Adversarial Attack and Defense

Stanley Chan and Guanzhe Hong
What is Defense?

Many attacks of security concern
• Digital attacks on image classifiers
• Misclassifying road signs [Evtimov et al. 2017]
• Incorrectly identifying ad as natural content [Biggio et al. 2017]
• Incorrectly recognizing voice commands [Carlini et al. 2018]
• Etc.

Current focus of defense research in CV
• Image classifier robustness
• Small $l_p$ distortion

But what does defense mean?
• What are the adversary’s goals?
• How much does the adversary know about the system?
• How much attack capability is allowed?
A defense system must be falsifiable!

Adversary goals
• To cause erroneous output
• E.g. targeted/untargeted image misclassification

Adversarial capabilities
• Need adversarial sample $x'$ to satisfy $d(x, x') < \epsilon$
• $d$ is commonly some $l_p$ norm
• Is this constraint reasonable?
  • Allows for rigorous definition of adversarial risk
    • Hard to compute it perfectly
  • Not suitable for many realistic scenarios [Gilmer et al. 2018]
Adversary knowledge

• Kerckhoff’s principle: unreasonable to assume defense system can be held secret, even in the black-box setting

• White-box
  • Full access to model

• Black-box
  • No knowledge about the model
  • Usually, attacker is allowed to probe the classifier a few times
Input transformations I

Existing attempts

1. Image cropping and rescaling + bit-depth reduction + JPEG compression + randomly drop pixels, restore with total variance minimization [Guo et al. 2018]
2. Image blurring [Li et al. 2016]
3. Color depth reduction + median filter [Xu et al. 2017]
4. Dimensionality reduction [Bhagoji et al. 2017]
5. Denoise before classification [Liao et al. 2018]
6. Upscale image with random zero paddings [Xie et al. 2018]
7. Project image to data manifold [Song et al. 2018, Samangouei et al. 2018]

Etc...
All found to be ineffective in their threat models [Athalye et al. 2018, Carlini et al. 2017]

1. Too “hard-coded”
   • E.g. Blurring filter
   • Vulnerable to slightly modified attacker
     • E.g. solve with gradient descent
       \[ \text{argmin}_{\delta} \|\delta\| \text{ s.t. } f(\text{blur}(x + \delta)) \neq f(\text{blur}(x)) \]

2. Obfuscated gradient
   • E.g. Image compression, bit-depth reduction, denoising, random zero padding, projection onto data manifold, etc.
   • Gradient information becomes nearly useless, obscuring iterative gradient-based attack
   • To attack
     • Compute expected gradient for network with randomization
     • Differentiable approximation to nondifferentiable layer

3. Ensemble of weak defenses is not strong
   • E.g. Color depth reduction + spatial smoothing, ensemble of detectors
   • For some defense, attacks transfer
   • For some, combine gradient info from each component
Adversarial Training I

Basic properties
• Aims to solve

$$\min_{\theta} \mathbb{E}_{(x,y) \in D} \left[ \max_{\delta \in S} L(\theta; x + \delta, y) \right]$$
• Inner maximum approximated by an attack algorithm
• Use natural and adversarial samples for training
  • The natural and the adversarial sample share the same label
• About 50% accuracy under strongest attack (white box) [Athalye et al. 2018]
Adversarial Training II

Adversarial training
• Current state-of-the-art
• Requires
  • A strong attack on the specific network
  • Network is sufficiently expressive [Madry et al. 2018]
• Why?
  • Seems to land on “flat” parts of the loss landscape
  • Decision boundary generally has low curvature near natural images [Dezfooli et al. 2019]
• Seems to lead to decrease in natural accuracy [Su et al. 2018]
Modified Network I

1. Network distillation [Papernot et al. 2016]
   • Train *teacher* network with temperature $T$
   • Evaluate teacher network on training instances, produce softmax vectors
   • Feed softmax vectors to *student* network as *soft labels*
1. Network distillation
   • Softmax output $q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$
     - As $T \to \infty$, $q_i \to \frac{1}{m}$
     - In usual NN, $T = 1$
1. Network distillation

