Supplementary Materials to
”DAT:Training Deep Networks Robust to Label-Noise by Matching the Feature Distributions”
Anonymous CVPR 2021 submission

APPENDIX A

CODE

The algorithmic description of DAT without clean set is shown in Algorithm 1. To illustrate how DAT works, we also provide the code on the MNIST and CIFAR-10 datasets. The provided code is in the DAT-master folder, and the github url will be released after the review procedure.

Algorithm 1 DAT-Algorithm without clean set

\textbf{Input}: noisy training set $D_\rho$, $\alpha$ and $\beta$, learning rate $\eta$, epoch $T$, iteration $N$.

\begin{algorithmic}[1]
\FOR{$t = 1, 2, 3, \ldots, T$}
\STATE Shuffle training set $D_\rho$
\STATE Sample a subset $D_s$ from $D_\rho$
\FOR{$n = 1, 2, 3, \ldots, N$}
\STATE Fetch mini-batch $\bar{\rho}$ from $D_\rho$
\STATE Fetch mini-batch $\bar{S}$ from $D_s$
\STATE Calculate $L_{\tilde{cce}}$ on $\bar{\rho}$, $L_{dis}$ on $\bar{S}$
\STATE Update $\theta_{h, h, g} = \theta_{h, h, g} - \nabla \theta_{h, h, g} L_{\tilde{cce}}$
\STATE Update $\theta_h = \theta_h + \alpha \nabla \theta_h L_{dis}$
\STATE Update $\theta_g = \theta_g - \beta \nabla \theta_g L_{dis}$
\ENDFOR
\ENDFOR

\textbf{Output}: $\theta_{h, h, g}$
APPENDIX B
THEORETICAL DERIVATION

In this section, we show the proof of Theorem 1 and the reason that \( h \Delta \mathcal{H} \)-divergence has a tighter upper bound. For ease of reference, we restate the definition of \( h \Delta \mathcal{H} \)-divergence and Theorem 1.

**Definition 1:** Given two feature distribution \( D^Z_\rho \) and \( D^Z_\tau \) extracted by a fixed \( g \), and a hypothesis class \( \mathcal{H} \) which is a set of binary classifiers. Through a given classifier \( h \), \( h \Delta \mathcal{H} \)-divergence between \( D^Z_\rho \) and \( D^Z_\tau \) is:

\[
d_{h \Delta \mathcal{H}}(D^Z_\rho, D^Z_\tau) = 2 \sup_{h \in \mathcal{H}} \left\{ \Pr_{z \sim D^Z_\rho} \left[ h(z) \neq \hat{h}(z) \right] - \Pr_{z \sim D^Z_\tau} \left[ h(z) \neq \hat{h}(z) \right] \right\}.
\]

The following Theorem 1 can be stated through the \( h \Delta \mathcal{H} \)-divergence.

**Theorem 1:** Let \( g \) be a fixed representation function from \( \mathcal{X} \) to \( \mathcal{Z} \), \( \mathcal{H} \) be the hypothesis class of Vapnik-Chervonenkis dimension \( d \). If random noisy samples of size \( m \) is generated by applying \( g \) from \( D^Z_{\rho-i.i.d.} \), then with probability at least \( 1 - \delta \), the generalized bound of the clean risk \( \epsilon_c(h) \):

\[
\epsilon_c(h) \leq \epsilon^m_c(h) + \frac{1}{2} d_{h \Delta \mathcal{H}}(D^Z_\rho, D^Z_\tau) + \lambda,
\]

where

\[
\lambda = \epsilon_c(h^\ast) + \epsilon_\rho(h^\ast) + \sqrt{\frac{4}{m}(d \log \frac{2em}{d} + \log \frac{4}{\delta})},
\]

\[
h^\ast = \arg\min_{h \in \mathcal{H}} \epsilon_c(h),
\]

\[
\epsilon^m_\rho(h) = \frac{1}{m} \sum_{i=1}^m |f_\rho(z) - h(z)|.
\]

**Proof 1:** For a classifier \( h \), let \( Z_h \subseteq \mathcal{Z} \) be the characteristic subset for whose characteristic function is \( h \). The parallel notation \( Z_h^\ast \) and \( \hat{Z}_h \) are used for classifier \( h^\ast \) and \( \hat{h} \). Through the characteristic subset, we make \( \Pr_c[Z_h \Delta Z_h^\ast] = \Pr_{z \sim D^Z}[h(z) \neq h^\ast(z)] \), and the parallel notation \( \Pr_\rho \) is used.

