Hard-label Black-box Universal Adversarial Patch Attack

Guanhong Tao, Shengwei An, Siyuan Cheng, Guangyu Shen, Xiangyu Zhang
Hard-label Black-box Universal Adversarial Patch Attack

Guanhong Tao, Shengwei An, Siyuan Cheng, Guangyu Shen, Xiangyu Zhang
Universal Black-box Induce misclassification for any given input

Hard-label
Universal Black-box

Induce misclassification for any given input

Black-box

No access to the model weight parameters

Hard-label
Universal  Induce misclassification for any given input

Black-box  No access to the model weight parameters

Hard-label  Only have the knowledge of the predicted label
Why Hard-label Black-box Universal Attack?

Machine learning as a service (MLaaS)

- Companies deploy ML models on online platforms
- Applications using MLaaS are susceptible to attacks: facial recognition, optical character recognition, etc.
Why Hard-label Black-box Universal Attack?

Machine learning as a service (MLaaS)
- Companies deploy ML models on online platforms
- Applications using MLaaS are susceptible to attacks: facial recognition, optical character recognition, etc.

ML Models are intellectual properties
- Only provide API access → black-box
- Only return the predicted result → hard-label
- Limited number of queries → universal
How To Generate?

White-box

\[ \nabla [ \text{Model (Input } \oplus \text{Trigger}) = \text{Bird} ] \]
How To Generate?

White-box

Black-box

\[ \nabla [ \text{Model (Input} \oplus \text{Trigger}) \rightarrow \text{Bird} ] \]

Confidence

Dog 0.9

Bird 0.1

Gradient
Let’s Approximate It!

Black-box

- For a single input, add a set of random noises on the trigger
- Inspect whether any noise leads to the target prediction
- Obtain the (estimated) gradient based on the noises

Misclassified

Bird

✓/✗?
Let’s Approximate It!

For a single input, add a set of random noises on the trigger
Inspect whether any noise leads to the target prediction
Obtain the (estimated) gradient based on the noises

Black-box

Trigger + Noise
Input
Model

Misclassified

Bird
✓/✗?
Let’s Approximate It!

For a single input, add a set of random noises on the trigger
Inspect whether any noise leads to the target prediction
Obtain the (estimated) gradient based on the noises
Aggregate the gradients for multiple inputs to mutate the trigger
Gradient Estimation for Multiple Inputs

- Leverage historical misclassified rate
- Dynamically adjust importance

Direct Estimation

Importance-aware Estimation
Is Grad Approx. Sufficient?

Additive noises may not increase the attack success rate
Is Grad Approx. Sufficient?

Additive noises may not increase the attack success rate

- Hard to determine the magnitude of the noise
Is Grad Approx. Sufficient?

Additive noises may not increase the attack success rate

- Hard to determine the magnitude of the noise
Is Grad Approx. Sufficient?

Additive noises may not increase the attack success rate

- Hard to determine the magnitude of the noise
- Limited number of queries
Is Grad Approx. Sufficient?

Additive noises may not increase the attack success rate

- Hard to determine the magnitude of the noise
- Limited number of queries

History is always instructive!

- Two close-by minima indicate a promising region
- Interpolation between them yields a better trigger
## Experiment Setup

### Datasets & Models
- **Datasets:** CIFAR-10, SVHN, STL-10, GTSRB
- **Models:** ResNet18, ResNet34, ResNet50, VGG11, GoogleNet, DenseNet121, MobileNet V2

### Commercial Services
- Microsoft Azure¹
- Clarifai²

### Baselines
- 3 hard-label black-box adversarial attacks: HSJA³, GRAPHITE⁴, SparseEvo⁵
- 3 soft-label black-box attacks: Bandits⁶, SPSA⁷, Sparse-RS⁸

---

² [https://www.clarifai.com/](https://www.clarifai.com/)
Attack Performance

- Generate a trigger for each pair of classes
  - Size: 7x7 (4.79% of the input)
  - # Queries: 50k
- Count the number of pairs above a certain attack success rate (ASR)
Attacking Online Services

Two online commercial services: Microsoft Azure and Clarifai

- Upload data for training (not deployed)
- Use the prediction API for attack
- Size: 7x7  # Queries: 240

Results (averaged on 10 pairs)

- Azure: 74% (vs. 60% by HSJA)
- Clarifai: 74% (vs. 53% by HSJA)
Countermeasures

Certifiable Defense: PatchCleanser¹
- Produce correct predictions no matter whether inputs are adversarially perturbed
- Average certified robust accuracy: 0.17%

Query-based Defense: Blacklight²
- Identify malicious queries by black-box attacks
- Average detection rate: 0.2%

Universal Adversarial Patch Detection: SentiNet³
- Reject adversarially perturbed inputs
- Average detection accuracy: 50.53%

Related Work

Propose a novel **hard-label black-box universal adversarial patch attack**, obtaining **more than twice high-ASR patch triggers (>90%)** than eight baselines

Successfully **attack two online commercial services**, Microsoft Azure and Clarifai, with an average **ASR of 74%**

Effectively **evade three state-of-the-art defense techniques**

**Conclusion**
Thank You

Guanhong Tao

taog@purdue.edu

https://www.cs.purdue.edu/homes/taog/