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 \( f(X; \Theta) \in \mathbb{R}^C \)
  – \( X \): input region, \( \Theta \): weights (fixed), \( C \): # of classes

• Goal: modifying \( X \) to make wrong prediction

• Optimization \textit{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 $f(\mathbf{X}; \Theta) \in \mathbb{R}^C$
  – $\mathbf{X}$: input region, $\Theta$: weights (fixed), $C$: # of classes
• Goal: modifying $\mathbf{X}$ to make wrong prediction
• Optimization $\textit{target}$: 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 $X$;
      the classifier $f(\cdot, \cdot) \in \mathbb{R}^C$;
      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 $r$;

1. $X_0 \leftarrow X$, $r \leftarrow 0$, $m \leftarrow 0$, $\mathcal{T}_0 \leftarrow \mathcal{T}$;
2. while $m < M_0$ and $\mathcal{T}_m \neq \emptyset$ do
3.   $\mathcal{T}_m = \{t_n \mid \arg \max_c \{f_c(X_m, t_n)\} = l_n\}$
4.   $r_m \leftarrow \sum_{t_n \in \mathcal{T}_m} \left[ \nabla X_m f'_n(X_m, t_n) - \nabla X_m f_n(X_m, t_n) \right]$;
5.   $r'_m \leftarrow \frac{r_m}{\|r_m\|_\infty} r_m$;
6.   $r \leftarrow r + r'_m$;
7.   $X_{m+1} \leftarrow X_m + r'_m$;
8.   $m \leftarrow m + 1$;
9. end

