ICCV 2017

Adversarial Examples for Semantic Segmentation and Object Detection

Cihang Xie, Jianyu Wang, Zhishuai Zhang,
Yuyin Zhou, Lingxi Xie, Alan Yuille

Department of Computer Science
The Johns Hopkins University

Outline

- Introduction
- Adversarial Examples in Computer Vision
- Dense Adversarial Generation (DAG)
- Experiments: White-box Attack
- Experiments: Black-box Attack
- Fancy Examples
- Conclusions and Future Work

Outline

- Introduction
- Adversarial Examples in Computer Vision
- Dense Adversarial Generation (DAG)
- Experiments: White-box Attack
- Experiments: Black-box Attack
- Fancy Examples
- Conclusions and Future Work

Introduction

- Deep Learning
  - The state-of-the-art machine learning theory
  - Using a cascade of many layers of non-linear neurons for feature extraction and transformation
  - Learning multiple levels of feature representation
    - Higher-level features are derived from lower-level features to form a hierarchical architecture
    - Multiple levels of representation correspond to different levels of abstraction

Introduction (cont.)

- The Convolutional Neural Networks
  - A fundamental machine learning tool
  - Good performance in a wide range of problems in computer vision as well as other research areas
  - Evolutions in many real-world applications
  - Theory: a multi-layer, hierarchical network often has a larger capacity, also requires a larger amount of data to get trained

Outline

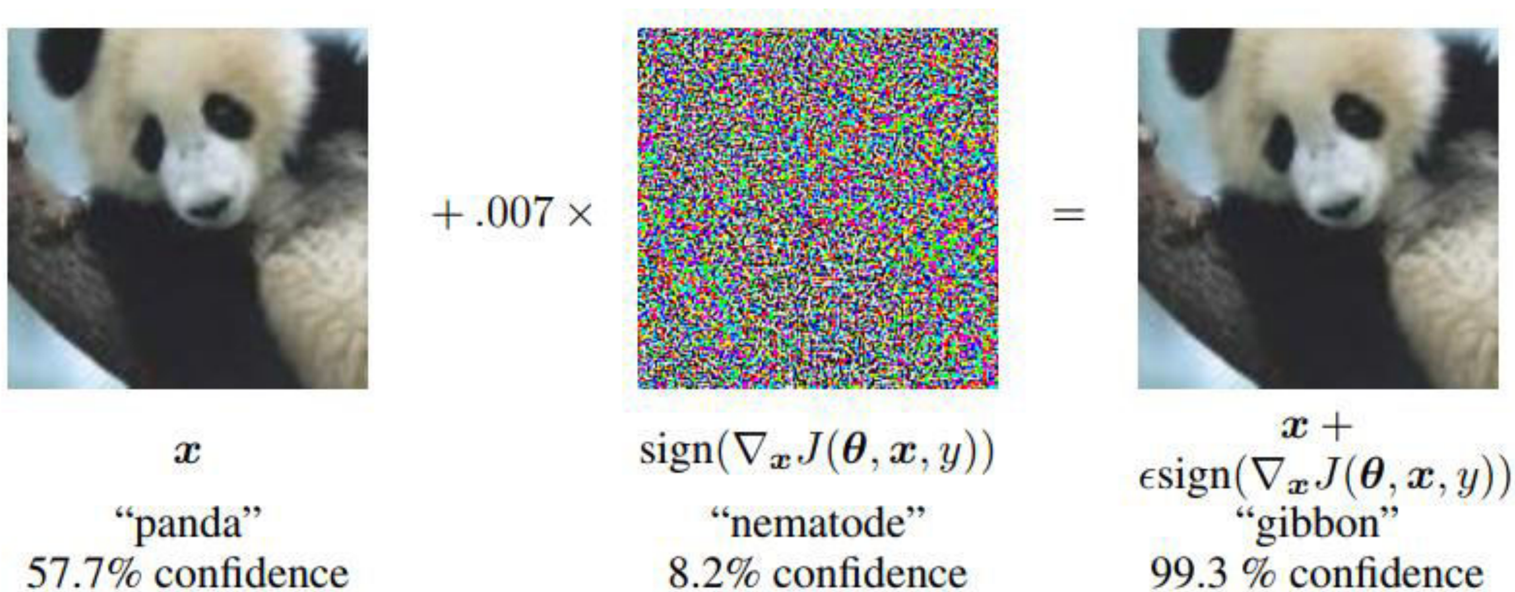
- Introduction
- Adversarial Examples in Computer Vision
- Dense Adversarial Generation (DAG)
- Experiments: White-box Attack
- Experiments: Black-box Attack
- Fancy Examples
- Conclusions and Future Work

Adversarial Examples: Introduction

- What is an adversarial example (in this work)?
  - An image, with a small perturbation added, which can still be recognized by humans, but not by the computers (*deep neural networks*)
  - Type 1: an image with clear visual contents is recognized incorrectly
  - Type 2: an image with no visual contents is recognized as a non-understandable class

