# ENSEMBLE ADVERSARIAL TRAINING: ATTACKS AND DEFENSES

Florian Tram`er, Alexey Kurakin, Nicolas Papernot,
Ian Goodfellow, Dan Boneh, Patrick McDaniel,
April 2020

# INTRODUCTION

- ML models are vulnerable to adversarial examples
- Adversarial Examples are transferable across models enabling *Black-box attacks* (attacks performed with no prior knowledge of the model)
- Solution: Adversarial Training
  - Augmenting training data with adversarial examples
  - One suggested paper by Madry et al. 2017 tries to implement this but it was not scalable to ImageNet
- Is it possible to have robust models against *black-box* adversaries?

# INTRODUCTION

- What the paper proposes:
  - Show that adversarially trained models using *single-step* methods remain vulnerable to simple attacks i.e., *fast-single step methods* that maximize the model's loss converge to a degenerate global minimum
  - *Ensemble Adversarial Training,* a technique that arguments training data with perturbations transferred from other models

# THE ADVERSARIAL TRAINING FRAMEWORK

- Threat Model
- **Adversarial Training**

$$h^* = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{(x, y_{\text{true}}) \sim \mathcal{D}} \left[ \max_{\|x^{\text{adv}} - x\|_\infty \leq \epsilon} L(h(x^{\text{adv}}), y_{\text{true}}) \right] . \quad (1)$$

# THE ADVERSARIAL TRAINING FRAMEWORK

- Threat Model
- Adversarial Training
  - **Fast Gradient Sign Method (FGSM)**: linearizing the inner maximization problem

$$x^{\text{adv}}_{\text{FGSM}} := x + \varepsilon \cdot \texttt{sign}\left(\nabla_x L(h(x), y_{\text{true}})\right) \ . \tag{2}$$

# THE ADVERSARIAL TRAINING FRAMEWORK

- Threat Model
- Adversarial Training
  - Fast Gradient Sign Method (FGSM)
  - **Single-Step Least-Likely Class Method (Step-LL):** variant of FGSM, targets the least-likely class

$$x_{\mathrm{LL}}^{\mathrm{adv}} := x - \varepsilon \cdot \mathtt{sign}\left(\nabla_x L(h(x), y_{\mathrm{LL}})\right) \ . \tag{3}$$

# THE ADVERSARIAL TRAINING FRAMEWORK

- Threat Model
- Adversarial Training
  - Fast Gradient Sign Method (FGSM)
  - Single-Step Least-Likely Class Method (Step-LL)
  - **Iterative Attack (I-FGSM or Iter-LL):** iteratively applies the FGSM or Step-LL k number of times

# THE ADVERSARIAL TRAINING FRAMEWORK

- A Degenerate Global Minimum for Single Step Adversarial Training
  - On FGSM and Step-LL we approximate Equation (1) by replacing the solution to the inner with output of the attacks

$$h^* = \arg\min_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{(x, y_{\text{true}}) \sim \mathcal{D}} \left[ L(h(x^{\text{adv}}_{\text{FGSM}}), y_{\text{true}}) \right] . \tag{4}$$

$$L(h^*(x^{\text{adv}}_{\text{FGSM}}), y_{\text{true}}) \ll \max_{\|x^{\text{adv}} - x\|_\infty \leq \epsilon} L(h^*(x^{\text{adv}}), y_{\text{true}})] . \tag{6}$$

# THE ADVERSARIAL TRAINING FRAMEWORK

- Ensemble Adversarial Training
  - Augmenting a model's training data with adversarial examples crafted on other *static pre-trained models* **(decouple the generation of adversarial examples from the model being trained)**
  - Since adversarial examples are transferable between models, perturbations crafted on an external model are good approximations for the maximization problem on (1)
- Domain Adaptation with Multiple Sources

# EXPERIMENTS

- Attacks against adversarially trained networks

Table 1: **Error rates (in %) of adversarial examples transferred between models.** We use Step-LL with $\epsilon = 16/256$ for 10,000 random test inputs. Diagonal elements represent a white-box attack. The best attack for each target appears in bold. Similar results for MNIST models appear in Table 7.

