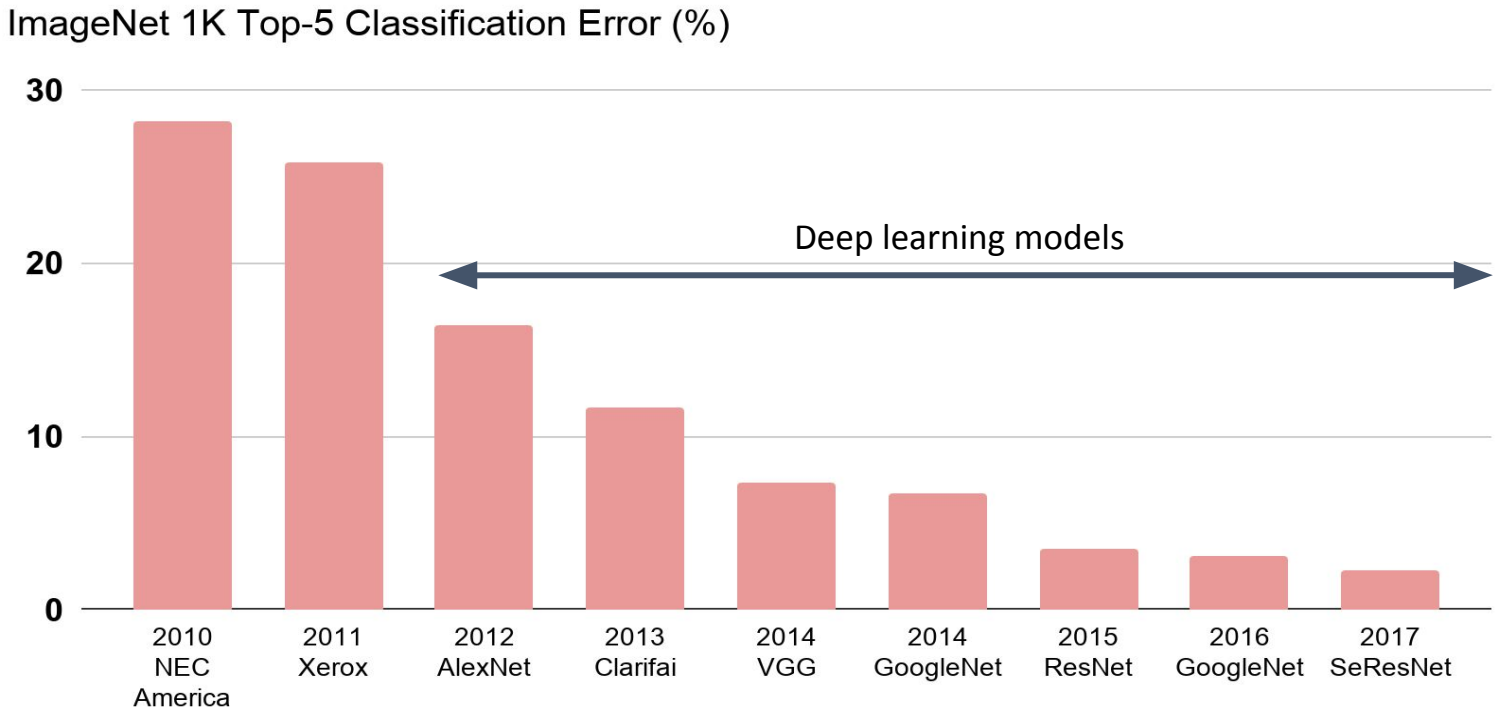
ICML 2020 - Workshop on Uncertainty & Robustness in Deep Learning (UDL)

# Regional Image Perturbation Reduces $L_p$ Norms of Adversarial Examples While Maintaining Model-to-model Transferability

Utku Ozbulak, Jonathan Peck, Wesley De Neve,
Bart Goossens, Yvan Saeys, and Arnout Van Messem

# Impact of deep learning models on computer vision problems

Deep learning methods drastically improved the state-of-the-art results obtained for computer vision problems

