PeerAiD: Improving Adversarial Distillation from a Specialized Peer Tutor

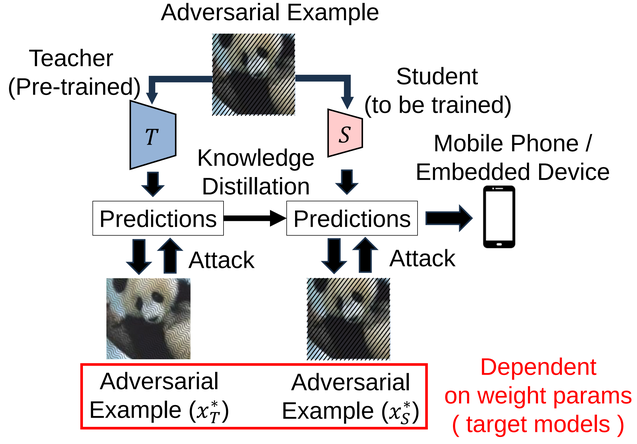
Jaewon Jung, Hongsun Jang, Jaeyong Song and Jinho Lee
Department of Electrical and Computer Engineering, Seoul National University, Seoul, Korea

Accelerated Intelligent Systems Lab.

CVPR SEATTLE, WA JUNE 17-21, 2024

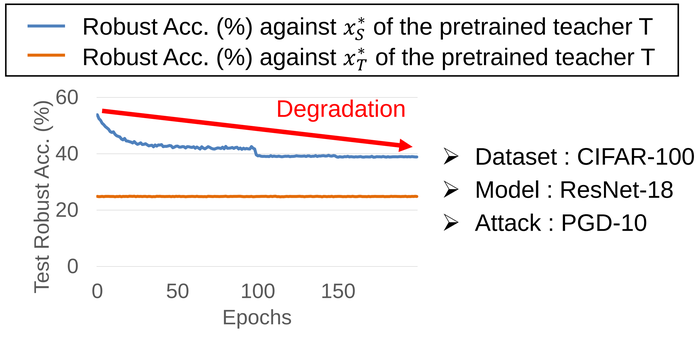
Background

- Adversarial Distillation (AD)
  - Adversarial distillation transfers the knowledge of the teacher to make the small student robust against adversarial examples.



Motivation

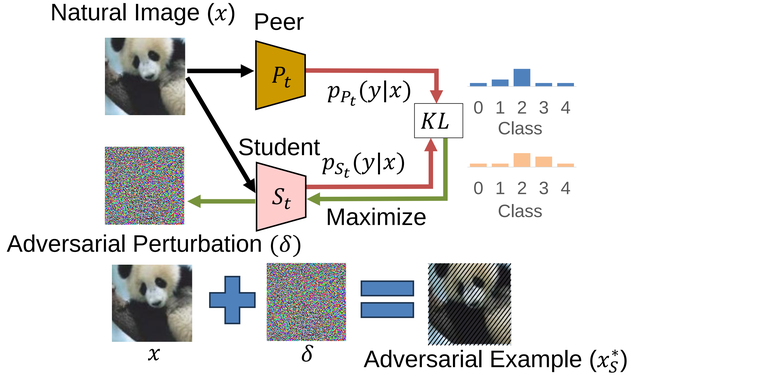
- The pretrained robust teacher model keeps losing its ability to defend against adversarial examples of the student model.



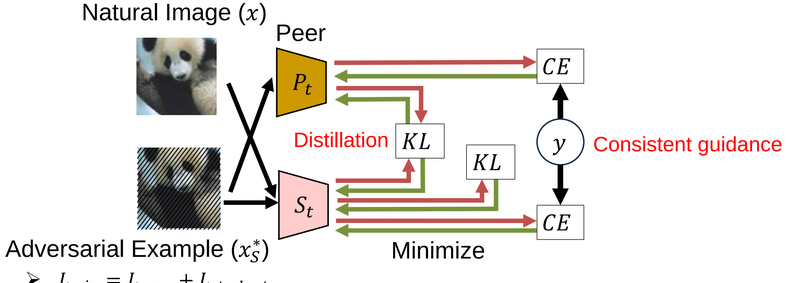
- Dataset : CIFAR-100
- Model : ResNet-18
- Attack : PGD-10

Method

- Peer Tutoring
  - PeerAiD proposes using the peer, which interactively learns with the student during adversarial distillation.
- Adversarial Example Generation (Inner Maximization)
  - The student model uses the predictions of **the peer model** as guidance.



