NEURAL INFORMATION PROCESSING SYSTEMS

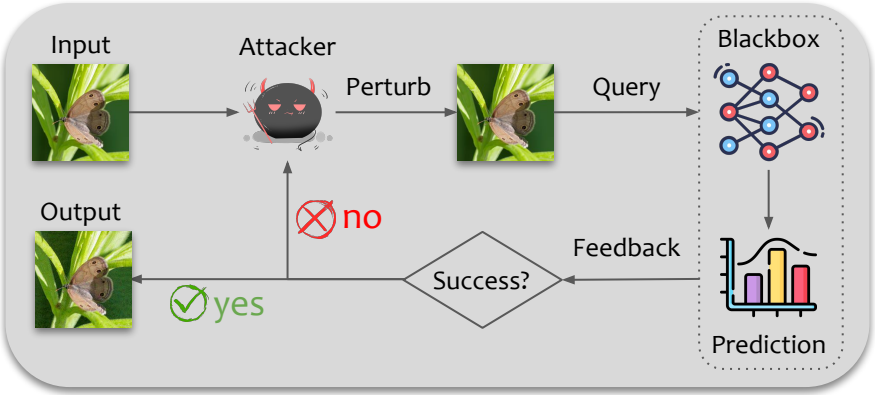
UC RIVERSIDE

Blackbox Attacks via Surrogate Ensemble Search

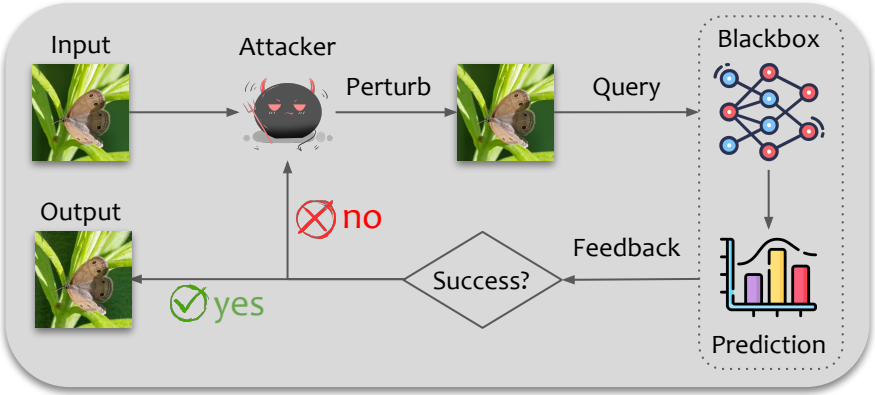
Zikui Cai, Chengyu Song, Srikanth V. Krishnamurthy, Amit K. Roy-Chowdhury, M. Salman Asif

University of California, Riverside

Blackbox Attacks



Blackbox Attacks and Current Approaches

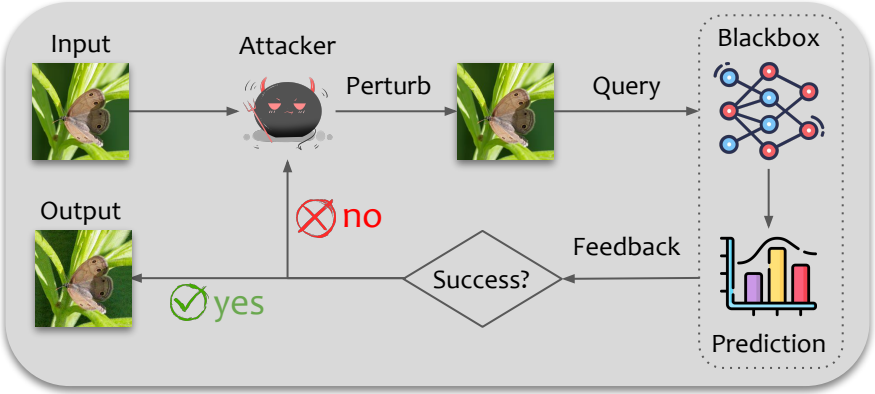


Transfer-based attacks

Attack surrogate model(s) and test the perturbation on the victim models.

