Contract And Conquer: An Iterative Approach to Assess the Robustness of Black-Box Models

Mikhail Pautov

AIRI, ISP RAS

May 13, 2026

Presentation based on a joint work with Anna Chistyakova (ISP RAS)

Outline

1. Introduction & Motivation
2. Method: Contract And Conquer
3. Experimental Evaluation
4. Results
5. Conclusion & Future Work

Motivation: Why Certified Attacks?

The Problem

- Black-box adversarial attacks test model robustness but lack **success guarantees**
- Regulatory frameworks (EU AI Act, US AI Initiative) require robustness standards
- Empirical defenses $\leftrightarrow$ attacks create an "arms race" with no formal assurances

Research Question

How to guarantee that a given black-box model is **not robust***?*

Our Contribution

CAC: First transfer-based black-box attack with **provable convergence guarantees**

Problem Formulation

Definition (Hard-label target model)

Let $T : \mathcal{X} \subset R^d \to [1, \ldots, K]$ be the hard-label neural network that maps input object to class indices.

Definition (Soft-label surrogate model)

Let $S : \mathcal{X} \subset R^d \to \Delta^{K-1}$ be the hard-label neural network that maps input object to probability vectors.

Here, d is the dimension of the input space, K is the number of classes,

$$\Delta^{K-1} = \left\{ (p_1, p_2, \ldots, p_K) : \sum_j p_j = 1 \ \wedge \ p_j \geq 0 \right\}.$$

Problem Formulation

Definition (Adversarial Example)

Let x be correctly classified by black-box model T, $y = T(x)$, and $\delta > 0$.
Then x' with $\|x - x'\|_\infty \leq \delta$ is an **adversarial example** if $T(x') \neq T(x)$.

Definition (Transferability)

An adversarial example x' computed for surrogate S is **transferable** to T if:

$$\begin{cases} \arg\max_i S(x)_i = T(x) \\ \arg\max_i S(x')_i = T(x') \end{cases}$$

Goal

Compute adversarial examples for black-box model T with **mathematical guarantee** of success within a fixed number of iterations.

Overview

Key Idea: On iteration j, alternate between:

1. **Knowledge Distillation**: Train surrogate model S_j on expanding dataset $\mathcal{D}(S)$. The goal is to mimic the predictions of target model T on a finite set of points $\mathcal{D}(S) = \{x_i, T(x_i)\}_{i=1}^{m}$.

Overview

Key Idea: On iteration j, alternate between:

1. **Knowledge Distillation**: Train surrogate model S_j on expanding dataset $\mathcal{D}(S)$. The goal is to mimic the predictions of target model T on a finite set of points $\mathcal{D}(S) = \{x_i, T(x_i)\}_{i=1}^{m}$.
2. **White-box Attack**: Find an adversarial example z_j for S_j using MI-FGSM. The goal is to find a single adversarial example within current attack search space, $U_\delta(x)_j$.

Overview

Key Idea: On iteration j, alternate between:

1. **Knowledge Distillation**: Train surrogate model S_j on expanding dataset $\mathcal{D}(S)$. The goal is to mimic the predictions of target model T on a finite set of points $\mathcal{D}(S) = \{x_i, T(x_i)\}_{i=1}^{m}$.
2. **White-box Attack**: Find adversarial example z_j for S_j using MI-FGSM. The goal is to find a single adversarial example within current attack search space, $U_\delta(x)_j$.
3. **Contraction**: If the point z_j is not transferable to T, add it to $\mathcal{D}(S)$ and contract the search space:

$$\mathcal{U}_\delta(x)_j \leftarrow \mathcal{U}_\delta(x) \cap \mathcal{U}_{\rho_j}(z_j), \quad \text{where} \quad \rho_j = t\|z_j - z_{j-1}\|_\infty \quad \text{and} \quad t \in (0, 1).$$