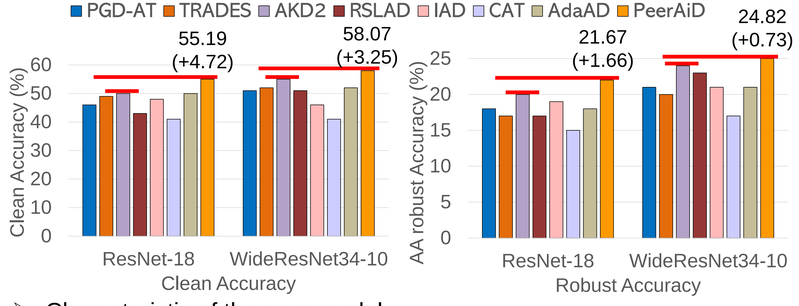
- $L_{max} = KL(P_t(x)||S_t(\tilde{x}))\ where\ \ \tilde{x} \in B(x,\epsilon),\ B(x,\epsilon) = \{\tilde{x} \mid \|x - \tilde{x}\|_\infty \leq \epsilon\}$
- Weight Optimization (Outer Minimization)
  - The student and the peer transfer their own knowledge to each other.



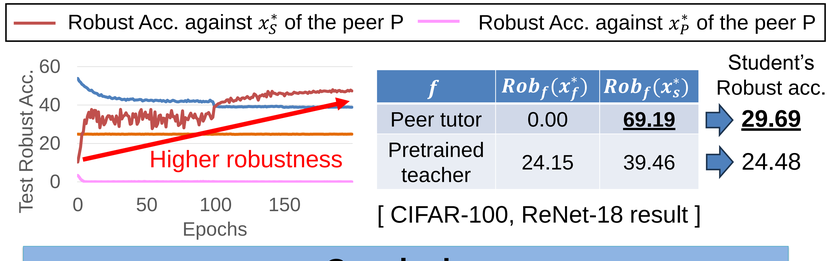
- $L_{min} = L_{peer} + L_{student}$
- $L_{peer} = \gamma_1 H(y, P_t(x_S^*)) + \gamma_2 \tau^2 KL(S_t^\tau(x_S^*)||P_t^\tau(x_S^*))$, τ : temperature
- $L_{student} = \lambda_1 H(y, S_t(x_S^*)) + \lambda_2 \tau^2 KL(P_t^\tau(x_S^*)||S_t^\tau(x_S^*)) + \lambda_3 \tau^2 KL(S_t^\tau(x)||S_t^\tau(x_S^*))$

Results

- TinyImageNet result
  - PeerAiD shows the highest AutoAttack robust accuracy compared to other baselines, while also providing higher clean accuracy.



- Characteristic of the peer model
  - ① Specialist who defends against adversarial examples of the student.
    - No tradeoff between the robustness and clean accuracy.
  - ② High clean accuracy
    - 75.63 (Peer) > 75.48 (Naturally trained)



f	$Rob_f(x_f^*)$	$Rob_f(x_S^*)$	Student's Robust acc.
Peer tutor	0.00	**69.19**	**29.69**
Pretrained teacher	24.15	39.46	24.48

[CIFAR-100, ReNet-18 result]

Conclusion

- We propose a novel online adversarial distillation method, PeerAiD
- The peer model specializes in defending against the student model's attack samples.
- PeerAiD improves AA robust accuracy by 1.66%p and clean accuracy by 4.72%p.

Profile