Hard-label Black-box Universal Adversarial Patch Attack

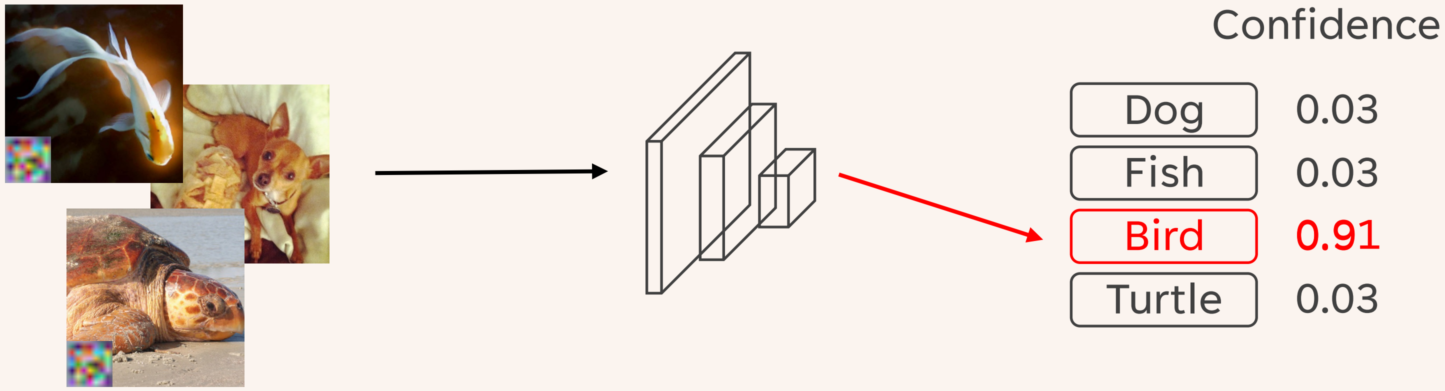
Guanhong Tao, Shengwei An, Siyuan Cheng, Guangyu Shen, Xiangyu Zhang

PURDUE UNIVERSITY

Hard-label Black-box Universal Adversarial Patch Attack

Guanhong Tao, Shengwei An, Siyuan Cheng, Guangyu Shen, Xiangyu Zhang

PURDUE UNIVERSITY

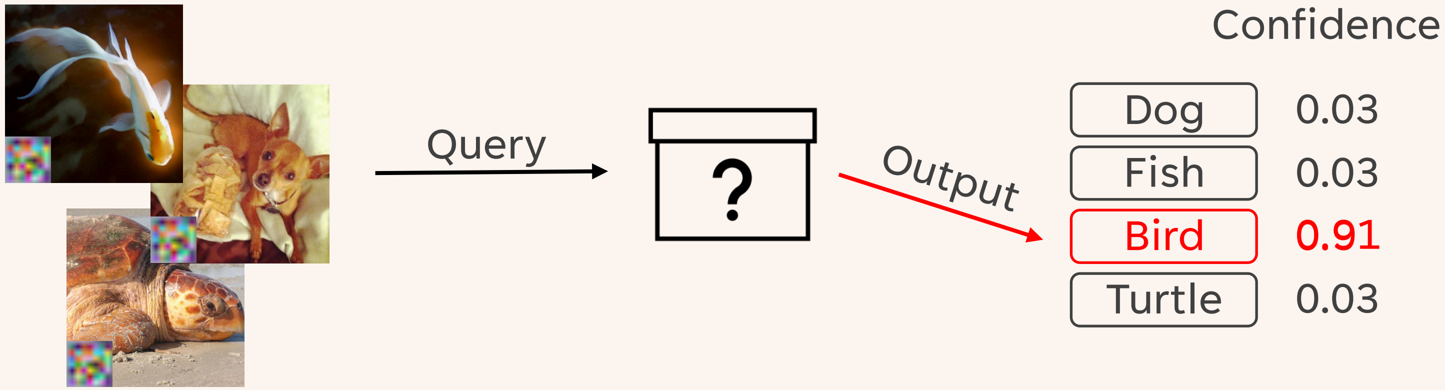


Universal

Induce misclassification for any given input

Black-box

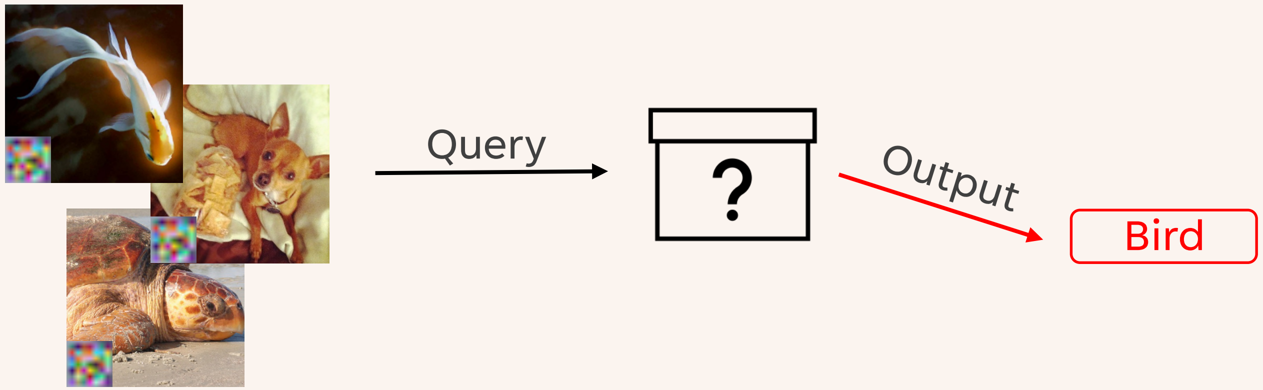
Hard-label



Universal — Induce misclassification for any given input

Black-box — No access to the model weight parameters

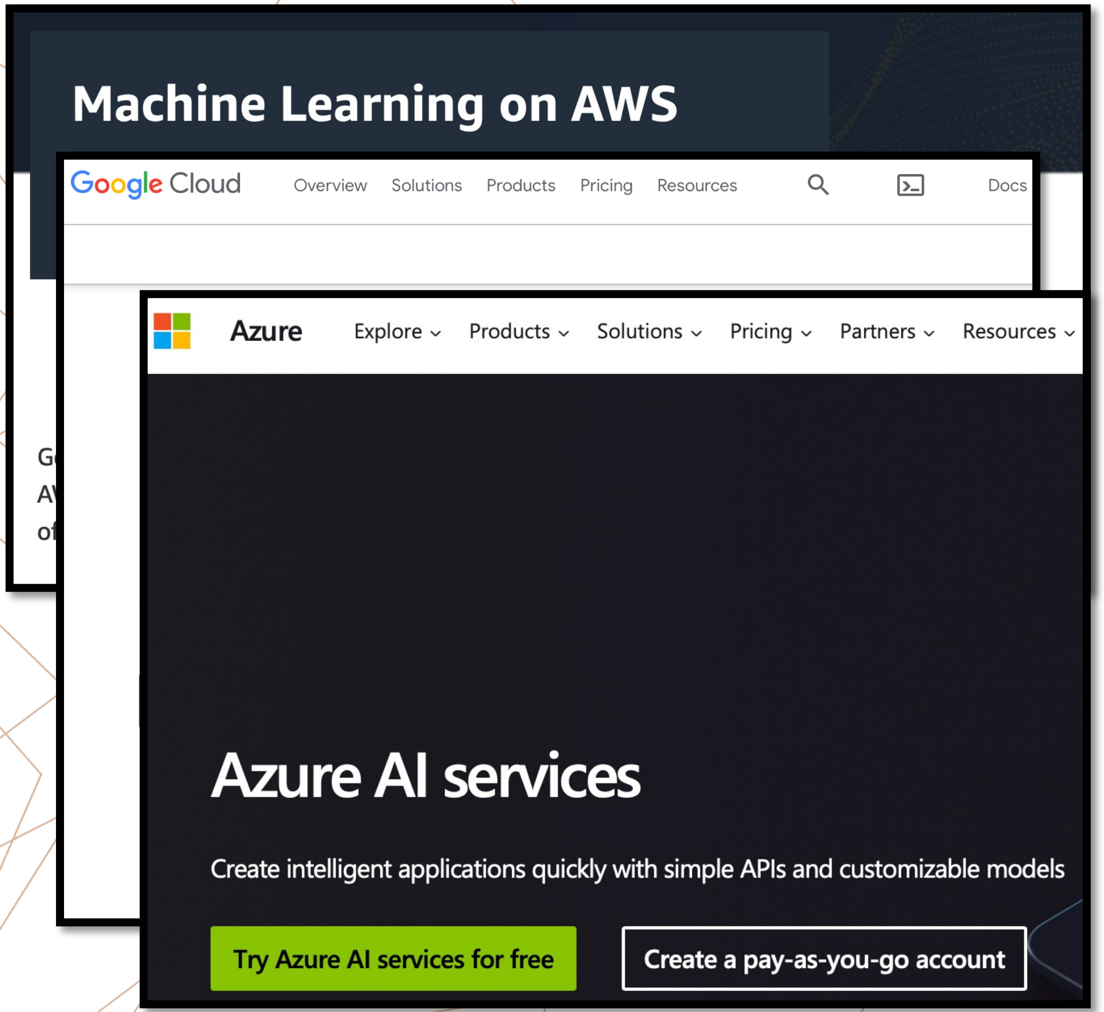
Hard-label



Universal — Induce misclassification for any given input

Black-box — No access to the model weight parameters

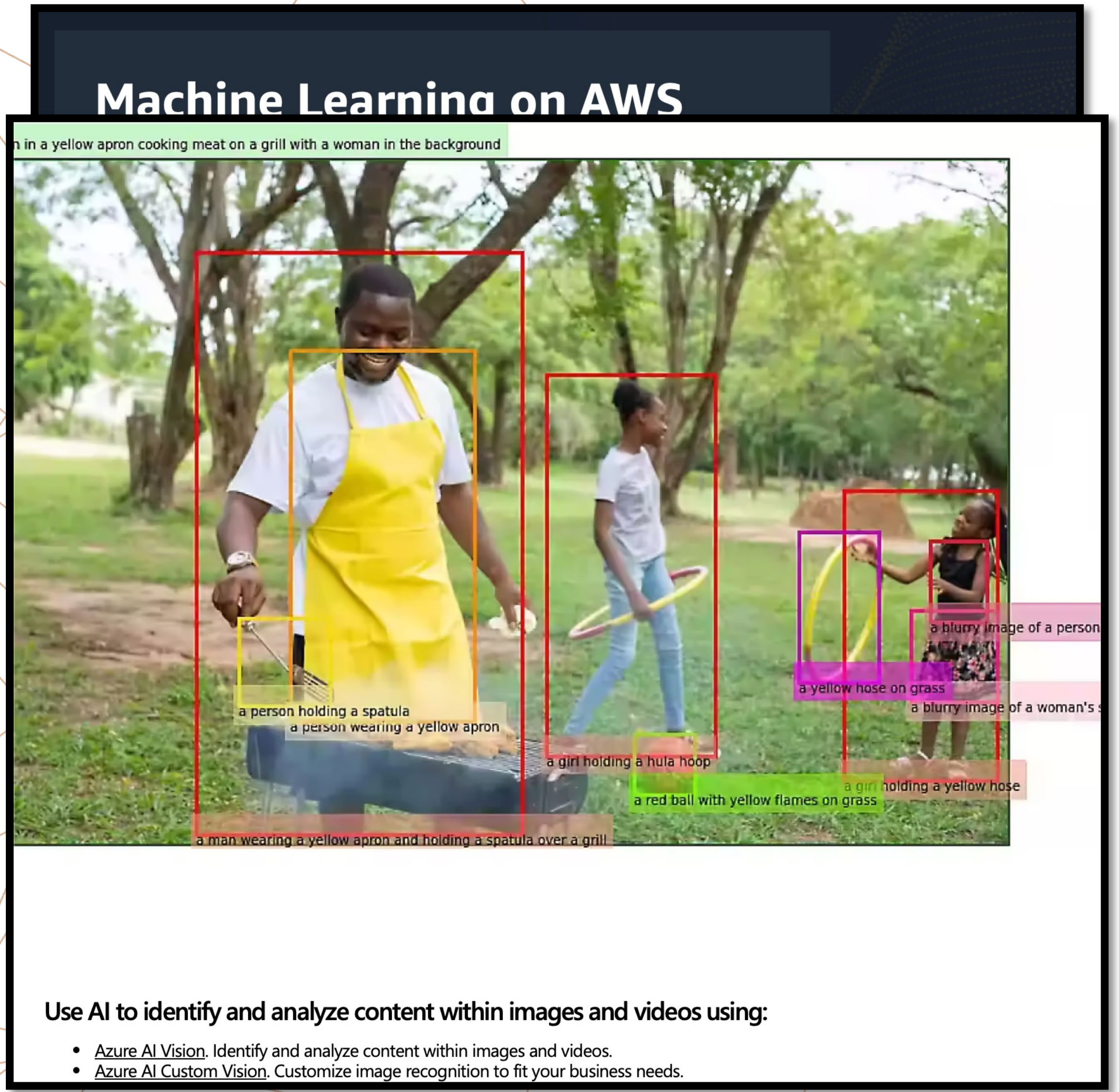
Hard-label — Only have the knowledge of the predicted label



Why Hard-label Black-box Universal Attack?

Machine learning as a service (MLaaS)

- Companies deploy ML models on online platforms
- Applications using MLaaS are suspectable to attacks: facial recognition, optical character recognition, etc.



Why Hard-label Black-box Universal Attack?

Machine learning as a service (MLaaS)

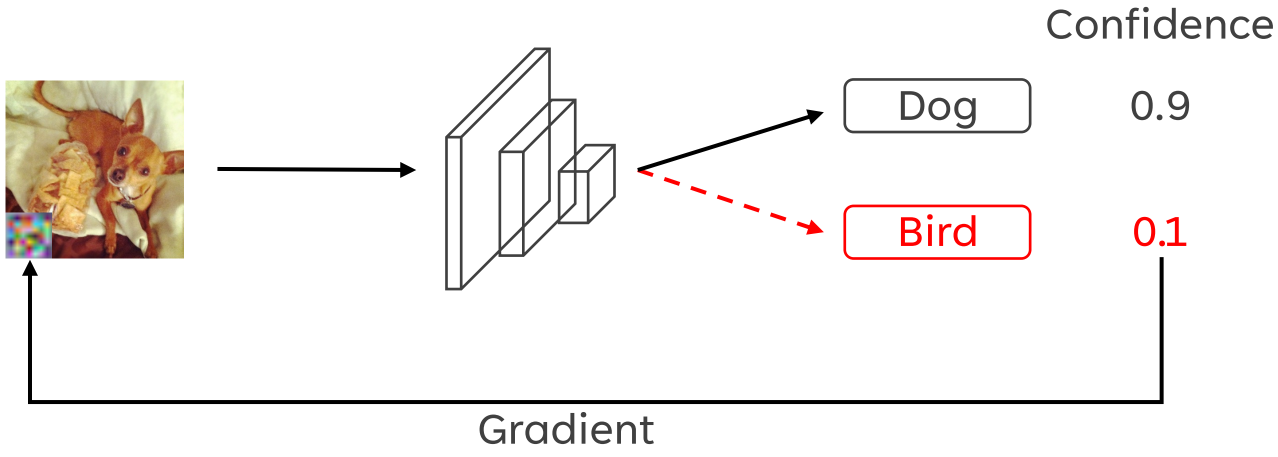
- Companies deploy ML models on online platforms
- Applications using MLaaS are suspectable to attacks: facial recognition, optical character recognition, etc.