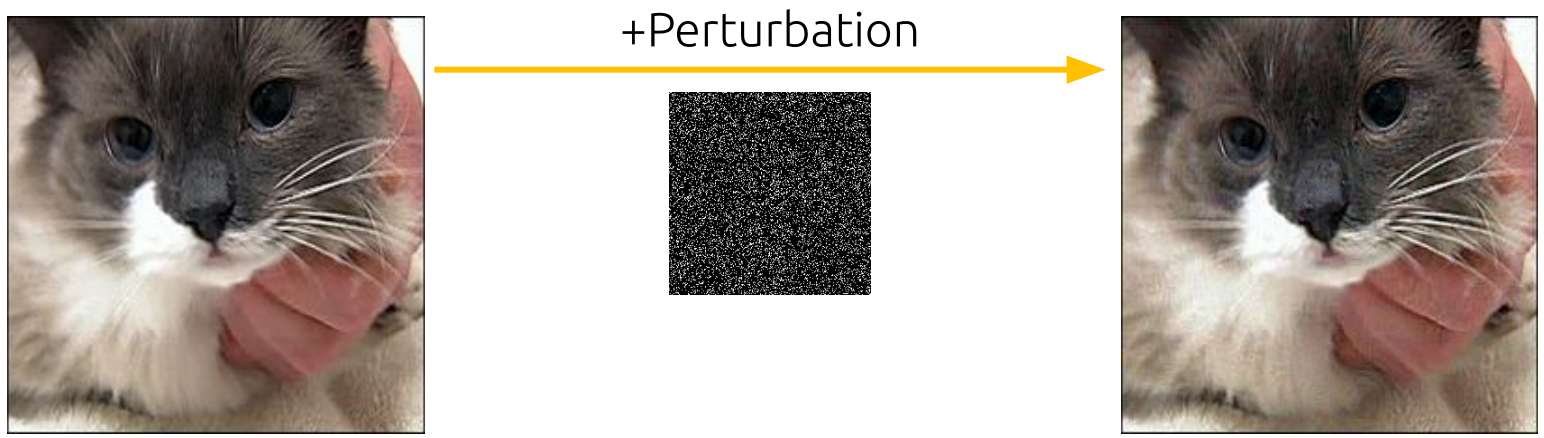


# Impact of deep learning models on solving complex problems

- Medical imaging problems (e.g., tumor detection)
- Facial recognition problems (e.g., person identification)
- Problems related to self-driving cars (e.g., lane detection)
- Problems related to smart-housing (e.g., voice commands)

# Vulnerability of deep learning models against adversarial attacks

Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake

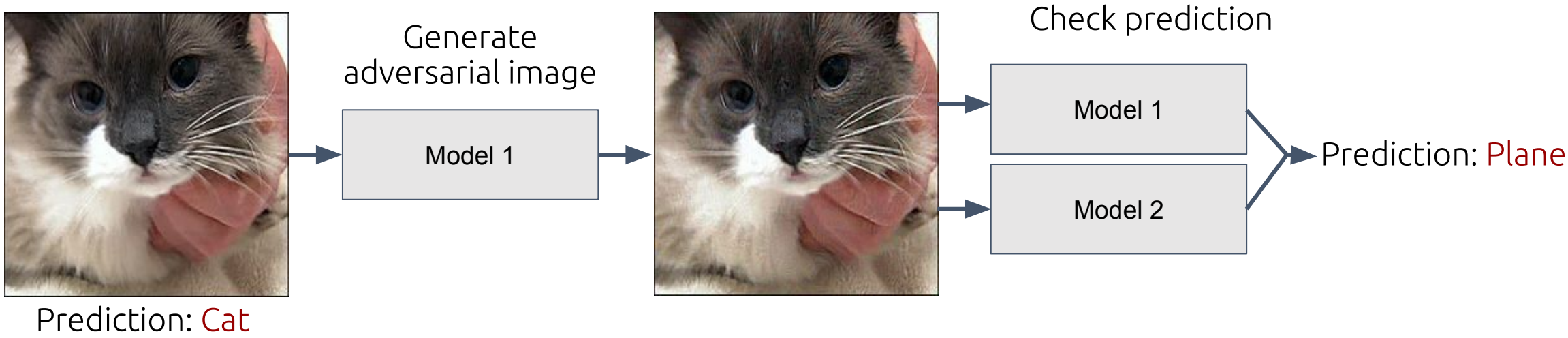



Original image
Prediction: Cat
Confidence: 96%

Adversarial image
Prediction: Plane
Confidence: 99%

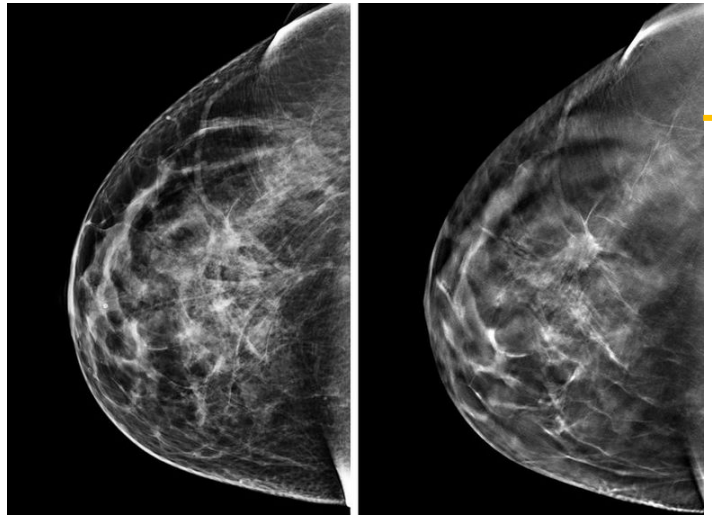
# Transferability of adversarial examples between models

