Researchers leverage new machine learning methods to learn from noisy labels for image classification

October 12 2022, by Zhuowei Wang

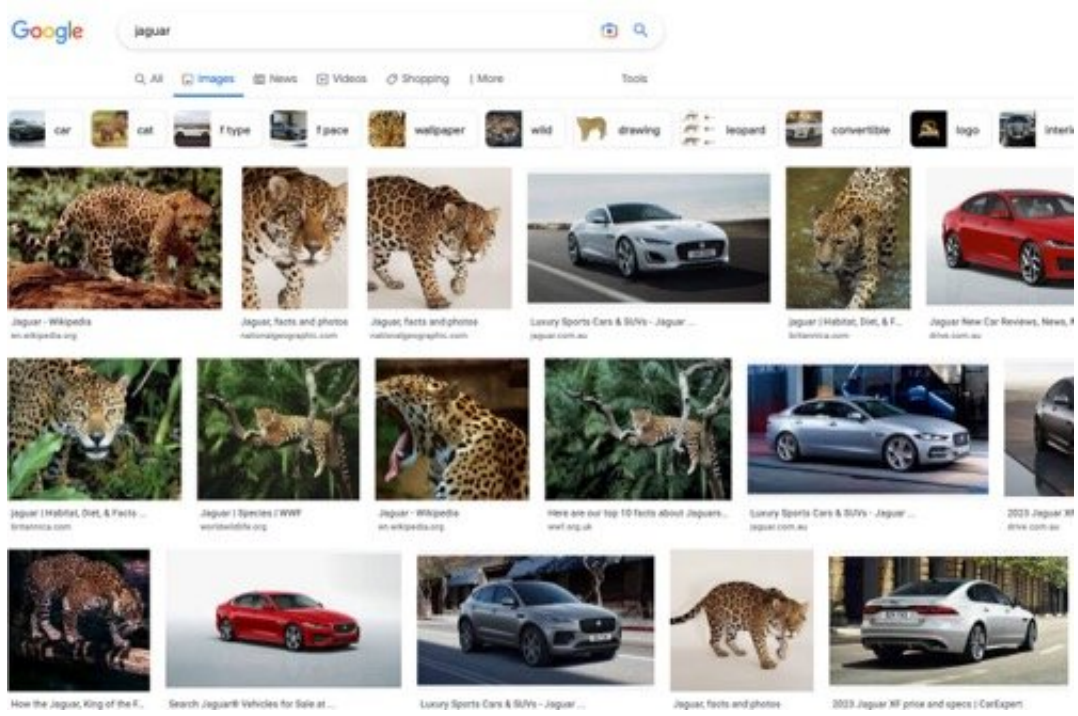

Figure 1. A normal process to generate large-scale labeled image datasets is to download the images returned by querying a keyword in the Google search engine. However, this kind of process usually generates noise labels for images. Credit: Google

The rapid development of deep learning in recent years is largely due to the rapid increase in the scale of data. The availability of large amounts

of data is revolutionary for model training by the deep learning community. With the increase in the amount of data, the scale of mainstream datasets in deep learning is also increasing. For example, the ImageNet dataset contains more than 14 million samples. In this case, larger and more complex models, such as deeper and wider convolutional layers or special network structures, are needed. This motivation basically marks the beginning of a new era of deep learning.

It is worth noting that although large-scale datasets with precise labels (all images have correct labels) are of great importance, the annotation process is rather tedious, which requires long-term human labor and huge financial investment. A common and less expensive way to collect large datasets is through online search engines. As shown in Figure 1, using the search engine to enter tag keywords can generate a series of corresponding pictures. In addition, there will be some explanatory text near the picture. By extracting the keywords in the explanatory text, one can get the category information presented in the image. Using the above methods, a large number of labeled images can be obtained quickly, allowing large-scale datasets to be built with lower overhead.

What are noisy labels?

However, using search engines for labeling will inevitably bring noisy labels to the collected datasets. These noisy labels can not provide strong supervision information for deep neural networks, leading to severe performance deterioration. For example, in medical analysis, domain expertise is required to label medical data, which may suffer from high inter- and intra-observer variability, resulting in noisy labels. These noisy labels will deteriorate the model performance, which might affect the decision-making process that impacts human health negatively. Thus it is necessary to study how to deal with noisy labels using NLL. The example of the noisy dataset is shown in Figure 2.

Figure 2. Two lines of images represent two datasets and all images are labeled as horse. The first line is a clean dataset since all the samples have the correct labels. While the second line is defined as the “noisy” dataset since the deer (in the purple box) is wrongly labeled as horse. Credit: Zhuowei Wang

How to combat noisy labels in model training?

The research problem becomes how to train deep neural networks in datasets with noisy labels. How could we improve the robustness of the deep neural networks if the image labels in the datasets are highly corrupted? To distinguish clean samples, a typical strategy is to conduct sample selection (SS) and train the models with selected samples. Many methods use the small-loss trick, where the samples with smaller losses are taken as clean ones. However, these methods simply discard other large-loss samples which may contain potentially useful information for the training process. Thus, they might not achieve satisfactory performance in highly corrupted datasets with fewer clean samples by training with only selected small-loss samples.

