

APPENDIX

A PROOF FOR PROPOSITION 1

Since f^t is δ -robust, the prediction of $f^t(\mathbf{x})$ is invariant to the input perturbations smaller than the certified robust radius by definition, i.e.,

$$\arg \max f^t(\mathbf{x} + \epsilon) = \arg \max f^t(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{D}, \quad \forall \epsilon \in (0, \delta)^D, \quad (12)$$

where \mathcal{D} is the task-specific data set. Denote the student model distilled from the teacher model using normal knowledge distillation as $f^{KD}(\mathbf{x}) : \mathbb{R}^D \rightarrow \mathbb{R}^N$. The loss of the normal knowledge distillation can be formulated as

$$\mathcal{L}_{KD}(\mathbf{x}, y) = \lambda_{CE} \mathcal{L}_{CE}(f^{KD}(\mathbf{x}), y) + \lambda_{KL} T^2 \mathcal{L}_{KL}(f^{KD}(\mathbf{x})/T, f^t(\mathbf{x})/T), \quad \forall (\mathbf{x}, y) \in \mathbb{D}, \quad (13)$$

where \mathcal{L}_{CE} is the cross-entropy loss, \mathcal{L}_{KL} is the KL-divergence loss which is also called the soft loss in knowledge distillation, T is the temperature factor, and $\lambda_{CE}, \lambda_{KL}$ are hyper-parameters to balance the effects of the two losses. The loss of KDIGA is calculated by

$$\begin{aligned} \mathcal{L}_{IGA}(\mathbf{x}, y) = & \lambda_{CE} \mathcal{L}_{CE}(f^{IGA}(\mathbf{x}), y) + \lambda_{KL} T^2 \mathcal{L}_{KL}(f^{IGA}(\mathbf{x})/T, f^t(\mathbf{x})/T) \\ & + \lambda_{IGA} \|\nabla_{\mathbf{x}} \mathcal{L}_{CE}(f^{IGA}(\mathbf{x}), y) - \nabla_{\mathbf{x}} \mathcal{L}_{CE}(f^t(\mathbf{x}), y)\|_2, \quad \forall (\mathbf{x}, y) \in \mathbb{D}, \end{aligned} \quad (14)$$

where f^{IGA} is the student model, $\lambda_{CE}, \lambda_{KL}$ and λ_{IGA} are hyper-parameters.

Without loss of generality, we set the temperature factor $T = 1$ for both KD and KDIGA. According to the perfect student assumption, f^{IGA} satisfies the following equations:

$$\begin{cases} \nabla_{\mathbf{x}} \mathcal{L}_{IGA}(\mathbf{x}, y) - \nabla_{\mathbf{x}} \mathcal{L}_{CE}(\mathbf{x}, y) = 0 & (15) \\ f^{IGA}(\mathbf{x}) - f^t(\mathbf{x}) = 0 & (16) \end{cases}$$

$$\begin{cases} f^{IGA}(\mathbf{x}) = y, & \forall (\mathbf{x}, y) \in \mathcal{D}. \end{cases} \quad (17)$$

The cross-entropy loss is defined as

$$\mathcal{L}_{CE}(f(\mathbf{x}), y) = -\log\left(\frac{\exp(f(\mathbf{x})_y)}{\sum_j \exp(f(\mathbf{x})_j)}\right) = -f(\mathbf{x})_y + \log\left(\sum_j \exp(f(\mathbf{x})_j)\right), \quad (18)$$

where $f(\cdot)$ is a classifier and $f(\mathbf{x})_j$ is the j -th prediction of the output. Then the gradient of the cross-entropy loss with respect to the input is