Adversarial examples generated from a model are found to transfer to other models

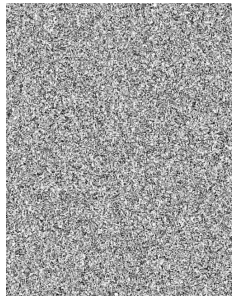


# Real-world consequences of adversarial examples

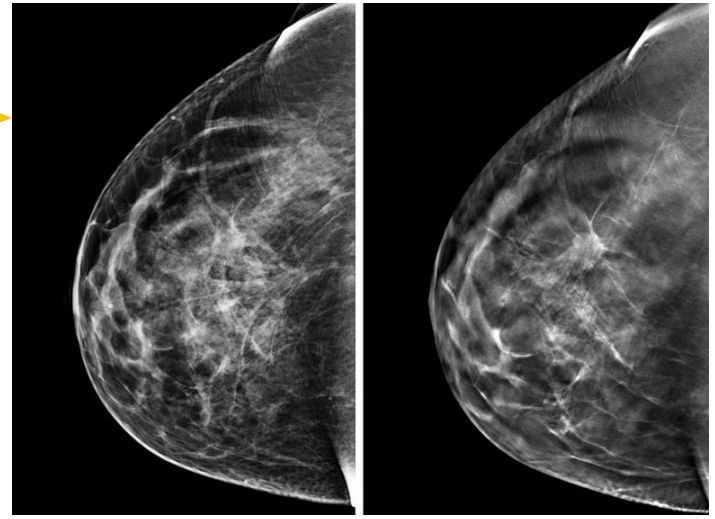
Original Mammogram
Prediction: Cancer
Confidence: 75%



+Perturbation

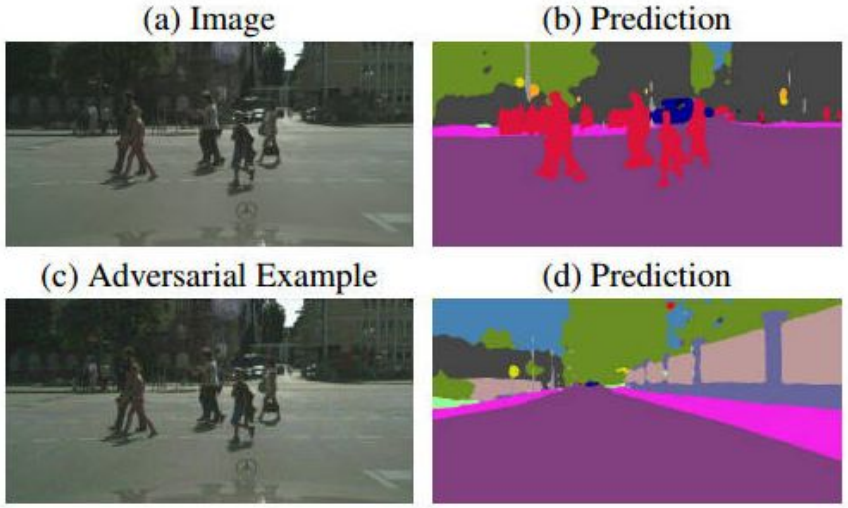


Adversarial Mammogram
Prediction: Healthy
Confidence: 99%

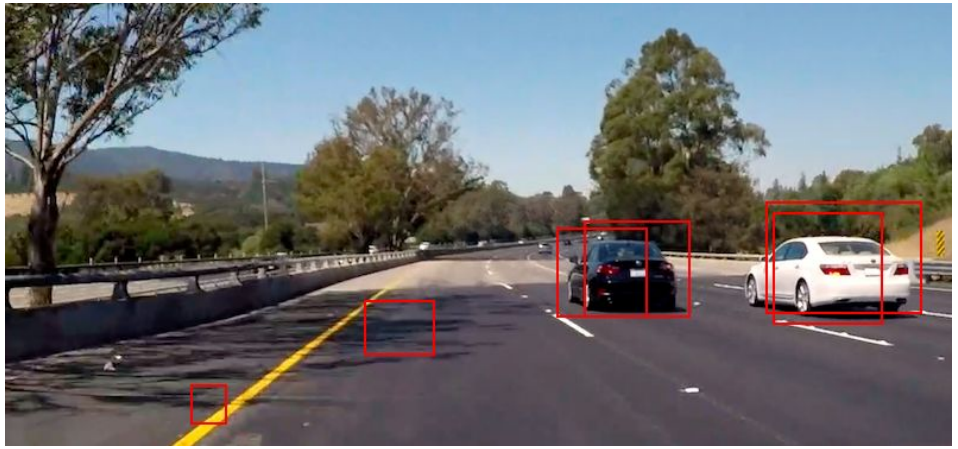


# Real-world consequences of adversarial examples

## Roadway Segmentation




## Vehicle Detection



# “Global” adversarial attacks

A number of popular methods for generating adversarial examples include:

- Fast-gradient sign (FGS)
- Iterative fast-gradient sign (IFGS/BIM)
- Projected gradient descent (PGD)
- Jacobian-based saliency map attack (JSMA)
- Carlini & Wagner’s attack (CW)

Initial image
Prediction: Cheetah

Adversarial images
Prediction: Elephant

# “Regional” adversarial attacks

A number of popular methods for generating adversarial examples include:

- One pixel attack
- Localized and visible adversarial noise (LaVAN)
- Adversarial patch
- Structured adversarial attack

One-pixel attack

LaVAN

Adversarial patch

Structured adv. attack

# Problems with “regional” adversarial attacks

- Proposing a completely new optimization method, making it harder to compare against well-known attacks
- Experiments are often presented in permissive white-box settings
- Might be hard to employ for datasets other than the one studied in their respective papers
- “Regional” adversarial attacks are often not studied in defense evaluations

# A typical adversarial optimization (for PGD, FGS, and BIM)

$$\boldsymbol{X}_{n+1} = \boldsymbol{X}_n + \boldsymbol{P}_n$$

- $\mathbf{X}_0$ represents the original image
- $\mathbf{X}_n$ represents the perturbed image at $n$th iteration
- $\mathbf{P}_n$ represents the added perturbation at $n$th iteration

$$\boldsymbol{P}_n = \alpha \operatorname{sign}\left(\nabla_x J(g(\theta, \boldsymbol{X}_n)_c)\right)$$

- $\boldsymbol{\alpha}$ represents the perturbation multiplier
- $\mathtt{sign}()$ represents the signature operation
- $\nabla x J()_{\mathbf{c}}$ represents the gradient of the Cross-Entropy from class **c** w.r.t. input **X**
- $g(\theta, \mathbf{X_n})$ represents a forward pass from a neural network with parameters $\theta$

# Converting a “global” attack to a “regional” attack

$$\boldsymbol{X}_{n+1} = \boldsymbol{X}_n + \boldsymbol{P}_n \odot \boldsymbol{L}$$

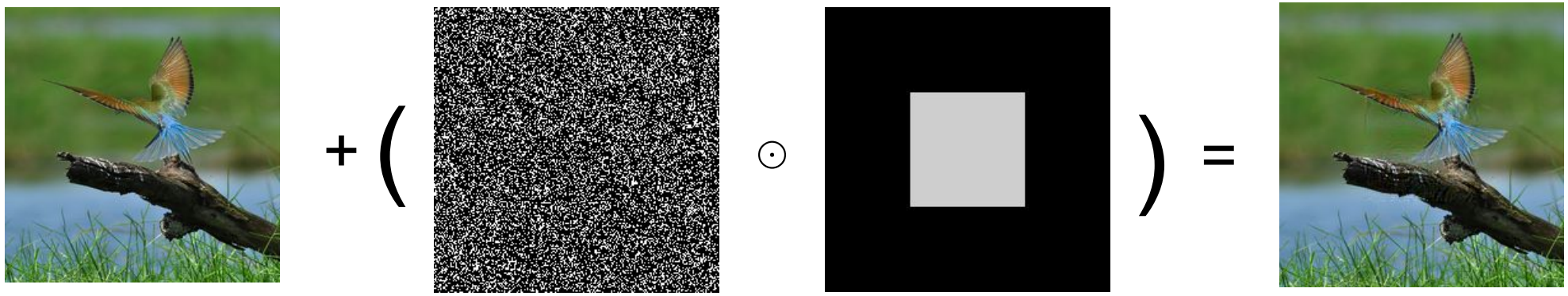
- $\mathbf{X}_0$ represents the original image
- $\mathbf{X}_n$ represents the perturbed image at $n$th iteration
- $\mathbf{P}_n$ represents the added perturbation at $n$th iteration
- $\odot \mathbf{L}$ represents pixel-wise multiplication (i.e., Hadamard product)

$$\boldsymbol{P}_n = \alpha \operatorname{sign}\left(\nabla_x J(g(\theta, \boldsymbol{X}_n)_c)\right)$$

- $\boldsymbol{\alpha}$ represents the perturbation multiplier
- $\mathtt{sign}()$ represents the signature operation
- $\nabla x J()_{\mathbf{c}}$ represents the gradient of the Cross-Entropy from class **c** w.r.t. input **X**
- $g(\theta, \mathbf{X_n})$ represents a forward pass from a neural network with parameters $\theta$

