Learning From Massive Noisy Labeled Data for Image Classification

Tong Xiao[1], Tian Xia[2], Yi Yang[2], Chang Huang[2], and Xiaogang Wang[1]
[1]The Chinese University of Hong Kong. [2]Baidu Research.

Deep learning from large-scale supervised training dataset has shown very impressive improvement on image classification challenge [4]. However, it requires reliable annotations from millions of images which are often expensive and time-consuming to obtain [3], preventing deep models from being quickly trained on new image recognition problems. Thus it is necessary to develop new efficient labeling and training frameworks for deep learning.

One possible solution is to automatically collect annotations from the Internet web images (*i.e.* extracting tags from the surrounding texts) and directly use them as ground truth to train deep models. Unfortunately, these labels are extremely unreliable due to various types of noise (*e.g.* labeling mistakes from annotators or computing errors from extraction algorithms). Many works have shown that noisy labels could adversely impact the classification accuracy of the induced classifiers [7]. Various label noise-robust algorithms [6] and data cleansing algorithms [2] are developed but experiments show that performances are still affected by label noise [1].

Although annotating all the data is costly, it is often easy to obtain a small amount of clean labels. Semi-supervised learning methods [5] can be employed by simply discarding all the noisy labels. Researchers have also studied the transferability of Convolutional Neural Networks (CNNs) by finetuning an ImageNet pretrained model on smaller datasets of specific tasks [8]. However, these methods cannot fully utilize the massive noisy labeled data, which may lead to suboptimal results.

Our goal is to build an end-to-end deep learning system that is capable of training with both limited clean labels and massive noisy labels more effectively. Figure 1 shows the framework of our approach. We collect $1,000,000$ clothing images from online shopping websites. Each image is automatically assigned with a noisy label according to the keywords in its surrounding text. We manually refine $72,409$ image labels, which constitute a clean sub-dataset. All the data are then used to train CNNs, while the major challenge is to identify and correct wrong labels during the training process.

To cope with this challenge, we extend CNNs with a novel probabilistic model, which infers the true labels and uses them to supervise the training of the network. Our work is inspired by [9], which modified a CNN by inserting a linear layer on top of the softmax layer to map clean labels to noisy labels. However, [9] assumed noisy labels are conditionally independent with input images given clean labels. By examining the collected dataset, we find that this assumption is too strong to fit real-world data well. For example, in Figure 2, all the images should belong to "Hoodie". The top five are correct while the bottom five are either mislabeled as "Windbreaker" or "Jacket". Since different sellers have their own bias on different categories, they may provide wrong keywords for similar clothes. We can capture these visual patterns to estimate how likely an image is mislabeled. Based on these observations, we further introduce two types of label noise:

- **Confusing noise** makes the noisy label reasonably wrong. It usually occurs when the image content is confusing (*e.g*, the samples with "?" in Figure 1).

- **Pure random noise** makes the noisy label totally wrong. It is often caused by either the mismatch between an image and its surrounding text, or the false conversion from the text to label (*e.g.*, the samples with "×" in Figure 1).

Our proposed probabilistic model builds the relations among images, noisy labels, ground truth labels, and noise types, where the latter two are treated as latent variables. We use the Expectation-Maximization (EM) algorithm to solve the problem and integrate it into the deep learning framework. Experiments on the collected dataset show that our model can better detect and correct the wrong labels, which benefits the training of underlying CNNs.

Figure 1: Overview of our approach. Labels of web images often suffer from different types of noise. A label noise model is proposed to detect and correct the wrong labels. The corrected labels are used to train underlying CNNs.

Figure 2: Mislabeled images often share similar visual patterns.

[1] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 2006.

[2] Carla E Brodley and Mark A Friedl. Identifying mislabeled training data. *arXiv:1106.0219*, 2011.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[5] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, 2013.

[6] Naresh Manwani and PS Sastry. Noise tolerance under risk minimization. *Cybernetics, IEEE Transactions on*, 43(3):1146–1151, 2013.

[7] David F Nettleton, Albert Orriols-Puig, and Albert Fornells. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial intelligence review*, 33(4):275–306, 2010.

[8] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.

[9] Sainbayar Sukhbaatar and Rob Fergus. Learning from noisy labels with deep neural networks. *arXiv:1406.2080*, 2014.