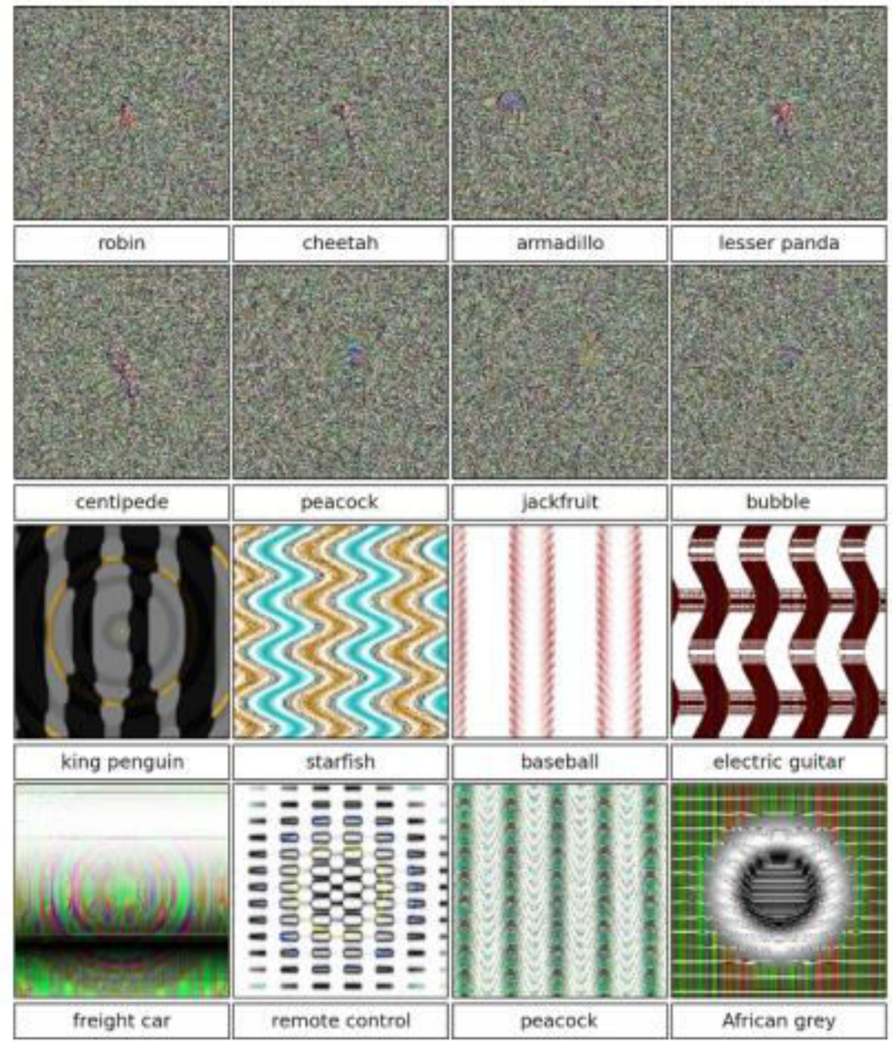
Adversarial Examples: Type 1

- Slightly perturbed natural images that are completely wrongly recognized
  - Example from [Goodfellow, ICLR'15]



Adversarial Examples: Type 2

- Meaningless patterns that are recognized as object classes with a very high confidence
  - Examples from [Nguyen, CVPR'14]



Previous Work

- Generating adversarial examples
  - Steepest gradient descent [Szegedy, ICLR'14], gradient sign [Goodfellow, ICLR'15], universal adversarial attack [M-Dezfooli, CVPR'17], *etc.*
- Defending adversarial examples
  - Distillation [Papernot, IEEE-SSP'16], large-scale learning [Kukarin, ICLR'17], ensemble [Tramer, arXiv'17], detection [Metzen, ICLR'17], randomization [Xie, arXiv’17], *etc.*

Why Adversaries Exist?

- Opinion 1: deep networks are too complicated so that the high-dimensional space contains many non-linear or unexplainable structures, or they are too sensitive to small noise
- Opinion 2: deep networks are still too simple to defend these malignant attacks
- Opinion 3: deep networks are not the model we want!

Our Contribution

- We extend the adversarial examples to both semantic segmentation and object detection
  - We are the first to achieve this goal systematically
- We evaluated both *white-box* attack and *black-box* attack tasks
  - White-box: the network parameters are known
  - Black-box: the network parameters are unknown (transferring the adversarial perturbations)

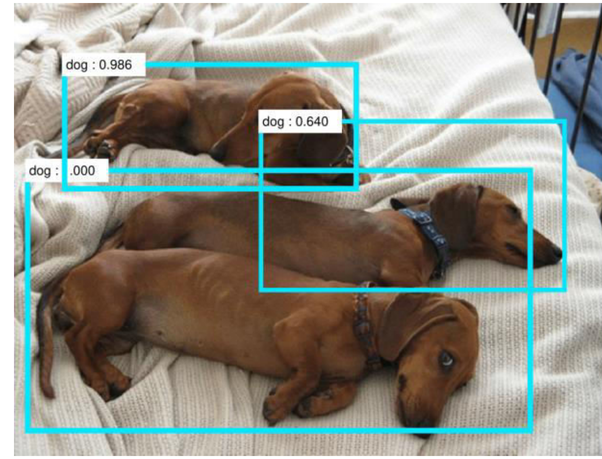
Outline

- Introduction
- Adversarial Examples in Computer Vision
- Dense Adversarial Generation (DAG)
- Experiments: White-box Attack
- Experiments: Black-box Attack
- Fancy Examples
- Conclusions and Future Work