# Converting a “global” attack to a “regional” attack

$$X_{n+1} = X_n + (P_n \odot L)$$



$X_n$: Initial image

$P_n$: Perturbation

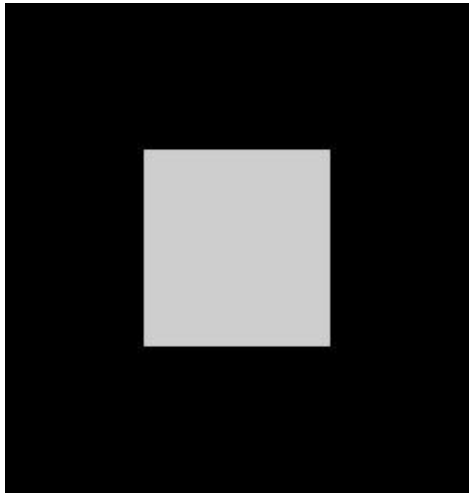
$L$: Localization

$X_{n+1}$: Adversarial image

$(P_n \odot L)$: Performed at each iteration

# Converting a “global” attack to a “regional” attack

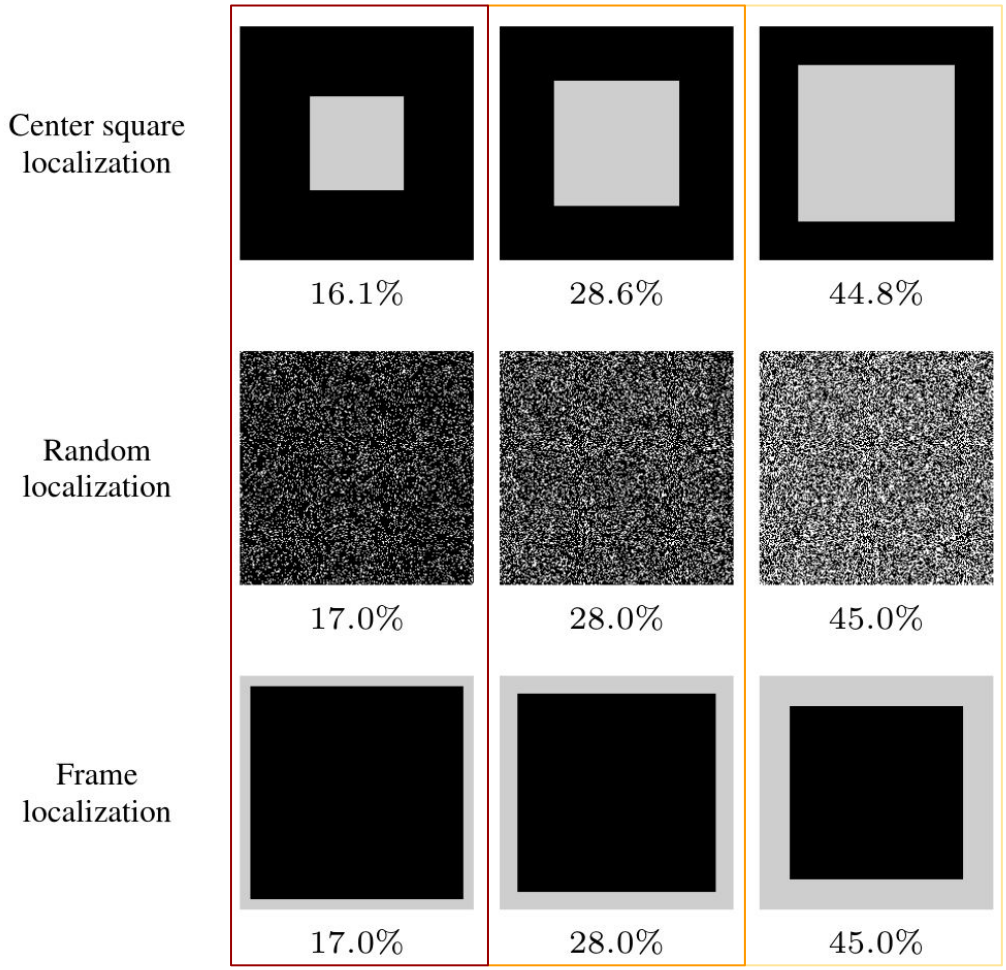
Initial image



Localization

Adversarial image

# Selected regions for evaluation




We evaluate the use of three different perturbation regions, each with three different settings

- Center square localization (90px, 120px, 150px of length)
- Random localization (17%, 28%, 45% of pixels)
- Outer frame localization (20px, 34px, 58px of length)

# How to calculate Lp norms of perturbation?

Adversarial image — Initial image — Total perturbation

$L_0$ norm: number of changed pixels
- 0-> None
- 1-> All

$L_2$ norm: Euclidean norm of the perturbation

$L_\infty$ norm: intensity of the largest change
- 0-> None
- 1-> Black to white or vice versa

