Query-Efficient Black -Box Adversarial Attack With Customized Iteration and Sampling

Yucheng Shi, Yahong Han, Qinghua Hu, Yi Yang, Qi Tian

Presented By: Mehul Sen

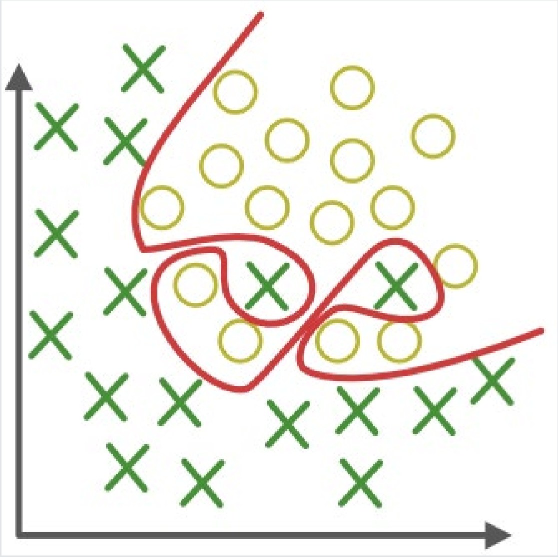
Introduction

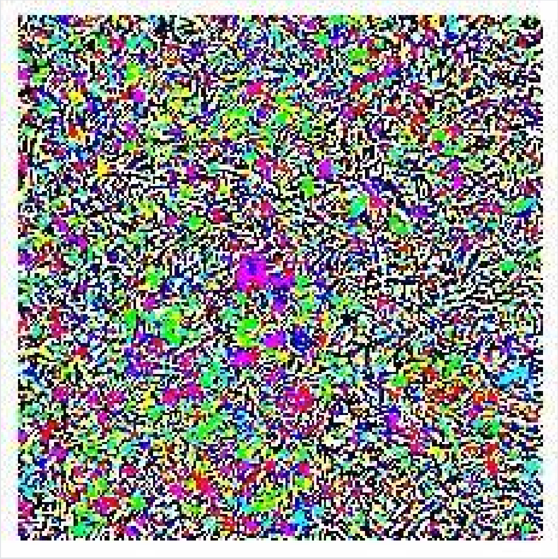
Adversarial Attacks generate input designed to fool a model into misclassifying the input.

- **White-box Attack**
- **Grey-box Attack**
- **Black-box Attack**
  - Transfer-based Attacks
  - Decision-based Attacks

Problems with Transfer-based Attacks

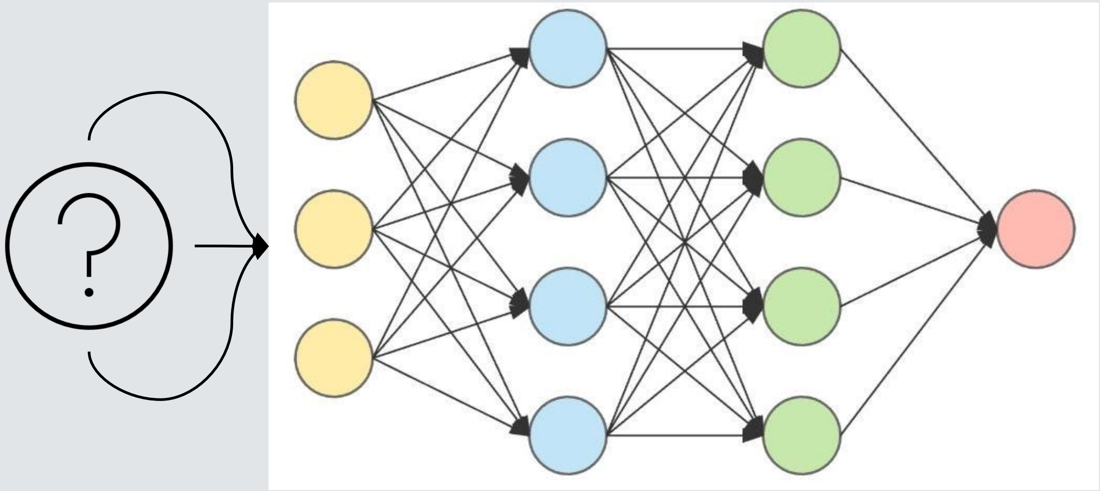
- Overfitting
- Noise Compression

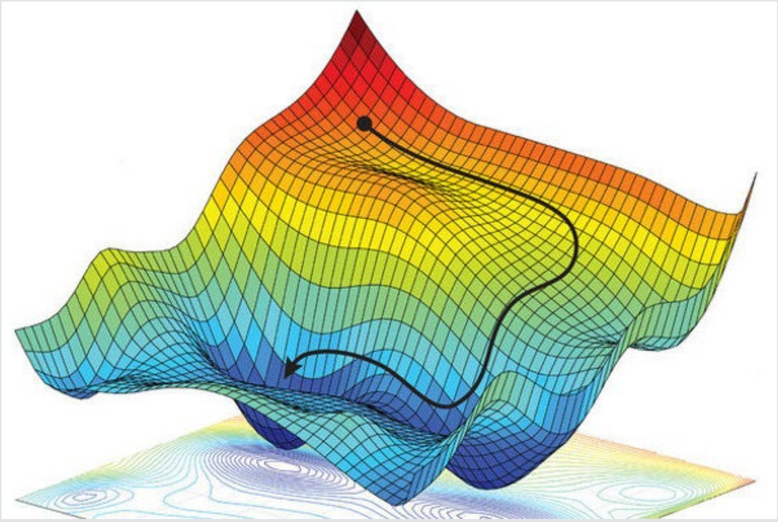




Problems with Decision-based Attacks

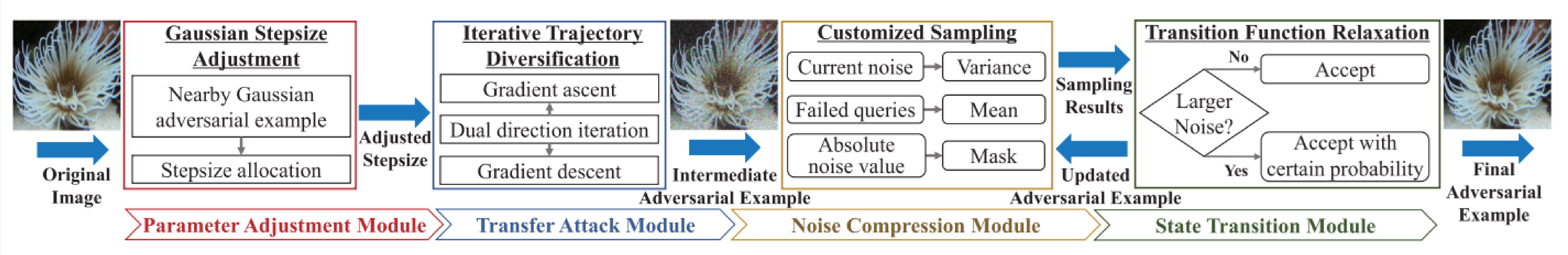
- Query Efficiency
- Local Optimum





Solution

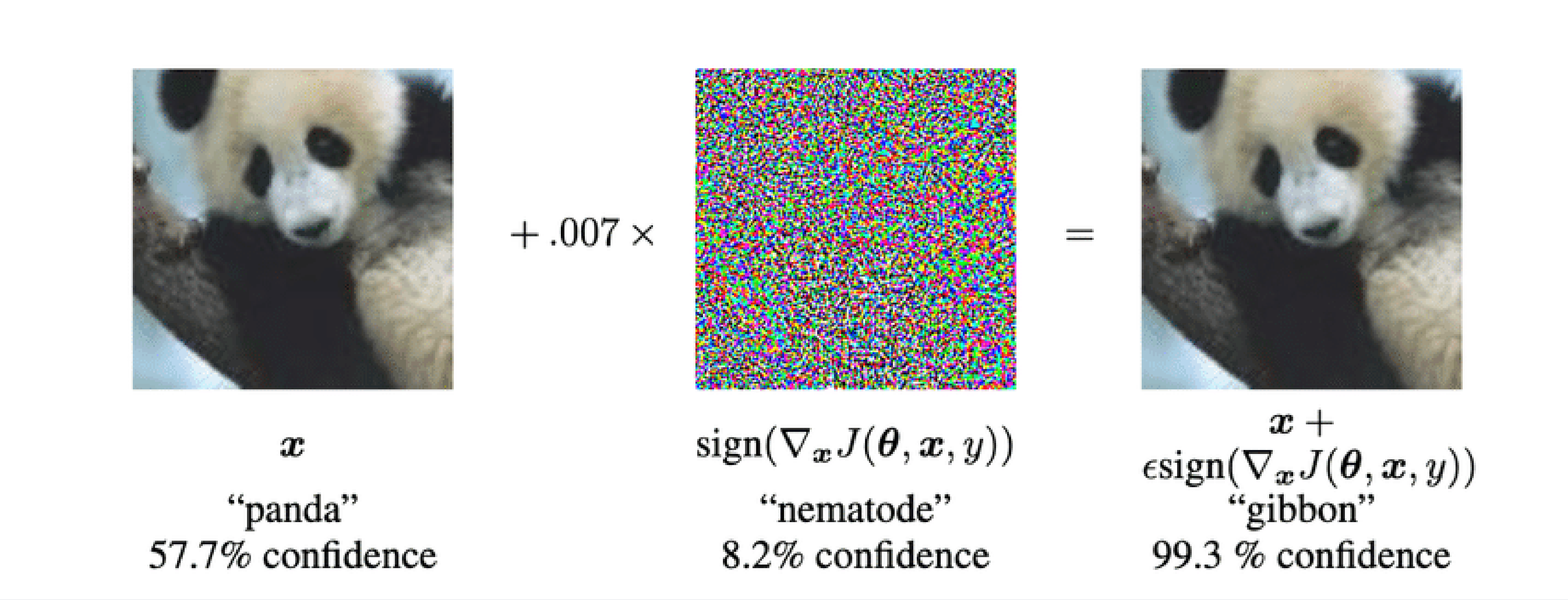
- Black-box Adversarial Attack Framework
- Customized Iteration and Sampling Attack (CISA)



Related Work

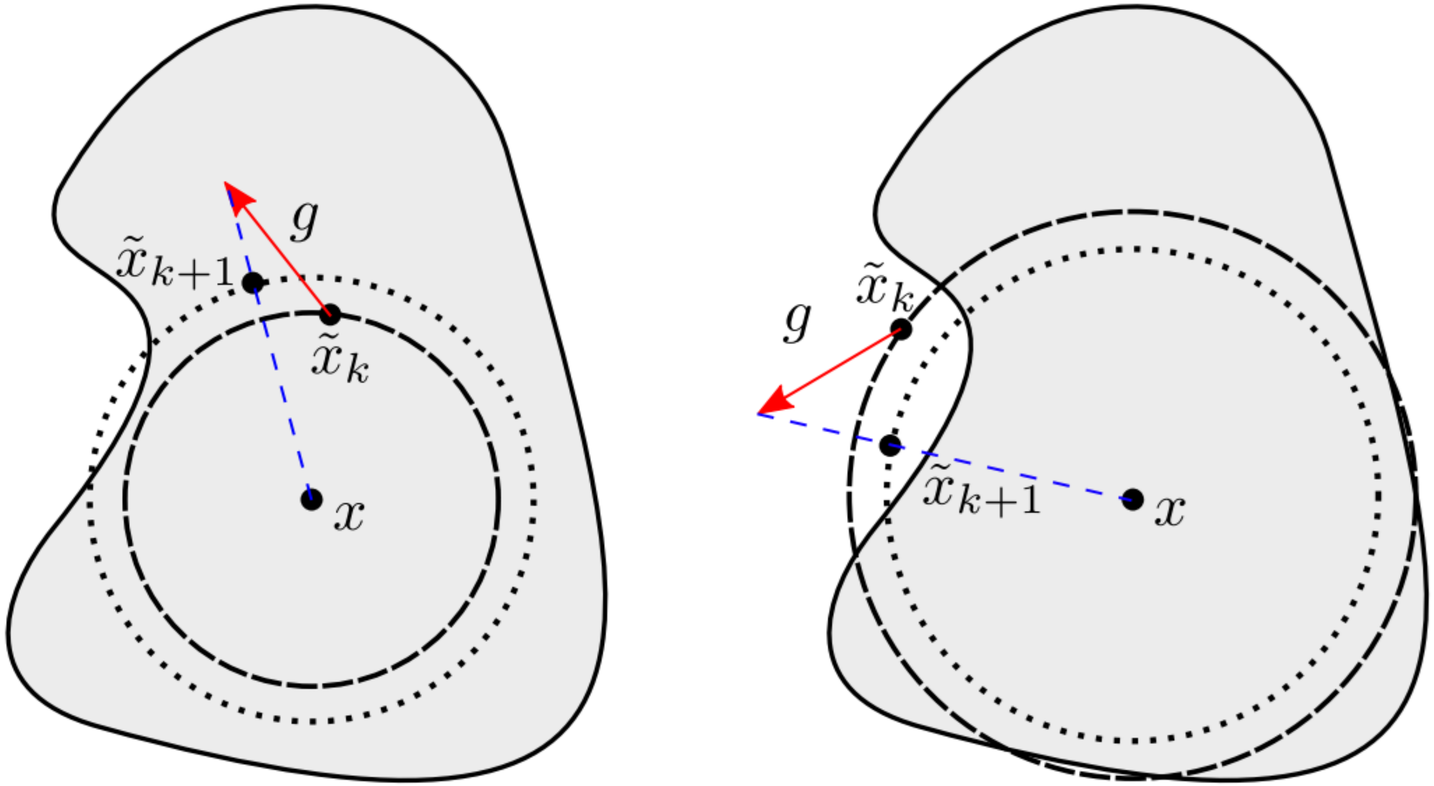
Transfer-based Attacks (TRA)
Decision-based Attacks (DEA)
Combination Attacks (TRA + DEA)

TRA: FGSM, I-FGSM, MI-FGSM, Vr-IGSM



I. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples”, 2015

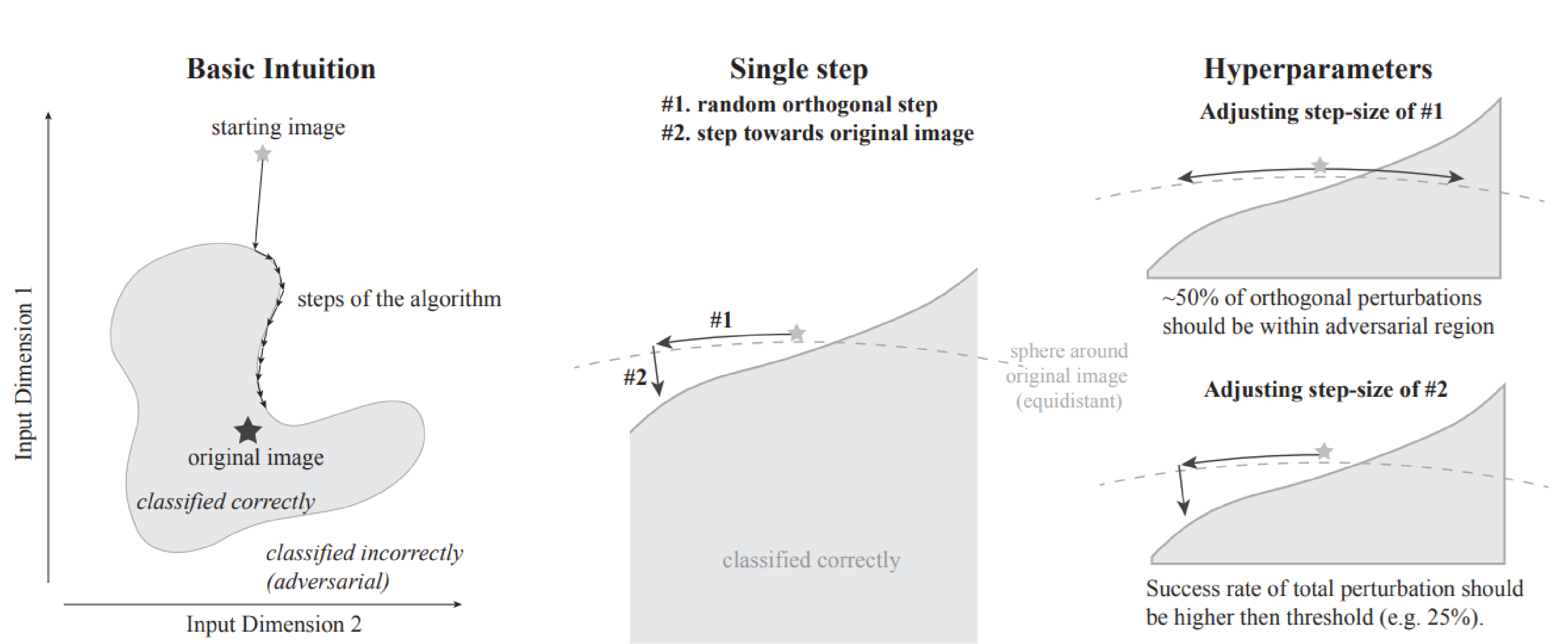
TRA: DDN, C&W, EAD



(a) $\tilde{x}_k$ not adversarial (b) $\tilde{x}_k$ adversarial

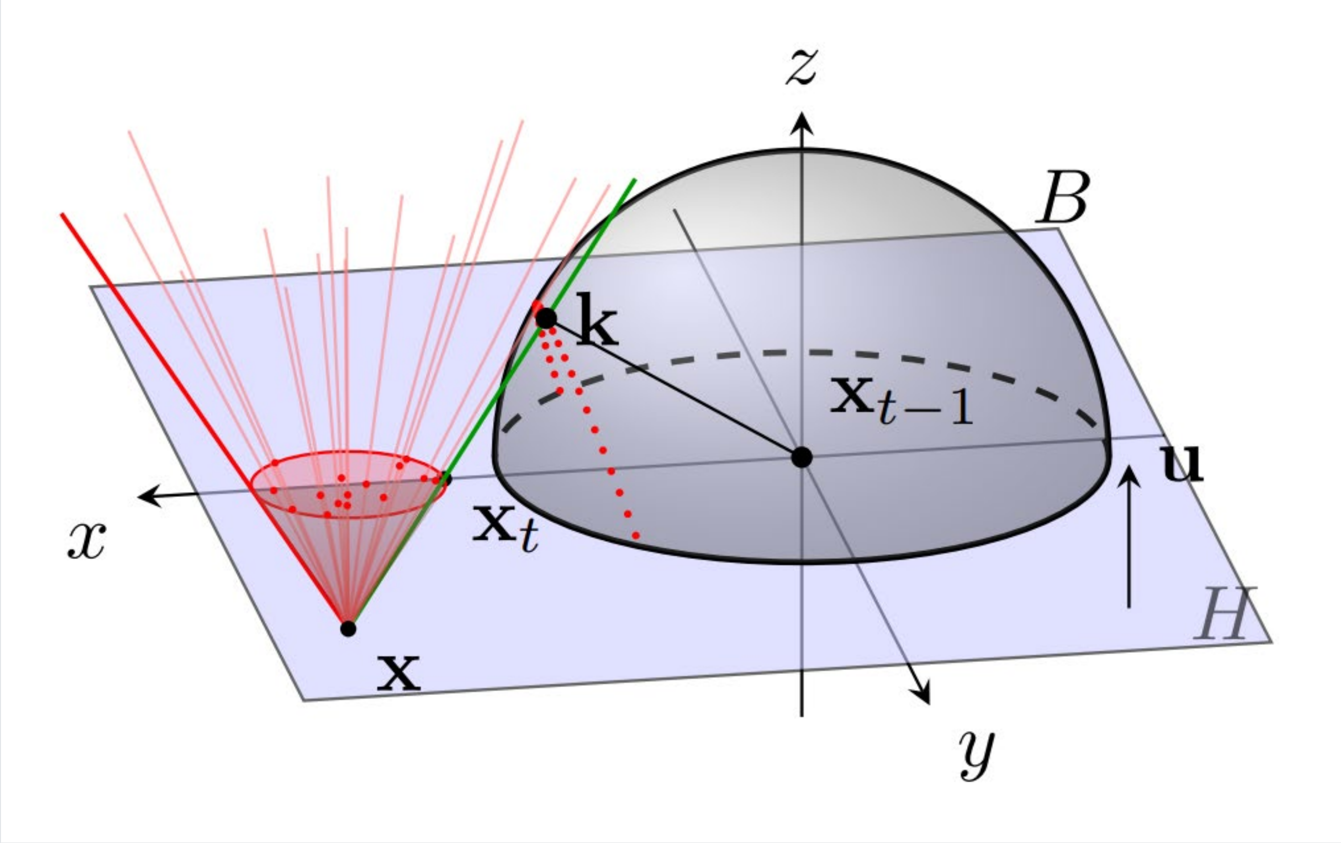
Rony et al. “Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses”, 2019

DEA: Boundary, QEBA, HSJA, Whey



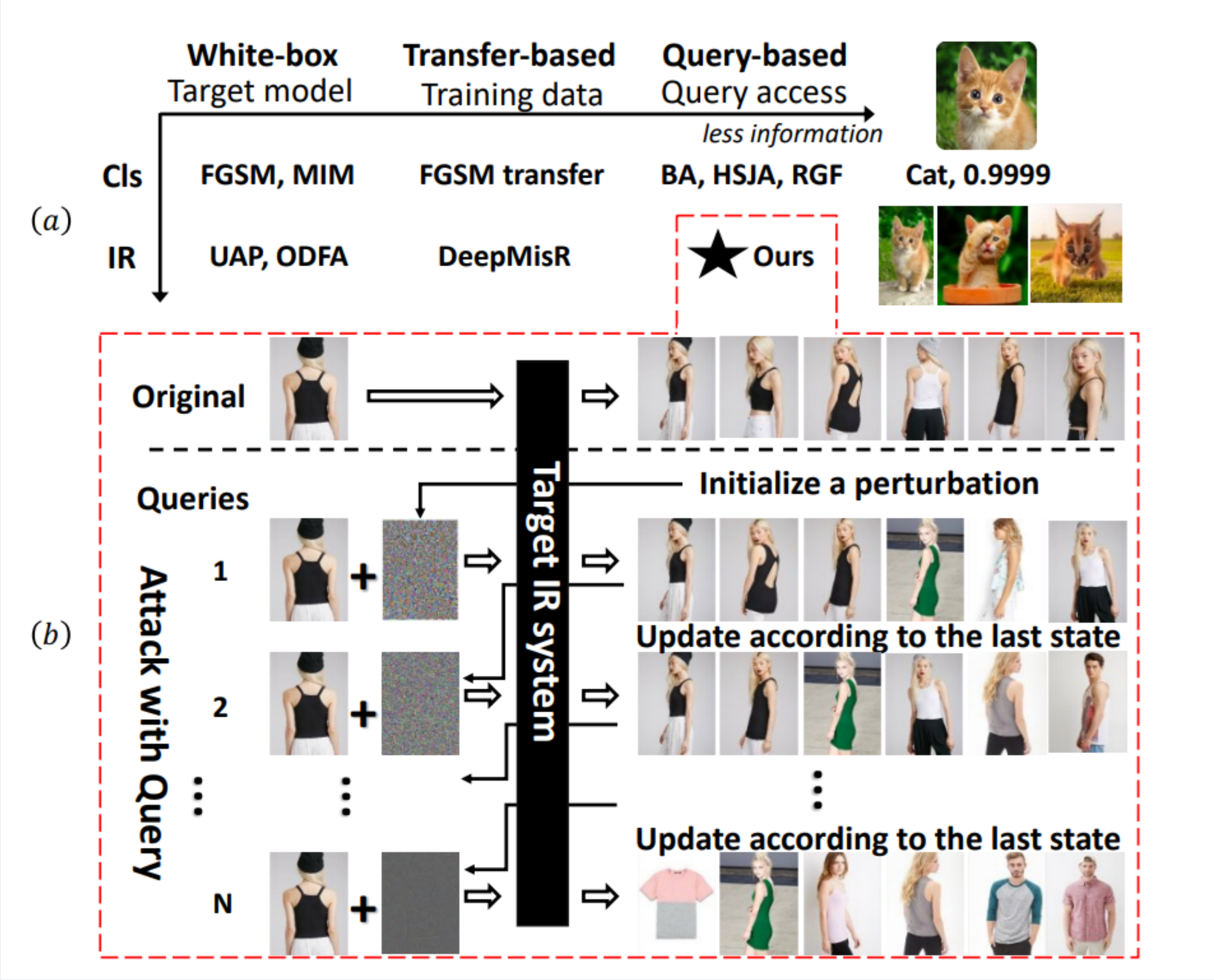
J. R. Wieland Brendel and M. Bethge, “Decision-based adversarial attacks: Reliable attacks against black-box machine learning models”, 2018

DEA: Tangent, Sign-OPT, EVO

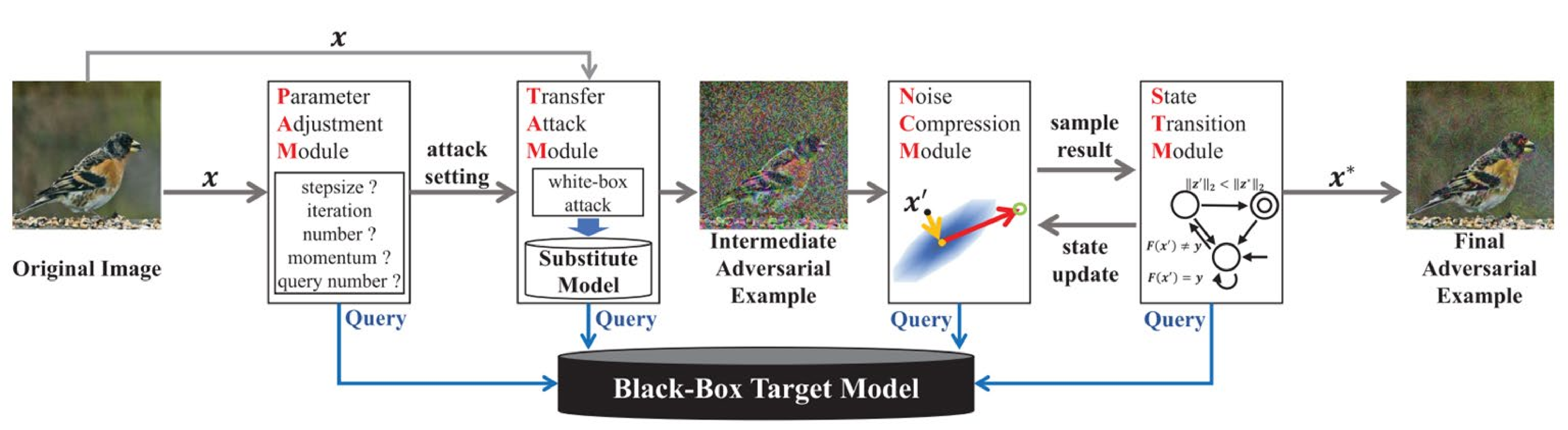


Ma et al. “Finding optimal tangent points for reducing distortions of hard-label”, 2021

TRA + DEA: BBA, LeBA, QAIR



Black-box Adversarial Framework



Parameter Adjustment Module

Uses a combination of estimation and query to adapt the attack parameters for each image exploiting the feedback from the target model

Fixed Stepsize

- Less Queries
- More Noise

Stepsize(μ)

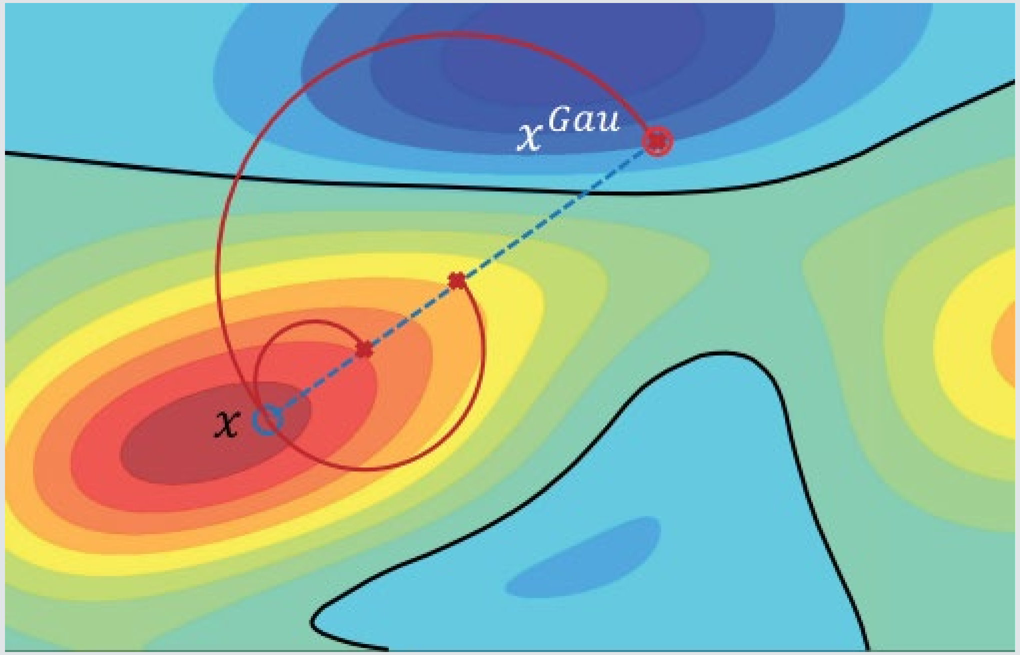
Binary Search

- Less Noise
- More Queries

Parameter Adjustment Module

Gaussian Stepsize Adjustment

- Add Gaussian Noise to Image
- If successful:
    - stop search
- If unsuccessful:
    - double the variance



Transfer Attack Module

Generates intermediate adversarial examples based on the substitute model

Gradient-based
Gradient ascent to maximize the loss function

Differences in classification space

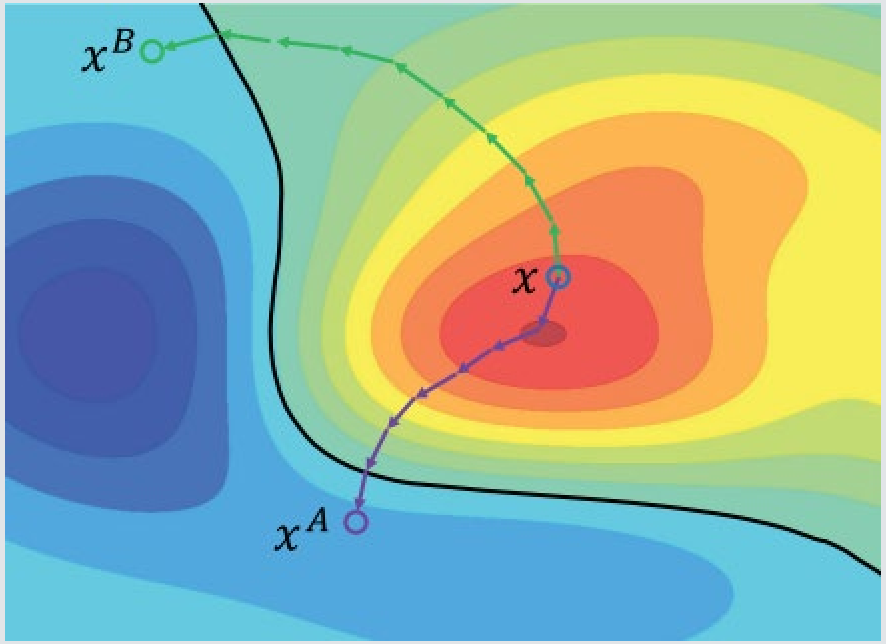
Don't adapt to limited query black-box attacks

Optimization-based
Optimization techniques to maximize the loss function

Transfer Attack Module

Iterative Trajectory Diversification

- Update along gradient descent until loss is declining
- If loss < previous step: Switch to gradient ascent and continue until the last step



Noise Compression Module

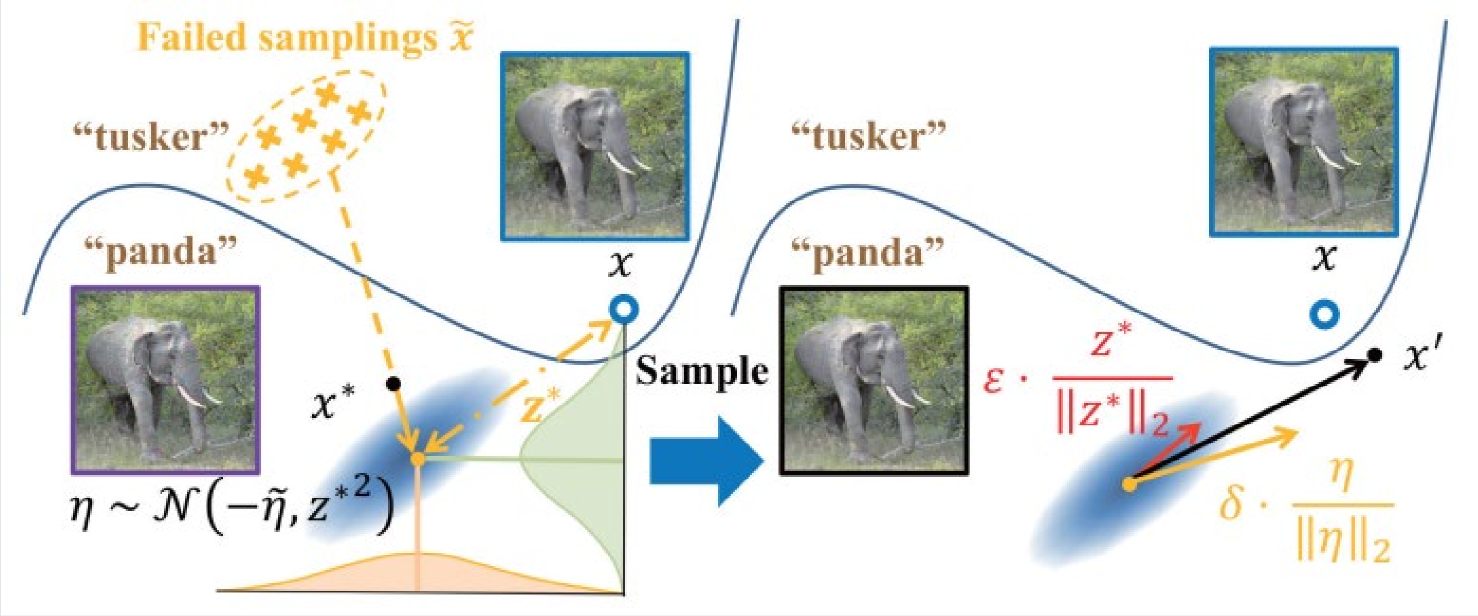
Uses queries to search for adversarial examples with smaller noise magnitude by sampling in the neighborhood

- Final noise is positively correlated with initial noise
- Noise compression reduces the probability of misclassification

Noise Compression Module

Customized Sampling

- Adaptive Variance Assignment
- Utilizing Failed Samplings
- Exponential Scheduling

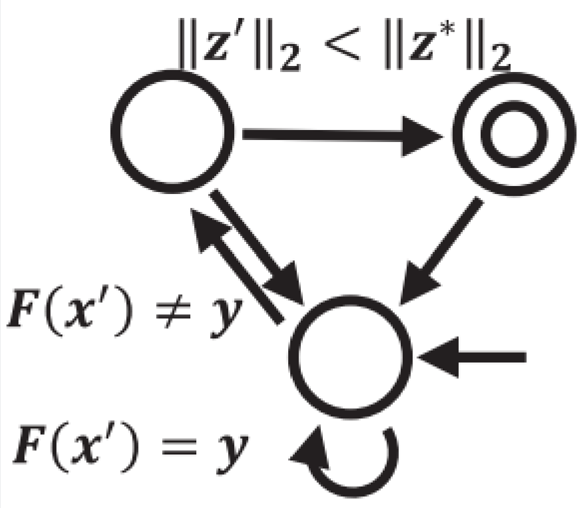


State Transition Module

Decides how to update the adversarial example for the next sampling of NCM

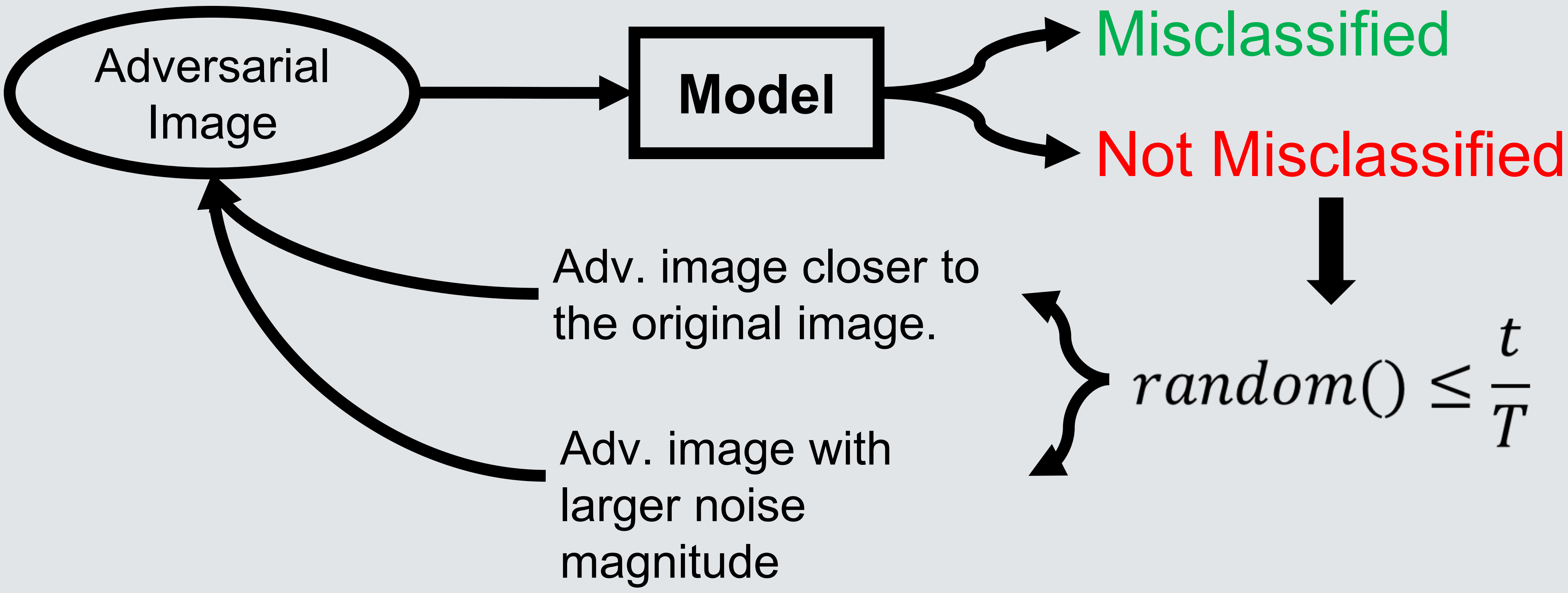
Existing DEA : If example fools the model and is closer, then replace current example.

Leads to falling into local optimum



State Transition Module

Transition Function Relaxation



Experimental Setup

8 Models: *resnet-18, resnet-101, inceptionv3, inception-resnetv2, nasnet, densenet-161, vgg19-bn, senet-154*
3 Datasets: *ImageNet, Tiny-Imagenet, CIFAR-10*
7 Transfer-based Attacks: *FGSM, I-FGSM, MI-FGSM, vr-IGSM, DDN, C&W, EAD*
8 Decision-based Attacks: *Boundary, Whey, BBA, EVO, HSJA, Tangent, Sign-OPT, QEBA*

Experimental Results

Target Model: inc-res Substitute Model: nasnet

6.42	1.474	2.035	1.209	1.389	3.444	1.502	1.922	1.340	0.944
6.9	2.303	2.87	2.048	2.271	4.78	2.350	3.274	2.306	1.998
1.407	1.148	1.009	0.983	1.165	1.34	1.243	1.311	1.147	0.75
3.971	2.284	2.631	2.074	2.266	3.123	2.207	2.673	2.024	1.985
3.893	1.755	2.03	1.895	1.767	3.538	1.944	2.030	1.585	1.106
8.014	3.218	4.289	3.258	3.294	5.307	2.955	3.607	2.714	2.685
1.947	1.527	1.369	1.189	1.437	1.79	1.444	1.652	1.266	0.974
3.942	3.988	3.888	3.874	4.028	4.667	3.891	4.163	3.745	3.542
2.245	1.494	1.388	1.376	1.437	2.048	1.613	1.896	1.541	1.01
5.095	3.169	3.806	3.34	3.102	4.177	3.159	3.718	2.868	2.741
1.235	1.131	0.832	0.785	1	1.163	0.968	1.112	0.866	0.61
3.488	1.814	1.737	1.603	1.855	2.718	2.066	2.412	1.856	1.489
1.375	1.912	1.842	1.571	1.888	2.144	1.809	1.854	1.525	1.483
28.329	39.867	39.863	39.831	39.864	39.899	38.624	38.662	38.601	39.812
1.116	0.988	0.838	0.81	1.145	1.373	0.922	1.100	0.827	0.575
2.055	1.799	1.604	1.502	1.791	2.529	1.624	1.877	1.494	1.307
1.705	0.974	0.767	0.75	0.983	1.091	0.772	0.792	0.691	0.568
3.07	1.436	1.259	1.315	1.418	1.789	1.529	1.662	1.457	1.057

