Supplementary to “Instance-dependent Label-noise Learning under a Structural Causal Model”

Yu Yao\textsuperscript{1} Tongliang Liu\textsuperscript{1}\textsuperscript{†} Mingming Gong\textsuperscript{2} Bo Han\textsuperscript{3} Gang Niu\textsuperscript{4} Kun Zhang\textsuperscript{5}
\textsuperscript{1}TML Lab, University of Sydney; \textsuperscript{2}University of Melbourne; \textsuperscript{3}Hong Kong Baptist University; \textsuperscript{4}RIKEN AIP; \textsuperscript{5}Carnegie Mellon University

Appendix

Appendix A: Derivation Details of evidence lower-bound (ELBO)

In this section, we show the derivation details of ELBO\((x, \hat{y})\).

Recall that the causal decomposition of the instance-dependent label noise is

\[
\]

Our encoders model following distributions 

\[
q_\phi(Z, Y|X) = q_{\phi_2}(Z|Y, X)q_{\phi_1}(Y|X),
\]

and decoders model the following distributions

\[
p_\theta(X, \hat{Y}|Y, Z) = p_{\theta_1}(X|Y, Z)p_{\theta_2}(\hat{Y}|Y, X).
\]

Now, we start with maximizing the log-likelihood \(p_\theta(x, \hat{y})\) of each datapoint \((x, \hat{y})\).

\[
\log p_\theta(x, \hat{y}) = \log \int_z \int_y p_\theta(x, \hat{y}, z, y) dy dz
\]

\[
= \log \int_z \int_y p_\theta(x, \hat{y}, z, y) \frac{q_\phi(z, y|x)}{q_\phi(z, y|x)} dy dz
\]

\[
= \log \mathbb{E}_{(z, y) \sim q_\phi(Z, Y|x)} \left[ \frac{p_\theta(x, \hat{y}, z, y)}{q_\phi(z, y|x)} \right]
\]

\[
\geq \mathbb{E}_{(z, y) \sim q_\phi(Z, Y|x)} \left[ \log \frac{p_\theta(x, \hat{y}, z, y)}{q_\phi(z, y|x)} \right]:= \text{ELBO}(x, \hat{y})
\]

\[
= \mathbb{E}_{(z, y) \sim q_\phi(Z, Y|x)} \left[ \log \frac{p(z)p(y)p_{\theta_1}(x|y,z)p_{\theta_2}(\hat{y}|y,x)}{q_\phi(z, y|x)} \right]
\]

\[
= \mathbb{E}_{(z, y) \sim q_\phi(Z, Y|x)} \left[ \log \frac{p(z)p(y)}{q_{\phi_2}(z|y,x)q_{\phi_1}(y|x)} \right] + \mathbb{E}_{(z, y) \sim q_\phi(Z, Y|x)} \left[ \log (p_{\theta_2}(\hat{y}|y,x)) \right]
\]

\[
\geq \mathbb{E}_{(z, y) \sim q_\phi(Z, Y|x)} \left[ \log \left( \frac{p(z)p(y)}{q_{\phi_2}(z|y,x)q_{\phi_1}(y|x)} \right) \right]
\]

The ELBO\((x, \hat{y})\) above can be further simplified. Specifically,

\[
\mathbb{E}_{(z, y) \sim q_\phi(Z, Y|x)} \left[ \log (p_{\theta_2}(\hat{y}|y,x)) \right] = \mathbb{E}_{y \sim q_{\phi_2}(Y|\hat{y})} \mathbb{E}_{z \sim q_{\phi_2}(Z|y,x)} \left[ \log (p_{\theta_2}(\hat{y}|y,x)) \right]
\]

\[
= \mathbb{E}_{y \sim q_{\phi_2}(Y|\hat{y})} \left[ \log (p_{\theta_2}(\hat{y}|y,x)) \right], \quad (2)
\]

\textsuperscript{†}Correspondence to Tongliang Liu (tongliang.liu@sydney.edu.au).

and similarly,
\[
\mathbb{E}_{(z,y) \sim q_\phi(Z,Y|x)} \left[ \log \left( \frac{p(z)p(y)}{q_{\phi_2}(z|y,x)q_{\phi_1}(y|x)} \right) \right] \\
= \mathbb{E}_{y \sim q_{\phi_1}(Y|x)} \mathbb{E}_{z \sim q_{\phi_2}(Z|y,x)} \left[ \log \left( \frac{p(z)p(y)}{q_{\phi_2}(z|y,x)q_{\phi_1}(y|x)} \right) \right] \\
= \mathbb{E}_{y \sim q_{\phi_1}(Y|x)} \mathbb{E}_{z \sim q_{\phi_2}(Z|y,x)} \left[ \log \left( \frac{p(z)}{q_{\phi_2}(z|y,x)} \right) \right]
\]
\[
= \mathbb{E}_{y \sim q_{\phi_1}(Y|x)} \left[ \log \left( \frac{p(y)}{q_{\phi_1}(y|x)} \right) \right] + \mathbb{E}_{y \sim q_{\phi_1}(Y|x)} \mathbb{E}_{z \sim q_{\phi_2}(Z|y,x)} \left[ \log \left( \frac{p(z)}{q_{\phi_2}(z|y,x)} \right) \right]
\]
\[
= \mathbb{E}_{y \sim q_{\phi_1}(Y|x)} \left[ \log \left( \frac{p(y)}{q_{\phi_1}(y|x)} \right) \right] + \mathbb{E}_{y \sim q_{\phi_1}(Y|x)} \left[ k\ell(q_{\phi_2}(Z|y,x)\|p(Z)) \right],
\]
(3)