# Locally perturbed adversarial examples maintain transferability

Percentage of adversarial examples with localized perturbation that transfer from source model (generated from) to target model (tested against) when 17%, 28%, and 45% of pixels are selected

| Source Model \ Target Model | AlexNet | VGG-16 | ResNet-50 |
|---|---|---|---|
| AlexNet | 100% | 73% | 66% |
| VGG-16 | 60% | 100% | 56% |
| ResNet-50 | 52% | 59% | 100% |

(a) 17% of pixels selected

| Source Model \ Target Model | AlexNet | VGG-16 | ResNet-50 |
|---|---|---|---|
| AlexNet | 100% | 76% | 67% |
| VGG-16 | 70% | 100% | 63% |
| ResNet-50 | 67% | 65% | 100% |

(b) 28% of pixels selected

| Source Model \ Target Model | AlexNet | VGG-16 | ResNet-50 |
|---|---|---|---|
| AlexNet | 100% | 78% | 75% |
| VGG-16 | 83% | 100% | 75% |
| ResNet-50 | 78% | 77% | 100% |

(c) 45% of pixels selected

# Locally perturbed adversarial examples have reduced Lp norms

Mean (standard deviation) Lp distances calculated between genuine images and their adversarial counterparts for the adversarial examples that transfer from the source model to the target model

| Localization | Source: | AlexNet | | | | VGG-16 | | | | ResNet-50 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Target: | VGG-16 | | ResNet-50 | | AlexNet | | ResNet-50 | | AlexNet | | VGG-16 | |
| | Norm: | $L_2$ | $L_\infty$ | $L_2$ | $L_\infty$ | $L_2$ | $L_\infty$ | $L_2$ | $L_\infty$ | $L_2$ | $L_\infty$ | $L_2$ | $L_\infty$ |
| No Localization | | 7.35 (5.37) | 0.07 (0.08) | 6.39 (4.50) | 0.05 (0.06) | 6.91 (4.17) | 0.07 (0.05) | 3.62 (3.16) | 0.02 (0.04) | 6.79 (4.31) | 0.07 (0.06) | 3.76 (2.20) | 0.02 (0.02) |
| Center | 90px | 6.55 (4.36) | 0.15 (0.14) | 5.33 (3.52) | 0.11 (0.10) | 4.01 (2.64) | 0.10 (0.09) | 3.41 (2.77) | 0.09 (0.10) | 3.54 (2.54) | 0.09 (0.09) | 2.79 (2.23) | 0.06 (0.07) |
| | 120px | 6.47 (4.48) | 0.11 (0.11) | 6.30 (4.45) | 0.10 (0.11) | 5.01 (2.99) | 0.10 (0.08) | 3.68 (3.05) | 0.06 (0.07) | 4.50 (2.93) | 0.09 (0.08) | 3.70 (3.09) | 0.06 (0.08) |
| | 150px | 6.80 (4.48) | 0.10 (0.10) | 6.46 (4.33) | 0.09 (0.09) | 6.71 (3.79) | 0.11 (0.08) | 3.92 (3.07) | 0.05 (0.06) | 6.64 (3.90) | 0.11 (0.08) | 4.65 (3.54) | 0.07 (0.07) |
| Frame | 20px | 9.86 (8.37) | 0.18 (0.20) | 10.1 (7.94) | 0.20 (0.21) | 6.07 (3.74) | 0.16 (0.14) | 4.64 (3.61) | 0.12 (0.15) | 4.77 (2.88) | 0.13 (0.12) | 4.32 (2.90) | 0.11 (0.11) |
| | 34px | 8.92 (6.60) | 0.13 (0.13) | 8.63 (6.52) | 0.13 (0.14) | 6.71 (4.04) | 0.15 (0.12) | 4.50 (2.96) | 0.08 (0.08) | 5.68 (3.25) | 0.12 (0.09) | 4.85 (3.44) | 0.10 (0.10) |
| | 58px | 8.44 (5.72) | 0.12 (0.13) | 7.23 (4.94) | 0.09 (0.11) | 7.79 (4.17) | 0.14 (0.09) | 5.44 (3.89) | 0.08 (0.09) | 7.02 (3.90) | 0.12 (0.09) | 5.78 (3.79) | 0.09 (0.08) |
| Random | 17% | 8.11 (5.63) | 0.15 (0.16) | 7.41 (4.63) | 0.13 (0.14) | 5.20 (3.14) | 0.13 (0.11) | 4.43 (3.30) | 0.10 (0.12) | 4.59 (2.65) | 0.10 (0.09) | 3.81 (2.92) | 0.08 (0.09) |
| | 28% | 6.82 (4.99) | 0.10 (0.12) | 7.51 (4.54) | 0.11 (0.11) | 5.97 (3.55) | 0.12 (0.10) | 4.29 (2.91) | 0.07 (0.07) | 5.50 (3.14) | 0.11 (0.09) | 4.35 (3.02) | 0.07 (0.08) |
| | 45% | 7.42 (4.80) | 0.10 (0.11) | 6.76 (4.20) | 0.09 (0.09) | 7.21 (3.98) | 0.12 (0.09) | 4.61 (3.41) | 0.06 (0.07) | 7.39 (4.03) | 0.12 (0.08) | 5.04 (3.53) | 0.07 (0.07) |