- Softmax output \( q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \)
  - As \( T \to \infty \), \( q_i \to \frac{1}{m} \)
  - In usual NN, \( T = 1 \)

- Claims
  - The student network should overfit less than the teacher
  - Robust against \( l_0 \) attacks
- Later found ineffective in white box setting [Carlini et al. 2016]
1. Network distillation
2. Detector network [Metzen et al. 2017]
   - Augment classifier with a detection network
   - Detection network takes input from intermediate layers
   - Procedure
     - Train classifier on natural examples
     - Freeze classifier, use natural and adversarial examples to train the detectors
   - Found ineffective against strong attacks [Carlini et al. 2017]
Defense Against Physical Attacks

Sentinet: detecting physical attacks [Chou et al. 2018]
• Assumes physical attacks are localized in space
• Rough procedure
  • Use Grad-CAM to generate saliency map of the NN
  • High-heat regions ≈ high focus from the NN
  • Extract these regions
  • Overlay each of them on test images
    • If many misclassified, this region is likely adversarial
Sentinet: detecting physical attacks [Chou et al. 2018]

- Assumes physical attacks are localized in space
- Rough procedure
  - Use Grad-CAM to generate saliency map of the NN
  - High-heat regions $\approx$ high focus from the NN
  - Extract these regions
  - Overlay each of them on test images
    - If many misclassified, this region is likely adversarial
- Claims
  - Robust against white box, data poisoning, and trojaning attacks
  - Somewhat long to run (3x longer to run than base NN)
  - Weak against nonlocalized attacks
Theory

Problems
1. Existence of adversarial examples
   • What are the causes?
   • Can we say something about terms like
     \[ \mathbb{P}(\exists \delta \in B_p(0, \epsilon) \text{ s.t. } f(x + \delta) \neq y) \]?
2. Adversarially robust classifiers
   • Can we construct provably robust classifiers?
   • Can we ever achieve high natural accuracy and adversarial robustness?
3. Is our current \( l_p \)-perturbation adversary model reasonable?
   • Realistic?
Existence of Adversarial Examples

Existing works

   - Isoperimetric inequalities
   - As dimension $\uparrow$, the volume of the band $\uparrow$
Existence of Adversarial Examples

Existing works
1. Limits on adversarial robustness
2. Linearity hypothesis [Goodfellow et al., 2014]
   • Local linearity of decision boundary
3. High nonlinearity
   • Result of overfitting
Existence of Adversarial Examples

Existing works
1. Limits on adversarial robustness [Fawzi, evasion, adv inevitable]
2. Linearity hypothesis [Goodfellow et al., 2014]
   • Local linearity of decision boundary
3. High nonlinearity
4. Explaining universal perturbations
   • Shared positively curved directions across images [Dezfooli et al. 2018]
   • Similar geometric observations for transfer attacks [Tramer et al. 2017]
Adversarially Robust Classifiers

Existing works
• A few provably robust classifiers, do not really scale to large datasets such as ImageNet [Wong et al. 2018]

Accuracy vs. robustness
• People observe empirically a tradeoff between the two on ImageNet [Su et al. 2018]
• Theoretical treatment suggests the same [Zhang et al. 2018]
This model is too restricted for many problems [Gilmer et al. 2018]

- Physical world computer vision
  - Object is imaged through an imaging system
  - Environmental factors affect the input image
- Content preservation constraint
  - E.g. adversarial video input
  - User can tolerate strong perturbation, cropping, etc.
- Spam detection
  - Attack sample can be outside of existing distribution
- Audio adversarial attack
  - E.g. weakly constrained attack on a voice assistant
  - E.g. no-constraint attack on stolen cell phone

......


References


References


Thank You