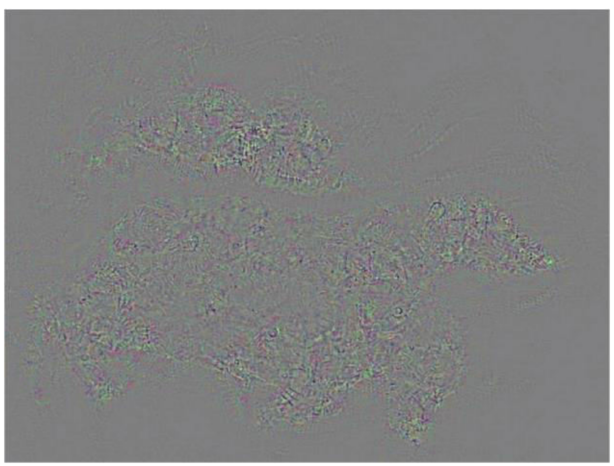
Some Typical Results

Original Image

Original Segmentation

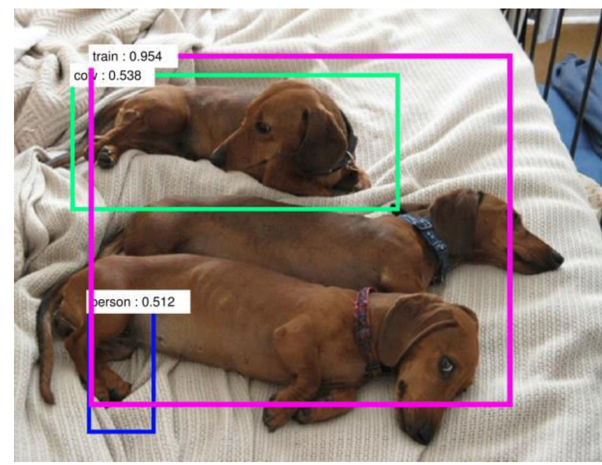


Original Detection



Added Perturbation (10x)

Attacked Segmentation



Attacked Detection

Formulation: Optimization Goal

- Let a deep network be $\mathbf{f}(\mathbf{X};\mathbf{\Theta}) \in \mathbb{R}^C$
  - $\mathbf{X}$: input region, $\mathbf{\Theta}$: weights (fixed), C: # of classes
- Goal: modifying $\mathbf{X}$ to make wrong prediction
- Optimization *target*: the basic unit
  - For classification: the entire image (previous work)
  - What about segmentation?
  - What about detection?

Formulation: Optimization Goal

- Let a deep network be $\mathbf{f}(\mathbf{X}; \boldsymbol{\Theta}) \in \mathbb{R}^C$
  - $\mathbf{X}$: input region, $\boldsymbol{\Theta}$: weights (fixed), C: # of classes
- Goal: modifying $\mathbf{X}$ to make wrong prediction
- Optimization *target*: the basic unit
  - For classification: the entire image (previous work)
  - For segmentation: all pixels in the image
  - For detection: densely distributed bounding boxes

Dense Adversarial Generation

- A white-box attack
  - Image and network dependent
- Flowchart
  - Defining the active set
  - Gradient descent
  - Until convergence

Algorithm 1: Dense Adversary Generation (DAG)

Input : input image $\mathbf{X}$;
the classifier $\mathbf{f}(\cdot,\cdot) \in \mathbb{R}^C$;
the target set $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$;
the original label set $\mathcal{L} = \{l_1, l_2, \ldots, l_N\}$;
the adversarial label set $\mathcal{L}' = \{l'_1, l'_2, \ldots, l'_N\}$;
the maximal iterations M_0;

Output: the adversarial perturbation $\mathbf{r}$;

$\mathbf{X}_0 \leftarrow \mathbf{X}$, $\mathbf{r} \leftarrow \mathbf{0}$, $m \leftarrow 0$, $\mathcal{T}_0 \leftarrow \mathcal{T}$;
while $m < M_0$ **and** $\mathcal{T}_m \neq \varnothing$ **do**
  $\mathcal{T}_m = \{t_n \mid \arg\max_c \{f_c(\mathbf{X}_m, t_n)\} = l_n\}$;
  $\mathbf{r}_m \leftarrow \sum_{t_n \in \mathcal{T}_m} \left[\nabla_{\mathbf{X}_m} f_{l'_n}(\mathbf{X}_m, t_n) - \nabla_{\mathbf{X}_m} f_{l_n}(\mathbf{X}_m, t_n)\right]$;
  $\mathbf{r}'_m \leftarrow \frac{\gamma}{\|\mathbf{r}_m\|_\infty} \mathbf{r}_m$;
  $\mathbf{r} \leftarrow \mathbf{r} + \mathbf{r}'_m$;
  $\mathbf{X}_{m+1} \leftarrow \mathbf{X}_m + \mathbf{r}'_m$;
  $m \leftarrow m + 1$;
end

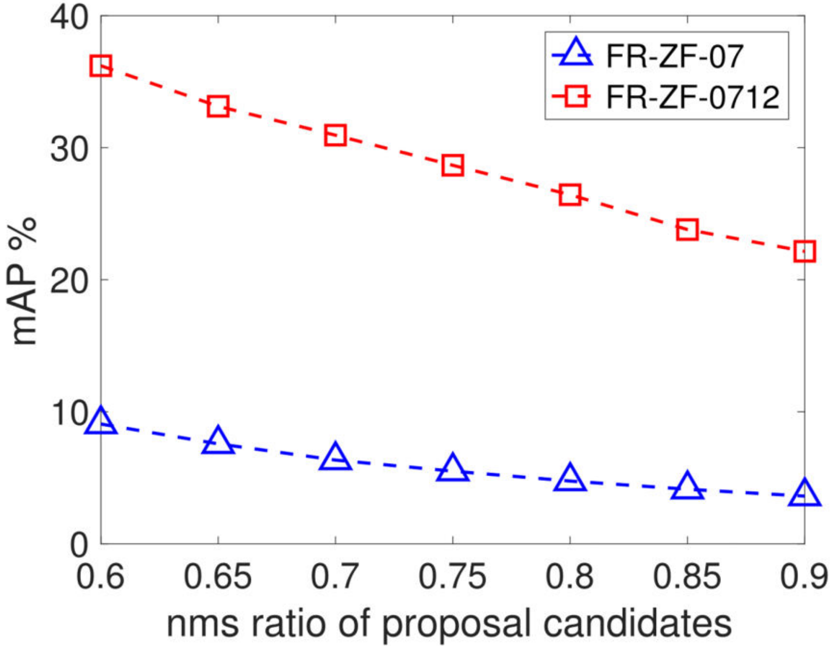
Return: $\mathbf{r}$

Comments on Object Detection

- We attacked a type of frameworks, which first extract a number of proposals, then assign a class label for each proposal
- A possibility: the adversarial perturbation changes the set of proposals, and our attack will not work on the new proposals
  - That is why we need to generate *dense* bounding boxes (see the next page)