# Locally perturbed adversarial examples have reduced Lp norms

Mean (standard deviation) Lp distances calculated between genuine images and their adversarial counterparts for the adversarial examples that transfer from the source model to the target model

| Localization | Source: | AlexNet | | | | VGG-16 | | | | ResNet-50 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Target: | VGG-16 | | ResNet-50 | | AlexNet | | ResNet-50 | | AlexNet | | VGG-16 | |
| | Norm: | $L_2$ | $L_\infty$ | $L_2$ | $L_\infty$ | $L_2$ | $L_\infty$ | $L_2$ | $L_\infty$ | $L_2$ | $L_\infty$ | $L_2$ | $L_\infty$ |
| No Localization | | 7.35 (5.37) | | 6.39 (4.50) | | 6.91 (4.17) | | 3.62 (3.16) | | 6.79 (4.31) | | 3.76 (2.20) | |
| Center | 90px | 6.55 (4.36) | | 5.33 (3.52) | | 4.01 (2.64) | | 3.41 (2.77) | | 3.54 (2.54) | | 2.79 (2.23) | |
| | 120px | 6.47 (4.48) | | 6.30 (4.45) | | 5.01 (2.99) | | 3.68 (3.05) | | 4.50 (2.93) | | 3.70 (3.09) | |
| | 150px | 6.80 (4.48) | | 6.46 (4.33) | | 6.71 (3.79) | | 3.92 (3.07) | | 6.64 (3.90) | | 4.65 (3.54) | |
| Frame | 20px | 9.86 (8.37) | | 10.1 (7.94) | | 6.07 (3.74) | | 4.64 (3.61) | | 4.77 (2.88) | | 4.32 (2.90) | |
| | 34px | 8.92 (6.60) | | 8.63 (6.52) | | 6.71 (4.04) | | 4.50 (2.96) | | 5.68 (3.25) | | 4.85 (3.44) | |
| | 58px | 8.44 (5.72) | | 7.23 (4.94) | | 7.79 (4.17) | | 5.44 (3.89) | | 7.02 (3.90) | | 5.78 (3.79) | |
| Random | 17% | 8.11 (5.63) | | 7.41 (4.63) | | 5.20 (3.14) | | 4.43 (3.30) | | 4.59 (2.65) | | 3.81 (2.92) | |
| | 28% | 6.82 (4.99) | | 7.51 (4.54) | | 5.97 (3.55) | | 4.29 (2.91) | | 5.50 (3.14) | | 4.35 (3.02) | |
| | 45% | 7.42 (4.80) | | 6.76 (4.20) | | 7.21 (3.98) | | 4.61 (3.41) | | 7.39 (4.03) | | 5.04 (3.53) | |

# Locally perturbed adversarial examples have reduced Lp norms

Mean (standard deviation) Lp distances calculated between genuine images and their adversarial counterparts for the adversarial examples that transfer from the source model to the target model

| Localization | Source: | AlexNet | | | | VGG-16 | | | | ResNet-50 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Target: | VGG-16 | | ResNet-50 | | AlexNet | | ResNet-50 | | AlexNet | | VGG-16 | |
| | Norm: | $L_2$ | $L_\infty$ | $L_2$ | $L_\infty$ | $L_2$ | $L_\infty$ | $L_2$ | $L_\infty$ | $L_2$ | $L_\infty$ | $L_2$ | $L_\infty$ |
| No Localization | | | 0.07 (0.08) | | 0.05 (0.06) | | 0.07 (0.05) | | 0.02 (0.04) | | 0.07 (0.06) | | 0.02 (0.02) |
| Center | 90px | | 0.15 (0.14) | | 0.11 (0.10) | | 0.10 (0.09) | | 0.09 (0.10) | | 0.09 (0.09) | | 0.06 (0.07) |
| | 120px | | 0.11 (0.11) | | 0.10 (0.11) | | 0.10 (0.08) | | 0.06 (0.07) | | 0.09 (0.08) | | 0.06 (0.08) |
| | 150px | | 0.10 (0.10) | | 0.09 (0.09) | | 0.11 (0.08) | | 0.05 (0.06) | | 0.11 (0.08) | | 0.07 (0.07) |
| Frame | 20px | | 0.18 (0.20) | | 0.20 (0.21) | | 0.16 (0.14) | | 0.12 (0.15) | | 0.13 (0.12) | | 0.11 (0.11) |
| | 34px | | 0.13 (0.13) | | 0.13 (0.14) | | 0.15 (0.12) | | 0.08 (0.08) | | 0.12 (0.09) | | 0.10 (0.10) |
| | 58px | | 0.12 (0.13) | | 0.09 (0.11) | | 0.14 (0.09) | | 0.08 (0.09) | | 0.12 (0.09) | | 0.09 (0.08) |
| Random | 17% | | 0.15 (0.16) | | 0.13 (0.14) | | 0.13 (0.11) | | 0.10 (0.12) | | 0.10 (0.09) | | 0.08 (0.09) |
| | 28% | | 0.10 (0.12) | | 0.11 (0.11) | | 0.12 (0.10) | | 0.07 (0.07) | | 0.11 (0.09) | | 0.07 (0.08) |
| | 45% | | 0.10 (0.11) | | 0.09 (0.09) | | 0.12 (0.09) | | 0.06 (0.07) | | 0.12 (0.08) | | 0.07 (0.07) |