Overview

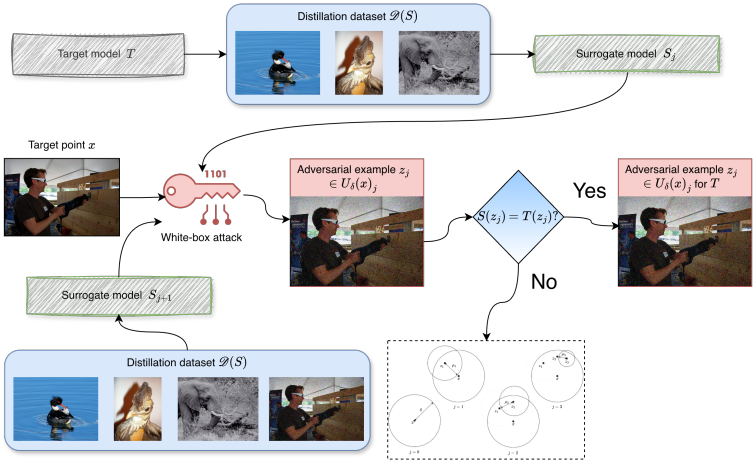

Figure: Schematic: Alternating distillation ↔ white-box attack ↔ search space contraction

Algorithm: Contract and Conquer

Require: Target model T, point x, radius δ, momentum μ, max MI-FGSM iters M, max queries N
Ensure: Adversarial example $(z, T(z))$ for T

$\mathcal{D}(S) \leftarrow \{(x_k, T(x_k))\}_{k=1}^{m-1} \cup \{(x, y)\}$ {Init distillation dataset}
$z_0 \leftarrow x$, $\mathcal{U}_\delta(x)_0 \leftarrow \mathcal{U}_\delta(x)$, $\alpha \leftarrow \delta / M$
$j \leftarrow 1$
while $N \geq 0$ **do**
    Train S on $\mathcal{D}(S)$
    $z_j \leftarrow \text{MI-FGSM}(S, \alpha, \mu, \mathcal{U}_\delta(x)_{j-1}, M, (x, y))$
    if $\arg\max_i S(z_j)_i = T(z_j) \neq y$ **then**

        return $S, (z_j, T(z_j))$ {Transferable!}
    else
        $\mathcal{D}(S) \leftarrow \mathcal{D}(S) \cup \{(z_j, T(z_j))\}$
        $\rho_j \leftarrow t\|z_j - z_{j-1}\|_\infty$
        $\mathcal{U}_\delta(x)_j \leftarrow \mathcal{U}_\delta(x) \cap \mathcal{U}_{\rho_j}(z_j)$ {Contract space}
        $\alpha \leftarrow \rho_j / M$
    end if
    $N \leftarrow N - 1$, $j \leftarrow j + 1$
end while

Theoretical Guarantee

Theorem (Convergence Guarantee)

Under assumptions for all j:

1. *Adversarial examples exist in contracted spaces for both S_j and T*
2. *Gradient of $h_j = S_j - T$ is bounded:* $\boxed{\gamma = \sup \|\nabla h_j\|_{op,\infty}}$
3. *At least one "good" distillation point exists for sufficiently large $k^* < j$:* $\boxed{\|h_j(z_{k^*})\|_\infty \leq E/2}$
4. *White-box attacks are successful:* $\boxed{S_j(z_j)_c - \max_{i \neq c} S_j(z_j)_i \geq 2E}$

CAC yields a transferable adversarial example within at most:

$$\tilde{k} + 1 \quad \textit{iterations, where} \quad \tilde{k} = \left\lceil \ln\left(\frac{E(1-t)}{2\gamma\delta}\right)(\ln t)^{-1} \right\rceil$$

Key Insight

Search space contraction ($t \in (0,1)$) ensures geometric convergence; distillation ensures surrogate fidelity.

Experimental Setup

Datasets & Models

- **Datasets**: CIFAR-10, ImageNet
- **Target models**:
  - ResNet-50 (ImageNet: 80.13%, CIFAR: 94.65%)
  - ViT-B (ImageNet: 85.21%, CIFAR: 96.89%)
- **Surrogate**: ResNet-18

Hyperparameters

- Distillation: up to 100 epochs, AdamW optimizer, lr=10^{-3}
- MI-FGSM: $M = 3$, $\mu = 1.0$, $\alpha = \delta/M$
- Contraction: $t = 0.99$, $\delta = 0.250$
- Queries: $N = 500$ (ImageNet), $N = 300$ (CIFAR-10)