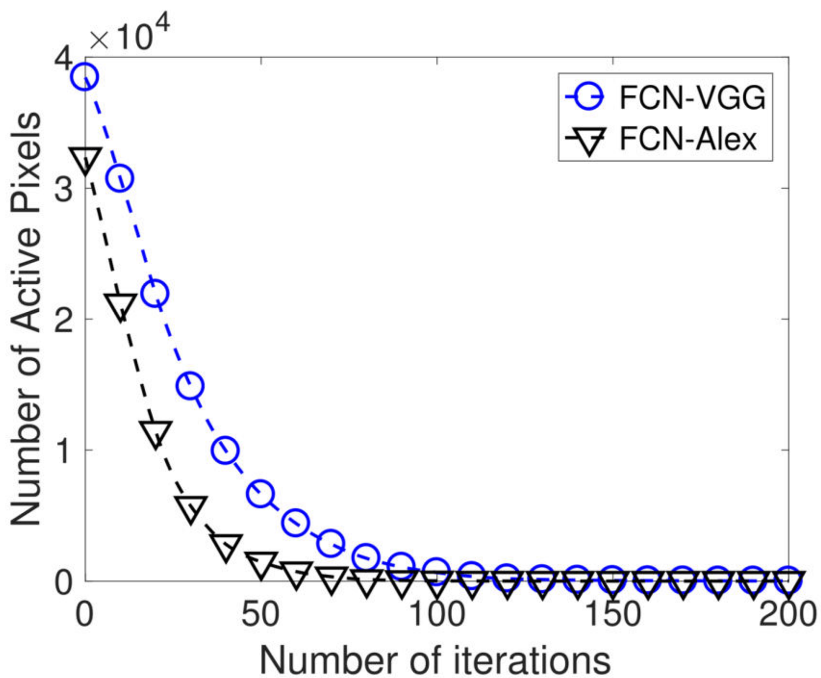
Diagnosis: Denseness

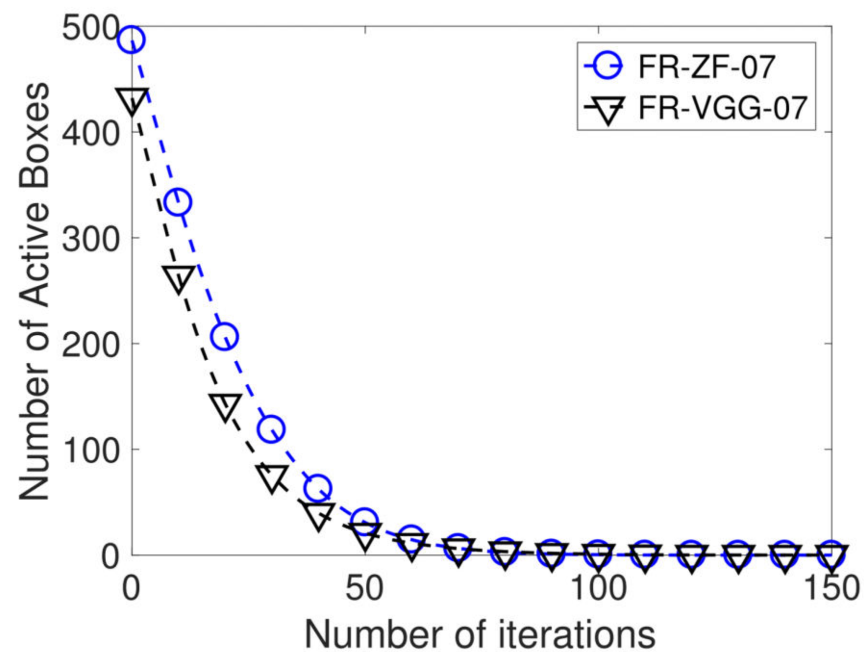
- Denseness: the number of generated boxes in *object detection* task (the more the better)
  - Controlled by the non-maximum-suppression ratio



Diagnosis: Convergence

- Convergence is mostly guaranteed
  - Failed to converge in a fixed # of rounds: $< 1\%$
    - Even in these cases, generated perturbations work well





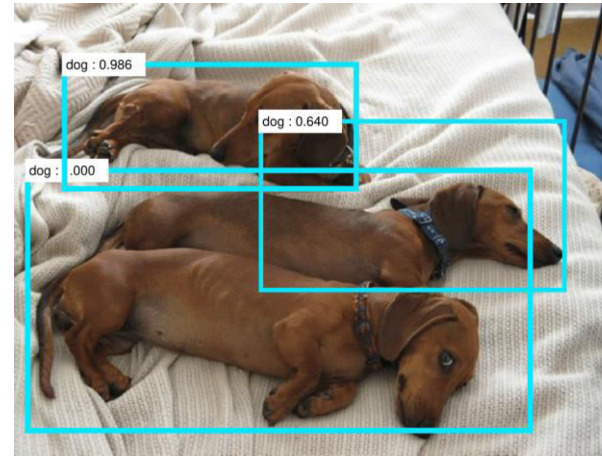
Diagnosis: Perceptibility

- Low intensity of adversarial perturbations
- Perceptibility: $p = \left(\frac{1}{K}\sum_k \|\mathbf{r}_k\|_2^2\right)^{1/2}$
  - K: # of image pixels
  - $\mathbf{r}_k$: RGB vector of perturbation ($[0,1]$-normalized)
- Typical value of p is $[1.0,3.0] \times 10^{-3}$