$$\begin{aligned} \nabla_{\mathbf{x}} \mathcal{L}_{CE}(f(\mathbf{x}), y) &= -\nabla_{\mathbf{x}} f(\mathbf{x})_y + \nabla_{\mathbf{x}} \log\left(\sum_j \exp(f(\mathbf{x})_j)\right) \\ &= -\nabla_{\mathbf{x}} f(\mathbf{x})_y + \frac{\nabla_{\mathbf{x}} (\sum_i \exp(f(\mathbf{x})_i))}{\sum_j \exp(f(\mathbf{x})_j)} \\ &= -\nabla_{\mathbf{x}} f(\mathbf{x})_y + \frac{\sum_i \nabla_{\mathbf{x}} \exp(f(\mathbf{x})_i)}{\sum_j \exp(f(\mathbf{x})_j)} \\ &= -\nabla_{\mathbf{x}} f(\mathbf{x})_y + \frac{\sum_i \exp(f(\mathbf{x})_i) \nabla_{\mathbf{x}} f(\mathbf{x})_i}{\sum_j \exp(f(\mathbf{x})_j)} \end{aligned} \quad (19)$$

Denote $\mathbf{g} = g(\mathbf{x}) = \nabla_{\mathbf{x}} f(\mathbf{x})$, $\boldsymbol{\alpha} = \text{softmax}(f(\mathbf{x}))$, then

$$\begin{aligned} \nabla_{\mathbf{x}} \mathcal{L}_{CE}(f(\mathbf{x}), y) &= -g(\mathbf{x})_y + \frac{\sum_i \exp(f(\mathbf{x})_i) g(\mathbf{x})_i}{\sum_j \exp(f(\mathbf{x})_j)} \\ &= -g(\mathbf{x})_y + \boldsymbol{\alpha} \cdot \mathbf{g} \\ &= (\boldsymbol{\alpha} - \mathbf{i}_y) \cdot \mathbf{g}. \end{aligned} \quad (20)$$

where $\mathbf{i}_y = (0, \dots, 0, 1, 0, \dots, 0)$ is a unit vector of which the y -th element equals one. According to Eq. 16, $\boldsymbol{\alpha}^t = \boldsymbol{\alpha}^{IGA} = \boldsymbol{\alpha}$. The third term in Eq. 14 for input gradient alignment is

$$\begin{aligned} & \|\nabla_{\mathbf{x}} \mathcal{L}_{CE}(f^t(\mathbf{x}), y) - \nabla_{\mathbf{x}} \mathcal{L}_{CE}(f^{IGA}(\mathbf{x}), y)\| \\ &= \|(\boldsymbol{\alpha}^t - \mathbf{i}_y) \cdot \mathbf{g}^t - (\boldsymbol{\alpha}^{IGA} - \mathbf{i}_y) \cdot \mathbf{g}^{IGA}\| \\ &= \|(\boldsymbol{\alpha} - \mathbf{i}_y) \cdot (\mathbf{g}^t - \mathbf{g}^{IGA})\|. \end{aligned} \quad (21)$$

Given $\alpha - i_y \neq \mathbf{0}$, $\mathbf{g}^t - \mathbf{g}^{IGA}$ must be $\mathbf{0}$ since $\alpha - i_y$ and $\mathbf{g}^t - \mathbf{g}^{IGA}$ are not strictly orthogonal unless $\mathbf{g}^t - \mathbf{g}^{IGA} = \mathbf{0}$. According to Eq. 15, we have $\mathbf{g}^t - \mathbf{g}^{IGA} = \mathbf{0}$.

According to the local linearity assumption, $\forall \mathbf{x} \in \mathbb{D}, \forall \epsilon \in [0, \delta)^{H \times W \times C}$,

$$\begin{aligned} f^{IGA}(\mathbf{x} + \epsilon) &= f^{IGA}(\mathbf{x}) + \epsilon^T \cdot \mathbf{g}^{IGA}(\mathbf{x}) \\ &= f^t(\mathbf{x}) + \epsilon^T \cdot \mathbf{g}^t(\mathbf{x}) \\ &= f^t(\mathbf{x} + \epsilon) = f^t(\mathbf{x}) = f^{IGA}(\mathbf{x}). \end{aligned} \quad (22)$$

Therefore, the certified robust radius of f^{IGA} is at least δ , which proves Proposition 1.

