Weakly Supervised Contrastive Adversarial Training for Learning Robust Features from Semi-supervised Data

Supplementary Material

A. Proofs of Theorems

A.1. Useful Lemmas

Lemma 1. *Let* $z = f(x)$*,* $z_1 = f(x_1)$ *and* $\bar{s}(z) = -\frac{1}{|\mathcal{N}_x^+|}\sum_{x_p \in \mathcal{N}_x^+} \log \frac{\exp s(z,z_p)}{\sum_{x_n \in \mathcal{D}} \exp s(z,z_n)}$*. If* $\Delta(z, z_1)=|\bar{s}(z) - \bar{s}(z_1)|$ *is valid when* $x_1 \in \mathcal{B}_\epsilon(x)$*,* $\Delta(\cdot, \cdot)$ *is a distance metric.*

Proof of Lemma 1. It is obvious that for any x, $\Delta(z, z) = 0$. And for any x and $x_1 \in \mathcal{B}_\epsilon(x)$ the symmetry and non-negativity are clearly satisfied by $\Delta(z, z_1)$. We only need to justify that $\Delta(\cdot, \cdot)$ satisfies the triangle inequality. For any x, x_1 and x_2, since we have

$$\begin{aligned}
&\Delta(z, z_1) + \Delta(z, z_2) \\
=&\Delta(z_1, z_2) + \Delta(z, z_2) - \Delta(z_1, z_2) + \Delta(z, z_1) \\
=&|\bar{s}(z_1) - \bar{s}(z_2)| + |\bar{s}(z) - \bar{s}(z_2)| - \Delta(z_1, z_2) + \Delta(z, z_1) \\
\geq&|\bar{s}(z_1) - 2\bar{s}(z_2) + \bar{s}(z)| - \Delta(z_1, z_2) + \Delta(z, z_1) \\
=&|2\bar{s}(z_1) - 2\bar{s}(z_2) + \bar{s}(z) - \bar{s}(z_1)| - \Delta(z_1, z_2) + \Delta(z, z_1) \\
\geq&|2\bar{s}(z_1) - 2\bar{s}(z_2)| - |\bar{s}(z) - \bar{s}(z_1)| - \Delta(z_1, z_2) + \Delta(z, z_1) \\
=&2\Delta(z_1, z_2) - \Delta(z, z_1) - \Delta(z_1, z_2) + \Delta(z, z_1) \\
=&\Delta(z_1, z_2),
\end{aligned}$$

i.e, the triangle inequality $\Delta(z, z_1) + \Delta(z, z_2) \geq \Delta(z_1, z_2)$ always holds. Therefore, $\Delta(\cdot, \cdot)$ is a distance metric. □

Lemma 2. $|l_{\text{con}}(z', z) - l_{\text{con}}(z, z)| = \Delta(z', z)$

Proof of Lemma 2. According to Eq. (7),

$$\begin{aligned}
&l_{\text{con}}(z', z) - l_{\text{con}}(z, z) \\
=&-\frac{1}{|\mathcal{N}_x^+|}\sum_{x_p \in \mathcal{N}_x^+} \log \frac{\exp(s(z', z_p))}{\sum_{x_n \in \mathcal{D}^*} \exp(s(z', z_n))} \\
&+\frac{1}{|\mathcal{N}_x^+|}\sum_{x_p \in \mathcal{N}_x^+} \log \frac{\exp(s(z, z_p))}{\sum_{x_n \in \mathcal{D}^*} \exp(s(z, z_n))} \\
=&\bar{s}(z') - \bar{s}(z).
\end{aligned}$$

