Attacks Meet Interpretability: Attribute-steered Detection of Adversarial Samples

Guanhong Tao, Shiqing Ma, Yingqi Liu, Xiangyu Zhang
Understanding Adversarial Samples

Pixel-wise Differences (x50 times)

Legitimate input

C&W$_2$ attack

Model

Isla Fisher

A.J. Buckley
Understanding Adversarial Samples

Model

Legitimate input

C&W attack

Pixel-wise Differences (x50 times)

Isla Fisher

A.J. Buckley

Human
Understanding Adversarial Samples

- Legitimate input
- C&W$_2$ attack
- Model
- Human
- Isla Fisher
- A.J. Buckley

Pixel-wise Differences ($\times$50 times)
Understanding Adversarial Samples

Pixel-wise Differences (×50 times)

Legitimate input

C&W\(_2\) attack

Model

Human

Isla Fisher

A.J. Buckley

Model
Understanding Adversarial Samples

- Legitimate input
- C&W$_2$ attack
- Model
- Human
- Isla Fisher
- A.J. Buckley

Pixel-wise Differences ($\times$50 times)
Understanding Adversarial Samples

Pixel-wise Differences (×50 times)

Legitimate input

C&W₂ attack

Model

Human

Isla Fisher

A.J. Buckley
Understanding Adversarial Samples

- Idea: is the classification result of a model mainly based on human perceptible attributes?
Architecture of Aml
Architecture of Aml
Architecture of Aml

Input → Landmark generation
Architecture of Aml

1. Input
2. Landmark generation

✓ Left eye
✓ Right eye
✓ Nose
✓ Mouth
✓ …

Attribute annotation
Architecture of Aml

1. Input
2. Landmark generation
3. Attribute annotation
   ✓ Left eye
   ✓ Right eye
   ✓ Nose
   ✓ Mouth
   ✓ …
3. Attribute witness extraction
Architecture of Aml

1. Input
2. Landmark generation
   - ✓ Left eye
   - ✓ Right eye
   - ✓ Nose
   - ✓ Mouth
   - ✓ …
3. Attribute annotation
4. Attribute witness extraction
   - Attribute-steered model
Architecture of Aml

1. Input
2. Landmark generation
   ✓ Left eye
   ✓ Right eye
   ✓ Nose
   ✓ Mouth
   ✓ ...
3. Attribute annotation
4. Attribute witness extraction
   Attribute-steered model
   Original model
Architecture of Aml

1. Input
2. Landmark generation
   ✓ Left eye
   ✓ Right eye
   ✓ Nose
   ✓ Mouth
   ✓ …
3. Attribute annotation
4. Attribute witness extraction
5. Attribute-steered model
   Original model
   Consistency observer
Architecture of Aml

1. Input
2. Landmark generation
   - Left eye
   - Right eye
   - Nose
   - Mouth
   - …
   - Attribute annotation
3. Attribute witness extraction
4. Attribute-steered model
5. Consistency observer

Original model
Challenges

• Are there correspondences between attributes and neurons?

• If yes, how to extract corresponding neurons?
Challenges

• Are there correspondences between attributes and neurons?

• If yes, how to extract corresponding neurons?

• Propose: Bi-directional reasoning
Challenges

• Are there correspondences between attributes and neurons?

• If yes, how to extract corresponding neurons?

• **Propose: Bi-directional reasoning**
  ‣ Forward: attribute changes —> neuron activation changes
Challenges

• Are there correspondences between attributes and neurons?

• If yes, how to extract corresponding neurons?

• **Propose: Bi-directional reasoning**
  ‣ Forward: attribute changes $\rightarrow$ neuron activation changes

  ‣ Backward: neuron activation changes $\rightarrow$ attribute changes
Challenges

• Are there correspondences between attributes and neurons?

• If yes, how to extract corresponding neurons?

• **Propose: Bi-directional reasoning**
  ‣ Forward: attribute changes $\rightarrow$ neuron activation changes
  ‣ Backward: neuron activation changes $\rightarrow$ attribute changes
  ‣ Backward: no attribute changes $\rightarrow$ no neuron activation changes
Attribute Witness Extraction
Attribute Witness Extraction

Input
Attribute Witness Extraction

Model → Attribute substitution → Model

Input

Feature variants
Attribute Witness Extraction

Input → Model

A: Attribute substitution

B: Attribute preservation

Model → Model

Feature variants

C: Feature invariants

D: Feature invariants
Attribute Witness Extraction

Input → Model

Attribute substitution

Feature variants

Feature invariants

Attribute witnesses
Experimental Results
Experimental Results

• Attribute witnesses
Experimental Results

- Attribute witnesses
  - The number of witnesses extracted is **smaller than 20**, although there are **64-4096** neurons in each layer.
Experimental Results

• Attribute witnesses
  ‣ The number of witnesses extracted is smaller than 20, although there are 64-4096 neurons in each layer

• Adversary detection
Experimental Results

• Attribute witnesses
  ‣ The number of witnesses extracted is smaller than 20, although there are 64-4096 neurons in each layer

• Adversary detection
  ‣ Achieve 94% detection accuracy for 7 different kinds of attacks with 9.91% false positives on benign inputs
Experimental Results

• Attribute witnesses
  ‣ The number of witnesses extracted is smaller than 20, although there are 64-4096 neurons in each layer

• Adversary detection
  ‣ Achieve 94% detection accuracy for 7 different kinds of attacks with 9.91% false positives on benign inputs
  ‣ A state-of-the-art technique Feature Squeezing (NDSS ’18) can only achieve 55% accuracy with 23.3% false positives for face recognition systems
Thank you!

Please visit our poster #99
05:00-07:00 PM @ Room 210 & 230 AB