Pros: do not need feedback
Cons: low success rate

Blackbox Attacks and Current Approaches



Transfer-based attacks

Attack surrogate model(s) and test the perturbation on the victim models.

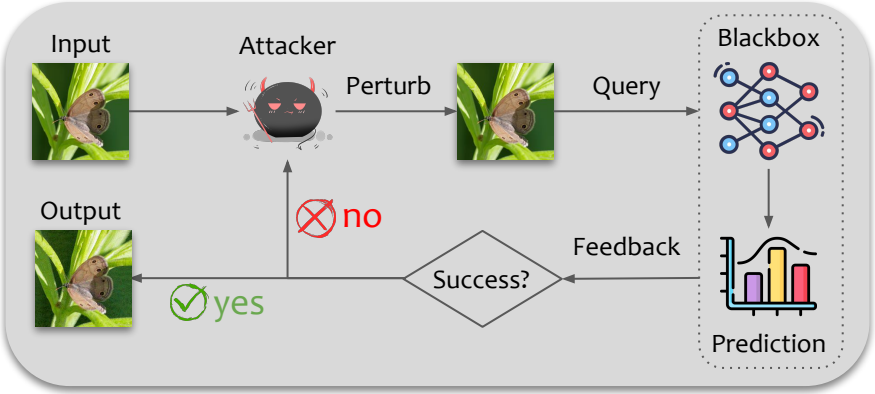
Pros: do not need feedback
Cons: low success rate

Query-based attacks

Update perturbations by iteratively querying the victim model.

Pros: high success rate
Cons: require a large number of queries

Blackbox Attacks and Current Approaches



Transfer-based attacks

Attack surrogate model(s) and test the perturbation on the victim models.

Pros: do not need feedback
Cons: low success rate

Transfer-based combined with query

Query in a potentially low dimensional transferable search space.

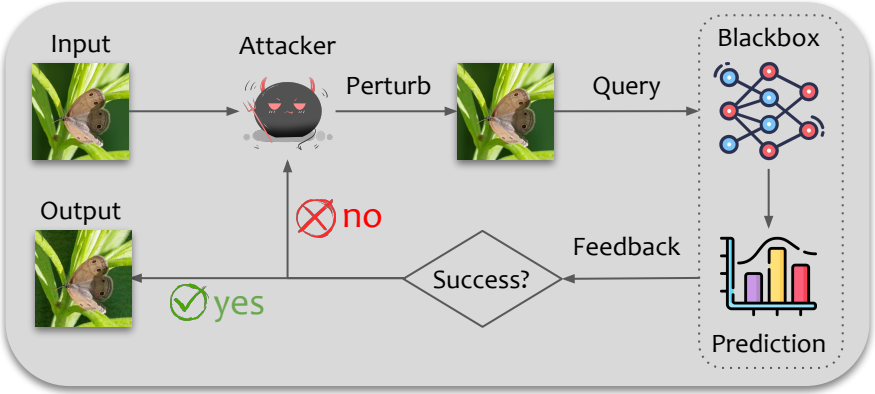
Pros: better overall performance

Query-based attacks

Update perturbations by iteratively querying the victim model.

Pros: high success rate
Cons: require a large number of queries

Blackbox Attacks and Current Approaches



Transfer-based attacks

Attack surrogate model(s) and test the perturbation on the victim models.

Pros: do not need feedback
Cons: low success rate

Our approach (BASES)

Blackbox Attacks via Surrogate Ensemble Search

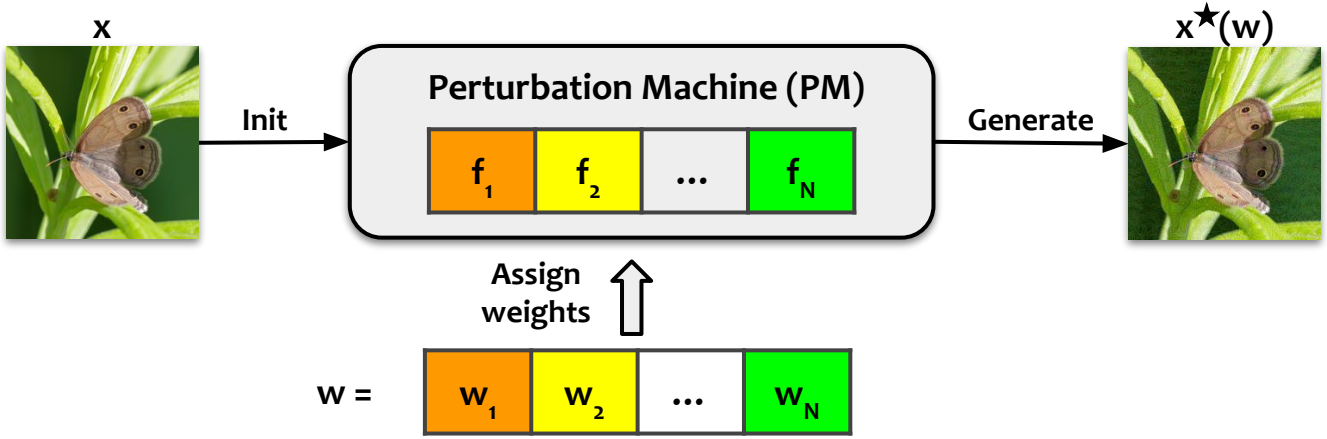
- **high successful rate**
- **extremely small query counts**

Query-based attacks

Update perturbations by iteratively querying the victim model.

Pros: high success rate
Cons: require a large number of queries

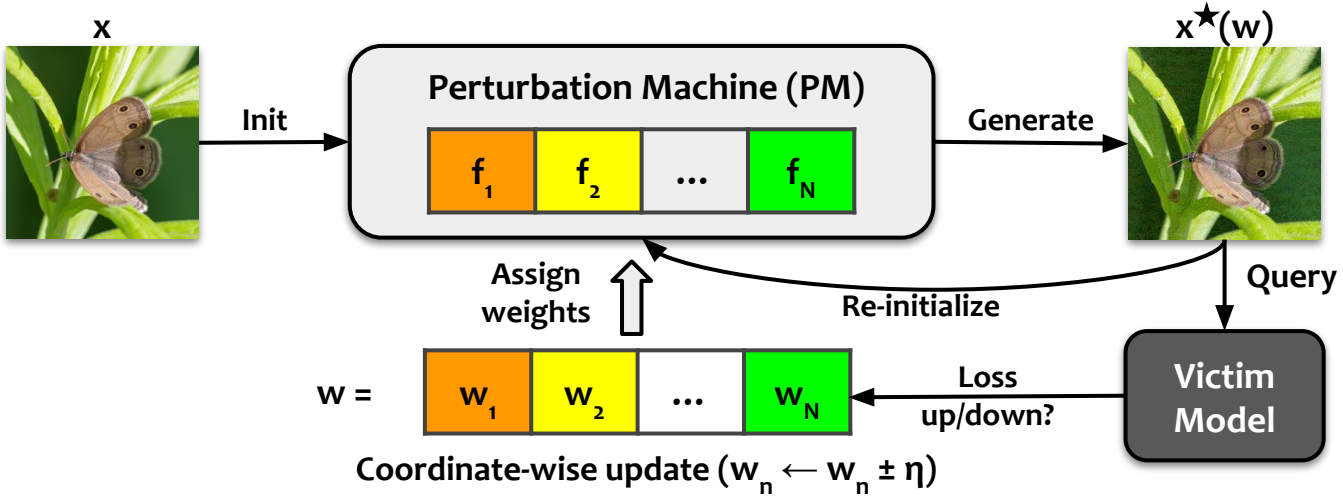
BASES Framework



$$x^\star(\mathbf{w}) = \arg\min_x \sum_{i=1}^{N} w_i \mathcal{L}(f_i(x), y^\star)$$