Therefore, $|l_{\text{con}}(z, z) - l_{\text{con}}(z', z)| = \Delta(z', z)$. □

A.2. Detailed Proofs

Proof of Theorem 1. By the LogSumExp operation, i.e., $\log(e^{x_1} + e^{x_2} + ... + e^{x_n}) \approx \max\{x_1, x_2, ..., x_n\}$, we can transform the contrastive loss in Eq. (7) to

$$\begin{aligned}
&l_{\text{con}}(z', z) \\
=&\frac{1}{|\mathcal{N}_x^+|}\sum_{x_p \in \mathcal{N}_x^+} \log \frac{\sum_{x_n \in \mathcal{D}^*} e^{s(z', z_n)}}{e^{s(z', z_p)}} \\
=&\frac{1}{|\mathcal{N}_x^+|}\sum_{x_p \in \mathcal{N}_x^+} \log \Big(e^0 + \sum_{x_n \in \mathcal{D}, x_n \neq x_p} e^{s(z', z_n) - s(z', z_p)}\Big) \\
\approx&\frac{1}{|\mathcal{N}_x^+|}\sum_{x_p \in \mathcal{N}_x^+} \max\big\{0, \{s(z', z_n) - s(z', z_p)\}_{x_n \in \mathcal{D}, x_n \neq x_p}\big\},
\end{aligned}$$

from which we can see that maximizing $l_{\text{con}}(z', z)$ is approximately maximizing the last line. □

Proof of Theorem 2. 1) According to Hoeffding's Inequality [11], with probability at least $1-\delta$ the following inequality holds:

$$\begin{aligned}
&\mathrm{E}_{P_{X,Y}}[l_{\text{nat}}(X, Y)] - \frac{1}{|\mathcal{D}^*|}\sum_{(x,y) \in \mathcal{D}^*} l_{\text{nat}}(x, y) \\
\leq&\mathrm{E}_{P_{X,Y}}[l_{\text{nat}}(X, Y)] - \frac{1}{|\mathcal{D}_l|}\sum_{(x,y) \in \mathcal{D}_l} l_{\text{nat}}(x, y) \\
\leq&l_{\text{m}}\sqrt{\frac{\log\frac{1}{\delta}}{2|\mathcal{D}_l|}},
\end{aligned}$$

where the second line holds because existing works [7, 34] have theoretically proven that pseudo-labeled data generated by self-training can decrease the generalization gap.

Then according to the Definition 1,

$$\begin{aligned}
\rho_{l_{\text{nat}}} &= \inf_g \mathrm{E}_{P_{X,Y}}[l_{\text{nat}}(X, Y)] \\
&\leq A_1 + l_{\text{m}}\sqrt{\frac{\log\frac{1}{\delta}}{2|\mathcal{D}_l|}}
\end{aligned}$$

with probability at least $1 - \delta$.

2) Again according to Hoeffding's Inequality, with probability at least $1 - \delta$ the following inequality holds:

$$\mathrm{E}_{P_X}[\Delta(f(X'), f(X))] - \frac{1}{n}\sum_{x \in \mathcal{D}} \Delta(x', x) \leq \Delta_{\text{m}}\sqrt{\frac{\log\frac{1}{\delta}}{2|\mathcal{D}|}},$$

where Δ_{m} is the supremum of the distance $\Delta(\cdot, \cdot)$ over $\{(z', z)|x \sim P_X \wedge x' \in \mathcal{B}_\epsilon(x) \wedge z = f(x) \wedge z' = f(x')\}$.

And thus we have the following inequalities:

$$
\begin{aligned}
&\mathrm{E}_{P_X}[\sup_{X'\in\mathcal{B}_\epsilon(X)} \Delta(f(X'),f(X))] - \Delta_{\mathrm{m}}\sqrt{\frac{\log\frac{1}{\delta}}{2|D|}} \\
\le&\frac{1}{|D|}\sum_{x\in\mathcal{D}}\sup_{x'\in\mathcal{B}_\epsilon(x)} \Delta(z',z) \\
\le&\frac{1}{|D|}\sum_{x\in\mathcal{D}}\sup_{x'\in\mathcal{B}_\epsilon(x)} \left\{l_{\mathrm{con}}(z',z)-l_{\mathrm{con}}(z,z)\right\} \\
\le&\frac{1}{\beta|D|}\sum_{x\in\mathcal{D}}\sup_{x'\in\mathcal{B}_\epsilon(x)} \big\{\mathrm{KL}(C(x)\|C(x')) \\
&+\beta(l_{\mathrm{con}}(z',z)-l_{\mathrm{con}}(z,z))\big\} \\
=&\frac{1}{\beta|D|}\sum_{x\in\mathcal{D}}\sup_{x'\in\mathcal{B}_\epsilon(x)} l_{\mathrm{adv}}(x',x) \\
=&\frac{2}{\beta}A_2,
\end{aligned}
$$

where the first inequality holds with probability at least $1-\delta$. Then according to Definition 1, we can get that

$$
\begin{aligned}
\gamma_\Delta &= \mathrm{E}_{P_X}[\sup_{X'\in\mathcal{B}_\epsilon(X)} \Delta(f(X'),f(X))] \\
&\le \frac{2}{\beta}A_2 + \Delta_{\mathrm{m}}\sqrt{\frac{\log\frac{1}{\delta}}{2|D|}}.
\end{aligned}
$$

Therefore, feature f captured by the target model $C = g\circ f$ trained by WSCAT is $\rho_{l_{\mathrm{nat}}}$-γ_Δ-robust, where $\rho_{l_{\mathrm{nat}}} \le l_{\mathrm{m}}\sqrt{\frac{\log\frac{1}{\delta}}{2|\mathcal{D}_l|}}$ with probability at least $1-\delta$, and $\gamma_\Delta \le \frac{2}{\beta}A_2 + \Delta_{\mathrm{m}}\sqrt{\frac{\log\frac{1}{\delta}}{2|D|}}$ with probability at least $1-\delta$. □

B. Additional Experimental Results

B.1. Performance Comparison (RQ1)

The performance of WSCAT-sup and TRADES under fully-supervised setting across various model architectures is shown in Tab. 1.

B.2. Ablation Study (RQ3)

The performance of WSCAT and WSCAT's different variants is shown in Tabs. 2 to 4.

B.3. Training Time (RQ5)

To show WSCAT does not excessively increases the training time than existing semi-supervised AT methods, we compare WSCAT's epoch time with that of RST, which is an efficient semi-supervised AT method [43]. The result is shown in Tab. 5, from which one can observe that WSCAT does not bring additional training time cost overall. The result is reasonable since during a batch of the training, the loss defined in Eq. (7) can be calculated just based on points in that batch instead of the entire dataset.

Table 1. Performance of models trained by Standard, TRADES and WSCAT-sup (a variant of our WSCAT that uses only labeled data) under fully-supervised setting.

Dataset (Model)	Method	Nat.	FGSM	PGD	CW	AA	Mean	NRF
CIFAR10 (ResNet50)	Standard	95.15	42.37	0.02	0.01	0.00	0.00	0.00
	TRADES	80.34	56.05	51.74	49.27	47.99	55.10	43.92
	WSCAT-sup	82.37	59.84	57.84	52.50	51.41	59.07	45.86
CIFAR10 (ResNet152)	Standard	95.26	49.42	0.01	0.00	0.00	0.00	0.00
	TRADES	81.52	56.56	51.55	49.96	48.15	55.48	57.01
	WSCAT-sup	80.98	59.92	58.46	52.89	52.09	59.35	58.56
CIFAR10 (WRN28-10)	Standard	96.23	43.34	0.01	0.02	0.00	0.00	0.00
	TRADES	84.65	60.92	56.34	54.12	52.85	59.97	50.44
	WSCAT-sup	84.18	61.42	59.72	54.02	52.91	60.74	55.12
CIFAR100 (WRN28-10)	Standard	78.62	16.27	0.34	0.09	0.00	0.00	-
	TRADES	58.69	33.74	30.77	28.31	27.02	33.00	-
	WSCAT-sup	59.71	34.61	32.63	28.80	27.46	33.92	-

Table 2. Performance of different variants on **CIFAR10**.

