Transferring Adversarial Robustness Through Robust Representation Matching

Pratik Vaishnavi, Kevin Eykholt, Amir Rahmati

August 2022

Intro

- ML models can be fooled by adversarial examples
- Need ways to make models robust to such adversarial attacks
- Existing methods are not practical for real-world use
- This work proposes a method to transfer robustness

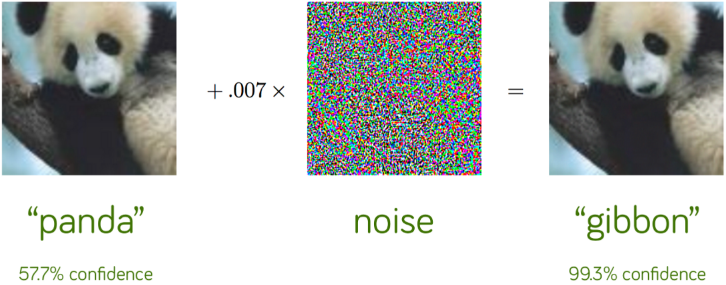

Figure 1: ML algorithms and especially DNNs are often brittle.

Standard Training

- Empirical Risk Minimization (ERM) updates the parameters, θ, of a ANN, F_θ, to minimize the learning model's loss, L

$$\min_{\theta} L(F_\theta(x), y)$$

Adversarial Attacks

- Adversarial Evasion Attacks (AEA) attempt to imperceptibly perturb inputs to cause misclassification
- Adversaries objective is to add a small perturbation, $\delta < \epsilon$, that maximizes the model's loss

$$\max_{\delta} L(F_{\theta}(x + \delta), y)$$

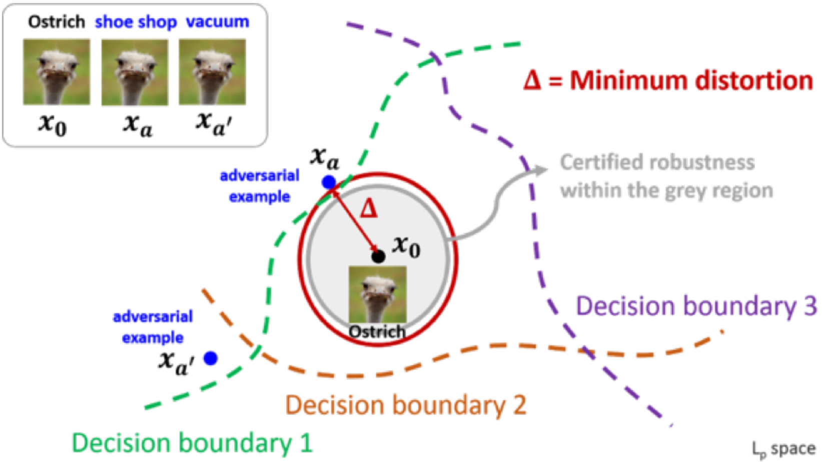

Figure 2: AEAs minimally perturb inputs to attain incorrect classification.

Adversarial Defense

- Adversarial Training (AT) [1] is the most natural defense
- Attempts to find parameters that minimize the adversary's expected attempts to increase loss

$$\min_{\theta} \max_{\delta} L(F_{\theta}(x + \delta), y)$$

- Essentially, augments the training data with adversarial inputs

Adversarial Training

- Several forward-backward passes each iteration vs single pass
- Slows training down by order of magnitude
- Requires knowledge of attacker's perturbation space

Explaining Robustness

- Adversarial examples are effective because of a model's tendency to learn non-robust features [2]
- Robust models must learn to focus on robust features that are strongly correlated with the input label
- Knowledge of robust features could be transferred between models

Transferring Adversarial Robustness

- Model robustification should not:
  1. reduce performance on non-adversarial examples
  2. be cost prohibitive
- Transferring robustness can eliminate the need to perform AT during retraining and make robustification cost efficient

Robust Representation Matching

- Robust Representation Matching (RRM) uses a student-teacher framework to transfer the knowledge of feature importance between models
- Trains a teacher model with AT
- Trains a student model with combined objective:
    1. Minimize the cross-entropy loss, L_C
    2. Minimize the robust representation loss, L_R

Robust Representation Matching (cont)

- Formally, the training objective for determining the parameters, θ, of the student NN S_θ is

$$\min_\theta \left[\lambda \cdot L_C(S_\theta(x), y) + L_R(x) \right]$$

- where the robust representation loss is the distance, e.g., cosine similarity, between output of the penultimate layers of the student and teacher models

$$L_R(x) = d(g_S(x), g_T(x))$$

- and λ weighs the contribution of the two different objectives

Why Match the Penultimate Layer?

