Policy-Driven Attack: Learning to Query for Hard-label Black-box Adversarial Examples

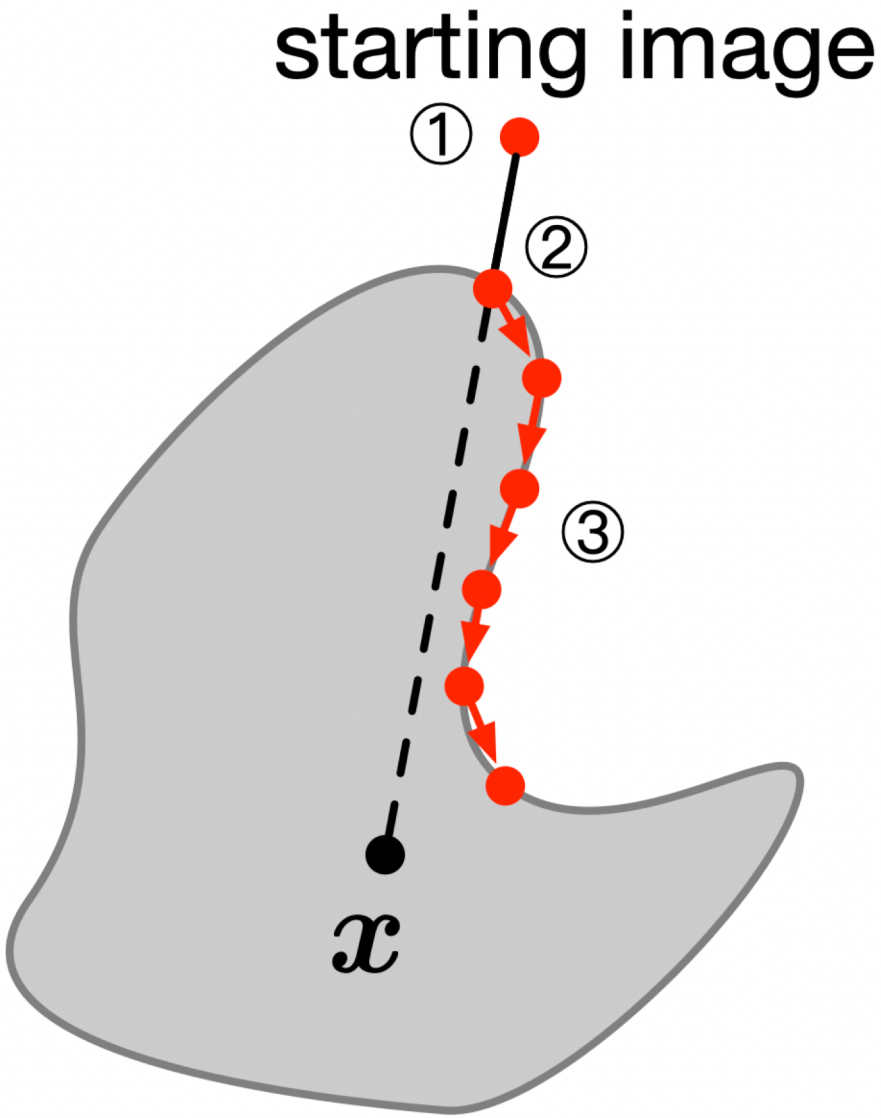
(ICLR 2021)

Ziang Yan[1,2*], Yiwen Guo[2*], Jian Liang[3], and Changshui Zhang[1]
[1]Tsinghua University
[2]Bytedance AI Lab
[3]Alibaba Group

Motivation

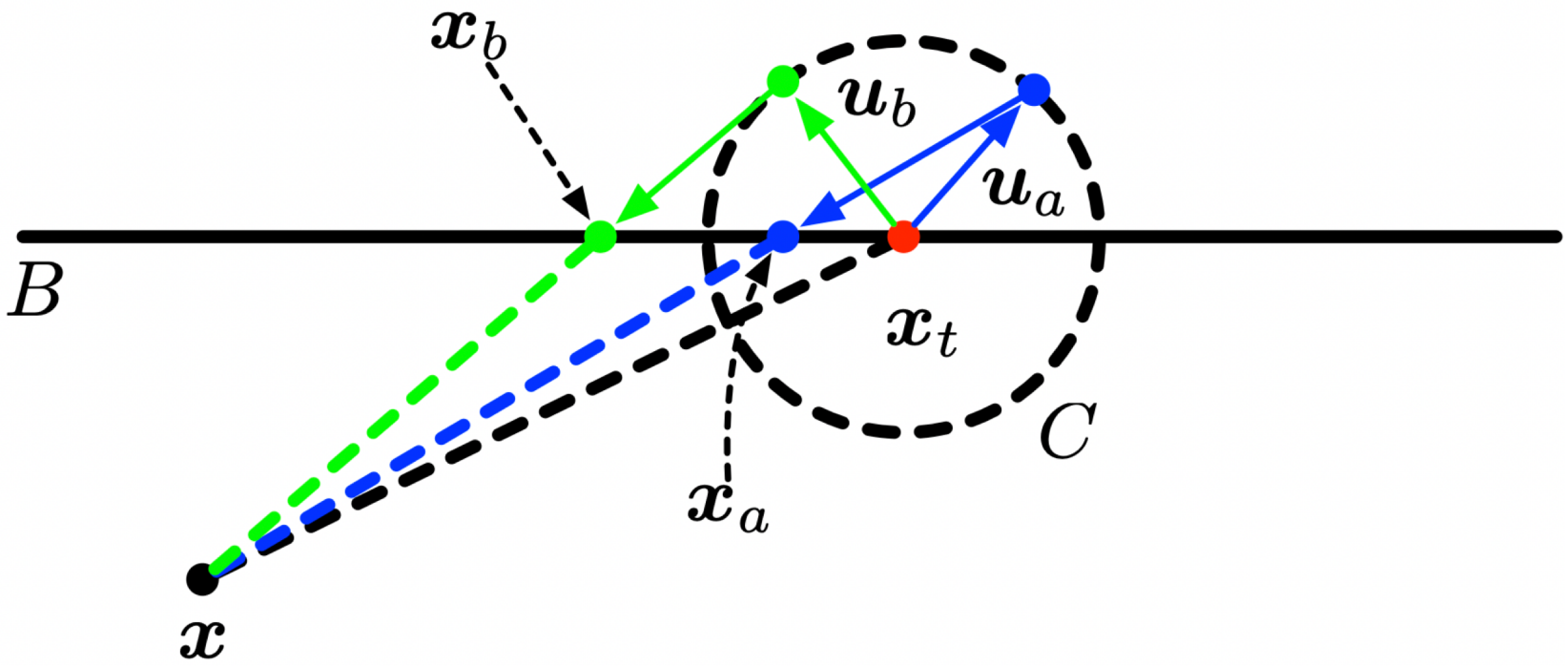
- A common hard-label attack framework:

    1. Find an initialization adversarial image
    2. Project it into the decision boundary
    3. Update it along the boundary to reduce the distortion from the clean image



Motivation

- Update step: jump by distance δ, then project into the boundary
- $\|\boldsymbol{x}_b - x\|_2 < \|\boldsymbol{x}_a - x\|_2 \Longrightarrow$ the update direction matters

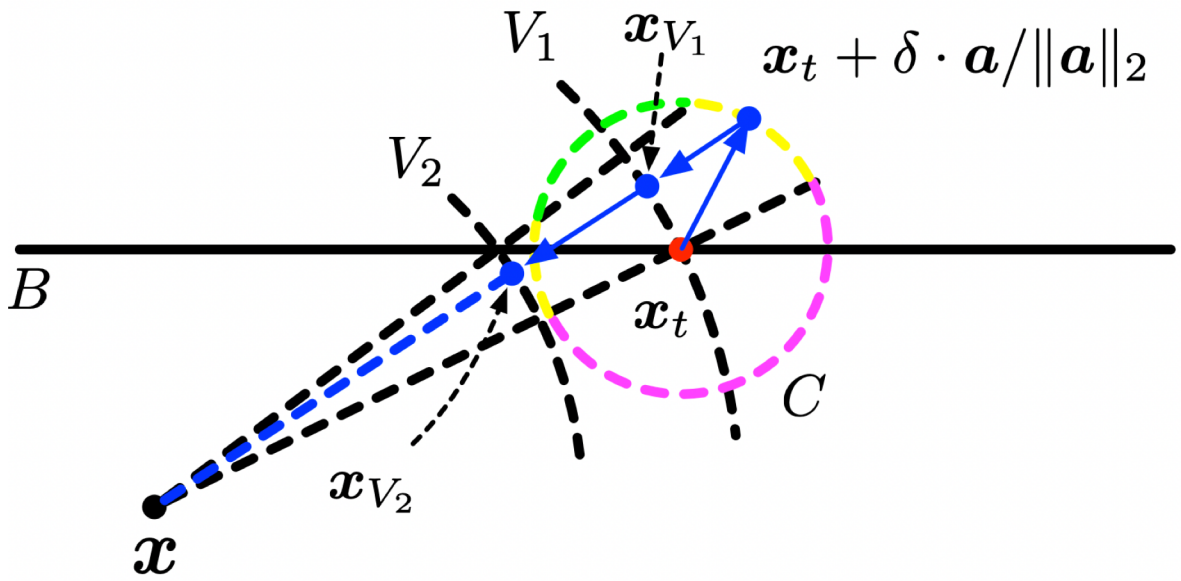


Question: Can we learn the quality of an update direction?

Policy-Driven Attack

- Cast the hard-label attack problem into a reinforcement learning task:
  - Environment: the victim model
  - State: current adversarial image
  - Agent: the attacker
  - Action: search direction
  - Reward: a scalar which assesses the quality of a given search direction

Policy-Driven Attack



- Query-efficient reward assignment mechanism:

1. Determine distortion baseline level V_1 and V_2
2. Jump along the direction by distance δ
3. Project resulted image into $V_1$and V_2 then check if they are still adversarial
4. If both, reward is 2; if only x_{V_1}, reward is 1; if neither, reward is 0

Policy-Driven Attack

- Policy network:
  - Input: current adversarial image, label, target label; output: direction
  - U-Nets and Auto-Encoders do not provide the best performance in this task
  - Our design: Use the gradient (C&W loss w.r.t input) of an internal classifier:

$$g(\boldsymbol{x}'_t, y, y') = \nabla_{\boldsymbol{x}'_t} h(\boldsymbol{x}'_t)_{y'} - \nabla_{\boldsymbol{x}'_t} h(\boldsymbol{x}'_t)_y + \boldsymbol{b}$$

Policy-Driven Attack

- Overall view:
  - (Optional) Pre-train the policy network using previous attack trajectories
  - When attacking an image, in each iteration:
    - The agent submits predicted directions as actions to the environment
    - The environment assigns a reward to each action, and update the adversarial image
    - Fine-tune the policy network using one-step REINFORCE

Empirical Results

Table 1: Mean ℓ_2 distortions for performing untargeted attacks with different query budgets.