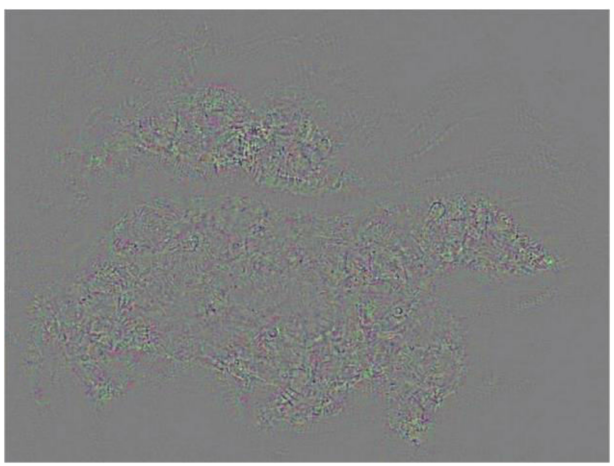
Some Typical Results

Original Image

Original Segmentation

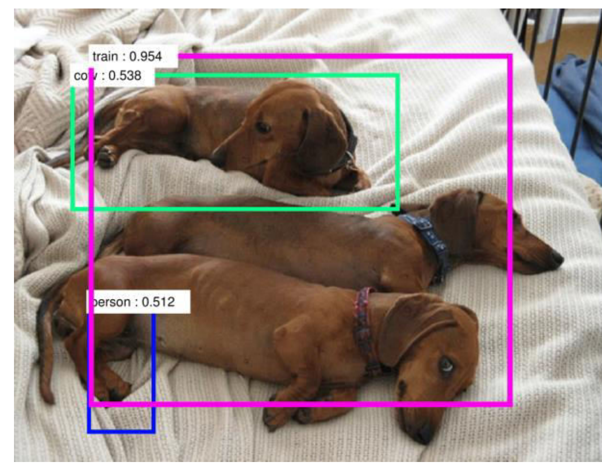


Original Detection



Added Perturbation (10x)

Attacked Segmentation



Attacked Detection

Outline

- Introduction
- Adversarial Examples in Computer Vision
- Dense Adversarial Generation (DAG)
- Experiments: White-box Attack
- Experiments: Black-box Attack
- Fancy Examples
- Conclusions and Future Work

White-Box Attack: Definition

- Given an image $\mathbf{X}$ and a network $\mathbf{f}(\mathbf{X};\mathbf{\Theta})$ in which the structure and weights are *known*
  - This is the same setting as in the algorithm
  - Adversarial examples are easily generated, given that our algorithm converges (mostly guaranteed)

White-Box Attack: Results

- Semantic segmentation part
  - FCN and DeepLab were evaluated
  - Bold numbers indicate white-box attacks

Adversarial Perturbations from	**FCN-Alex**	**FCN-Alex***	**FCN-VGG**	**FCN-VGG***	**DL-VGG**	**DL-RN101**
None	48.04	48.92	65.49	67.09	70.72	76.11
FCN-Alex ($\mathbf{r}_5$)	**3.98**	7.94	64.82	66.54	70.18	75.45
FCN-Alex* ($\mathbf{r}_6$)	5.10	**3.98**	64.60	66.36	69.98	75.52
FCN-VGG ($\mathbf{r}_7$)	46.21	47.38	**4.09**	16.36	45.16	73.98
FCN-VGG* ($\mathbf{r}_8$)	46.10	47.21	12.72	**4.18**	46.33	73.76
$\mathbf{r}_5 + \mathbf{r}_7$	**4.83**	8.55	**4.23**	17.59	43.95	73.26
$\mathbf{r}_5 + \mathbf{r}_7$ (permuted)	48.03	48.90	65.47	67.09	70.69	76.04
$\mathbf{r}_6 + \mathbf{r}_8$	5.52	**4.23**	13.89	**4.98**	44.18	73.01
$\mathbf{r}_6 + \mathbf{r}_8$ (permuted)	48.03	48.90	65.47	67.05	70.69	76.05

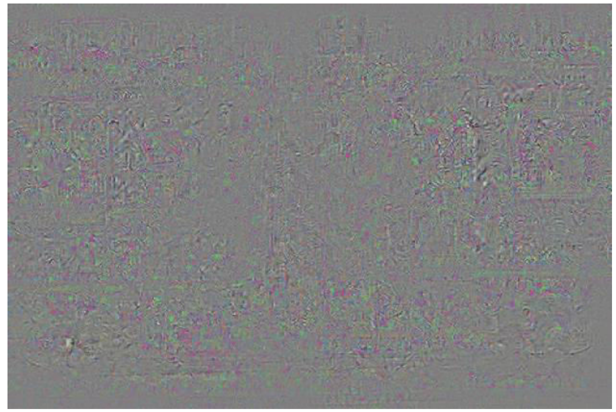
White-Box Attack: Results

- Object recognition part
  - Faster-RCNN and R-FCN were evaluated
  - Bold numbers indicate white-box attacks

