doi:10.1017/S1431927621009004

Automated Data Labeling and Label Cleaning for Nanoparticle Classification in Electron Microscopy

Kate Groschner[1], Assaf Ben-Moshe[2], Alexander Pattinson[3], Wolfgang Theis[3] and Mary Scott[4]

[1]UC Berkeley, Oakland, California, United States, [2]UC Berkeley, United States, [3]University of Birmingham, United States, [4]UC Berkeley, Berkeley, California, United States

While there have been a large number of advances for rapid data acquisition in electron microscopy, gaining meaningful insights from this data remains challenging. This is often due to the challenges in creating a robust way to analyze the large amount of data generated. Manually labeling large datasets is often too time consuming, while automated techniques are more error-prone. However, recent advances in the machine learning community have focused on how to develop accurate classifiers from datasets with erroneous labels like those generated by automated techniques, yielding neural networks which are able to accurately analyze the data without the time investment from researchers [1-7]. Here, we analyze two approaches to outperforming accuracy limits imposed by label noise: noise robust techniques, which are immune to erroneous labels, and noise cleaning techniques, network training procedures which work to remove erroneous examples.

The case we use to test these methods is classifying handedness of chiral nanoparticles from scanning electron microscopy (SEM) data [8]. This dataset provides a good test case in which data labeling can be easily automated but is likely to have a certain fraction of erroneous labels. Synthesis routes for chiral particles often lead to skew in the population to one handedness or the other, to create labels we can assume all one handedness and mirror the images to create the opposite handedness. To control for the error rate we use a set of single handed nanoparticles and synthetically introduce error, as shown in Figure 1. In this way we create an electron microscopy dataset with controlled error rates to test how methods respond across varying error rates.

Using this dataset, we study which methods yield the best performance. We will first analyze the basic noise robustness of convolutional neural networks (CNNs) [2] to the binary label flip noise and the limits of relying on this robustness for network development. We will then discuss a two step training procedure in which we train basic CNNs initially with an extremely small clean dataset [7], before introducing noise, leading to a more noise robust network. Finally, we test a co-teaching method, developed by Han *et al* [5], which uses two networks trained in tandem to try and filter erroneous labels while training. We find that the two step training procedure leads to the highest accuracy networks and is robust to noise at higher error rates than other methods. Finally, we demonstrate these methods as part of a complete analysis pipeline, automating particle segmentation and handedness classification for large SEM datasets of chiral Te nanoparticles.

