Appendix

Yulin Jin[1] , Xiaoyu Zhang[1*], Jian Lou[2] , Xu Ma[3] , Zilong Wang[1] Xiaofeng Chen[1] ,
[1]State Key Laboratory of Integrated Service Networks (ISN), Xidian University
[2]ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University
[3]School of Cyber Science and Engineering, Qufu Normal University
jyl1990903@163.com xiaoyuzhang@xidian.edu.cn jian.lou@zju.edu.cn
xma@qfnu.edu.cn zlwang@xidian.edu.cn xfchen@xidian.edu.cn

A.1 Proof of Theorem 1.

Proposition 1 *If T is a K-dimension random variable with finite mean vector μ and covariance matrix Σ, then the maximum entropy distribution of T is $\mathcal{N}(\mu, \Sigma)$.*

Proposition 2 *If T is a K-dimension Gaussian random variable with mean vector μ and finite covariance matrix Σ, then $Klog2\pi + \frac{1}{2}\sum_{i=1}^{K}\Sigma_{ii}$ is an upper bound of $H(T)$.*

Proof 1 *Due to that Σ is a real symmetric matrix, we diagonalize Σ as $U^{-1}\Lambda U$, where Λ is the eigen matrix of Σ and any $\Lambda_{ii} > 0, 1 \leq i \leq K$, thus*

$$log|\Sigma| = log|U^{-1}\Lambda U| = log\prod_{i=1}^{K}\Lambda_{ii}$$

$$= \sum_{i=1}^{K} log\Lambda_{ii} \leq \sum_{i=1}^{K}(\Lambda_{ii} - 1) = \sum_{i=1}^{K}(\Sigma_{ii} - 1).$$

To sum up, T is a K-dimension random variable with finite mean vector μ and covariance matrix Σ, then a upper bound of $H(T)$ could be $Klog2\pi + \frac{1}{2}\sum_{i=1}^{K}\Sigma_{ii}$.

Definition 1 *Given the input variable X and its corresponding ground-true label variable Y, the loss function L_{IB} of the Information Bottleneck principle on the neural network f is defined as:*

$$L_{IB} = I(X;T) - \gamma I(T;Y), \gamma > 1, \tag{1}$$

where T is the output of the middle-layer representation of X, which means T can be formalized as $T(X, \theta)$,and the function $I(;)$ represents the mutual information between the two input variables. We reformulate the Information Bottleneck loss function with the relationship between mutual information and entropy, where $L_{IB} = (1-\gamma)H(T) + \gamma H(T|Y)$.

Theorem 1 *Given a series of continuous K-dimension probability density distributions $\{p_i(t)|1 \leq i \leq N\}$ with their corresponding finite covariance matrices $\{\Sigma^i|1 \leq i \leq N\}$ and mean vectors $\{\mu^i|1 \leq i \leq N\}$, with $\lambda > 0$, the following two minimization problems (2) and (3) have the same optimal solution:*

$$\min_{p_i(t),1\leq i\leq N} -\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\|\mu^i - \mu^j\|_2 + \lambda\sum_{i=1}^{N}\sum_{k=1}^{K}\Sigma_k^i, \tag{2}$$

$$\min_{p_i(t),1\leq i\leq N} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\int p_i(t)p_j(t)dt + \lambda\sum_{i=1}^{N}\sum_{k=1}^{K}\Sigma_k^i, \tag{3}$$

*Corresponding Author

Proof 2 *For one of the optimal solutions of the minimization problem 2., $p_i(t)$ will converge to a Dirac distribution, $1 \leq i \leq N$, and there exists the $\epsilon \in R^+$, where $1 \leq i, j \leq N, \epsilon < \|\mu^i - \mu^j\|_2$. Thus, for any $1 \leq i, j \leq N$, as the ϵ grows larger during the minimization of the problem 2, $\int p_i(t)p_j(t)dt$ converges to zero progressively, that is to say,*

$$\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\int p_i(t)p_j(t)dt + \lambda\sum_{i=1}^{N}\sum_{k=1}^{K}\Sigma_k^i \to 0.$$