Adversarial Perturbations from	FR-ZF-07	FR-ZF-0712	FR-VGG-07	FR-VGG-0712	R-FCN-RN50	R-FCN-RN101
None	58.70	61.07	69.14	72.07	76.40	78.06
FR-ZF-07 ($\mathbf{r}_1$)	**3.61**	22.15	66.01	69.47	74.01	75.87
FR-ZF-0712 ($\mathbf{r}_2$)	13.14	**1.95**	64.61	68.17	72.29	74.68
FR-VGG-07 ($\mathbf{r}_3$)	56.41	59.31	**5.92**	48.05	72.84	74.79
FR-VGG-0712 ($\mathbf{r}_4$)	56.09	58.58	31.84	**3.36**	70.55	72.78
$\mathbf{r}_1 + \mathbf{r}_3$	**3.98**	21.63	**7.00**	44.14	68.89	71.56
$\mathbf{r}_1 + \mathbf{r}_3$ (permuted)	58.30	61.08	68.63	71.82	76.34	77.71
$\mathbf{r}_2 + \mathbf{r}_4$	13.15	**2.13**	28.92	**4.28**	63.93	67.25
$\mathbf{r}_2 + \mathbf{r}_4$ (permuted)	58.51	61.09	68.68	71.78	76.23	77.71

White-Box Attack: Examples

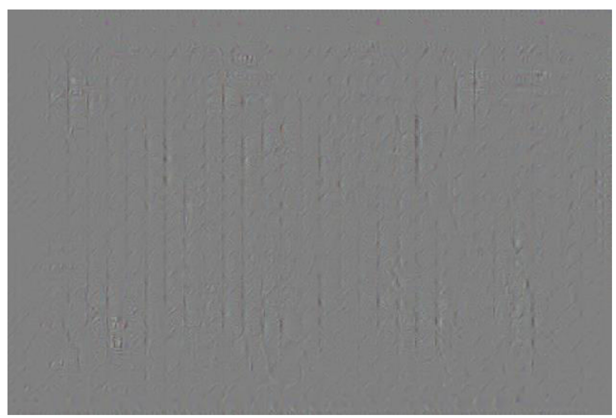
Original Image



Added Perturbation (10x)

Attacked Segmentation

Original Image



Added Perturbation (10x)

Attacked Segmentation

Outline

- Introduction
- Adversarial Examples in Computer Vision
- Dense Adversarial Generation (DAG)
- Experiments: White-box Attack
- Experiments: Black-box Attack
- Fancy Examples
- Conclusions and Future Work

Black-Box Attack: Definition

- Given an image $\mathbf{X}$ and a network $\mathbf{f}(\mathbf{X};\boldsymbol{\Theta})$ in which the structure and weights are *unknown*
  - It is even possible that the task is unknown
  - This setting is much more challenging
  - The difficulty goes up with the difference between the source network (the white box) and the target network (the black box)

Black-Box Attack: Results

- Semantic segmentation part
  - Transfer across different training sets
  - Transfer across different networks

Adversarial Perturbations from	FCN-Alex	FCN-Alex*	FCN-VGG	FCN-VGG*	DL-VGG	DL-RN101
None	48.04	48.92	65.49	67.09	70.72	76.11
FCN-Alex ($\mathbf{r}_5$)	**3.98**	7.94	64.82	66.54	70.18	75.45
FCN-Alex* ($\mathbf{r}_6$)	5.10	**3.98**	64.60	66.36	69.98	75.52
FCN-VGG ($\mathbf{r}_7$)	46.21	47.38	**4.09**	16.36	45.16	73.98
FCN-VGG* ($\mathbf{r}_8$)	46.10	47.21	12.72	**4.18**	46.33	73.76
$\mathbf{r}_5 + \mathbf{r}_7$	**4.83**	8.55	**4.23**	17.59	43.95	73.26
$\mathbf{r}_5 + \mathbf{r}_7$ (permuted)	48.03	48.90	65.47	67.09	70.69	76.04
$\mathbf{r}_6 + \mathbf{r}_8$	5.52	**4.23**	13.89	**4.98**	44.18	73.01
$\mathbf{r}_6 + \mathbf{r}_8$ (permuted)	48.03	48.90	65.47	67.05	70.69	76.05

Black-Box Attack: Results

- Object recognition part
  - Transfer across different training sets
  - Transfer across different networks

Adversarial Perturbations from	**FR-ZF-07**	**FR-ZF-0712**	**FR-VGG-07**	**FR-VGG-0712**	**R-FCN-RN50**	**R-FCN-RN101**
None	58.70	61.07	69.14	72.07	76.40	78.06
FR-ZF-07 ($\mathbf{r}_1$)	**3.61**	22.15	66.01	69.47	74.01	75.87
FR-ZF-0712 ($\mathbf{r}_2$)	13.14	**1.95**	64.61	68.17	72.29	74.68
FR-VGG-07 ($\mathbf{r}_3$)	56.41	59.31	**5.92**	48.05	72.84	74.79
FR-VGG-0712 ($\mathbf{r}_4$)	56.09	58.58	31.84	**3.36**	70.55	72.78
$\mathbf{r}_1 + \mathbf{r}_3$	**3.98**	21.63	**7.00**	44.14	68.89	71.56
$\mathbf{r}_1 + \mathbf{r}_3$ (permuted)	58.30	61.08	68.63	71.82	76.34	77.71
$\mathbf{r}_2 + \mathbf{r}_4$	13.15	**2.13**	28.92	**4.28**	63.93	67.25
$\mathbf{r}_2 + \mathbf{r}_4$ (permuted)	58.51	61.09	68.68	71.78	76.23	77.71

Black-Box Attack: Results

- Transfer across different tasks
  - This is the most challenging task investigated
  - Ensemble is the only way of enhancing attack