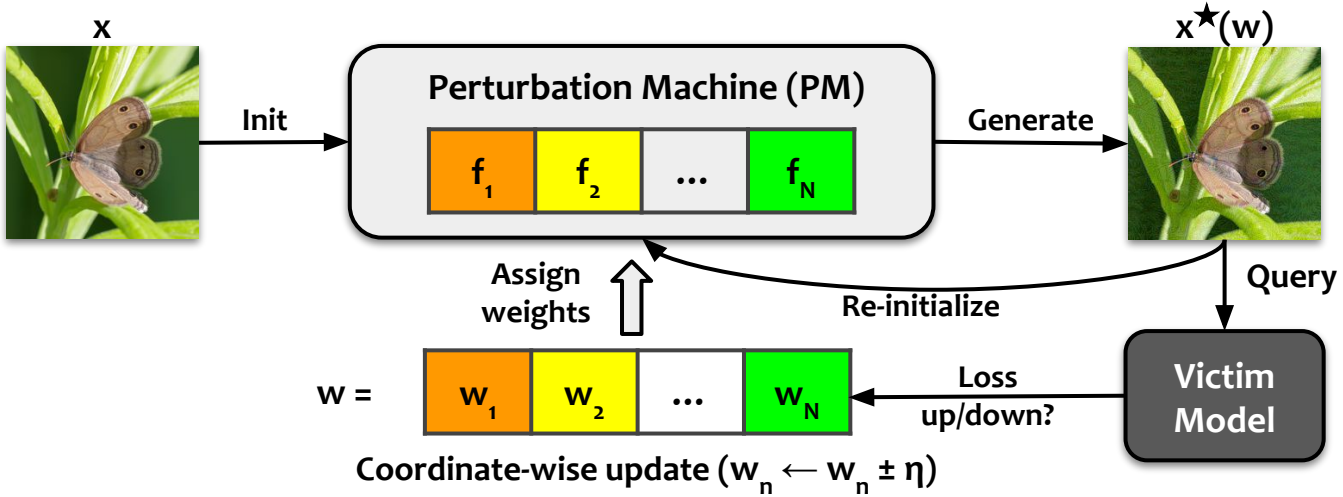
Minimize a weighted loss function to generate a perturbed image from the perturbation machine

BASES Framework



$$\mathbf{w} = \arg\min_{\mathbf{w}} \ \mathcal{L}_{\mathbf{v}}(f_{\mathbf{v}}(x^{\star}(\mathbf{w})), y^{\star})$$

BASES Framework



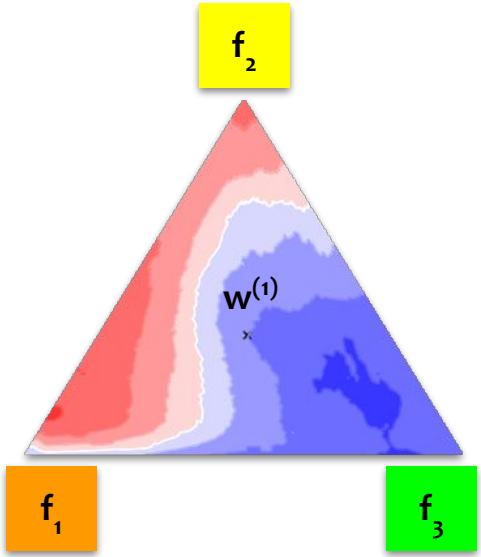
Iteration

1 $w^{(1)}$ = 0.33 0.33 ... 0.33

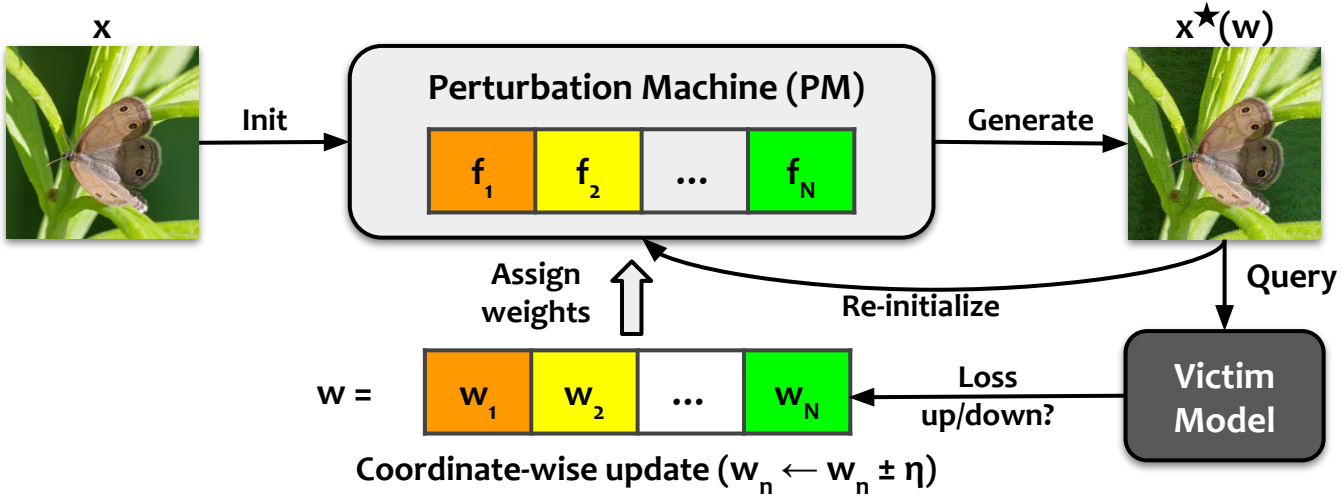
Fail
loss = 5.0
label: Butterfly
queries = 1

Loss landscape vs weights of (N=3) models

Red indicates high loss
Blue indicates low loss



BASES Framework



Iteration

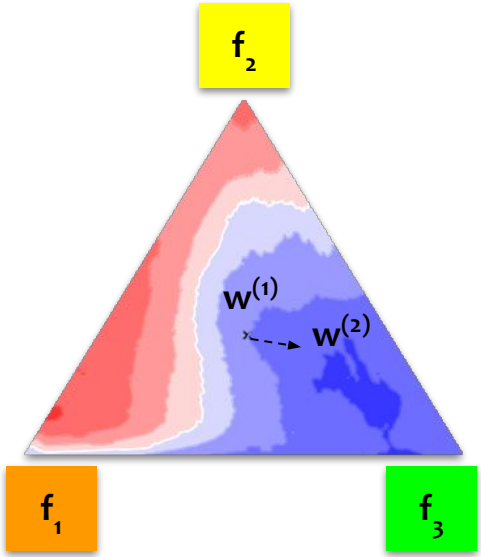
1 $w^{(1)}$ = 0.33 0.33 ... 0.33

2 $w^{(2)}$ = 0.3 0.3 ... 0.4

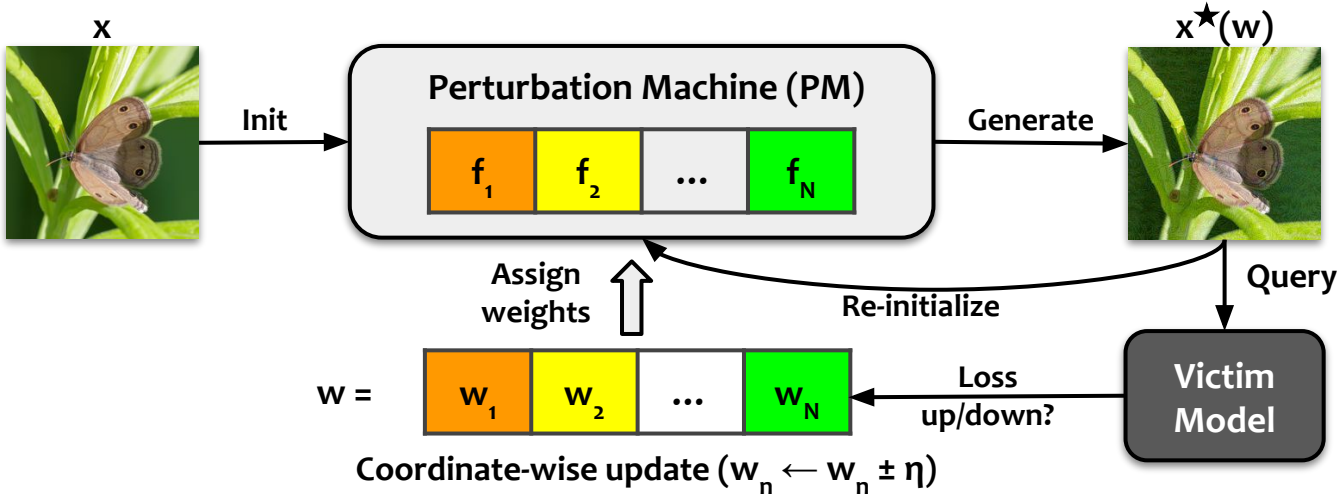
Fail
loss = 1.5
label: Butterfly
queries = 2

Loss landscape vs weights of (N=3) models

Red indicates high loss
Blue indicates low loss



BASES Framework



Iteration

1 $w^{(1)}$ = 0.33 0.33 ... 0.33

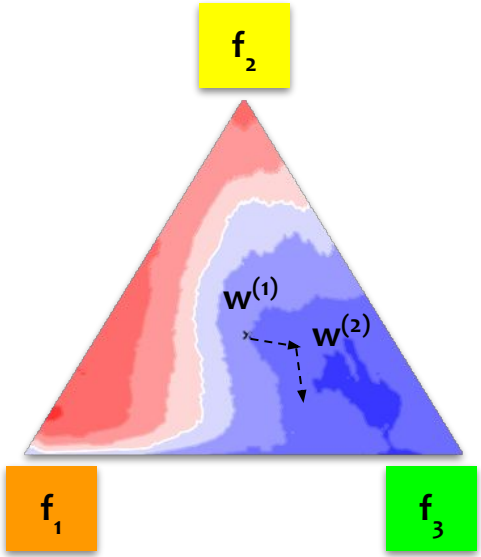
2 $w^{(2)}$ = 0.3 0.3 ... 0.4

⋮ ⋮

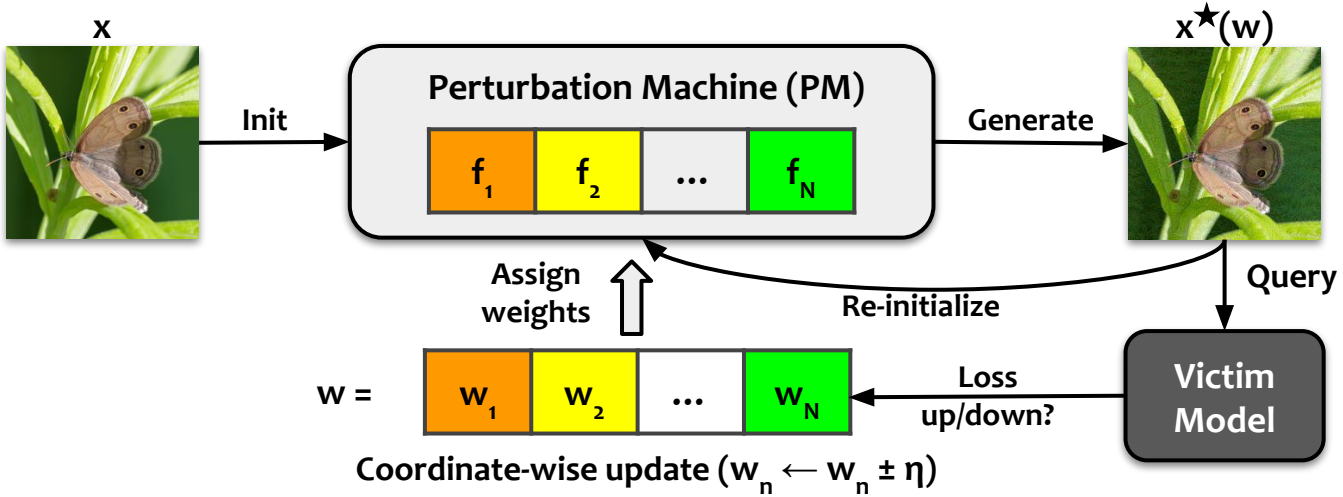
Fail
loss = 1.5
label: Butterfly
queries = 2

Loss landscape vs weights of (N=3) models

Red indicates high loss
Blue indicates low loss



BASES Framework



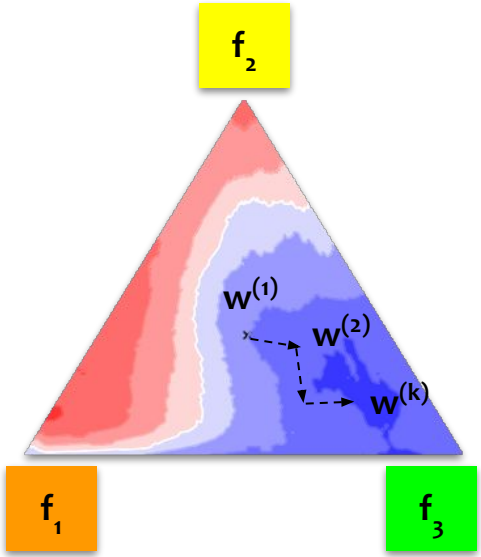
Iteration

1	$w^{(1)}$ =	0.33 0.33 ... 0.33
2	$w^{(2)}$ =	0.3 0.3 ... 0.4
⋮		⋮
k	$w^{(k)}$ =	0.1 0.1 ... 0.8

Success
loss = - 2.5
label: Primate
queries = 6

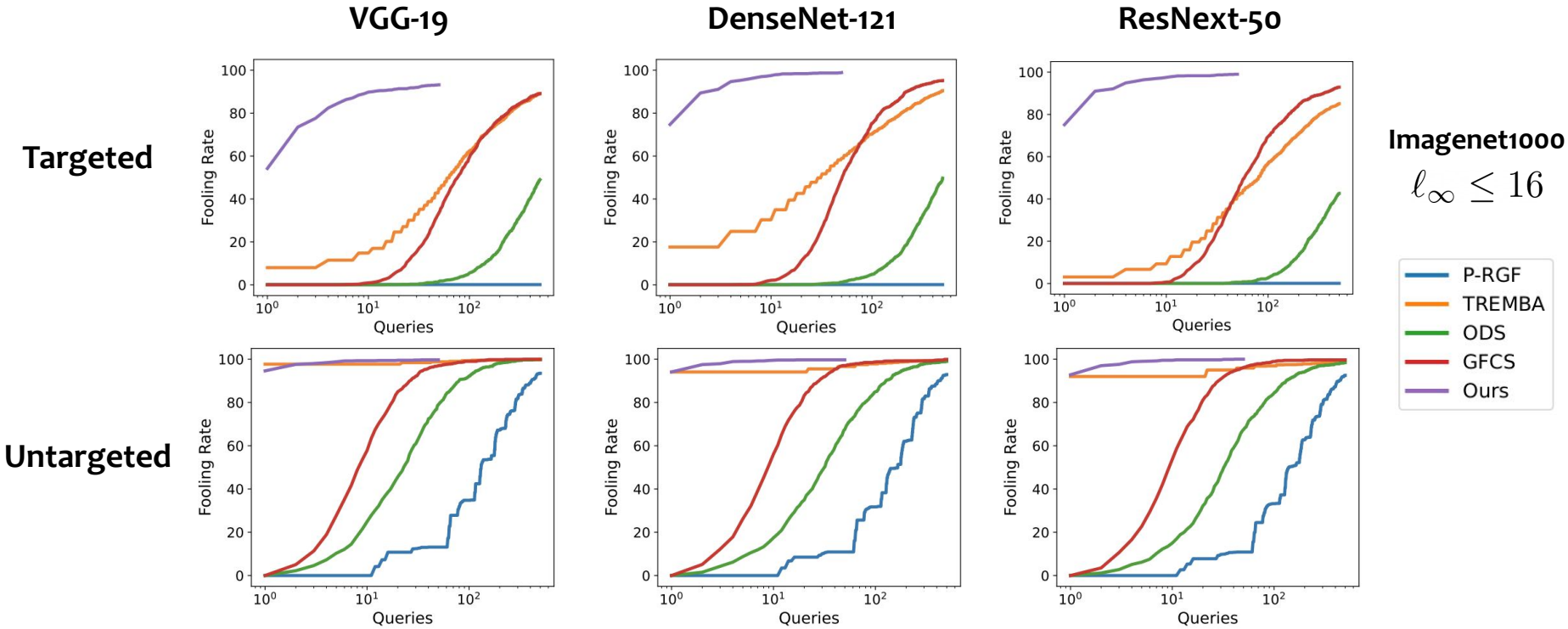
Loss landscape vs weights of (N=3) models

Red indicates high loss
Blue indicates low loss



Attacks on Classifiers and Comparison

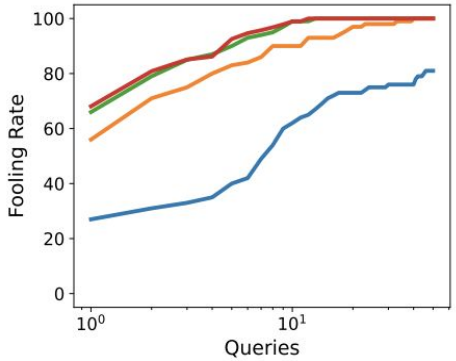
- Fooling rate vs Queries in Targeted (1st row) and Untargeted (2nd row) settings



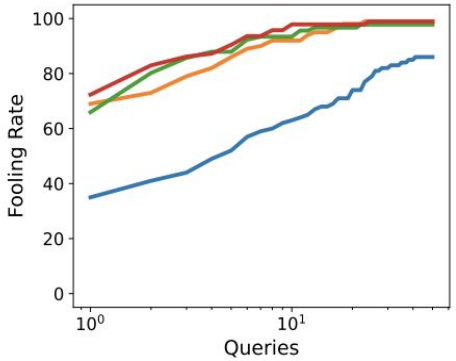
Attack on Object Detectors

- Fooling rates for vanishing attacks on three victim object detectors using different number (N $\in\{2, 3, 4, 5\}$) of surrogate models in PM
- surrogate models: {Faster R-CNN, YOLOv3, FreeAnchor, DETR, CenterNet}

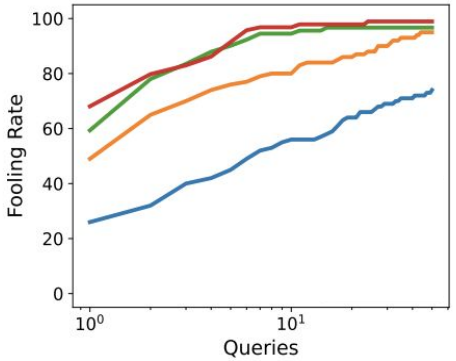
RetinaNet



RepPoints

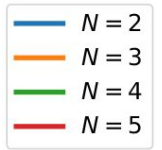


Deformable DETR



COCO 2014 val

$\ell_\infty \leq 16$



Attack Google Cloud Vision API

- Classification results of clean images (left) and perturbed images (right)

87a2147620d5e1cb.png

Bus - clean

87a2147620d5e1cb_iter00.png

Bus - attacked

6612fd36e6dd9534.png

Fly - clean

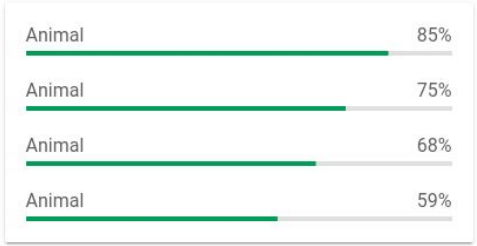
6612fd36e6dd9534_iter12.png

Fly - attacked

Attack Google Cloud Vision API - Object Detectors

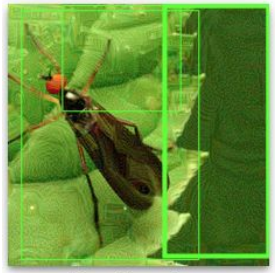
- Detection results of clean images (left) and perturbed images (right)

Bus - clean



Bus - attacked

Fly - clean



Fly - attacked

Conclusion

- Summary:
  - BASES can effectively perform blackbox attacks in a query-efficient manner, by searching over the weight space of ensemble models.
  - BASES demonstrates generalizable attacks in real life, such as Google Cloud Vision API
  - BASES is generic and can be applied to different tasks

Paper

https://arxiv.org/abs/2208.03610

Code

https://github.com/CSIPlab/BASES

Acknowledgements: DARPA under Agreement No. HR00112090096.
More information. Zikui Cai (zcai032@ucr.edu), M. Salman Asif (sasif@ucr.edu)

Thank you!

Stay safe and healthy!