- Including the robust representation loss term L_R forces the student to match the teacher's penultimate layer
- Matching the penultimate layer can transfer more knowledge than matching the output layer and is architecture-agnostic
- Intuitively, these hidden layer's encapsulate the network's understanding of the input and TLDR it seems to work
- Previous works used similar strategies [2, 3]

Adversarial Training Speedup

- When compared against other AT methods:
  - RRM achieves comparable performance to SAT/Fast AT in significantly less training time
  - RRM achieves greater performance to Free AT in almost the same training time
- Attacks conducted using AutoPGD [4], an iterative form of the FGSM attack [5]

Method	Training Time	Natural Accuracy	Adversarial Accuracy
SAT	1808	86%	48%
Fast AT	193	84%	50%
Free AT	29	71%	42%
RRM	30	76%	49%

Adversarial Robustness Transfer

- When compared against other transfer methods, RRM vastly outperforms its competitors

Method	Natural Accuracy	Adversarial Accuracy
RDT	80%	1%
KD	83%	3%
RRM	81%	46%

Tuning λ

- Recall RRM's optimization objective:

$$\min_{\theta}\left[\lambda \cdot L_C(S_\theta(x), y) + L_R(x)\right]$$

- L_C encourages the model to learn natural accuracy
- L_R encourages the model to learn robust representations
- λ balances the two training objectives

Tuning λ (cont)

- Increasing λ increases the importance of L_C and increases natural accuracy
- Decreasing λ increases the importance of L_R and increases adversarial accuracy (to an extent)

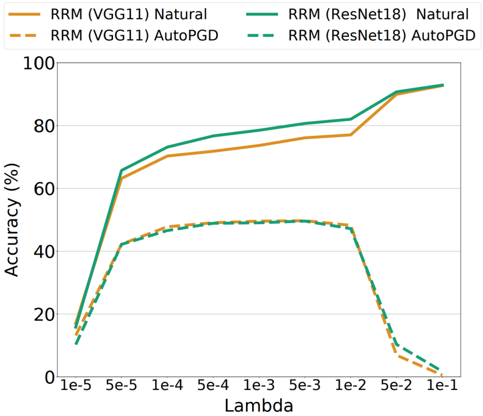

Figure 3

Limit Testing

- Hypothesize that training time per epoch roughly approximates a model's expressive power
- Found that simpler students struggle to learn from complex teachers because they are not complex enough to learn the robust features of the teacher

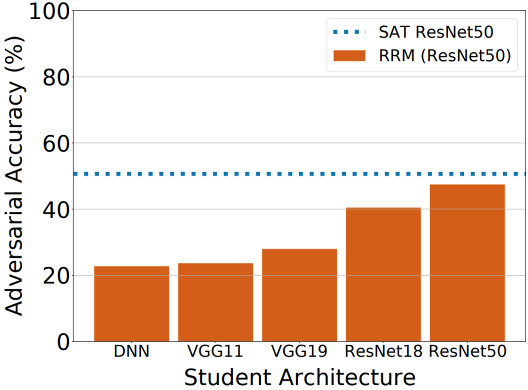

Figure 4

Limitations and Future Work

- RRM still depends on a teacher model and the difficulties that go along with using AT to attain one
- This work only studies RRM with respect to DNNs and image classification
- AT is not the silver bullet for adversarial defense nor is it the only defense strategy

Conclusions

- Introduced Robust Representation Matching (RRM) technique to transfer robustness between DNN models
- Demonstrated that RRM outperforms other adversarial training techniques and adversarial robustness transfer techniques

References

[1] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[2] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *Advances in neural information processing systems*, vol. 32, 2019.

[3] M. Goldblum, L. Fowl, S. Feizi, and T. Goldstein, "Adversarially robust distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3996–4003.

[4] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International conference on machine learning*. PMLR, 2020, pp. 2206–2216.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint*