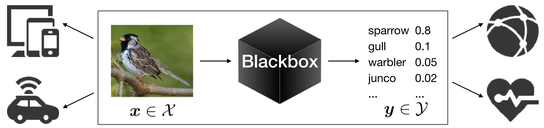
Knockoff Nets: Stealing Functionality of Black-Box Models

Tribhuvanesh Orekondy[1] Bernt Schiele[1] Mario Fritz[2]

[1] Max Planck Institute for Informatics [2] CISPA Helmholtz Centre for Information Security
Saarland Informatics Campus, Germany

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

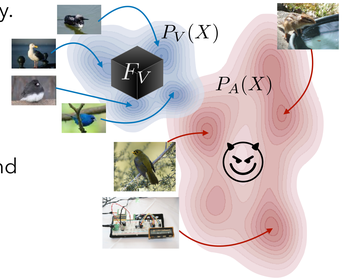
Motivation



- ML models deployed as blackboxes: inputs in, predictions out.
- Threat: Can the blackbox be 'knocked-off'?
- Could circumvent IP, monetary, and labour costs of training an ML model.

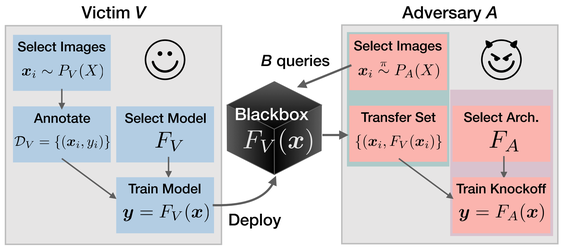
Problem: Model Functionality Stealing

- Data-limited adversary.
- Two-player game.



Adversary's objective:

1. Steal functionality of blackbox (when F_V and P_V are unknown).
2. Using minimum # of queries B.



Generating Knockoffs

Transfer set construction: Which images to query?

Method 1: π = Random

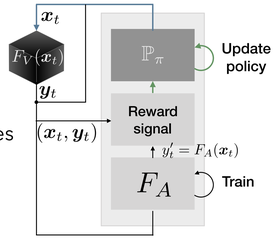
- Samples images randomly (without replacement).
- But: might query irrelevant images.

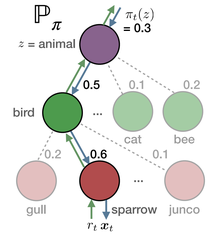
Training knockoff F_A: Which model to train?

- Any sufficiently complex architecture.

Method 2: π = Adaptive

- Motivation: (a) Improve sample efficiency (b) Interpretable policy π.
- Approach: Multi-armed hierarchical bandits.
- Actions (arms): supplemented labels of images in P_A (e.g., 1K classes for Imagenet).
- Rewards
  - a. Certainty: $R^{\text{cert}}(y_t) = P(y_{t,k_1}|x_t) - P(y_{t,k_2}|x_t)$
  - b. Diversity: $R^{\text{div}}(y_{1:t}) = \sum_k \max(0, y_{t,k} - \bar{y}_{t:t-\Delta,k})$
  - c. Loss: $R^{\mathcal{L}}(y_t, \hat{y}_t) = \mathcal{L}(y_t, \hat{y}_t)$



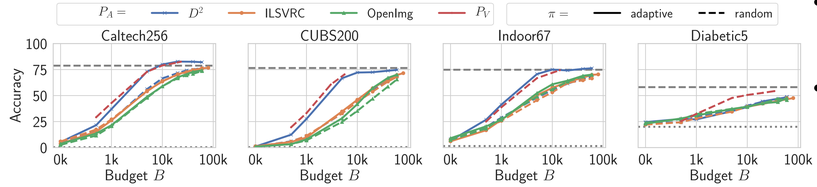


$$\pi_t(z) = \frac{e^{H_t(z)}}{\sum_{z'} e^{H_t(z')}} \qquad H_{t+1}(z_t) = H_t(z_t) + \alpha(r_t - \bar{r}_t)(1 - \pi_t(z_t)) \quad \text{and} \quad H_{t+1}(z') = H_t(z') + \alpha(r_t - \bar{r}_t)\pi_t(z') \quad \forall z' \neq z_t$$

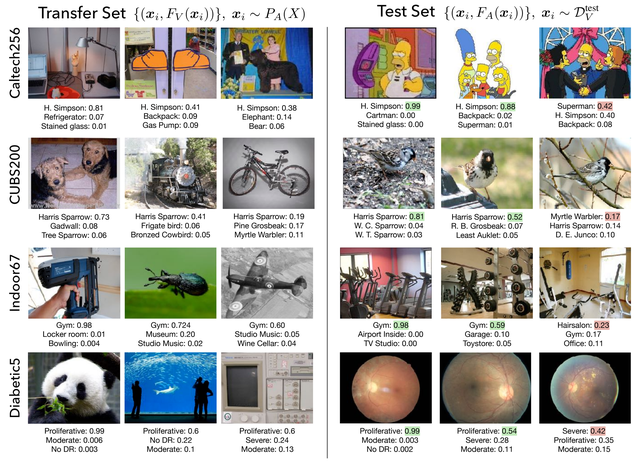
Experimental Results

- Four victim blackboxes (Resnet-34s): Caltech256, CUB-200, Indoor-Scene, Diabetic Retinopathy.
- Four choices of P_A: D^2, ImageNet, OpenImages, P_V (KD).

Key Results

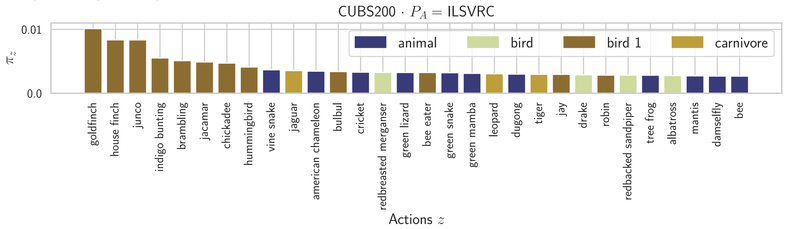


- Knockoffs are effective; recovers 0.81-0.97× victim blackbox's accuracy.
- Knockoffs also learn to classify images from unseen classes at test-time e.g., >90% classes in CUB-200.



- Adaptive improves sample-efficiency in closed-world settings (P_A = D^2). Difficult to significantly outperform random policy in open-world (ImageNet, OpenImages).
- Adaptive also improves performance of knockoffs (up to 4.5% test accuracy).

Inspecting Policy



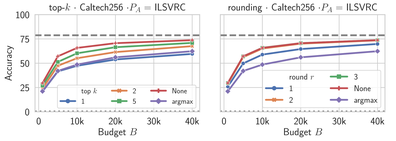
- Actions (arms of bandit) typically correspond to classes closely related to victim blackbox's output classes.

Acknowledgment This research was partially supported by the German Research Foundation (DFG CRC 1223)

Take-aways

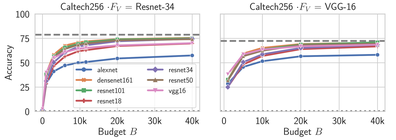
- Model functionality of complex blackbox CNNs can be stolen under minimal assumptions.
- Attacks are surprisingly effective and difficult to defend.
- Poses a threat to deployed ML APIs.

Victim's Defense: Information Truncation



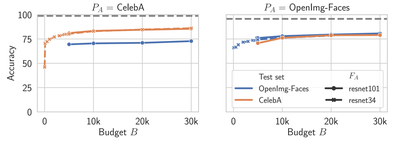
- Limited effect on knockoff attacks.

Vs. Choices of Architectures



- Robust to choices of F_V and F_A.

Real-world Experiment



- Knockoffs could be trained using $30 worth of queries.