Adversarial Perturbations from	FR-ZF-07	FR-VGG-07	FCN-Alex	FCN-VGG	R-FCN-RN101
None	56.83	68.88	35.73	54.87	80.20
FR-ZF-07 ($\mathbf{r}_1$)	**5.14**	66.63	31.74	51.94	76.00
FR-VGG-07 ($\mathbf{r}_3$)	54.96	**7.17**	34.53	43.06	74.50
FCN-Alex ($\mathbf{r}_5$)	55.61	68.62	**4.04**	54.08	77.09
FCN-VGG ($\mathbf{r}_7$)	55.24	56.33	33.99	**4.10**	73.86
$\mathbf{r}_1+\mathbf{r}_3+\mathbf{r}_5$	**5.02**	**8.75**	**4.32**	37.90	69.07
$\mathbf{r}_1+\mathbf{r}_3+\mathbf{r}_7$	**5.15**	**5.63**	28.48	**4.81**	65.23
$\mathbf{r}_1+\mathbf{r}_5+\mathbf{r}_7$	**5.14**	47.52	**4.37**	**5.20**	68.51
$\mathbf{r}_3+\mathbf{r}_5+\mathbf{r}_7$	53.34	**5.94**	**4.41**	**4.68**	67.57
$\mathbf{r}_1+\mathbf{r}_3+\mathbf{r}_5+\mathbf{r}_7$	**5.05**	**5.89**	**4.51**	**5.09**	64.52

Black-Box Attack: Facts

- Black-box attack is much more difficult
  - The difficulty goes up with the difference between the source and target networks
- "Difficulty levels" in transfer
  - Level 1: across different datasets
  - Level 2: across different network structures
    - Shallower networks are not easier to attack
  - Level 3: across different vision tasks
    - Same network structure makes things easier

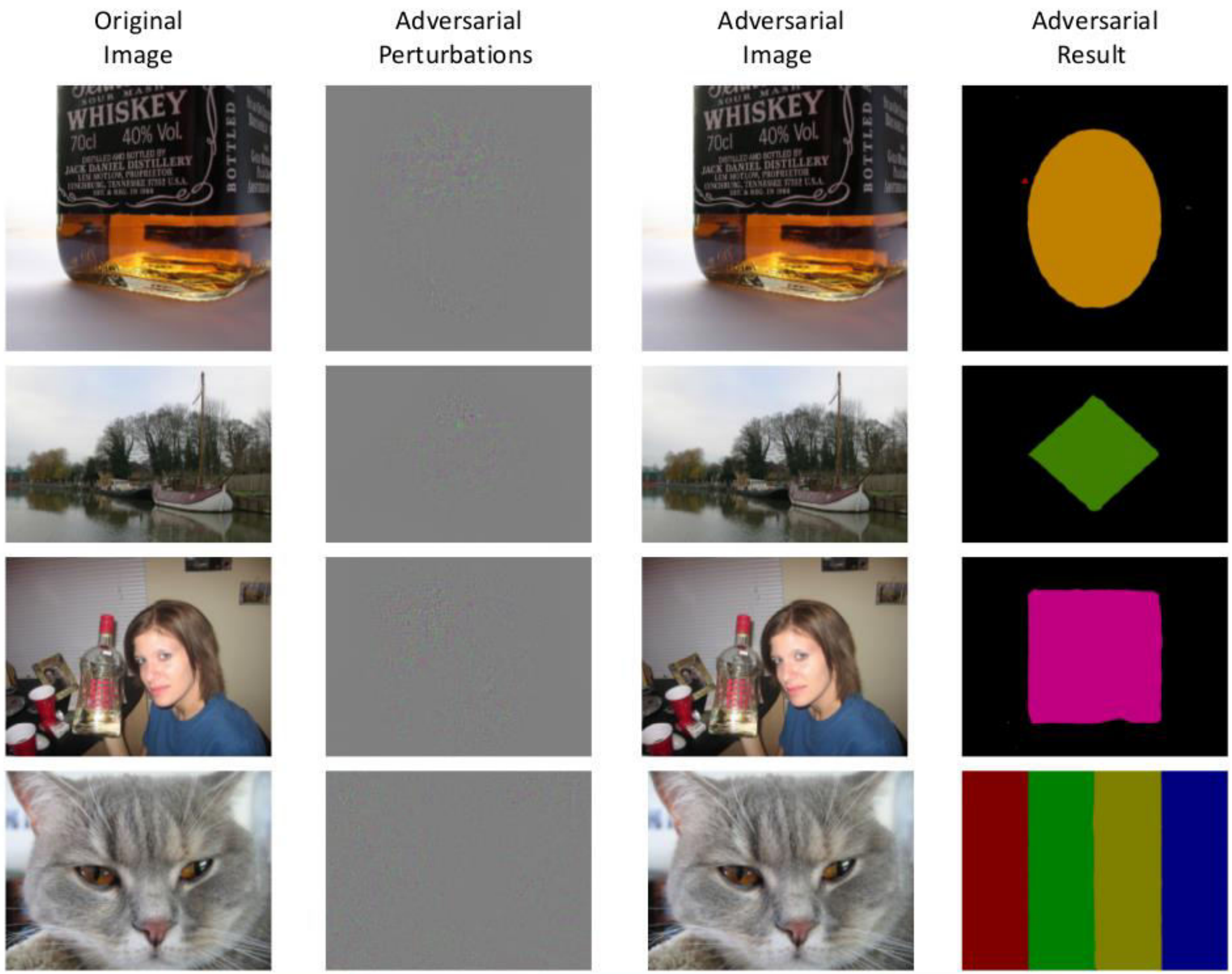
Black-Box Attack: Examples

	Original Image	Original Result from Network 2	Adversarial Result from Network 1	Adversarial Result from Network 2
Network 1: FR-ZF-0712 Network 2: FR-VGG-07				
Network 1: FCN-VGG Network 2: DL-VGG				
Network 1: FCN-VGG Network 2: FR-VGG-07				
Network 1: FR-VGG-07 Network 2: FCN-VGG				