| | Source | | | | |
|---|---|---|---|---|---|
| **Target** | v4 | v3 | $v3_{adv}$ | IRv2 | $IRv2_{adv}$ |
| v4 | **60.2** | 39.2 | 31.1 | 36.6 | 30.9 |
| v3 | 43.8 | **69.6** | 36.4 | 42.1 | 35.1 |
| $v3_{adv}$ | **36.3** | 35.6 | ~~26.6~~ | 35.2 | 35.9 |
| IRv2 | 38.0 | 38.0 | 30.8 | **50.7** | 31.9 |
| $IRv2_{adv}$ | **31.0** | 30.3 | 25.7 | 30.6 | ~~21.4~~ |

**Top 1**

| | Source | | | | |
|---|---|---|---|---|---|
| **Target** | v4 | v3 | $v3_{adv}$ | IRv2 | $IRv2_{adv}$ |
| v4 | **31.0** | 14.9 | 10.2 | 13.6 | 9.9 |
| v3 | 18.7 | **42.7** | 13.0 | 17.8 | 12.8 |
| $v3_{adv}$ | 13.6 | 13.5 | ~~9.0~~ | 13.0 | **14.5** |
| IRv2 | 14.1 | 14.8 | 9.9 | **24.0** | 10.6 |
| $IRv2_{adv}$ | 10.3 | **10.5** | 7.7 | 10.4 | ~~5.8~~ |

**Top 5**

# EXPERIMENTS

- Attacks against adversarially trained networks

Table 2: **Error rates (in %) for Step-LL, R+Step-LL and a two-step Iter-LL on ImageNet.** We use $\epsilon = 16/256$, $\alpha = \epsilon/2$ on 10,000 random test inputs. R+FGSM results on MNIST are in Table 7.

| | Top 1 | | | | | Top 5 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | v4 | v3 | $v3_{adv}$ | IRv2 | $IRv2_{adv}$ | v4 | v3 | $v3_{adv}$ | IRv2 | $IRv2_{adv}$ |
| **Step-LL** | 60.2 | 69.6 | 26.6 | 50.7 | 21.4 | 31.0 | 42.7 | 9.0 | 24.0 | 5.8 |
| **R+Step-LL** | 70.5 | 80.0 | **64.8** | 56.3 | 37.5 | 42.8 | 57.1 | **37.1** | 29.3 | 15.0 |
| **Iter-LL(2)** | **78.5** | **86.3** | ~~58.3~~ | **69.9** | **41.6** | **56.2** | **70.2** | ~~29.6~~ | **45.4** | **16.5** |

# EXPERIMENTS

- Attacks against Ensemble Adversarial Training

Table 4: **Error rates (in %) for Ensemble Adversarial Training on ImageNet.** Error rates on clean data are computed over the full test set. For 10,000 random test set inputs, and $\epsilon = 16/256$, we report error rates on white-box Step-LL and the *worst-case error* over a series of black-box attacks (*Step-LL, R+Step-LL, FGSM, I-FGSM, PGD*) transferred from the holdout models in Table 3. For both architectures, we mark methods tied for best in bold (based on 95% confidence).
**The subsequent work of Wu et al. (2020) proposes more powerful black-box attacks that result in error rates of at least 78% for all models.**

| | Top 1 | | | Top 5 | | |
|---|---|---|---|---|---|---|
| **Model** | **Clean** | **Step-LL** | **Max. Black-Box** | **Clean** | **Step-LL** | **Max. Black-Box** |
| v3 | **22.0** | 69.6 | 51.2 | **6.1** | 42.7 | 24.5 |
| $v3_{adv}$ | **22.0** | **26.6** | 40.8 | **6.1** | **9.0** | 17.4 |
| $v3_{adv\text{-}ens3}$ | 23.6 | 30.0 | **34.0** | 7.6 | **10.1** | **11.2** |
| $v3_{adv\text{-}ens4}$ | 24.2 | 43.3 | **33.4** | 7.8 | 19.4 | **10.7** |
| IRv2 | **19.6** | 50.7 | 44.4 | **4.8** | 24.0 | 17.8 |
| $IRv2_{adv}$ | **19.8** | **21.4** | 34.5 | **4.9** | **5.8** | 11.7 |
| $IRv2_{adv\text{-}ens}$ | **20.2** | 26.0 | **27.0** | **5.1** | 7.6 | **7.9** |

# CONCLUSION AND FUTURE WORK

- This paper has showed that adversarial training can be improved by decoupling generation of adversarial examples from the model being trained
- Experiments show that robustness attained to attacks from some models transfers to attacks from other models

# RESEARCH PLANS

- Aid the generation of **adversarial examples** that are used in the training process of an **adversarially robust model** by the use  of **Generative Models**
- Explore more into **domain adaptation** in order to improve results from current paper