)	25.2(1.4)	30.3(2.1)	31.2(1.2)



Attributes	Predicted Values
① has_underparts_color: grey	0.0415
② has_throat_color: grey	0.2470
③ has_eye_color: brown	0.1638
④ has_eye_color: grey	0.1740
⑤ has_eye_color: yellow	0.1780
⑥ has_belly_color: grey	0.0804

Attributes	Predicted Values
① has_breast_pattern: spotted	0.0745
② has_eye_color: brown	0.1435
③ has_eye_color: grey	0.1673
④ has_back_pattern: spotted	0.0468
⑤ has_belly_pattern: spotted	0.2293
⑥ has_wing_pattern: spotted	0.0641

Attributes	Predicted Values
① has_breast_pattern: spotted	0.4730
② has_eye_color: brown	0.1698
③ has_bill_length: longer-than-head	0.1003
④ has_back_pattern: spotted	0.4351
⑤ has_belly_pattern: spotted	0.4003
⑥ has_wing_pattern: spotted	0.3246

Fig. 1. Examples of the propagated attributes (right) of web images (left) on the CUB+Web dataset. For clear illustration, we manually annotate the propagated attributes for each web image.

2 Qualitative Results for SAP

We provide qualitative results to show why our Sparse Attribute Propagation (SAP) is important under our new FSZSL setting. Fig. 1 presents three examples of propagated attributes (obtained by our SAP) of web images. Note that each web image originally has an all-zero attribute vector, and its propagated attributes take the values in the range (0, 1). We can observe that the attributes of unannotated web images are often predicted correctly by our SAP.

3 Parameter Sensitivity Test

Since only a few images are annotated with attributes from each seen class under our new FSZSL setting, it is impossible to select the parameters by cross-validation. In Section 5.1.1, we have mentioned that our algorithm has five parameters to tune: k_g, m, λ_1, λ_2, λ_3. In Fig. 2, we study how sensitive our model

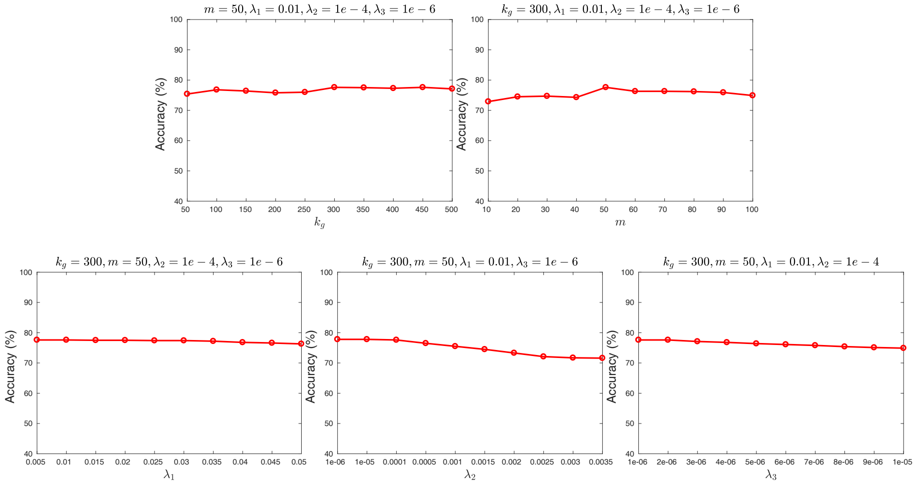


Fig. 2. The effect of different parameter settings on our algorithm for the AwA dataset.