Outline

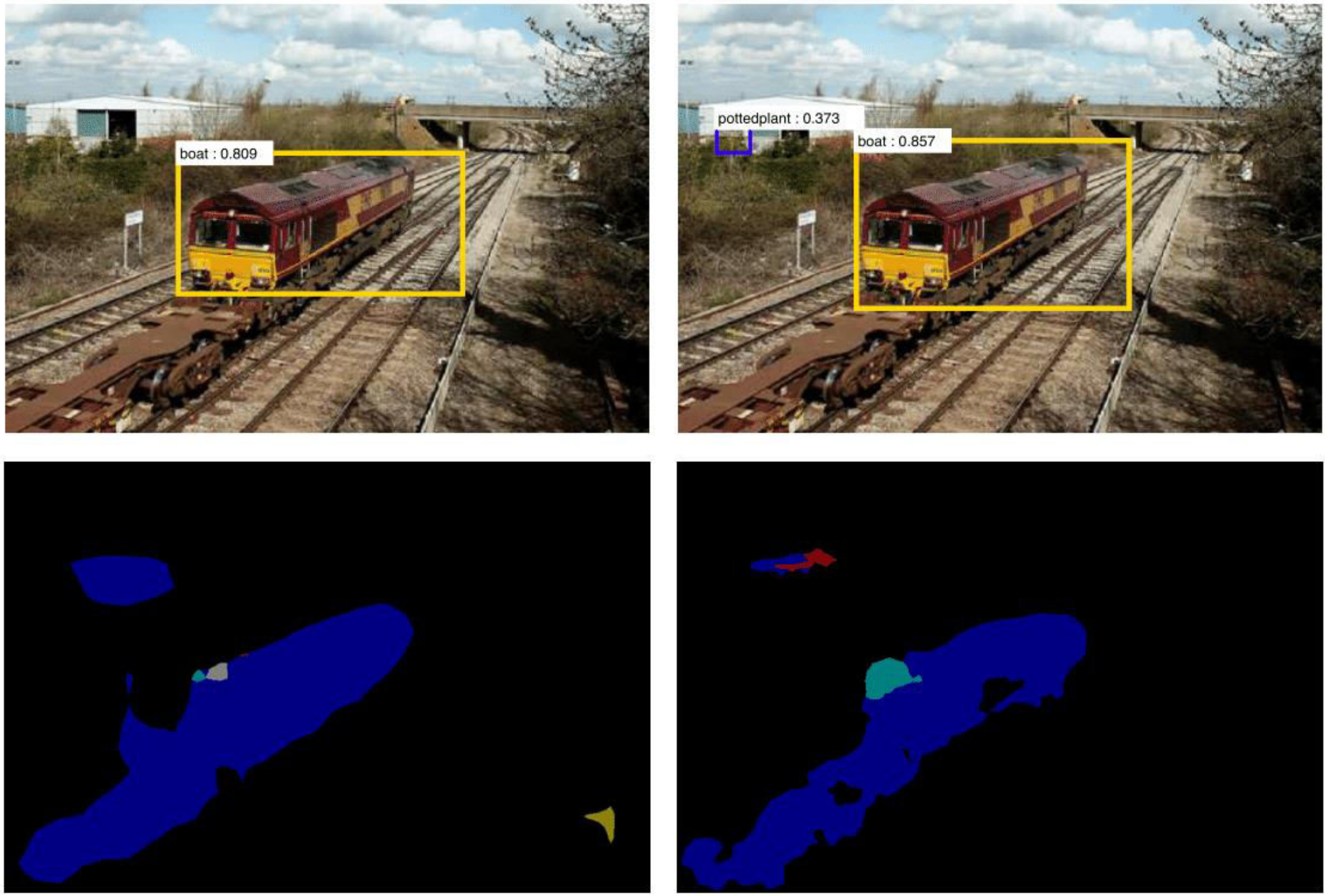
- Introduction
- Adversarial Examples in Computer Vision
- Dense Adversarial Generation (DAG)
- Experiments: White-box Attack
- Experiments: Black-box Attack
- Fancy Examples
- Conclusions and Future Work

Different Geometric Patterns



B-ground	Aero plane	Bicycle	Bird	Boat	Bottle	Bus
Car	Cat	Chair	Cow	Dining-Table	Dog	Horse
Motorbike	Person	Potted-Plant	Sheep	Sofa	Train	TV/Monitor

An adversarial example for both detection and segmentation



The top row shows FR-VGG-07 and FR-ZF-07 detection results, and the bottom row shows FCN-Alex and FCN-VGG segmentation results. The blue in segmentation results corresponds to boat.

Same adversarial example, Completely different Outputs

Original Image | Adversarial Perturbations | Adversarial Image | Adversarial Result from FCN-Alex | Adversarial Result from FCN-VGG

We add one adversarial perturbation (magnified by 10) to the same original image to generate different pre-specified segmentation masks on two deep segmentation networks (FCN-Alex and FCN-VGG). This is a more difficult task compared to that shown in previous figure, where two different adversarial perturbations are used to generate two pre-specified segmentation masks. The blue regions in the segmentation masks are predicted as bus, a randomly selected class.

Outline

- Introduction
- Adversarial Examples in Computer Vision
- Dense Adversarial Generation (DAG)
- Experiments: White-box Attack
- Experiments: Black-box Attack
- Fancy Examples
- Conclusions and Future Work

Conclusions

- Adversarial examples exist in both semantic segmentation and object detection
  - A simple algorithm based on gradient descent
  - The target can be arbitrary to some extents
- White-box attack: efficient and effective
- Black-box attack: a more challenging problem
  - Transfer across datasets, networks and tasks
  - Ensemble is an effective solution

Future Work

- Defending adversarial attacks
  - Attack vs. defense: which one is stronger?
- Finding out the reason of adversarial examples in the context of deep neural networks
- Integrating adversarial examples in training deep neural networks

Thank you!

Questions please?