We then proof that solve the problem 3 is equal to minimize an upper bound of $\mathcal{L}_{IB}$. 1) Firstly, apply proposition 1 to every $p_{T|Y=i}, 1 \leq i \leq |Y|$, we have

$$\begin{aligned}
H(T|Y) &= \sum_{i=1}^{|Y|} p(Y=i)H(T|i) \\
&\leq \sum_{i=1}^{|Y|} p(Y=i)\left(d_l log 2\pi + \frac{1}{2}\sum_{k=1}^{d_l}\Sigma_{kk}^i\right) \\
&= \sum_{i=1}^{|Y|}\frac{1}{|Y|}\left(d_l log 2\pi + \frac{1}{2}\sum_{k=1}^{d_l}\Sigma_{kk}^i\right) \\
&= \sum_{i=1}^{|Y|}\frac{d_l log 2\pi}{|Y|} + \frac{1}{|Y|}\sum_{i=1}^{|Y|}\sum_{k=1}^{d_l}\Sigma_{kk}^i.
\end{aligned}$$

Therefore,

$$\min_\theta \sum_{i=1}^{|Y|}\sum_{k=1}^{d_l}\sqrt{\Sigma}_{kk}^i \Leftrightarrow \min_\theta \sum_{i=1}^{|Y|}\sum_{k=1}^{d_l}\Sigma_{kk}^i \Rightarrow \min_\theta UpperBound(H(T|Y)).$$

2)

$$\begin{aligned}
H(T) &= -\int p_T(t) log p_T(t)dt \\
&= -\int \sum_{i=1}^{|Y|} p_{T|Y=i}(t)p(Y=i) log \sum_{i=j}^{|Y|} p_{T|Y=j}(t)dt \\
&= -\sum_{i=1}^{|Y|}\int p_{T|Y=i}(t)p(Y=i) log \sum_{i=j}^{|Y|} p_{T|Y=j}(t)dt \\
&\geq -\sum_{i=1}^{|Y|}\int p_{T|Y=i}(t)p(Y=i)\left(\sum_{i=j}^{|Y|} p_{T|Y=j}(t) - 1\right)dt \\
&= -\sum_{i=1}^{|Y|}\sum_{j=1}^{|Y|}\int p_{T|Y=i}(t)p_{T|Y=j}(t)dt + 1 \\
&= -2\sum_{i=1}^{|Y|-1}\sum_{j=i+1}^{|Y|}\int p_{T|Y=i}(t)p_{T|Y=j}(t)dt - \sum_{i=1}^{|Y|}\int \left(p_{T|Y=i}(t)\right)^2 dt + 1 \\
&\geq -2\sum_{i=1}^{|Y|-1}\sum_{j=i+1}^{|Y|}\int p_{T|Y=i}(t)p_{T|Y=j}(t)dt - |Y| + 1.
\end{aligned}$$

Hence,

$$L_{IB} \leq \gamma\left(\sum_{i=1}^{|Y|}\frac{d_l log 2\pi}{|Y|} + \frac{1}{|Y|}\sum_{i=1}^{|Y|}\sum_{k=1}^{d_l}\Sigma_{kk}^i\right)$$

$$+(1-\gamma)\left(-2\sum_{i=1}^{|Y|-1}\sum_{j=i+1}^{|Y|}\int p_{T|Y=i}(t)p_{T|Y=j}(t)dt-|Y|+1\right),\gamma>1,$$

where we obtain a function greater than or equal to L_{IB} everywhere. In spite of the constant term, the minimization of this function is equal to minimize

$$\gamma\sum_{i=1}^{|Y|}\sum_{k=1}^{d_l}\Sigma_{kk}^{i}-2(1-\gamma)\sum_{i=1}^{|Y|-1}\sum_{j=i+1}^{|Y|}\int p_{T|Y=i}(t)p_{T|Y=j}(t)dt.$$

From the beginning result of the relationship between Problem 2 and 3, we know that above minimization optimal problem can solved by

$$\min_{\theta}\lambda\sum_{i=1}^{N}\sum_{k=1}^{K}\Sigma_{kk}^{i}-\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\|\mu_l^i-\mu_l^j\|_2,\lambda>0.$$

The Problem 2 is suitable to characterize the Clustering Effect of an intermediate layer of the model, which minimizes the distance from extracted features to the centroid while maximizes the distance between centroids. Therefore, the formation of the Clustering Effect is equivalent to the minimization of an upper bound of $\mathcal{L}_{IB}$.

