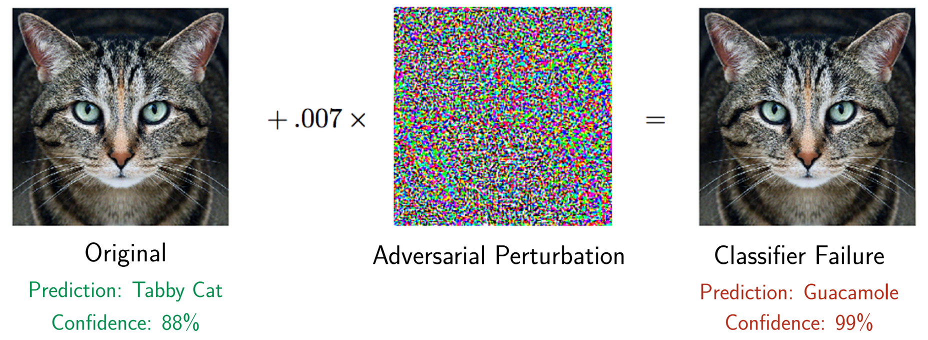
Black-Box Adversarial Attacks with Limited Queries and Infomation

Andrew Ilyas*, Logan Engstrom*, Anish Athalye*, Jessy Lin*

Massachussets Institute of Technology, LabSix

Black-box attacks

– Attacker goal: produce an *adversarial example* for a classifier



– Unrealistic for attacker to see weights or even model architecture

– Usually only valid operation is querying model f

– **Query-limited.** Limited number of queries to the model. Example. Attacking the Clarifai NSFW classifier API (which gives $P(\text{nsfw}|x)$) with 1,000,000 queries would cost $2,400.

– **Partial-information.** Access to $P(y|x)$ for k most likely labels. Example. The Google Cloud Vision API only outputs "confidence scores" for a dynamically determined number of classes.

– **Label-only.** Access to only k most likely labels, **no** probabilities. Example. Google Photos adds labels to images without probabilities.

Contributions

1. We systemize and define three black-box threat models that correspond to realistic black-box adversarial attack situations.
2. We demonstrate that targeted black-box attacks are feasible even under constraints of real-world threat models:
   – *When queries are limited,* with Natural Evolution Strategies
   – *When we only have access to the top k probabilities,* with a new alternating projections-based algorithm
   – *When only the top label(s) are given,* with a new surrogate loss-based algorithm
3. We perform a **targeted** black-box adversarial attack on Google Cloud Vision, a commercially deployed system.

Query-efficient attacks with NES

Goal: Construct black-box examples with limited # queries.

Idea: Use a more efficient unbiased gradient estimator

Finite differences approximate directional derivatives:

$$\lim_{h\to 0}\frac{f(\mathbf{x}+\epsilon\mathbf{u})-f(\mathbf{x})}{h} = D_{\mathbf{u}}f(x) = \langle\nabla_{\mathbf{x}}f(\mathbf{x}),\mathbf{u}\rangle$$

Previous work (Chen et. al, 2017) uses this to estimate $\nabla_x f$ **pixel-wise**: Accurately estimates $\nabla_x f$, but requires O(#pixels) queries of $f(\cdot)$

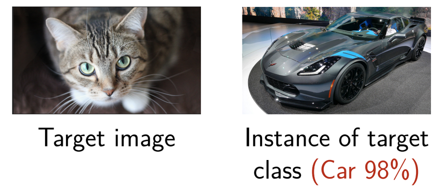
Natural Evolution Strategies (Wierstra et. al, 2014):

$$\nabla_x f \approx \mathbb{E}\left[f(\mathbf{x}+h\mathbf{u_i})\right] \quad \text{for } \mathbf{u_i} \text{ Gaussian vectors}$$

– Equivalent to Johnson-Lindenstrauss approximation (random projection) of $\nabla_x f(x)$

Partial-information attacks

Goal: Generate a targeted adversarial example with access to only the top-k classes and relative scores.