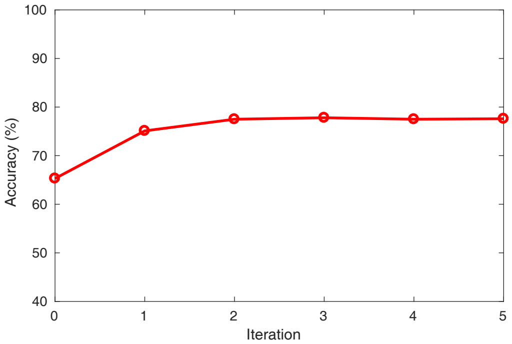


Fig. 3. Convergence analysis of our algorithm for standard FSZSL on the AwA dataset.

is to different values of the five parameters on the AwA [7] dataset, where 25 annotated images per seen class are provided. The results show that our algorithm is insensitive to these parameters when their values are in certain ranges. We thus uniformly set $k_g = 300$, $m = 50$, $\lambda_1 = 0.01$, $\lambda_2 = 1e - 4$, and $\lambda_3 = 1e - 6$.

4 Convergence Analysis Results

We further give a convergence analysis of our algorithm for standard FSZSL. Concretely, we compute the recognition accuracies on the test set at each training iteration before our algorithm stops. Fig. 3 shows the results on AwA [7] (with 25 annotated images per seen class). It can be seen that our algorithm converges very quickly (≤ 5 iterations). Moreover, the improvements over iteration 0 also clearly show the effectiveness of our algorithm under the new FSZSL setting.

Table 2. Comparative results of social image annotation on the NUS-WIDE dataset.

Method	Label Correlation	C-P (%)	C-R (%)	C-F1 (%)	I-P (%)	I-R (%)	I-F1 (%)
Ours (Refined Tags+CNN+SVM)	no	64.79	**68.16**	**66.43**	**77.71**	70.69	74.03
Ours (Refined Tags+SVM)	no	56.70	58.79	57.73	70.65	61.31	65.65
Ours (Noisy Tags+SVM)	no	14.13	53.75	22.38	22.75	47.91	30.85
SR-CNN-RNN [11]	yes	**71.73**	61.73	66.36	77.41	76.88	**77.15**
SINN [8]	yes	58.30	60.30	59.44	57.05	**79.12**	66.30
TagNeighboor [10]	no	54.74	57.30	55.99	53.46	75.10	62.46
RIA [9]	yes	52.92	43.62	47.82	68.98	66.75	67.85
CNN-RNN [12]	yes	40.50	30.40	34.70	49.90	61.70	55.20
CNN+WARP [14]	no	31.65	35.60	33.51	48.59	60.49	53.89
CNN+softmax [14]	no	31.68	31.22	31.45	47.82	59.52	53.03
CNN+logistic [8]	no	45.60	45.03	45.31	51.32	70.77	59.50

5 Generalization to Social Image Annotation

5.1 Methodology

Our SAP model can be easily extended to other partially labelled image classification tasks. To demonstrate that, we choose the social image annotation (SIA) task [8–12]. In this task, the side information collected from social media websites is typically exploited to improve the performance on image annotation. The user-provided side information can be extracted from the noisy tags [10, 11] and group labels [8]. Since the noisy tags of social images are analogous to the attributes in ZSL, by forming the semantic space using the social tags, our algorithm originally developed for ZSL can also be generalized to SIA.

However, the user-provided social tags are rather noisy and sparse, offering only a partial semantic description of the image content. With the proposed SAP model, the SIA problem can be solved as follows: (1) The noise in social tags is reduced by sparse coding; (2) The noise reduction problem is further regularized by BPL to obtain better results. Although the label correlation is not considered in our model, it is shown to generally outperform the state-of-the-art alternatives [8, 9, 11, 12] that employ the well-known recurrent neural network (RNN) [13] to model the label correlation for SIA (see Table 2).

Concretely, we first replace $\mathbf{Y}^{(s)}$ with the semantic representation defined with noisy tags, and then refine $\mathbf{Y}^{(s)}$ by running our algorithm. By concatenating $\mathbf{Y}^{*}$ and $\mathbf{X}^{(s)}$, we finally train multi-class classifiers (i.e., one-vs-all SVM) for SIA. Note that it is still time consuming to find the m smallest eigenvectors of $\mathbf{L}$ on an extremely large dataset (e.g., NUS-WIDE [15]). To keep the scalability of our model, we thus employ nonlinear approximation to find m smallest eigenvectors as in [16]. In this application, our model can be fairly compared to the state-of-the-art alternatives [8, 9, 11, 12].