Dataset	Victim Model	Method	@100	@500	@1K	@5K	10K	@25K
MNIST	CNN	Boundary Attack	10.179	9.970	9.692	8.960	3.670	1.850
		HopSkipJumpAttack	7.267	3.576	2.709	1.843	1.721	1.641
		Ours	**2.204**	**2.053**	**1.994**	**1.674**	**1.643**	**1.639**
CIFAR-10	CNN	Boundary Attack	9.625	8.649	8.463	5.442	1.408	0.395
		HopSkipJumpAttack	4.567	1.520	0.911	0.378	0.304	**0.261**
		Ours	**0.640**	**0.546**	**0.507**	**0.358**	**0.298**	0.263
	WRN	Boundary Attack	9.268	8.095	7.972	3.639	0.664	0.287
		HopSkipJumpAttack	3.041	1.052	0.649	**0.268**	**0.213**	0.179
		Ours	**0.700**	**0.515**	**0.453**	0.273	0.217	**0.174**
	ResNet50 Adv.	Boundary Attack	10.672	10.273	10.217	9.064	4.998	2.378
		HopSkipJumpAttack	7.980	5.453	4.353	2.521	2.097	1.793
		Ours	**2.638**	**2.454**	**2.365**	**1.926**	**1.733**	**1.570**
ImageNet	ResNet-18	Boundary Attack	65.810	60.861	60.618	39.359	13.887	4.877
		HopSkipJumpAttack	36.594	20.881	14.295	4.846	**2.883**	**1.479**
		Ours	**11.998**	**8.200**	**6.751**	**4.093**	2.921	1.481

Empirical Results

Table 2: Mean ℓ_2 distortions for performing targeted attacks with different query budgets.

Dataset	Victim Model	Method	@100	@500	@1K	@5K	10K	@25K
MNIST	CNN	Boundary Attack	5.559	5.486	5.480	5.417	4.087	2.638
		HopSkipJumpAttack	5.291	4.480	3.786	2.644	2.423	**2.268**
		Ours	**3.072**	**2.772**	**2.712**	**2.423**	**2.355**	2.314
CIFAR-10	CNN	Boundary Attack	11.194	10.075	9.381	7.933	2.778	0.750
		HopSkipJumpAttack	8.864	3.763	2.118	0.724	**0.549**	**0.448**
		Ours	**1.426**	**1.207**	**0.930**	**0.691**	0.589	0.481
	WRN	Boundary Attack	10.021	8.269	7.903	5.368	1.223	0.467
		HopSkipJumpAttack	7.770	3.191	1.573	**0.398**	**0.295**	**0.240**
		Ours	**2.759**	**1.518**	**1.203**	0.680	0.494	0.321
	ResNet50 Adv.	Boundary Attack	11.268	11.086	11.002	10.677	7.386	3.702
		HopSkipJumpAttack	10.316	8.592	7.021	3.766	3.030	2.524
		Ours	**4.327**	**3.843**	**3.704**	**3.119**	**2.789**	**2.419**