\[
\epsilon_c(h) \leq \epsilon_c(h^\ast) + \Pr_c[Z_h \Delta Z_h^\ast]
\]

\[
\leq \epsilon_c(h^\ast) + \Pr_\rho[Z_h \Delta Z_h^\ast] + \{ \Pr_c[Z_h \Delta Z_h^\ast] - \Pr_\rho[Z_h \Delta Z_h^\ast] \}
\]

\[
\leq \epsilon_c(h^\ast) + \epsilon_\rho(h^\ast) + \epsilon_\rho(h) + \{ \Pr_c[Z_h \Delta Z_h^\ast] - \Pr_\rho[Z_h \Delta Z_h^\ast] \}
\]

\[
\leq \epsilon_c(h^\ast) + \epsilon_\rho(h^\ast) + \epsilon_\rho(h) + \sup_{h \in \mathcal{H}} \{ \Pr_c[Z_h \Delta Z_h^\ast] - \Pr_\rho[Z_h \Delta Z_h^\ast] \}
\]

\[
\leq \epsilon_c(h^\ast) + \epsilon_\rho(h^\ast) + \epsilon_\rho(h) + \frac{1}{2} d_{h \Delta \mathcal{H}}(D^Z_\rho, D^Z_\tau)
\]

InEq. (6) and InEq. (8) relies on the triangle inequality for classification error \[1\]. According to the standard Vapnik-Chervonenkis theory \[2\], we can then bound the true \( \epsilon_\rho(h) \) by its empirical estimate \( \epsilon^m_\rho(h) \):

\[
\epsilon_\rho(h) \leq \sqrt{\frac{4}{m}(d \log \frac{2em}{d} + \log \frac{4}{\delta})} + \epsilon^m_\rho(h)
\]

in summary:

\[
\epsilon_c(h) \leq \epsilon^m_\rho(h) + \lambda + \frac{1}{2} d_{h \Delta \mathcal{H}}(D^Z_\rho, D^Z_\tau)
\]

Before explaining why \( h \Delta \mathcal{H} \)-divergence has a tighter upper bound, we give a definition of \( \mathcal{H} \Delta \mathcal{H} \)-divergence \[3\] (the same analysis type is suitable for \( \mathcal{H} \)-divergence):

**Definition 2:** Given two feature distribution \( D^Z_\rho \) and \( D^Z_\tau \) extracted by a fixed \( g \), and a hypothesis class \( \mathcal{H} \) which is a set of binary classifiers. Through a given classifier \( h, \hat{h} \), \( h \Delta \mathcal{H} \)-divergence between \( D^Z_\rho \) and \( D^Z_\tau \) is:

\[
d_{h \Delta \mathcal{H}}(D^Z_\rho, D^Z_\tau) = 2 \sup_{h, \hat{h} \in \mathcal{H}} \left| \Pr_{z \sim D^Z_\rho} \left[ h(z) \neq \hat{h}(z) \right] - \Pr_{z \sim D^Z_\tau} \left[ h(z) \neq \hat{h}(z) \right] \right|.
\]

Assuming that the \( h \Delta \mathcal{H} \)-divergence is replaced by the \( \mathcal{H} \Delta \mathcal{H} \)-divergence in Theorem 1, the proof becomes of the following form.

**Proof 2:**
\[ \epsilon_c(h) \leq \epsilon_c(h^*) + \Pr_c[Z_h \Delta Z_{h^*}] \]
\[ \leq \epsilon_c(h^*) + \Pr_\rho[Z_h \Delta Z_{h^*}] + |\Pr_c[Z_h \Delta Z_{h^*}] - \Pr_\rho[Z_h \Delta Z_{h^*}]| \]
\[ \leq \epsilon_c(h^*) + \epsilon_\rho(h^*) + \epsilon_\rho(h) + |\Pr_c[Z_h \Delta Z_{h^*}] - \Pr_\rho[Z_h \Delta Z_{h^*}]| \]
\[ \leq \epsilon_c(h^*) + \epsilon_\rho(h^*) + \epsilon_\rho(h) + \sup_{h, h' \in H} |\Pr_c[Z_h \Delta Z_{h'}] - \Pr_\rho[Z_h \Delta Z_{h'}]| \]
\[ \leq \epsilon_c(h^*) + \epsilon_\rho(h^*) + \epsilon_\rho(h) + \frac{1}{2} d_{\Delta H} (D_Z^c, D_Z^\rho) \]

Compared to InEq. (7), InEq. (15) add an additional absolute value, which is an absolute value inequality that allows the upper bound of the clean error rate \( \epsilon_c(h) \) to be amplified. In addition, InEq. (17) searches both \( h \) and \( h' \) in \( H \) to maximize the probability difference, which also amplifies the upper bound of \( \epsilon_c(h) \) even more compared to InEq. (9). As a result, \( h \Delta H \)-divergence has a tighter generalized upper bound.

REFERENCES