5.2 Experiments

We make performance evaluation on the NUS-WIDE [15] benchmark dataset, which consists of 269,648 images and 81 class labels from Flickr image metadata. By removing invalid Flickr links and also images with no social tags, we obtain

162,806 training images and 57,202 test images. In this paper, we keep 1,000 most frequent tags to form the semantic space, and finetune ResNet101 [17] with the training set to extract the visual features.

The per-class and per-image metrics including precision and recall have been widely used in previous works. In the following, the per-class precision (C-P) and per-class recall (C-R) are obtained by computing the mean precision and recall over all the classes, while the overall per-image precision (I-P) and overall per-image recall (I-R) are computed by averaging over all the test images. Moreover, the per-class F1-score (C-F1) and overall per-image F1-score (I-F1) are used for comprehensive performance evaluation by combining precision and recall with the harmonic mean.

Our model has three variants: 1) Ours (Noisy Tags+SVM) – SVM trained with the original noisy tags; 2) Ours (Refined Tags+SVM) – SVM trained with the refined tags; 3) Ours (Refined Tags+CNN+SVM) – SVM trained with the refined tags and CNN features. Moreover, we also make comparison to the state-of-the-art models [14, 8–12]. The comparative results in Table 2 show that: (1) The refined tags obtained by our model yield over 30% improvements when C-F1 and I-F1 are concerned. (2) Our model achieves state-of-the-art performance according to C-F1 and competitive results according to I-F1. This is really impressive given that the label correlation is not exploited for SIA in our model.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (61976220 and 61832017), and Beijing Outstanding Young Scientist Program (BJJWZYJH012019100020098).

References

1. Berthelot, D., Carlini, N., Goodfellow, I.J., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems. (2019) 5050–5060
2. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology (2011)
3. Shigeto, Y., Suzuki, I., Hara, K., Shimbo, M., Matsumoto, Y.: Ridge regression, hubness, and zero-shot learning. In: ECML-PKDD. (2015) 135–151
4. Romera-Paredes, B., Torr, P.H.S.: An embarrassingly simple approach to zero-shot learning. In: ICML. (2015) 2152–2161
5. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. In: CVPR. (2017) 3174–3183
6. Zhang, H., Koniusz, P.: Zero-shot kernel learning. In: CVPR. (2018) 7670–7679
7. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. TPAMI **36** (2014) 453–465
8. Hu, H., Zhou, G.T., Deng, Z., Liao, Z., Mori, G.: Learning structured inference neural networks with label relations. In: CVPR. (2016) 2960–2968

9. Jin, J., Nakayama, H.: Annotation order matters: Recurrent image annotator for arbitrary length image tagging. In: ICPR. (2016) 2452–2457
10. Johnson, J., Ballan, L., Fei-Fei, L.: Love thy neighbors: Image annotation by exploiting image metadata. In: ICCV. (2015) 4624–4632
11. Liu, F., Xiang, T., Hospedales, T.M., Yang, W., Sun, C.: Semantic regularisation for recurrent image annotation. In: CVPR. (2017) 4160–4168
12. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: CNN-RNN: A unified framework for multi-label image classification. In: CVPR. (2016) 2285–2294
13. Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. TPAMI **31** (2009) 855–868
14. Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe, S.: Deep convolutional ranking for multilabel image annotation. arXiv preprint arXiv:1312.4894 (2013)
15. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: A real-world web image database from National University of Singapore. In: CIVR. (2009) 48:1–48:9
16. Chen, X., Cai, D.: Large scale spectral clustering with landmark-based representation. In: AAAI. (2011) 313–318
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016) 770–778