To make full use of all given samples, a prominent strategy is to consider selected samples as labeled "clean" data and other samples as unlabeled data and to perform semi-supervised learning (SSL). For example, some method detects clean samples by removing noisy samples whose self-ensemble predictions of the model do not match the given labels in each iteration. With the selected labeled and unlabeled data, the problem becomes an SSL problem, and a SSL backbone can be trained.

A framework that combines SS and SSL to combat noisy labels

As shown above, both methods rely on a specific SS strategy and a specific SSL model. The two components play a vitally important role in combating label noise, and stronger components are expected to achieve better performance. This motivates me to investigate a versatile algorithmic framework that can leverage various SS strategies and SSL models.

Thus, I propose a versatile framework to bridge the gap between SSL and NLL. Note that this is not just a mere combination of a semi-supervised learning algorithm with a noisy label detecting method. My framework can absorb various SS strategies and SSL backbones, utilizing their power to achieve promising performance. Guided by this framework, one can easily instantiate a specific learning algorithm for NLL, by specifying a commonly used SSL backbone with an SS strategy.

The idea behind the framework is that it effectively take advantage of the whole training set by trusting the labels of undoubtedly correct samples and utilizing only the image content of potentially corrupted samples. My framework makes use of those corrupted samples by ignoring their labels while keeping the associated image content, transforming the NLL problem into an SSL setup. The mechanism of

SSL that leverages labeled data to guide the learning of unlabeled data naturally fits well in training the model with the clean and noisy samples divided by our SS strategy. The illustration of my framework is in Figure 3.

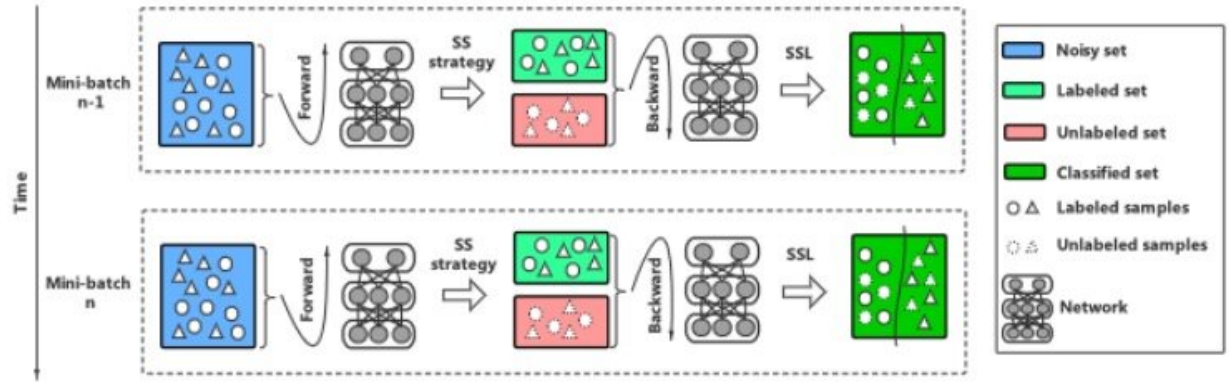

Figure 3. The schematic of SemiNLL. First, each mini-batch of data is forwarded to the network to conduct SS, which divides the original data into the labeled/unlabeled sets. Second, labeled/unlabeled samples are used to train the SSL backbone to produce accurate model output. Credit: Zhuowei Wang

The instantiations of the framework

Guided by our framework, one can easily instantiate a specific learning algorithm for NLL, by specifying a commonly used SSL backbone with an SS strategy. My framework can not only provide an important prototype in the NLL community for further exploration into SS and SSL but can also act as a conclusive work to prevent future researchers from simply reinventing the wheel.

To instantiate my framework, I propose DivideMix+ by replacing the epoch-level selection strategy with a mini-batch level one. I also propose GPL, another instantiation of the framework that leverages a two-component Gaussian mixture model to select labeled (unlabeled) data and uses Pseudo-Labeling as the SSL backbone. I have conducted extensive experiments on benchmark-simulated and real-world datasets with noisy labels. The instantiations outperform other state-of-the-art noisy-label learning methods.

Conclusion

I believe the superior performance of instantiations of the framework benefit from the conceptually simple idea of Pseudo-Labeling to reduce confirmation bias generated from the SS process. Please be noted that Pseudo-Labeling was not published at top conferences. That being said, I will keep exploring the possibility of more combinations of SS and SSL based on my framework to figure out the chemistry in between, instead of just simply piling up state-of-the-art SS and SSL methods published in top conferences.

This story is part of Science X Dialog*, where researchers can report findings from their published research articles.* Visit this page *for information about ScienceX Dialog and how to participate.*

More information: SemiNLL: A Framework of Noisy-Label Learning by Semi-Supervised Learning, Published in *Transactions on Machine Learning Research* (07/2022) openreview.net/pdf?id=qzM1Tw5i7N

Junnan Li, Richard Socher, Steven C.H. Hoi, Dividemix: learning with noisy labels as semi-supervised learning. arXiv:2002.07394v1 [cs.CV], arxiv.org/abs/2002.07394

Dr. Zhuowei Wang is with the Australian Artificial Intelligence Institute (AAII), University of Technology Sydney. His current research interests include federated learning, noisy label learning, weakly-supervised learning, few-shot learning, image classification, and the corresponding real-world applications.