Methods	WSCAT	WSCAT-fixed	WSCAT-self	WSCAT-std
Natural	$80.93_{\pm 0.14}$	$79.04_{\pm 0.45}$	$80.72_{\pm 0.12}$	$76.65_{\pm 0.30}$
FGSM	$59.62_{\pm 0.16}$	$57.56_{\pm 0.22}$	$58.71_{\pm 0.25}$	$55.33_{\pm 0.37}$
PGD	$58.52_{\pm 0.22}$	$54.55_{\pm 0.17}$	$54.58_{\pm 0.43}$	$53.75_{\pm 0.18}$
CW	$53.15_{\pm 0.08}$	$51.66_{\pm 0.03}$	$52.20_{\pm 0.34}$	$48.68_{\pm 0.10}$
AA	$52.23_{\pm 0.06}$	$50.77_{\pm 0.06}$	$51.20_{\pm 0.34}$	$48.00_{\pm 0.02}$
Mean	$59.40_{\pm 0.05}$	$57.20_{\pm 0.06}$	$57.80_{\pm 0.27}$	$54.88_{\pm 0.03}$

Table 3. Performance of different variants on **CIFAR100**.

Methods	WSCAT	WSCAT-fixed	WSCAT-self	WSCAT-std
Natural	$55.14_{\pm 0.52}$	$55.09_{\pm 0.08}$	$54.70_{\pm 1.48}$	$51.66_{\pm 0.18}$
FGSM	$28.41_{\pm 0.09}$	$27.43_{\pm 0.35}$	$27.55_{\pm 0.49}$	$25.22_{\pm 0.46}$
PGD	$25.26_{\pm 0.32}$	$23.84_{\pm 0.36}$	$24.08_{\pm 0.11}$	$21.89_{\pm 0.29}$
CW	$22.99_{\pm 0.41}$	$22.65_{\pm 0.03}$	$22.04_{\pm 0.58}$	$19.39_{\pm 0.40}$
AA	$21.82_{\pm 0.40}$	$21.83_{\pm 0.01}$	$20.77_{\pm 0.72}$	$18.70_{\pm 0.15}$
Mean	$27.43_{\pm 0.29}$	$26.97_{\pm 0.02}$	$26.36_{\pm 0.44}$	$23.83_{\pm 0.05}$

Table 4. Performance of different variants on **ImageNet32-100**.

Methods	WSCAT	WSCAT-fixed	WSCAT-self	WSCAT-std
Natural	$34.64_{\pm 2.76}$	$33.28_{\pm 0.00}$	$32.32_{\pm 0.00}$	$31.43_{\pm 0.11}$
FGSM	$12.63_{\pm 0.13}$	$12.54_{\pm 0.00}$	$12.62_{\pm 0.00}$	$8.61_{\pm 0.19}$
PGD	$9.89_{\pm 0.35}$	$9.94_{\pm 0.00}$	$9.80_{\pm 0.00}$	$6.90_{\pm 0.20}$
CW	$8.01_{\pm 0.37}$	$8.02_{\pm 0.00}$	$8.06_{\pm 0.00}$	$5.11_{\pm 0.23}$
AA	$7.27_{\pm 0.33}$	$7.14_{\pm 0.00}$	$7.06_{\pm 0.00}$	$4.61_{\pm 0.23}$
Mean	$10.59_{\pm 0.32}$	$10.52_{\pm 0.00}$	$10.46_{\pm 0.00}$	$7.09_{\pm 0.27}$

Table 5. Epoch time of WSCAT and RST.

Datasets	CIFAR10	CIFAR100	ImageNet32-100
WSCAT	5′15″	5′18″	13′02″
RST	5′14″	5′18″	13′20″