Supplementary Material: Graph convolutional networks for learning with few clean and many noisy labels

Ahmet Iscen[1], Giorgos Tolias[2], Yannis Avrithis[3], Ondřej Chum[2], and Cordelia Schmid[1]

[1] Google Research
[2] VRG, Faculty of Electrical Engineering, Czech Technical University in Prague
[3] Inria, Univ Rennes, CNRS, IRISA

A The role of base classes

The proposed method is applicable with any given feature extractor. Herein, we describe the learning of the feature extractor on a set of base classes according to a standard few-shot learning setup and benchmark [12]. Then, we describe the extended classifiers to the union of all classes, *i.e.* base classes and novel classes, which are the ones used in Section 5.

A.1 Representation learning on base classes

We are given a set $X_{\mathcal{B}} \subset \mathcal{X}$ of examples, each having a clean label in a set of *base classes* $C_{\mathcal{B}}$ with $|C_{\mathcal{B}}| = K_{\mathcal{B}}$. Base classes $C_{\mathcal{B}}$ are disjoint from C, which are also known as novel classes. These data are used to learn a feature representation, *i.e.* a feature extractor g_θ, by learning a $K_{\mathcal{B}}$-way base-class classifier for unseen data in $\mathcal{X}$. The parameters θ of the feature extractor and $W_{\mathcal{B}}$ of the classifier are jointly learned by minimizing the cross entropy loss

$$L_{\mathcal{B}}(C_{\mathcal{B}}, X_{\mathcal{B}}; \theta, W_{\mathcal{B}}) = -\sum_{c \in C_{\mathcal{B}}} \frac{1}{|X_{\mathcal{B}}^c|} \sum_{x \in X_{\mathcal{B}}^c} \log(\boldsymbol{\sigma}(s\hat{W}_{\mathcal{B}}^\top \hat{g}_\theta(x))_c). \tag{1}$$

The learned feature extractor parameters θ and the learned scale parameter s are used by our method as described Sections 4 and 5.

A.2 Classification on all classes

The classifier parameters $W_{\mathcal{B}}$ are used, combined with classifier parameters W learned as described in Section 5, for classification on *all classes* $C_{\mathcal{A}} = C \cup C_{\mathcal{B}}$.

Class prototypes. The concatenated parameter matrix $W_{\mathcal{A}} = [W_{\mathcal{B}}, W]$ is used for $K_{\mathcal{A}}$-way prediction on all (base and novel) classes by $\pi_{\theta, W_{\mathcal{A}}}$, where $K_{\mathcal{A}} = K + K_{\mathcal{B}}$. $W_{\mathcal{B}}$ is learned according to $L_{\mathcal{B}}(C_{\mathcal{B}}, X_{\mathcal{B}}; \theta, W_{\mathcal{B}})$ (1), while W is learned according to (5).

Method	Top-5 accuracy on all classes				
	k=1	2	5	10	20
	ResNet-10 – Few Clean Examples				
Proto.-Nets [33]†	49.5	61.0	69.7	72.9	74.6
Logistic reg. w/ H [41]†	54.4	61.0	69.0	73.7	76.5
PMN w/ H [41]†	40.8	49.9	64.2	71.9	76.9
Class proto. [9]	57.0±0.36	64.7±0.16	72.5±0.18	75.8±0.16	77.4±0.19
Class proto. w/ Att. [9]	58.1±0.48	65.2±0.15	72.9±0.25	76.6±0.18	78.8±0.16
	ResNet-10 – Few Clean & Many Noisy Examples				
Ours - class proto. (5)	**70.3±0.05**	**72.1±0.18**	**74.1±0.12**	**75.6±0.13**	**76.9±0.09**
Ours - cosine (6)	**72.4±0.07**	**73.4±0.21**	**77.2±0.20**	**78.8±0.21**	**79.2±0.17**
Ours - fine-tune	**76.0±0.10**	**77.3±0.13**	**78.7±0.19**	**80.7±0.25**	**82.2±0.14**
	ResNet-50 – Few Clean Examples				
Proto.-Nets [33]†	61.4	71.4	78.0	80.0	81.1
PMN w/ H [41] †	65.7	73.5	80.2	82.8	84.5
	ResNet-50 – Few Clean & Many Noisy Examples				
Ours - class proto. (5)	**73.8±0.33**	**76.6±0.36**	**78.9±0.19**	**80.8±0.21**	**82.2±0.14**
Ours - cosine (6)	**78.2±0.25**	**79.6±0.23**	**80.4±0.18**	**82.4±0.19**	**84.1±0.09**
Ours - fine-tune	**81.6±0.20**	**83.2±0.16**	**84.3±0.23**	**86.2±0.17**	**87.8±0.03**