A.2 Proof of Theorem 2.

Theorem 2 *Given a $L+1$ layers ReLu neural network f, and a input x with the p-norm ε of the constrained perturbation r, $\|r\|_p\leq\varepsilon$. Assume that the network classifies x as label y, $1\leq y\leq|Y|$, then if the inequality below holds,*

$$\Delta\leq\min_{r}\left\{\left\|\underset{r}{\operatorname{argmin}}\ \mathrm{ReLu}\left(\min_{j\neq y}\frac{f_{L+1}^{(y)}(x)-f_{L+1}^{(j)}(x)}{\left\|W_{L+1}^{y}-W_{L+1}^{j}\right\|_u}-\|f_L(x+r)-f_L(x)\|_v\right)\right\|_p,\varepsilon\right\},1=\frac{1}{u}+\frac{1}{v},$$

the classification of the set $\left\{x+r|\|r\|_p\leq\Delta\right\}$ will be the same.

Proof 3 *Assume that for a $d\in U_\Delta$, $\arg\max_i f_{L+1}^{(i)}(x+d)=j\neq y$, we note that $f_L(x+d)-f_L(x)=\tau\in R^{d_L}$, consider the l-th layer's output space as the input space of $f_{L+1}(\cdot)=W_{L+1}\sigma(\cdot)+b_{L+1}$, then we have the following inequalities,*

$$\begin{cases}f_{L+1}^{(y)}\left(f_L(x)\right)-f_{L+1}^{(j)}\left(f_L(x)\right)>0\\ f_{L+1}^{(j)}\left(f_L(x+d)\right)-f_{L+1}^{(y)}\left(f_L(x)\right)>0\end{cases}$$

Then we have a series of inequalities,

$$\begin{aligned}&f_{L+1}^{(y)}\left(f_L(x)\right)-f_{L+1}^{(j)}\left(f_L(x+d)\right)\\ <\ &f_{L+1}^{(j)}\left(f_L(x+d)\right)-f_{L+1}^{(y)}\left(f_L(x)\right)+f_{L+1}^{(y)}\left(f_L(x)\right)-f_{L+1}^{(j)}\left(f_L(x)\right)\\ <\ &f_{L+1}^{(j)}\left(f_L(x+d)\right)-f_{L+1}^{(j)}\left(f_L(x)\right)\\ &-\left(f_{L+1}^{(y)}\left(f_L(x+d)\right)-f_{L+1}^{(y)}\left(f_L(x)\right)\right)\\ =\ &\int_0^1\langle\nabla f^{(j)}(f_L(x)+t\tau),\tau dt\rangle-\int_0^1\langle\nabla f^{(y)}(f_L(x)+t\tau),\tau dt\rangle\\ =\ &\int_0^1\langle\nabla f^{(j)}(f_L(x)+t\tau)-\nabla f^{(y)}(f_L(x)+t\tau),\tau\rangle dt.\end{aligned}$$

According to Hölder inequality,

$$\int_0^1\langle\nabla f^{(j)}(f_L(x)+t\tau)-\nabla f^{(y)}(f_L(x)+t\tau),\tau\rangle dt$$

$$\leq \int_0^1 \left\| \nabla f^{(j)}(f_L(x)+t\tau) - \nabla f^{(y)}(f_L(x)+t\tau) \right\|_\tau dt \int_0^1 \|\tau\|_v dt$$
$$\leq \max_t \left\| \nabla f^{(j)}(f_L(x)+t\tau) - \nabla f^{(y)}(f_L(x)+t\tau) \right\|_u \|\tau\|_v$$

Further, from the property of ReLu function σ that $\sigma^{'}(x) \leq 1$, we have the following inequalities:

$$\left\| \nabla f^{(j)}(f_L(x)+t\tau) - \nabla f^{(y)}(f_L(x)+t\tau) \right\|_u$$
$$= \left\| \left(W_{L+1}^{(j)} - W_{L+1}^{(y)} \right) \times \sigma(f_L(x)+t\tau) \right\|_u$$
$$\leq \sqrt[q]{\sum_i^{d_L} \left\| \left(W_{L+1}^{(j,i)} - W_{L+1}^{(y,i)} \right) \sigma(f_L(x)+t\tau) \right\|_u^u}$$
$$\leq \sqrt[q]{\sum_i^{d_L} \left\| \left(W_{L+1}^{(j,i)} - W_{L+1}^{(y,i)} \right) \right\|_u^u}$$
$$= \left\| \left(W_{L+1}^{(j)} - W_{L+1}^{(y)} \right) \right\|_u$$

Therefore, we confirm that if the perturbation r mislead the network to classify $x+r$ differently to x, then we have the inequality

$$\|f_L(x+r) - f_L(x)\|_p > \frac{f^{(y)}\left(f_L(x)\right) - f^{(j)}\left(f_L(x)\right)}{\left\| W_{L+1}^{(j)} - W_{L+1}^{(y)} \right\|_q}$$

That means, if the

$$\max_{r \in U_\Delta} \|f_L(x+r) - f_L(x)\|_p \leq \min_j \frac{f^{(y)}\left(f_L(x)\right) - f^{(j)}\left(f_L(x)\right)}{\left\| W_{L+1}^{(j)} - W_{L+1}^{(y)} \right\|_q},$$

we can confirm that there is no perturbation in U_ε could classify $x+r$ differently to x. We can compute the minimum of r which satisfied $\|f_L(x+r) - f_L(x)\|_v \geq \frac{f^{(y)}(f_L(x)) - f^{(j)}(f_L(x))}{\left\| W_{L+1}^{(j)} - W_{L+1}^{(y)} \right\|_u}$ by a minimum optimization problem below,

$$\arg\min_r Relu\left(\min_{j \neq y} \frac{f_{L+1}^{(y)}(x) - f_{L+1}^{(j)}(x)}{\left\| W_{L+1}^{y} - W_{L+1}^{j} \right\|_u} - \|f_L(x+r) - f_L(x)\|_v \right)$$

And if

$$\Delta \leq \min_r \left\{ \left\| \arg\min_r Relu\left(\min_{j \neq y} \frac{f_{L+1}^{(y)}(x) - f_{L+1}^{(j)}(x)}{\left\| W_{L+1}^{y} - W_{L+1}^{j} \right\|_u} - \|f_L(x+r) - f_L(x)\|_p \right) \right\|_v, \varepsilon \right\}, 1 = \frac{1}{u} + \frac{1}{v}.$$

Then there is no perturbation in U_Δ could classify $x+r$ differently to x.

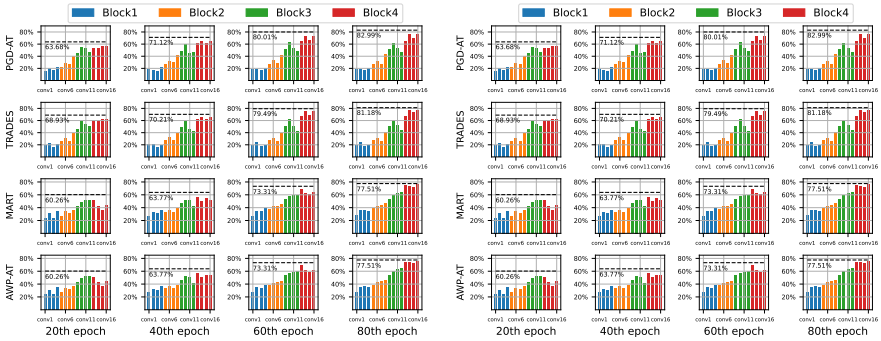
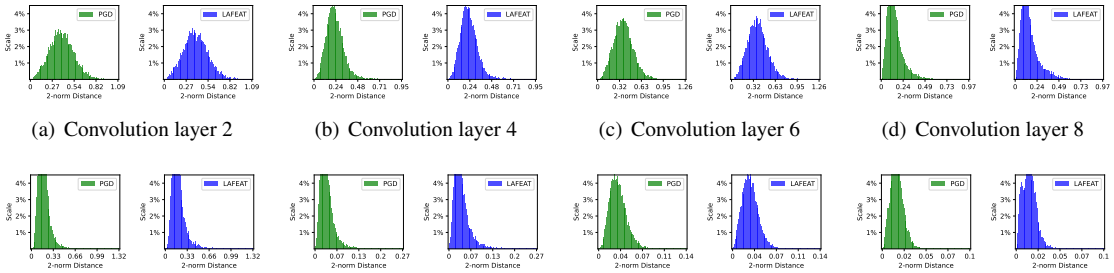


Figure 1. The figure shows the Clu.Acc for convolution layers of ResNet-18 trained by PGD-AT, TRADES, MART, and AWP-AT at different epochs on CIFAR10 testset. Clu.Acc is computed on 1-norm (left) and ∞-norm (right) respectively. The color of each column represents the residual block located. The black dashed line indicates the classification accuracy of the model at the present epoch.



(a) Convolution layer 2 (b) Convolution layer 4 (c) Convolution layer 6 (d) Convolution layer 8

(e) Convolution layer 10 (f) Convolution layer 12 (g) Convolution layer 14 (h) Convolution layer 16

Figure 2. The distribution of the size of the shift at intermediate layers caused by OLA and ILA. The result is derived from the last convolution layer of ResNet-18 trained by PGD-AT on CIFAR10 testset.

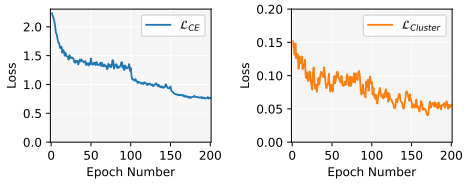


Figure 3. The variation of $\mathcal{L}_{CE}$ and $\mathcal{L}_{Cluster}$ of ResNet-18 during SAT on SVHN.

B.1 Clustering Effect in Different Norm.

Fig. 1 shows the comparison results of Clu.Acc of convolution layers in different convolution layers of models trained by PGD-AT, TRADES, MART, and AWP-AT. The results show the existence of Clustering Effect on different norm is similar to that of 2-norm. Therefore, we use 2-norm to represent the Clustering Effect in all measure methods.

B.2 OLA v.s. ILA on the Shift of Intermediate Layers.

As Fig. 2 shows, compared to OLA, ILA performs smaller distortion to the intermediate layer on average, which is compatible with the results in Fig. 2. The results show the difference of shift at intermediate layers induced by ILA and OLA increases with the depth of layers. Therefore we set the last two layer incorporating the computation of SAT.

B.3 The Convergence of SAT on SVHN.

We record the variation of $\mathcal{L}_{Cluster}$ during training on SVHN as the supplement. As shown in Fig. 3, $\mathcal{L}_{Cluster}$ converges rapidly during training, which is similar to the results on CIFAR10 dataset.

Dataset	Model	Method	Clean	FGSM	PGD_{100}	CW_∞	$AutoAttack_{100}$	$LAFEAT_{100}$
SVHN	ResNet-18	PGD-AT	85.37	70.92	50.97	50.24	48.91	46.14
		TRADES	86.81	74.05	53.65	53.38	53.07	49.42
		SAT	85.38	71.30	50.80	50.68	49.11	51.14
		SAT-TRADES	86.65	74.90	55.03	54.34	54.31	55.86
	WRN28×10	PGD-AT	87.02	73.12	53.05	53.01	52.24	48.39
		TRADES	88.13	73.95	53.24	52.17	51.97	48.54
		SAT	87.63	74.82	54.39	54.48	53.01	55.18
		SAT-TRADES	88.90	75.56	56.65	55.18	54.97	56.19

Table 1. Comparison of clean accuracy and robust accuracy against baseline attacks of neural networks across different defense mechanism on SVHN dataset(%).

Dataset	Model	Method	Clean	PGD_{100}	$AutoAttack_{100}$	$LAFEAT_{100}$
CIFAR100	WRN28×10	PGD-AT	63.66	28.72	28.60	26.23
		AWP-AT	64.79	30.75	30.64	28.24
		SAT	64.44	29.94	29.52	30.50
		SAT-AWP-AT	65.67	31.48	30.33	32.41

Table 2. Comparison of clean accuracy and robust accuracy against baseline attacks of neural networks across different defense mechanism on CIFAR100 dataset(%).

B.4 Performance Evaluation.

Table 1 and 2 describe the comparison of the adversarial robustness of neural networks on CIFAR100 and SVHN datasets. The model trained by baseline robust training methods and trained by the proposed SAT and its variants.