Empirical Results

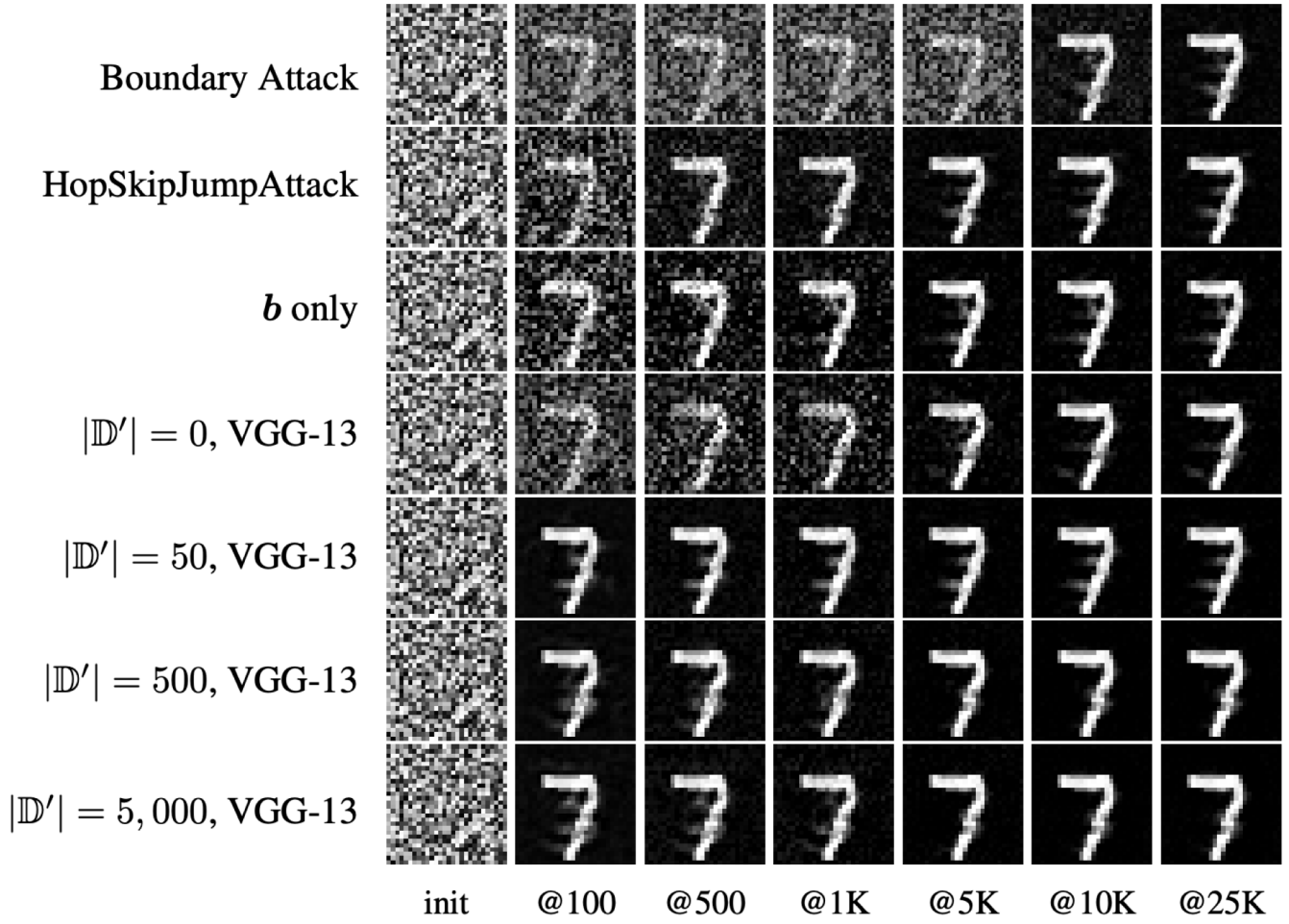


Figure 3: Visualization of generated adversarial images using different pre-training configurations. All shown images are classified as "3" by the victim model.

Empirical Results

Table 7: Transferability of pre-trained policy networks on CIFAR-10. Last six columns are median ℓ_2 distortions for performing untargeted attacks with different query budgets.

Victim Model for Evaluation	Victim Model for Pre-Training	@100	@500	@1K	@5K	10K	@25K
CNN	CNN	**0.640**	**0.546**	**0.507**	**0.358**	**0.298**	0.263
	WRN	2.719	1.683	1.322	0.565	0.399	0.278
	ResNet50 Adv.	1.405	1.087	0.895	0.422	0.315	**0.249**
WRN	CNN	2.060	1.299	1.035	0.471	0.325	0.216
	WRN	**0.700**	**0.515**	**0.453**	**0.273**	**0.217**	**0.174**
	ResNet50 Adv.	2.028	1.474	1.173	0.482	0.309	0.197
ResNet50 Adv.	CNN	6.324	5.257	4.790	3.146	2.449	1.827
	WRN	6.737	5.697	5.203	3.524	2.807	2.020
	ResNet50 Adv.	**2.638**	**2.454**	**2.365**	**1.926**	**1.733**	**1.570**

Thanks