Table 1. Comparison to the state of the art on the Low-shot ImageNet benchmark. We report top-5 accuracy on all classes. We use class prototypes (5), cosine classifier learning (6) and deep network fine-tuning for classification with our GCN-based data addition method. † denotes numbers taken from the corresponding papers. All other experiments are re-implemented by us.

Cosine classifier learning. Prediction on all classes is made as in the previous case, but W is learned according to (6).

Deep network fine-tuning. We now assume that base class examples are accessible too and, given all examples $X_{\mathcal{A}} = X_{\mathcal{B}} \cup X_{\mathcal{E}}$, we jointly learn the parameters θ of the feature extractor and $W_{\mathcal{A}} = [W_{\mathcal{B}}, W]$ of the $K_{\mathcal{A}}$-way cosine classifier for all classes by minimizing loss function

$$L_{\mathcal{A}}(C_{\mathcal{A}}, X_{\mathcal{A}}; \theta, W_{\mathcal{A}}) = L_{\mathcal{B}}(C_{\mathcal{B}}, X_{\mathcal{B}}; \theta, W_{\mathcal{B}}) + L(C, X_{\mathcal{E}}; \theta, W). \tag{2}$$

Note that in contrast to (6), the last term of (2) optimizes parameters θ too. As mentioned earlier, such learning is typically avoided in a few-shot learning setup. In few cases, it takes the form of fine-tuning including all base class data [26], or only lasts for a few iterations when the base class data is not accessible [6].

A.3 Results on all classes

We report the accuracy over all classes in Table 1. When fine-tuning the network by (2), the learned W is used to initialize the corresponding part of $W_{\mathcal{A}}$ and we train all layers for 10 epochs with learning rate 0.01. The results indicate that our method still brings significant improvements when all classes are used.

Method	k=1	k=5
	Few Clean Examples	
Class proto. [9]	54.2$\pm$0.77	71.2$\pm$0.61
Class proto. w/ Att. [9]	56.2$\pm$0.81	72.9$\pm$0.62
Few Clean & Many Noisy Examples - Class proto. (5)		
β-weighting, $\beta = 1$	63.5$\pm$0.77	65.2$\pm$0.81
Label Propagation	67.0$\pm$0.74	74.8$\pm$0.61
MLP	65.9$\pm$0.78	73.9$\pm$0.63
Ours	68.2$\pm$0.76	74.7$\pm$0.59

Table 2. Comparison with baselines using noisy examples on the Mini-ImageNet dataset. We report the accuracy for 5-way k-shot experiments where $k = 1$ and $k = 5$.

B Results on Mini-Imagenet

We evaluate the proposed method on another popular benchmark, *i.e.* few-shot learning on Mini-ImageNet [38]. The dataset is a subset of ImageNet [32], and contains 100 different classes, split into 64 base, 16 validation and 20 test classes [27]. Each class contains 600 images that are re-sized to a resolution of 84×84. We use the ConvNet-128 model with cosine classifier, following [9]. Novel categories are classified using class prototypes (5).

Table 2 shows the accuracy on Mini-Imagenet for the 5-way k-shot classification scenario with $k = 1$ and $k = 5$. We report the average accuracy over 600 trials along with the confidence interval. Our method brings significant improvements for $k = 1$, showing its generalization across different few-shot datasets and benchmarks.