However, the knowledge distillation without input gradient alignment cannot guarantee the adversarial robustness preservation. Suppose f^{KD} is a perfect student, we have

$$\begin{cases} f^{KD}(\mathbf{x}) - f^t(\mathbf{x}) = 0 \\ f^{KD}(\mathbf{x}) = y, \end{cases} \quad \forall (\mathbf{x}, y) \in \mathcal{D}. \quad (23)$$

$$(24)$$

We point out that f^{KD} can have different predictions around \mathbf{x} , for example, let $\tilde{\mathbf{x}} = \mathbf{x} + \epsilon \in \tilde{\mathcal{B}}(\mathbf{x}, \delta)$, denote $h(\mathbf{x}) = f^{KD}(\mathbf{x}) - f^t(\mathbf{x})$, then $h(\mathbf{x}) = 0, \forall (\mathbf{x}, y) \in \mathcal{D}$ according to Eq. 23. But $\exists h(\mathbf{x}), \exists \mathbf{x} \in \tilde{\mathcal{B}}(\mathbf{x}, \delta)$ s.t.

$$\arg \max f^{KD}(\mathbf{x}) \neq \arg \max f^t(\mathbf{x}) \quad (25)$$

since the first-order derivative of $h(\mathbf{x})$ is not constrained to be 0 in the neighbourhood of \mathbf{x} . This means the predictions of the student model distilled using knowledge distillation without input gradient alignment can be altered if we add perturbations to the input image.

B PROOF FOR PROPOSITION 2

$$\begin{aligned} & |\mathcal{L}_{CE}(f^s(\mathbf{x} + \epsilon), y) - \mathcal{L}_{CE}(f^t(\mathbf{x} + \epsilon), y)| \\ &= |\mathcal{L}_{CE}(f^s(\mathbf{x} + \epsilon), y) - \mathcal{L}_{CE}(f^s(\mathbf{x}), y) - \epsilon^T \nabla_{\mathbf{x}} \mathcal{L}_{CE}(f^s(\mathbf{x}), y) \\ &\quad - (\mathcal{L}_{CE}(f^t(\mathbf{x} + \epsilon), y) - \mathcal{L}_{CE}(f^t(\mathbf{x}), y) - \epsilon^T \nabla_{\mathbf{x}} \mathcal{L}_{CE}(f^t(\mathbf{x}), y)) \\ &\quad + (\mathcal{L}_{CE}(f^s(\mathbf{x}), y) - \mathcal{L}_{CE}(f^t(\mathbf{x}), y)) \\ &\quad + \epsilon^T (\nabla_{\mathbf{x}} \mathcal{L}_{CE}(f^s(\mathbf{x}), y) - \nabla_{\mathbf{x}} \mathcal{L}_{CE}(f^t(\mathbf{x}), y))| \\ &\leq \max_{\epsilon \in B(\delta)} |\mathcal{L}_{CE}(f^s(\mathbf{x} + \epsilon), y) - \mathcal{L}_{CE}(f^s(\mathbf{x}), y) - \epsilon^T \nabla_{\mathbf{x}} \mathcal{L}_{CE}(f^s(\mathbf{x}), y)| \\ &\quad + \max_{\epsilon \in B(\delta)} |\mathcal{L}_{CE}(f^t(\mathbf{x} + \epsilon), y) - \mathcal{L}_{CE}(f^t(\mathbf{x}), y) - \epsilon^T \nabla_{\mathbf{x}} \mathcal{L}_{CE}(f^t(\mathbf{x}), y)| \\ &\quad + \mathcal{L}_{CE}(f^s(\mathbf{x}), y) + \mathcal{L}_{CE}(f^t(\mathbf{x}), y) + \delta \|\nabla_{\mathbf{x}} \mathcal{L}_{CE}(f^s(\mathbf{x}), y) - \nabla_{\mathbf{x}} \mathcal{L}_{CE}(f^t(\mathbf{x}), y)\|. \end{aligned} \quad (26)$$