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

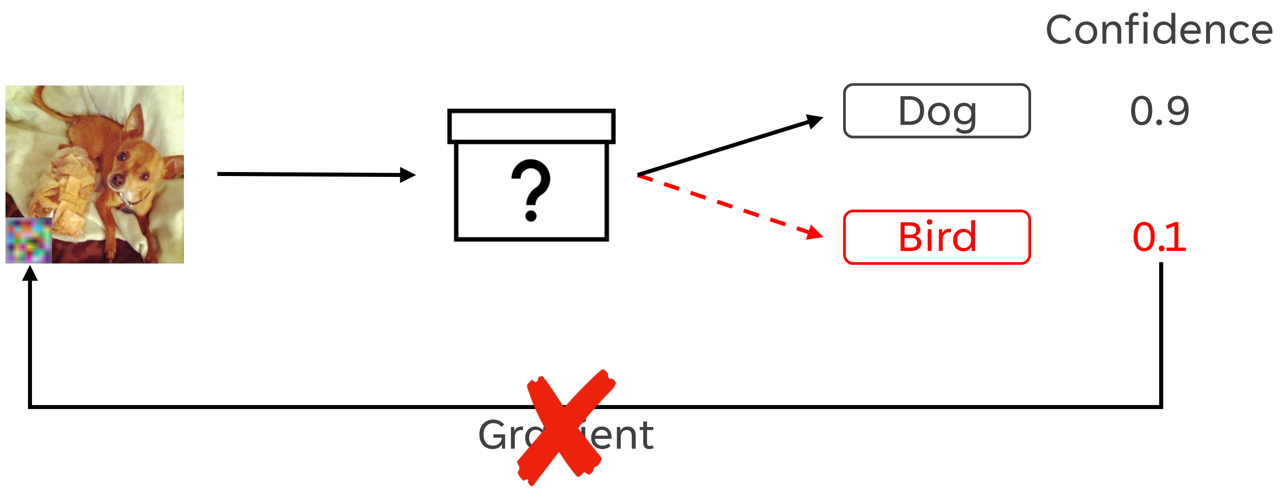


∇ [Model (Input ⊕ Trigger) = Bird]

How To Generate?

~~White-box~~

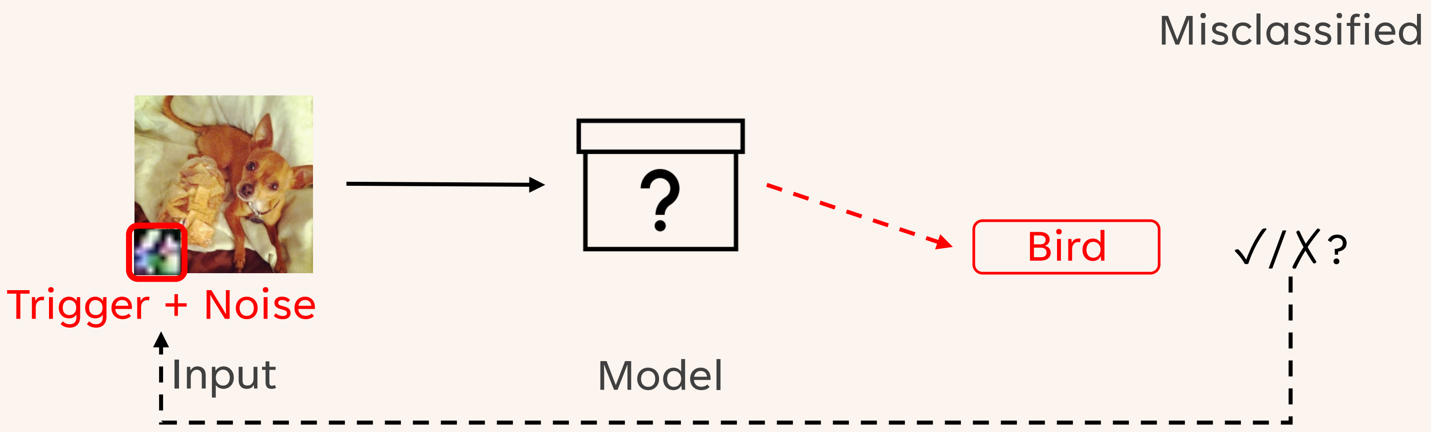
Black-box

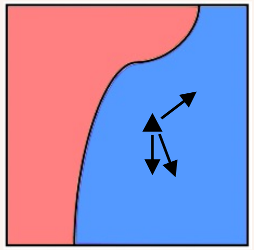


~~∇ [Model (Input ⊕ Trigger) = Bird]~~

Let's Approximate It!