References

1. Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P.: Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine **34**(4), 18–42 (2017)
2. Bruna, J., Zaremba, W., Szlam, A., Lecun, Y.: Spectral networks and locally connected networks on graphs. In: ICLR (2014)
3. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B.: A closer look at few-shot classification. ICLR (2019)
4. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: NeurIPS (2016)
5. Douze, M., Szlam, A., Hariharan, B., Jégou, H.: Low-shot learning with large-scale diffusion. In: CVPR (2018)
6. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML (2017)
7. Finn, C., Xu, K., Levine, S.: Probabilistic model-agnostic meta-learning. In: NeurIPS (2018)
8. Garcia, V., Bruna, J.: Few-shot learning with graph neural networks. In: ICLR (2018)

9. Gidaris, S., Komodakis, N.: Dynamic few-shot visual learning without forgetting. In: CVPR (2018)
10. Gidaris, S., Komodakis, N.: Generating classification weights with gnn denoising autoencoders for few-shot learning. In: CVPR (2019)
11. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: International Conference on Artificial Intelligence and Statistics (2010)
12. Hariharan, B., Girshick, R.: Low-shot visual recognition by shrinking and hallucinating features. In: CVPR (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
14. Henaff, M., Bruna, J., LeCun, Y.: Deep convolutional networks on graph-structured data. arXiv preprint arXiv:1506.05163 (2015)
15. Iscen, A., Tolias, G., Avrithis, Y., Furon, T., Chum, O.: Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations. In: CVPR (2017)
16. Joulin, A., van der Maaten, L., Jabri, A., Vasilache, N.: Learning visual features from large weakly supervised data. In: ECCV (2016)
17. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
18. Lee, K.H., He, X., Zhang, L., Yang, L.: CleanNet: Transfer learning for scalable image classifier training with label noise. In: CVPR (2018)
19. Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., Li, L.J.: Learning from noisy labels with distillation. In: ICCV (2017)
20. Liu, T., Tao, D.: Classification with noisy labels by importance reweighting. IEEE Trans. PAMI **38**(3) (2015)
21. Liu, Y., Lee, J., Park, M., Kim, S., Yang, E., Hwang, S.J., Yang, Y.: Learning to propagate labels: Transductive propagation network for few-shot learning. In: ICLR (2019)
22. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. ICLR (2017)
23. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML (2010)
24. Natarajan, N., Dhillon, I.S., Ravikumar, P.K., Tewari, A.: Learning with noisy labels. In: NeurIPS (2013)
25. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: CVPR (2017)
26. Qi, H., Brown, M., Lowe, D.G.: Low-shot learning with imprinted weights. In: CVPR (June 2018)
27. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: ICLR (2017)
28. Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A.: Training deep neural networks on noisy labels with bootstrapping. ICLR (2015)
29. Ren, M., Ravi, S., Triantafillou, E., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classification. In: ICLR (2018)
30. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. In: ICML (2018)
31. Rohrbach, M., Ebert, S., Schiele, B.: Transfer learning in a transductive setting. In: NeurIPS (2013)

32. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3) (2015)
33. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS (2017)
34. Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., Fergus, R.: Training convolutional networks with noisy labels. arXiv preprint arXiv:1406.2080 (2014)
35. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Commun. ACM **59**(2) (2016)
36. Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., Belongie, S.: Learning from noisy large-scale datasets with minimal supervision. In: CVPR (July 2017)
37. Vilalta, R., Drissi, Y.: A perspective view and survey of meta-learning. Artificial intelligence review **18**(2), 77–95 (2002)
38. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: NeurIPS (2016)
39. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: CVPR (2015)
40. Wang, Y., Liu, W., Ma, X., Bailey, J., Zha, H., Song, L., Xia, S.T.: Iterative learning with open-set noisy labels. In: CVPR (2018)
41. Wang, Y.X., Girshick, R., Hebert, M., Hariharan, B.: Low-shot learning from imaginary data. In: CVPR (2018)
42. Weston, J., Ratle, F., Collobert, R.: Deep learning via semi-supervised embedding. In: ICML (2008)
43. Wu, Z., Xiong, Y., Yu, S., Lin, D.: Unsupervised feature learning via non-parametric instance-level discrimination. CVPR (2018)
44. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Trans. PAMI **40**(6) (2017)
45. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: NeurIPS (2003)
46. Zhou, D., Weston, J., Gretton, A., Bousquet, O., Schölkopf, B.: Ranking on data manifolds. In: NeurIPS (2003)