# Summary of the results

- 76% of the adversarial examples generated with the regional attack used in this research maintain model-to-model transferability
- 99% of the “regional” adversarial examples have reduced $L_0$ norms
- 75% of the “regional” adversarial examples have reduced $L_2$ norms
- 43% of the “regional” adversarial examples have reduced $L_\infty$ norms

# Conclusions

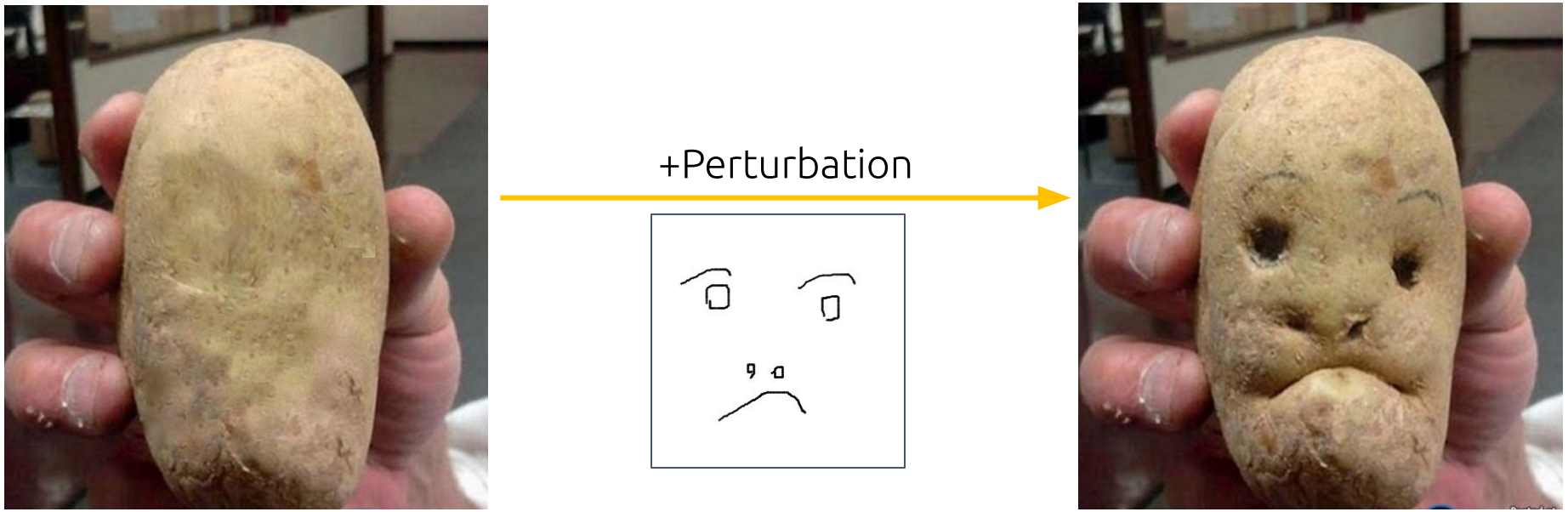
- We have proposed a simple and general method for localizing perturbations generated by adversarial attacks in specific regions
- Our method is experimentally confirmed to be effective, maintaining high black-box transferability at distortion levels that are significantly lower than the levels required by existing attacks
- The reduction in the amount of perturbation achieved by our method raises the concern that existing adversarial defenses may be undermined

# Directions for future work

- Investigate to what extent our localization method can fool state-of-the-art adversarial defenses

- Identify regions of importance where this localized perturbation can be made more effective, linking the observations made in this study to the interpretability of DNNs

# Any questions?




Regular potato

Adversarial potato