By combining Eq. 1, Eq. 2 and Eq. 3, we get
\[
\hat{\text{ELBO}}(x, \hat{y}) = \mathbb{E}_{(z,y) \sim q_\phi(Z,Y|x)} \left[ \log p_\theta_1(x|y,z) \right] + \mathbb{E}_{y \sim q_{\phi_1}(Y|x)} \left[ \log p_\theta_2(\hat{y}|y,x) \right] \\
- k\ell(q_{\phi_1}(Y|x)\|p(Y)) + \mathbb{E}_{y \sim q_{\phi_1}(Y|x)} \left[ k\ell(q_{\phi_2}(Z|y,x)\|p(Z)) \right],
\]
which is the ELBO in our main paper.

**Appendix B: Loss Functions**

In this section, we provide the empirical solution of the ELBO and co-teaching loss. Remind that our encoder networks and decoder networks in the the first branch are defined as follows
\[
Y_1 = \hat{q}_{\phi_1}(X), \quad Z_1 \sim \hat{q}_{\phi_2}(X, Y_1), \quad X_1 = \hat{p}_{\theta_1}(Y_1, Z_1), \quad \hat{Y}_1 = \hat{p}_{\theta_2}(X_1, Y_1).
\]

Let $S$ be the noisy training set, and $d^2$ be the dimension of an instance $x$. Let $y_1$ and $z_1$ be the estimated clean label and latent representation for the instance $x$, respectively, by the first branch. As mentioned in our main paper (see Section 3.2), the negative ELBO loss is to minimize 1), a reconstruction loss between each instance $x$ and $\hat{p}_{\theta_1}(x, y_1)$; 2) a cross-entropy loss between noisy labels $\hat{p}_{\theta_2}(x, y_1)$ and $\hat{y}_1$; 3) a cross-entropy loss between $\hat{q}_{\phi_2}(x, y_1)$ and uniform distribution $P(Y)$; 4) a cross-entropy loss between $\hat{q}_{\phi_2}(x, y_1)$ and Gaussian distribution $P(Z)$. Specifically, the empirical version of the ELBO for the first branch is as follows.
\[
\sum_{(x, \hat{y}) \in S} \hat{\text{ELBO}}^1(x, \hat{y}) = \sum_{(x, \hat{y}) \in S} \left[ \beta_0 \frac{1}{d^2} \| x - \hat{p}_{\theta_1}(y_1, z_1) \|_1 - \beta_1 \hat{y} \log \hat{p}_{\theta_2}(x_1, y_1) \\
+ \beta_2 \hat{q}_{\phi_1}(x) \log \hat{q}_{\phi_1} + \beta_3 \sum_{j=1}^J (1 + \log((\sigma_j)^2) - (\mu_j)^2 - (\sigma_j)^2) \right].
\]

The hyper-parameter $\beta_0$ and $\beta_1$ are set to 0.1, and $\beta_2$ and $\beta_3$ are set to 1 to 5 because encouraging the distribution to be uniform on a small mini-batch (i.e., 128) could have a large estimation error. The hyper-parameter $\beta_3$ are set to 0.01. The empirical version of the ELBO for the second branch shares the same settings as the first branch.

For co-teaching loss, we directly follow Han et al. [1]. Intuitively, in each mini-batch data, both encoders $\hat{q}_{\phi_1}(X)$ and $\hat{q}_{\phi_2}(X)$ select their small-loss instances as the useful knowledge and exchange the knowledge to each other by a cross-entropy loss. The number of the small-loss instances used for training decays with respect to the training epoch. The experimental settings for co-teaching loss are the same as the settings in the original paper [1].

**Appendix C: More Experimental Settings**

In this section, we summarize the network structures for different datasets. The network structure for modeling $q_{\phi_1}(Y|X)$ and the dimension of the latent representation $Z$ has been discussed in our main paper. For the optimization method, we use Adam with the default learning rate $1e-3$ in Pytorch. The source code has been included in our supplementary material.
For FashionMNIST [3], SVHN [2], CIFAR10 and CIFAR100, we use the same number of hidden layers and feature maps. Specifically, 1). we model $q_{\phi_2}(Z|Y,X)$ and $p_{\theta_2}(\tilde{Y}|Y,X)$ by two 4-hidden-layer convolutional networks, and the corresponding feature maps are 32, 64, 128 and 256; 2). we model $p_{\theta_1}(X|Y,Z)$ by a 4-hidden-layer transposed-convolutional network, and the corresponding feature maps are 256, 128, 64 and 32. We ran 150 epochs for each experiment on these datasets.

For Clothing1M [4], 1). we model $q_{\phi_2}(Z|Y,X)$ and $p_{\theta_2}(\tilde{Y}|Y,X)$ by two 5-hidden-layer convolutional networks, and the corresponding feature maps are 32, 64, 128, 256, 512; 2). we model $p_{\theta_1}(X|Y,Z)$ by a 5-hidden-layer transposed-convolutional network, and the corresponding feature maps are 512, 256, 128, 64 and 32. We ran 40 epochs on Clothing1M.

References