Black-box

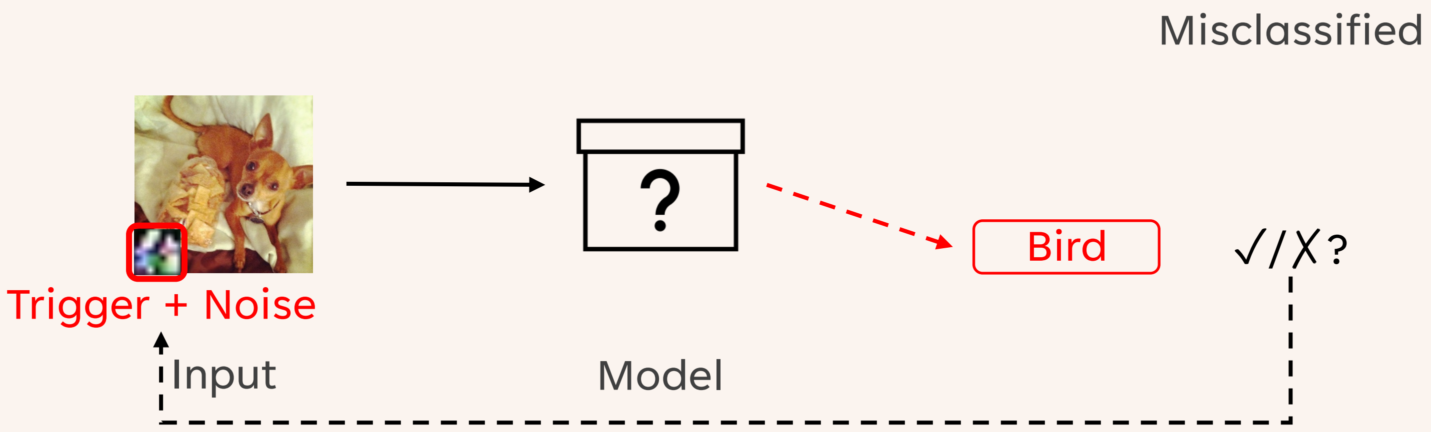


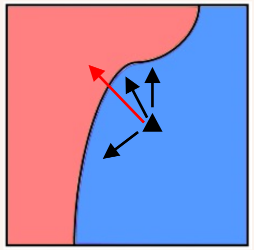


- For a single input, add a set of random noises on the trigger
- Inspect whether any noise leads to the target prediction
- Obtain the (estimated) gradient based on the noises

Let's Approximate It!

Black-box

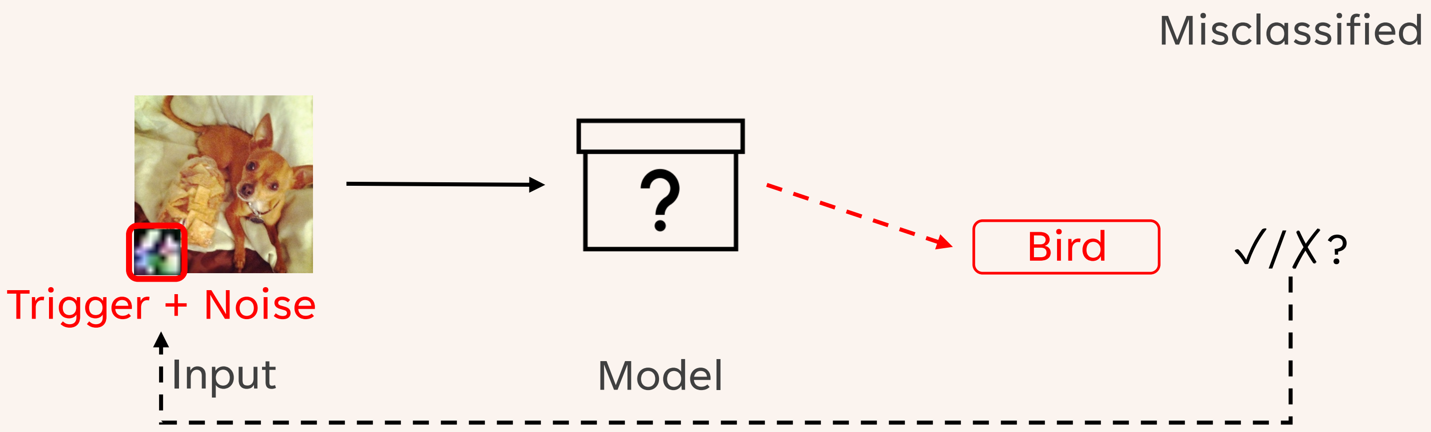


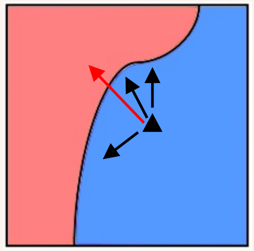


- For a single input, add a set of random noises on the trigger
- Inspect whether any noise leads to the target prediction
- Obtain the (estimated) gradient based on the noises

Let’s Approximate It!