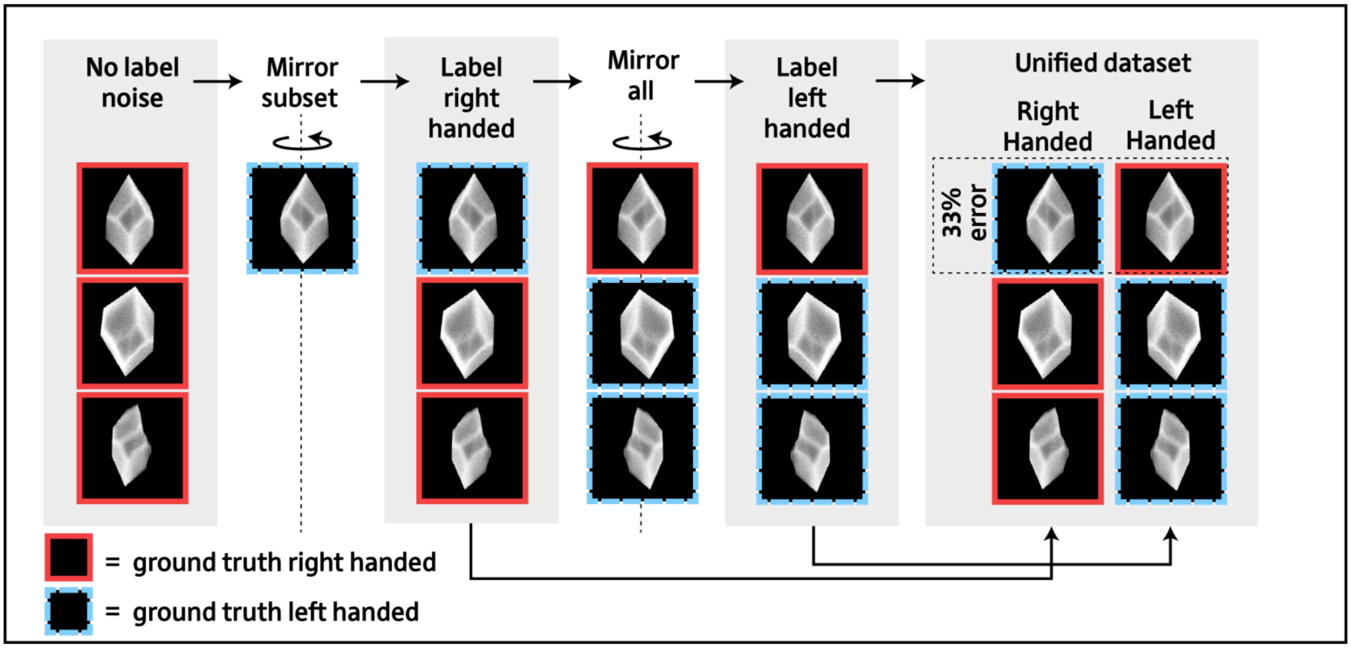
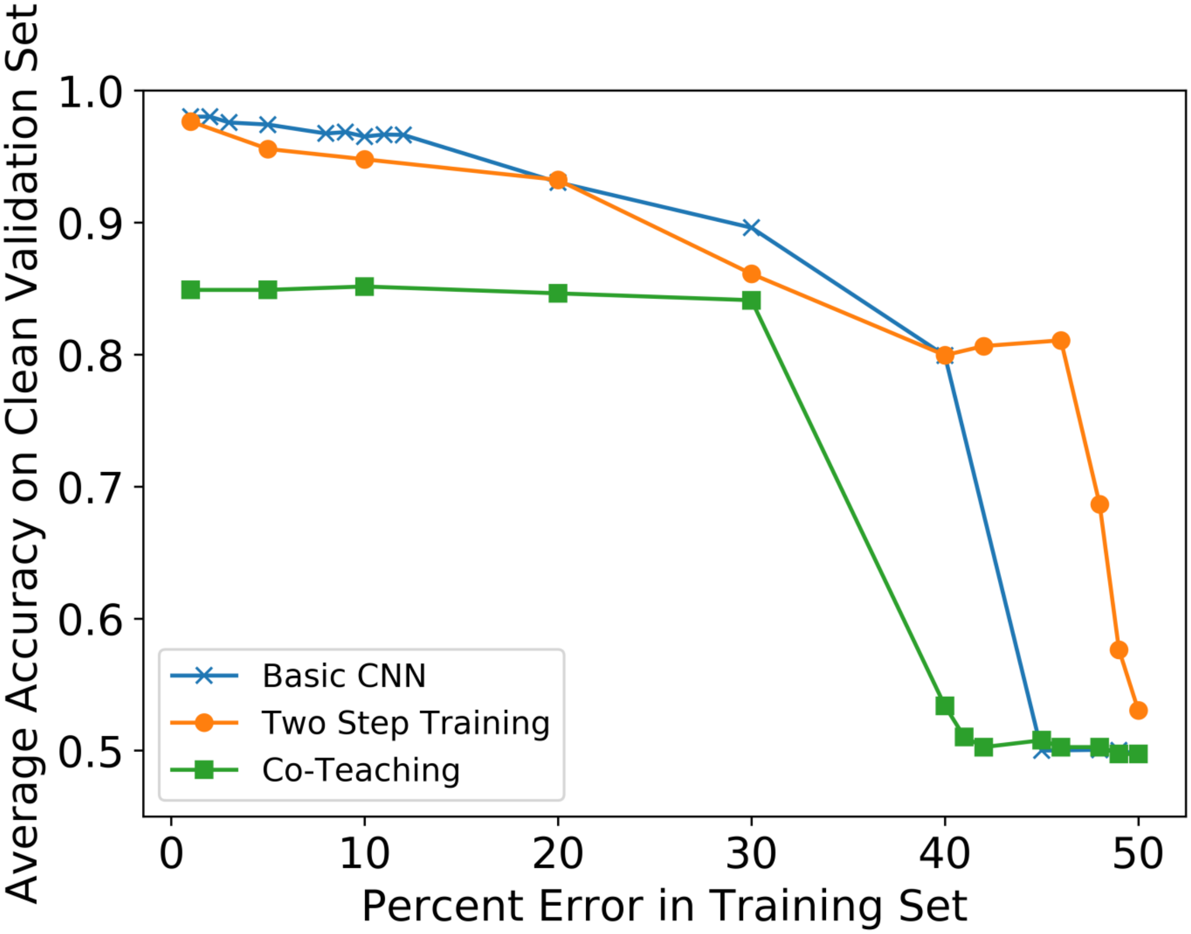


Figure 1. Process for noisy dataset creation.



https://doi.org/10.1017/S1431927621009004 Published online by Cambridge University Press

Figure 2. Average accuracy of each network type on predicting the ground truth handedness given clean validation data

References

1. Wang, Y. *et al.* Symmetric cross entropy for robust learning with noisy labels. *arXiv* 322–330 (2019).
2. Rolnick, D., Veit, A., Belongie, S. & Shavit, N. Deep learning is robust to massive label noise. *arXiv* (2017).
3. Thulasidasan, S., Bhattacharya, T., Bilmes, J., Chennupati, G. & Mohd-Yusof, J. Combating label noise in deep learning using abstention. *arXiv* (2019).
4. Arazo, E., Ortego, D., Albert, P., O'Connor, N. E. & McGuinness, K. Unsupervised label noise modeling and loss correction. *36th Int. Conf. Mach. Learn. ICML 2019* **2019–June,** 465–474 (2019).
5. Han, B. *et al.* Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Adv. Neural Inf. Process. Syst.* **2018–December,** 8527–8537 (2018).
6.Karimi, D., Dou, H., Warfield, S. K. & Gholipour, A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *arXiv* 1–22 (2019).
7. Hendrycks, D., Mazeika, M., Wilson, D. & Gimpel, K. Using trusted data to train deep networks on labels corrupted by severe noise. *Adv. Neural Inf. Process. Syst.* **2018–December,** 10456–10465 (2018).
8. Ben-Moshe, A. *et al.* Enantioselective control of lattice and shape chirality in inorganic nanostructures using chiral biomolecules. *Nat. Commun.* **5,** (2014).

https://doi.org/10.1017/S1431927621009004 Published online by Cambridge University Press