Summary of Read Papers

Sixu Li

- **"Using Trusted Data to Train Deep Networks on Labels Corrupted by Severe Noise"** [1]
  Preliminaries:
  We are given an untrusted dataset $\widetilde{\mathcal{D}}$ of u examples $(x, \widetilde{y})$, and we assume that these examples are *potentially* corrupted examples from the true data distribution $p(x, y)$ with K classes. Corruption is specified by a label noise distribution $p(\widetilde{y}|y, x)$. We are also given a trusted dataset $\mathcal{D}$ of t examples drawn from $p(x, y)$, where $t/(t + u) \ll 1$. Our method makes use of $\mathcal{D}$ to estimate the $K \times K$ matrix of corruption probabilities $C_{ij} = p(\widetilde{y} = j|y = i)$.
  Key Results:
  By Bayes' theorem, we have,

  $$p(\widetilde{y} \,|\, y, x)p(x \,|\, y) = p(\widetilde{y} \,|\, y)p(x \,|\, \widetilde{y}, y) \tag{1}$$

  Intergrating over all x on both sides gives us,

  $$\int p(\widetilde{y} \,|\, y, x)p(x \,|\, y)dx = p(\widetilde{y} \,|\, y) \int p(x \,|\, \widetilde{y}, y)dx = p(\widetilde{y} \,|\, y) \tag{2}$$

  We can approximate the intergral on the left with the expectation of $p(\widetilde{y} \,|\, y, x)$ over the empirical distribution of x given y. Assuming conditional independence of $\widetilde{y}$ and y given x, we have $p(\widetilde{y} \,|\, y, x) = p(\widetilde{y} \,|\, x)$, which can be directly approximated by $\hat{p}(\widetilde{y} \,|\, x)$, the classifier trained on $\widetilde{\mathcal{D}}$. More explicitly, let A_i be the subset of x in $\mathcal{D}$ with label i. Denote our estimate of C by $\hat{C}$. We have,

  $$\hat{C}_{ij} = \frac{1}{|A_i|} \sum_{x \in A_i} \hat{p}(\widetilde{y} = j \,|\, x) = \frac{1}{|A_i|} \sum_{x \in A_i} \hat{p}(\widetilde{y} = j \,|\, y = i, x) \approx p(\widetilde{y} = j \,|\, y = i) \tag{3}$$

  The accuracy of this approximation relies on three effects:
  (i). $\hat{p}(\widetilde{y} \,|\, x)$ being a good estimate of $p(\widetilde{y} \,|\, x)$.
  (ii). The number of trusted examples of each class, which effects the second approximation of Eq. (3).

(iii). Conditional independence assumption is satisfied.

- **"Combating Label Noise in Deep Learning Using Abstention"** [2]

  Preliminaries:

  We assume we are interested in training a k–class multi-class classifier with a deep neural network(DNN) where x is the input and y is the output. For a given x, we define $p_i = p_w(y = i|x)$ as the i^{th} output of the DNN that implements the probability model $p_w(y = i|x)$ where w is the parameters of the DNN.

  Key Results:

  We train the deep abstaining classifier(DAC) with following modified version of the k–class cross-entropy per–sample loss:

$$\mathcal{L}(x_j) = (1 - p_{k+1})\big(- \sum_{i=1}^{k} q_i \log \frac{p_i}{1 - p_{k+1}}\big) + \alpha \log \frac{1}{1 - p_{k+1}} \tag{4}$$

References

[1] Hendrycks, D., Mazeika, M., Wilson, D., and Gimpel, K. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in neural information processing systems* (2018), pp. 10456–10465.

[2] Thulasidasan, S., Bhattacharya, T., Bilmes, J., Chennupati, G., and Mohd-Yusof, J. Combating label noise in deep learning using abstention. *arXiv preprint arXiv:1905.10964* (2019).