Baselines: HopSkipJump, Sign-OPT, GeoDA, SquareAttack, SparseRS, PAR, AdvViT

Evaluation Metrics

ASR	Attack Success Rate: fraction of targets input points with successful adversarial example
AQN	Average Query Number: mean queries to target model per successful attack
AVG l_p	Mean perturbation norm (l_2 or l_∞) between x and x'
STD l_p	Standard deviation of perturbation norms
TIME	Average runtime (seconds) per attack

Comparison Protocol

- Fixed query budget and initial search space $\mathcal{U}_\delta(x)$ for all methods
- 100 correctly-classified target points sampled from test set
- Report mean $\pm$ std over successful attacks

Hard-Label Setting: ImageNet (ResNet-50)

Method	ASR	AQN	AVG l_2	STD l_2	AVG l_∞	TIME
CAC (Ours)	**1.00**	487.95	**35.07**	18.83	**0.153**	17.30
HopSkipJump l_2	1.00	500.31	48.84	29.12	0.539	6.28
HopSkipJump l_∞	1.00	500.01	73.26	35.86	0.361	4.58
Sign-OPT	1.00	548.24	48.05	28.47	0.551	12.70
GeoDA	1.00	524.98	49.66	31.12	0.180	20.10

Table: Hard-label attack on ResNet-50 (ImageNet). **Bold**: best perturbation closeness.

Key Observation

CAC achieves **100% success rate** with **smallest perturbations** (l_∞), at cost of higher query count.

Hard-Label Setting: ImageNet (ViT-B)

Method	ASR	AQN	AVG l_2	STD l_2	AVG l_∞	TIME
CAC (Ours)	**1.00**	488.91	49.28	26.49	**0.165**	17.85
HopSkipJump l_2	1.00	500.34	70.12	38.34	0.685	7.01
Sign-OPT	1.00	557.31	74.74	44.85	0.708	18.07
PAR	1.00	**322.38**	38.75	25.75	0.889	6.91
AdvViT	0.75	461.04	**34.52**	20.26	0.584	19.29

Table: Hard-label attack on ViT-B (ImageNet). CAC maintains guarantee for transformer architectures.

Takeaway

Even against transformer targets, CAC provides **provable success** with competitive perturbation magnitude.

Soft-Label Setting: ImageNet Results

ResNet-50 Target

Method	ASR	AVG l_∞
CAC	**1.00**	**0.122**
SquareAttack	0.98	0.250
SparseRS	0.94	0.994

ViT-B Target

Method	ASR	AVG l_∞
CAC	**1.00**	**0.144**
SquareAttack	0.26	0.250
SparseRS	0.79	0.993

Critical Result

In soft-label setting against ViT-B, CAC is the **only method** achieving 100% ASR with minimal perturbations.

CIFAR-10 Results Summary

Hard-Label (ResNet-50)

- CAC: ASR=1.00, AVG l_∞=**0.061**
- Best baseline: AVG l_∞=0.071 (GeoDA)
- Query efficiency is comparable

Soft-Label (ViT-B)

- CAC: ASR=1.00, AVG l_∞=**0.050**
- SquareAttack: ASR=0.85, AVG l_∞=0.250
- SparseRS: ASR=0.98, AVG l_∞=0.974

Consistent Pattern

Across all settings:

1. **100% Attack Success Rate** (guaranteed by theory)
2. **Smallest perturbation norm** (due to search space contraction)
3. Moderate query overhead

Summary of Contributions

1. **Simplicity**: CAC alternates knowledge distillation, white-box attack, and search space contraction
2. **Theoretical Guarantee**: Provable convergence to transferable adversarial example within the fixed number of iterations
3. **Empirical Efficiency**:
   - 100% ASR across all benchmarks (ResNet-50, ViT-B; CIFAR-10, ImageNet)
   - Smallest l_∞ perturbations among competing methods
   - Effective for both hard-label and soft-label black-box settings

Practical Impact

CAC enables **certified robustness evaluation** for black-box models—critical for AI regulation compliance.

Limitations & Future Work

Limitations

- Assumes bounded gradients, existence of adversarial examples, and success of distillation
- Distillation overhead (fine-tuning a surrogate model is required)

Future Directions

- Reduce computational complexity via efficient distillation
- Obtaining tighter guarantees
- Develop framework for **regulatory compliance testing**

Thank You!

Questions?

Paper: *Contract And Conquer: How to Provably Compute Adversarial Examples for a Black-Box Model?*

[@mikepautov]