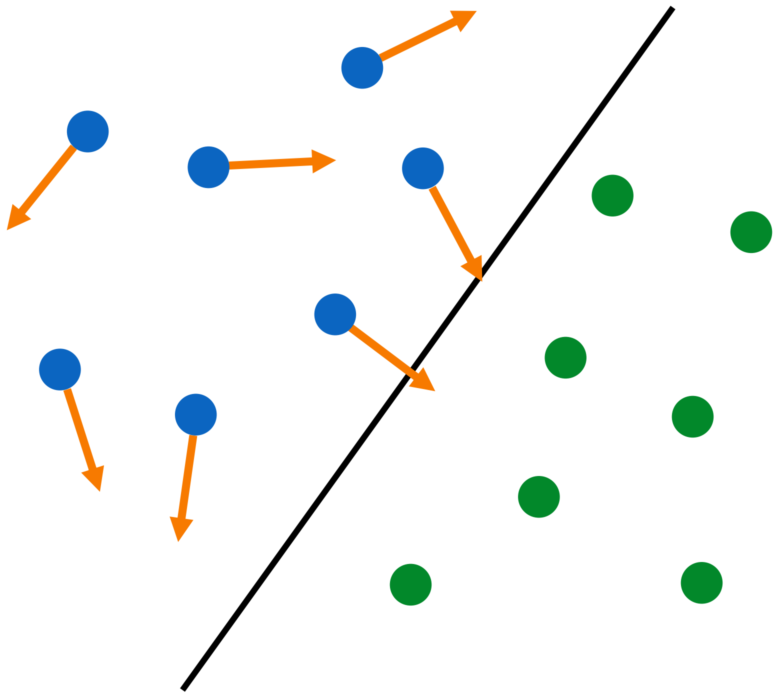
Black-box



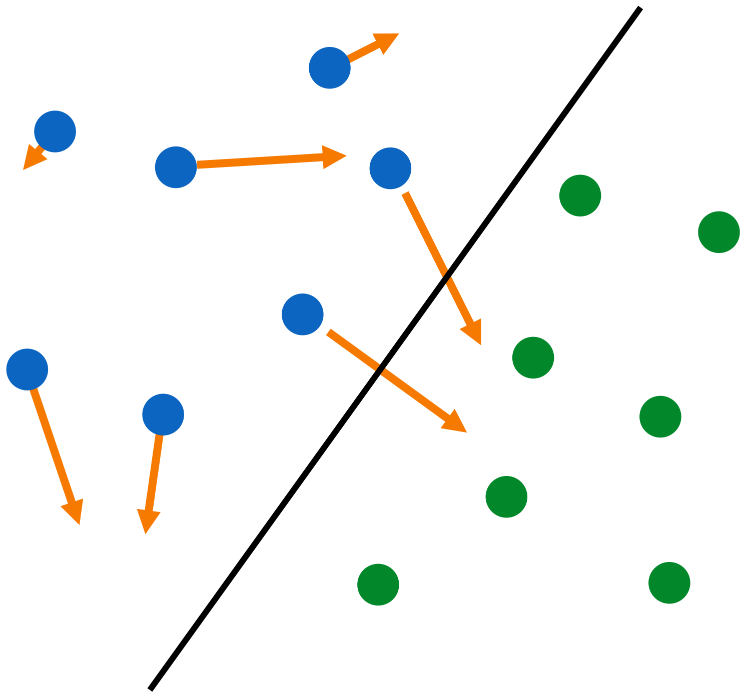


- For a single input, add a set of random noises on the trigger
- Inspect whether any noise leads to the target prediction
- Obtain the (estimated) gradient based on the noises
- Aggregate the gradients for multiple inputs to mutate the trigger

Gradient Estimation for Multiple Inputs

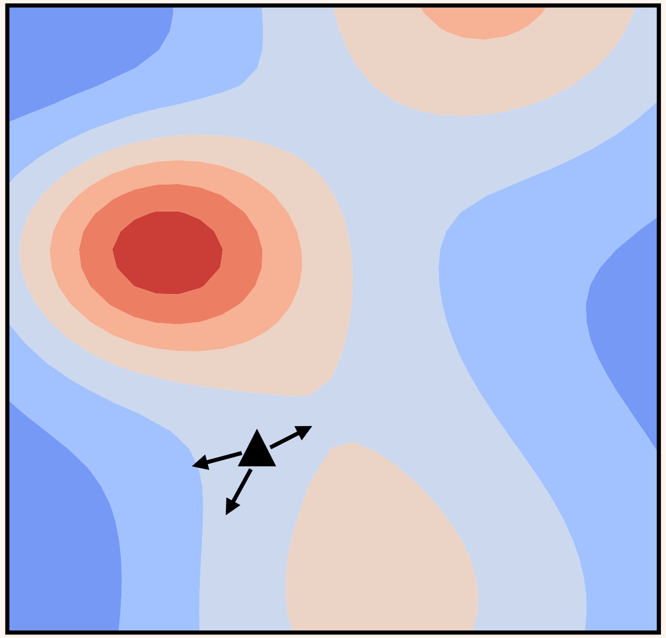


Direct Estimation



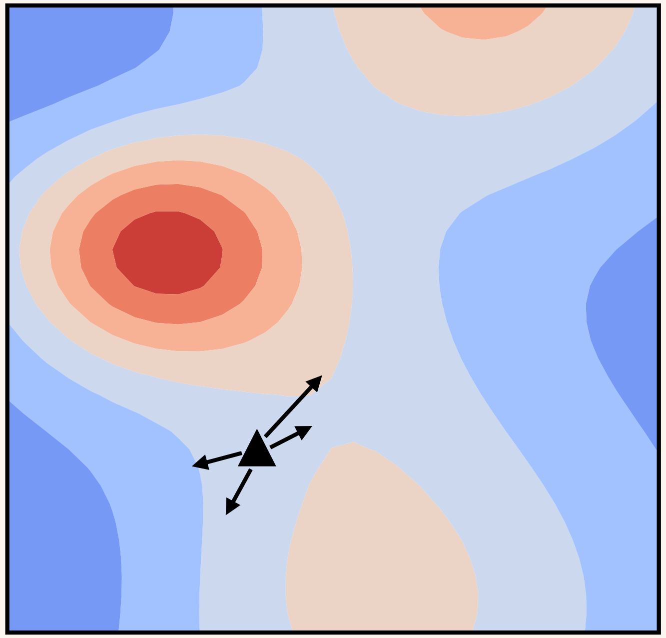
Importance-aware Estimation

- Leverage historical misclassified rate
- Dynamically adjust importance



Is Grad Approx. Sufficient?