Target image

Instance of target class (Car 98%)

Idea: Start with instance of target class and alternate between (1) blending with target image and (2) maintaining the target class:

(1) project onto l_∞ boxes of decreasing sizes ϵ_t centered at the original image x_0, maintaining that the adversarial class remains within the top-k at all times

Auto show 72%
Car 71%
Wheel 55%

$$\epsilon_t = \min \epsilon' \text{ s.t. rank}\left(y_{adv}|\Pi_{\epsilon'}(x^{(t-1)})\right) < k$$

(2) perturb the image to maximize the probability of the adversarial target class

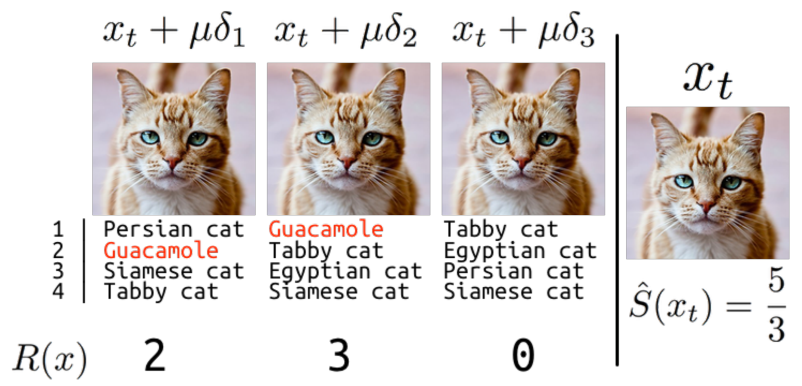
Car 91%
Auto show 75%
Wheel 61%

$$x^{(t)} = \arg\max_{x'} P(y_{adv}|\Pi_{\epsilon_{t-1}}(x'))$$

Label-only attacks

Goal: Generate a targeted adversarial example with access to only the ranking of the top-k classes.

Idea: Construct a proxy score with the rankings.



Discretized score $R(x^{(t)})$: quantify how adversarial an image is given the ranking of the target adversarial class y_{adv} in the top k classes.

$$R(x^{(t)}) = k - \text{rank}(y_{adv}|x^{(t)})$$

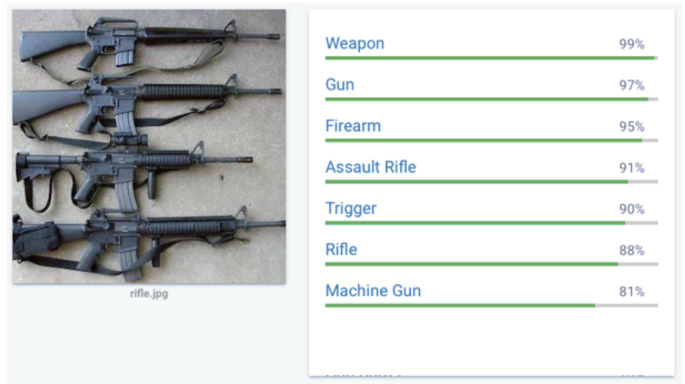
Proxy the softmax probability by considering the robustness of the adversarial image to random perturbations (uniformly chosen from a ℓ_∞ ball of radius μ), using the discretized score to quantify adversariality:

$$S(x^{(t)}) = \mathbb{E}_{\delta\sim\mathcal{U}[-\mu,\mu]}[R(x^{(t)}+\delta)]$$

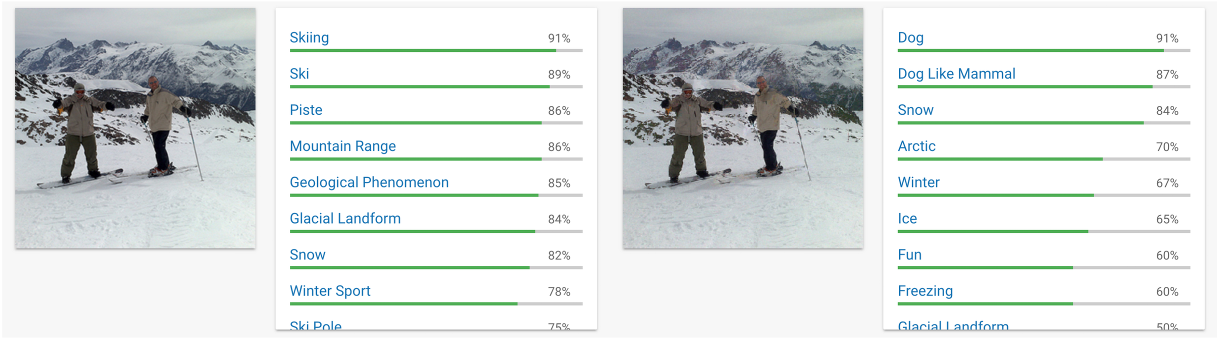
Google Cloud Vision (GCV)

Partial-information attack allows us to construct the first targeted adversarial example for GCV.

Large-scale, commercially deployed classifier with unkown number of classes and uninterpretable "confidence scores" instead of $P(y|x)$:



GCV Attack



Evaluation

Method:

1. Select 1000 source images $\{x_i\}$ from ImageNet validation set
2. Select 1000 random target classes to be used as "targets" $\{t_i\}$
3. Run query-efficient, partial-information, and label-only algorithms in respective threat models to get $\{x'_i\}$, a set of perturbed adversarial examples

We measure:

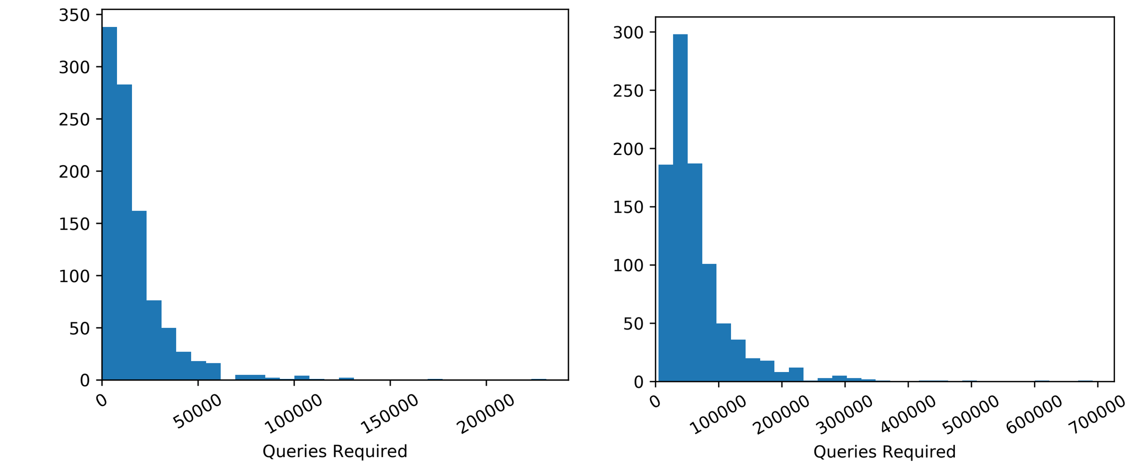
– **Success rate:** Fraction of the $\{x'_i\}$ classified as $\{t_i\}$

– **Query efficiency:** Median # queries required to construct successful examples

Threat model	Success rate	Median queries
Query-Limited	99.2%	11,550
Partial-information	93.6%	49,624
Label-only	90%	54,063

Attack is $> 100\times$ more efficient than prior work, with similar levels of distortion.

Generally, few queries are required (left: query-limited, right: partial-information).



2018 International Conference on Machine Learning, Stockholm, Sweden