Experimental Results

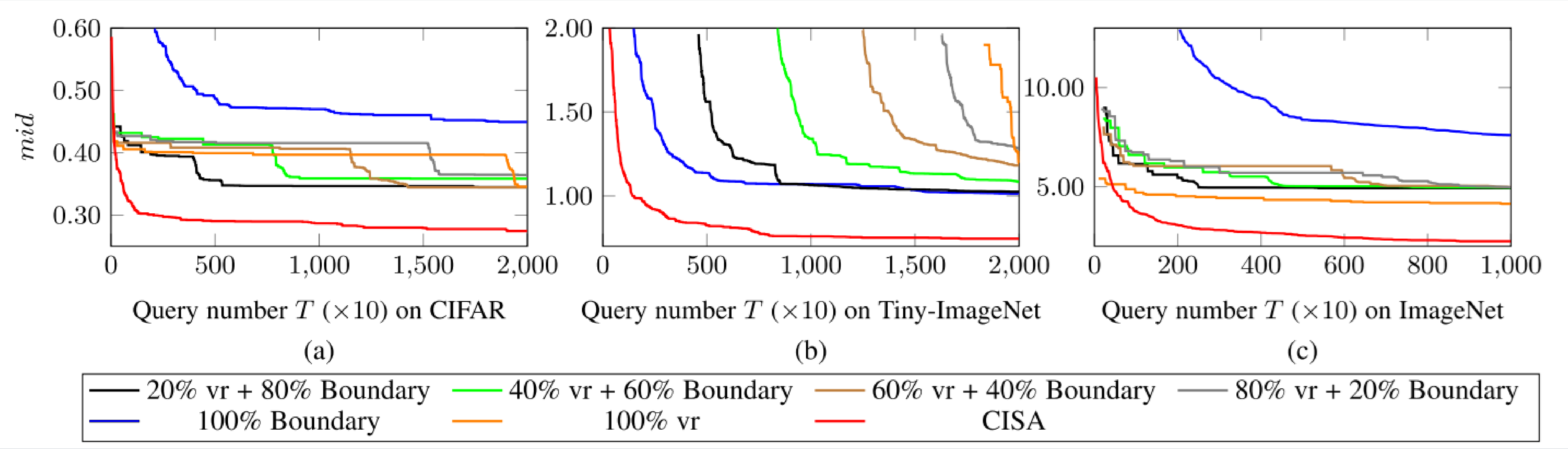
Lowest Noise Magnitude

Larger Noise Magnitude

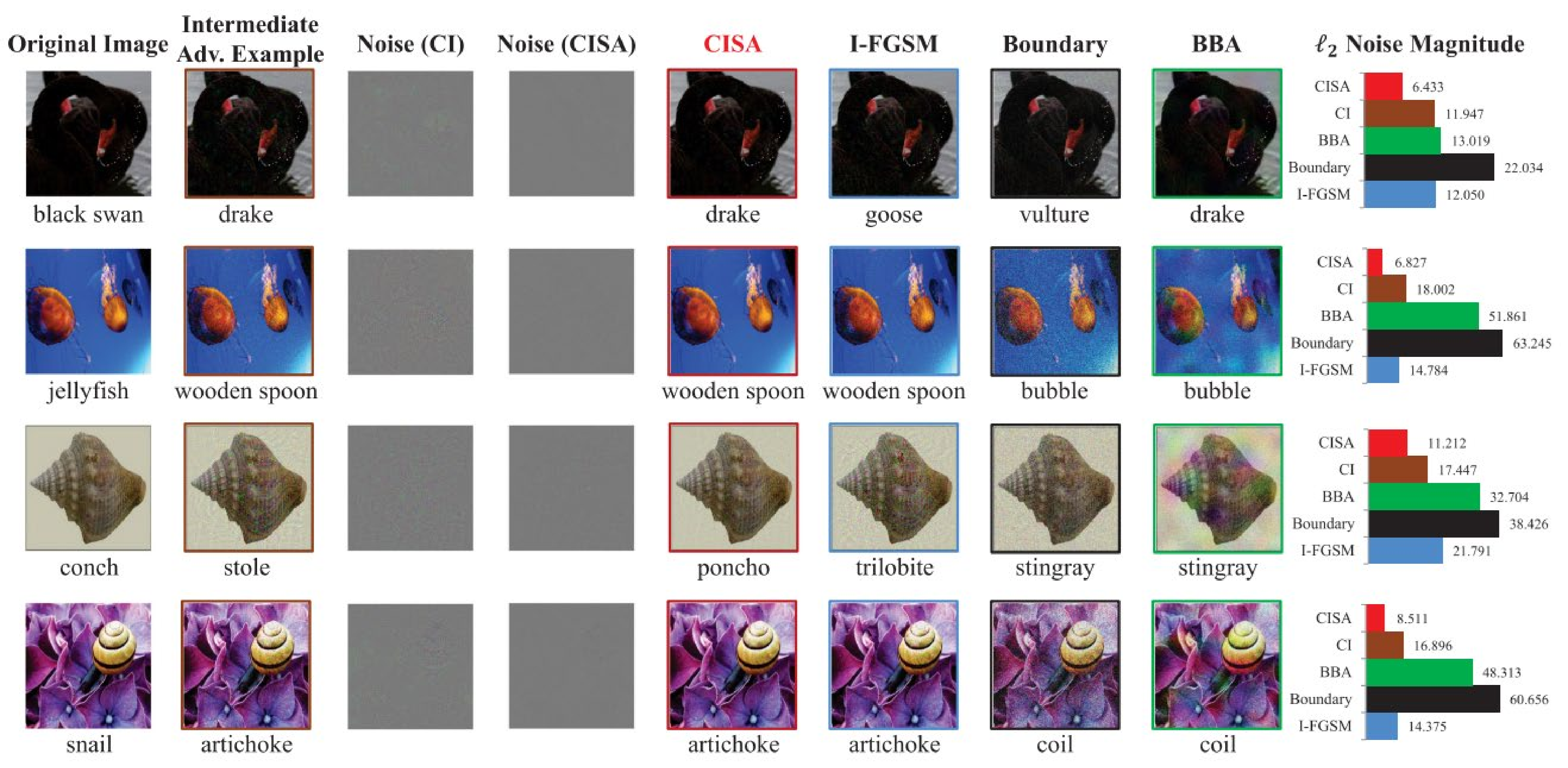
				Target Model: inc-res					Substitute Model: nasnet				
				NCM									
		STM		N/A	Greedy Search								Relax
	PAM			Vanilla	Boundary	Whey	BBA	EVO	HSJA	Tangent	Sign-OPT	QEBA	CS
TAM	N/A	Random	Mid	6.42	1.474	2.035	1.209	1.389	3.444	1.502	1.922	1.340	0.944
			Avg	6.9	2.303	2.87	2.048	2.271	4.78	2.350	3.274	2.306	1.998
	Binary Search	DDN	Mid	1.407	1.148	1.009	0.983	1.165	1.34	1.243	1.311	1.147	0.75
			Avg	3.971	2.284	2.631	2.074	2.266	3.123	2.207	2.673	2.024	1.985
		FGSM	Mid	3.893	1.755	2.03	1.895	1.767	3.538	1.944	2.030	1.585	1.106
			Avg	8.014	3.218	4.289	3.258	3.294	5.307	2.955	3.607	2.714	2.685
		I-FGSM	Mid	1.947	1.527	1.369	1.189	1.437	1.79	1.444	1.652	1.266	0.974
			Avg	3.942	3.988	3.888	3.874	4.028	4.667	3.891	4.163	3.745	3.542
		MI-FGSM	Mid	2.245	1.494	1.388	1.376	1.437	2.048	1.613	1.896	1.541	1.01
			Avg	5.095	3.169	3.806	3.34	3.102	4.177	3.159	3.718	2.868	2.741
		vr-IGSM	Mid	1.235	1.131	0.832	0.785	1	1.163	0.968	1.112	0.866	0.61
			Avg	3.488	1.814	1.737	1.603	1.855	2.718	2.066	2.412	1.856	1.489
		C&W	Mid	1.375	1.912	1.842	1.571	1.888	2.144	1.809	1.854	1.525	1.483
			Avg	28.329	39.867	39.863	39.831	39.864	39.899	38.624	38.662	38.601	39.812
		EAD	Mid	1.116	0.988	0.838	0.81	1.145	1.373	0.922	1.100	0.827	0.575
			Avg	2.055	1.799	1.604	1.502	1.791	2.529	1.624	1.877	1.494	1.307
	GSA	CI	Mid	1.705	0.974	0.767	0.75	0.983	1.091	0.772	0.792	0.691	**0.568**
			Avg	3.07	1.436	1.259	1.315	1.418	1.789	1.529	1.662	1.457	**1.057**

Lowest Noise Magnitude

Experimental Results



Experimental Results



Conclusion

Shi et al. presented:

- Black-box adversarial framework (PAM, TAM, NCM, STM)
- CISA Attack
- Solves Overfitting, Ensures robust noise compression, query efficiency, avoids local

Thank You
Questions?

Discussion

- Complexity
- Scalability
- Gradient masking defences