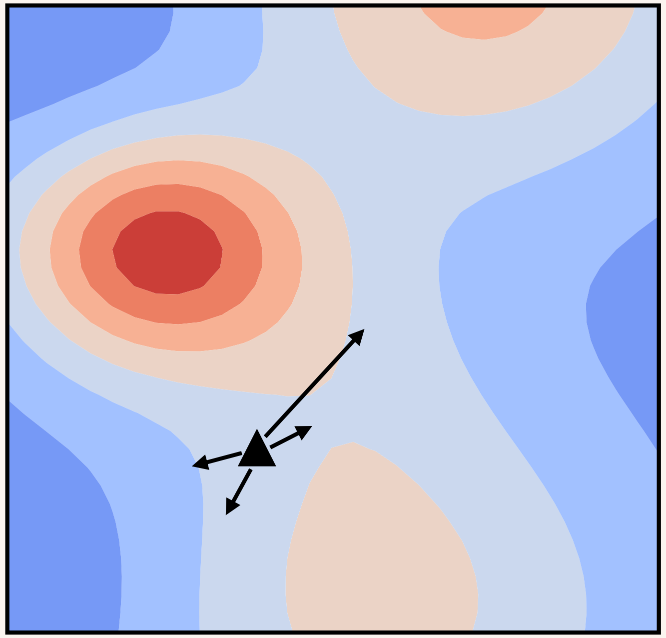
Additive noises may not increase the attack success rate



Is Grad Approx. Sufficient?

Additive noises may not increase the attack success rate

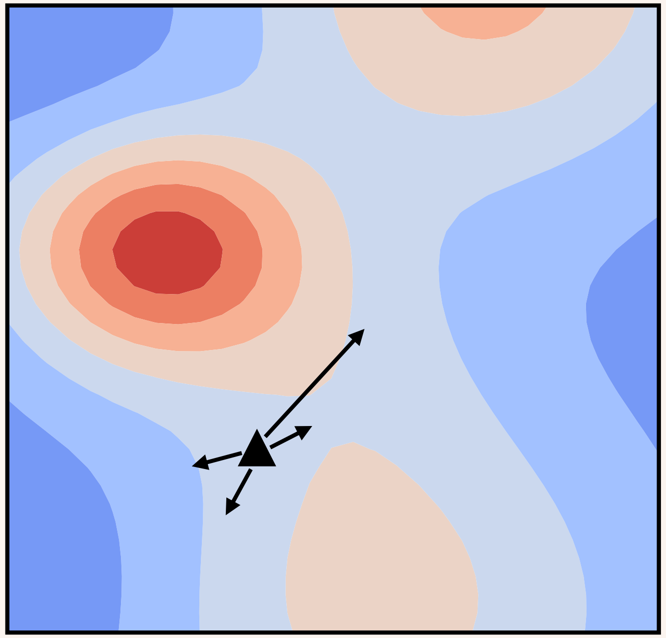
- Hard to determine the magnitude of the noise



Is Grad Approx. Sufficient?

Additive noises may not increase the attack success rate

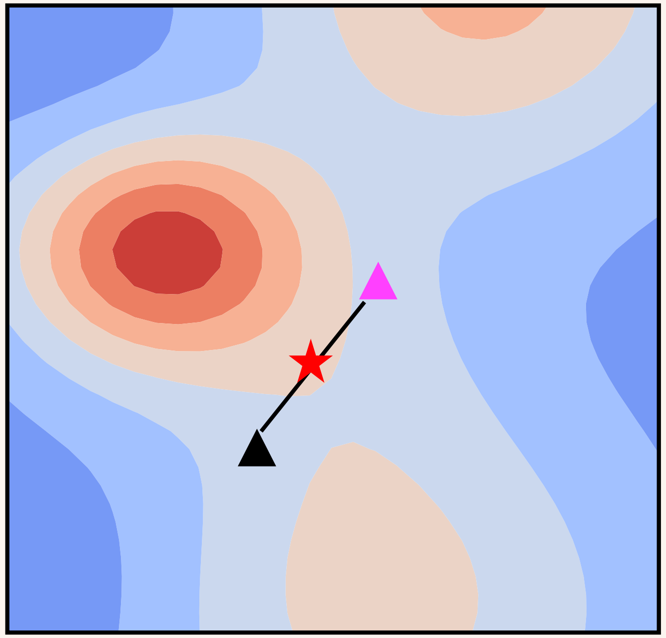
- Hard to determine the magnitude of the noise



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 Azure[1]
- Clarifai[2]

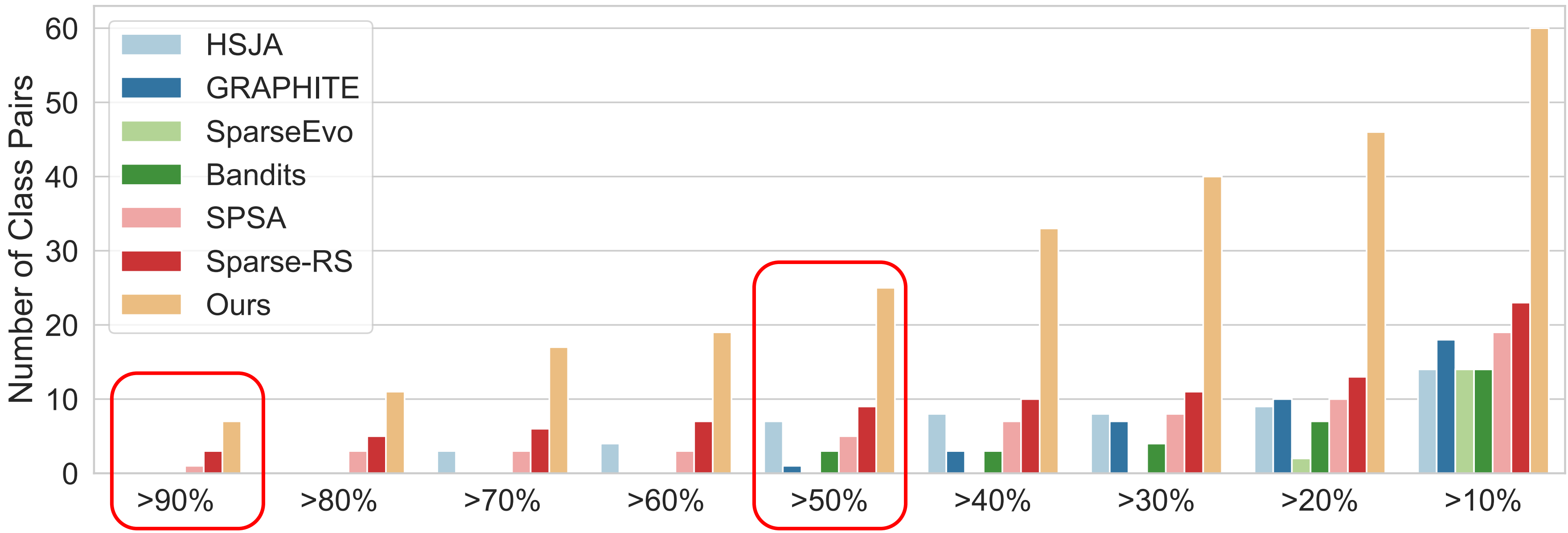
Baselines

- 3 hard-label black-box adversarial attacks: HSJA[3], GRAPHITE[4], SparseEvo[5]
- 3 soft-label black-box attacks: Bandits[6], SPSA[7], Sparse-RS[8]

[1] https://azure.microsoft.com/en-us/ services/cognitive- services/
[2] https://www.clarifai.com/
[3] Chen, Jianbo, et al. HopSkipJumpAttack: A query-efficient decision-based attack. S&P 2020.
[4] Feng, Ryan, et al. Graphite: Generating automatic physical examples for machine-learning attacks on computer vision systems. EuroS&P 2022.
[5] Vo, Viet, et al. Query efficient decision based sparse attacks against black-box deep learning models. ICLR 2022.
[6] Ilyas, Andrew, et al. Prior convictions: Black-box adversarial attacks with bandits and priors. ICLR 2019.
[7] James C Spall. A one-measurement form of simultaneous perturbation stochastic approximation. Automatica 1997.
[8] Croce, Francesco, et al. Sparse-RS: a versatile framework for query-efficient sparse black-box adversarial attacks. AAAI 2022.

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: PatchCleanser[1]

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

Query-based Defense: Blacklight[2]

- Identify malicious queries by black-box attacks
- Average detection rate: 0.2%

Universal Adversarial Patch Detection: SentiNet[3]

- Reject adversarially perturbed inputs
- Average detection accuracy: 50.53%

[1] Xiang, Chong, et al. PatchCleanser: Certifiably robust defense against adversarial patches for any image classifier. USENIX Security 2022.
[2] Li, Huiying, et al. Blacklight: Scalable defense for neural networks against query-based black-box attacks. USENIX Security 2022.
[3] Chou, Edward, et al. SentiNet: Detecting localized universal attack against deep learning systems. SPW 2020.

Related Work

[1] Chen, Jianbo, et al. HopSkipJumpAttack: A query-efficient decision-based attack. S&P 2020.
[2] Feng, Ryan, et al. Graphite: Generating automatic physical examples for machine-learning attacks on computer vision systems. EuroS&P 2022.
[3] Vo, Viet, et al. Query efficient decision based sparse attacks against black-box deep learning models. ICLR 2022.
[4] Ilyas, Andrew, et al. Prior convictions: Black-box adversarial attacks with bandits and priors. ICLR 2019.
[5] James C Spall. A one-measurement form of simultaneous perturbation stochastic approximation. Automatica 1997.
[6] Croce, Francesco, et al. Sparse-RS: a versatile framework for query-efficient sparse black-box adversarial attacks. AAAI 2022.
[7] Gilks, Walter R, et al. Markov chain Monte Carlo in practice. CRC press, 1995.
[8] Banzhaf, Wolfgang, et al. Genetic programming: an introduction: on the automatic evolution of computer programs and its applications. Morgan Kaufmann Publishers Inc., 1998.
[9] Xiang, Chong, et al. PatchCleanser: Certifiably robust defense against adversarial patches for any image classifier. USENIX Security 2022.
[10] Li, Huiying, et al. Blacklight: Scalable defense for neural networks against query-based black-box attacks. USENIX Security 2022.
[11] Chou, Edward, et al. SentiNet: Detecting localized universal attack against deep learning systems. SPW 2020.
...

Conclusion

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

Thank You

Guanhong Tao

taog@purdue.edu

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