Return: $r$
```
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: \( p = \left( \frac{1}{K} \sum_k \| r_k \|_2^2 \right)^{1/2} \)
  – \( K \): # of image pixels
  – \( r_k \): RGB vector of perturbation ([0,1]-normalized)

• Typical value of \( p \) is \([1.0,3.0] \times 10^{-3}\)
Some Typical Results
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 $X$ and a network $f(X; \Theta)$ 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

<table>
<thead>
<tr>
<th>Adversarial Perturbations from</th>
<th>FCN-Alex</th>
<th>FCN-Alex*</th>
<th>FCN-VGG</th>
<th>FCN-VGG*</th>
<th>DL-VGG</th>
<th>DL-RN101</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>48.04</td>
<td>48.92</td>
<td>65.49</td>
<td>67.09</td>
<td>70.72</td>
<td>76.11</td>
</tr>
<tr>
<td>FCN-Alex (r₅)</td>
<td>3.98</td>
<td>7.94</td>
<td>64.82</td>
<td>66.54</td>
<td>70.18</td>
<td>75.45</td>
</tr>
<tr>
<td>FCN-Alex* (r₆)</td>
<td>5.10</td>
<td>3.98</td>
<td>64.60</td>
<td>66.36</td>
<td>69.98</td>
<td>75.52</td>
</tr>
<tr>
<td>FCN-VGG (r₇)</td>
<td>46.21</td>
<td>47.38</td>
<td>4.09</td>
<td>16.36</td>
<td>45.16</td>
<td>73.98</td>
</tr>
<tr>
<td>FCN-VGG* (r₈)</td>
<td>46.10</td>
<td>47.21</td>
<td>12.72</td>
<td>4.18</td>
<td>46.33</td>
<td>73.76</td>
</tr>
<tr>
<td>r₅ + r₇</td>
<td>4.83</td>
<td>8.55</td>
<td>4.23</td>
<td>17.59</td>
<td>43.95</td>
<td>73.26</td>
</tr>
<tr>
<td>r₅ + r₇ (permuted)</td>
<td>48.03</td>
<td>48.90</td>
<td>65.47</td>
<td>67.09</td>
<td>70.69</td>
<td>76.04</td>
</tr>
<tr>
<td>r₆ + r₈</td>
<td>5.52</td>
<td>4.23</td>
<td>13.89</td>
<td>4.98</td>
<td>44.18</td>
<td>73.01</td>
</tr>
<tr>
<td>r₆ + r₈ (permuted)</td>
<td>48.03</td>
<td>48.90</td>
<td>65.47</td>
<td>67.05</td>
<td>70.69</td>
<td>76.05</td>
</tr>
</tbody>
</table>
White-Box Attack: Results

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>58.70</td>
<td>61.07</td>
<td>69.14</td>
<td>72.07</td>
<td>76.40</td>
<td>78.06</td>
</tr>
<tr>
<td>FR-ZF-07 (r₁)</td>
<td>3.61</td>
<td>22.15</td>
<td>66.01</td>
<td>69.47</td>
<td>74.01</td>
<td>75.87</td>
</tr>
<tr>
<td>FR-ZF-0712 (r₂)</td>
<td>13.14</td>
<td>1.95</td>
<td>64.61</td>
<td>68.17</td>
<td>72.29</td>
<td>74.68</td>
</tr>
<tr>
<td>FR-VGG-07 (r₃)</td>
<td>56.41</td>
<td>59.31</td>
<td>5.92</td>
<td>48.05</td>
<td>72.84</td>
<td>74.79</td>
</tr>
<tr>
<td>FR-VGG-0712 (r₄)</td>
<td>56.09</td>
<td>58.58</td>
<td>31.84</td>
<td>3.36</td>
<td>70.55</td>
<td>72.78</td>
</tr>
<tr>
<td>r₁ + r₃</td>
<td>3.98</td>
<td>21.63</td>
<td>7.00</td>
<td>44.14</td>
<td>68.89</td>
<td>71.56</td>
</tr>
<tr>
<td>r₁ + r₃ (permuted)</td>
<td>58.30</td>
<td>61.08</td>
<td>68.63</td>
<td>71.82</td>
<td>76.34</td>
<td>77.71</td>
</tr>
<tr>
<td>r₂ + r₄</td>
<td>13.15</td>
<td>2.13</td>
<td>28.92</td>
<td>4.28</td>
<td>63.93</td>
<td>67.25</td>
</tr>
<tr>
<td>r₂ + r₄ (permuted)</td>
<td>58.51</td>
<td>61.09</td>
<td>68.68</td>
<td>71.78</td>
<td>76.23</td>
<td>77.71</td>
</tr>
</tbody>
</table>
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 $X$ and a network $f(X; \Theta)$ 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

<table>
<thead>
<tr>
<th>Adversarial Perturbations from</th>
<th>FCN-Alex</th>
<th>FCN-Alex*</th>
<th>FCN-VGG</th>
<th>FCN-VGG*</th>
<th>DL-VGG</th>
<th>DL-RN101</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>48.04</td>
<td>48.92</td>
<td>65.49</td>
<td>67.09</td>
<td>70.72</td>
<td>76.11</td>
</tr>
<tr>
<td>FCN-Alex ($r_5$)</td>
<td>3.98</td>
<td>7.94</td>
<td>64.82</td>
<td>66.54</td>
<td>70.18</td>
<td>75.45</td>
</tr>
<tr>
<td>FCN-Alex* ($r_6$)</td>
<td>5.10</td>
<td>3.98</td>
<td>64.60</td>
<td>66.36</td>
<td>69.98</td>
<td>75.52</td>
</tr>
<tr>
<td>FCN-VGG ($r_7$)</td>
<td>46.21</td>
<td>47.38</td>
<td>4.09</td>
<td>16.36</td>
<td>45.16</td>
<td>73.98</td>
</tr>
<tr>
<td>FCN-VGG* ($r_8$)</td>
<td>46.10</td>
<td>47.21</td>
<td>12.72</td>
<td>4.18</td>
<td>46.33</td>
<td>73.76</td>
</tr>
<tr>
<td>$r_5 + r_7$</td>
<td>4.83</td>
<td>8.55</td>
<td>4.23</td>
<td>17.59</td>
<td>43.95</td>
<td>73.26</td>
</tr>
<tr>
<td>$r_5 + r_7$ (permuted)</td>
<td>48.03</td>
<td>48.90</td>
<td>65.47</td>
<td>67.09</td>
<td>70.69</td>
<td>76.04</td>
</tr>
<tr>
<td>$r_6 + r_8$</td>
<td>5.52</td>
<td>4.23</td>
<td>13.89</td>
<td>4.98</td>
<td>44.18</td>
<td>73.01</td>
</tr>
<tr>
<td>$r_6 + r_8$ (permuted)</td>
<td>48.03</td>
<td>48.90</td>
<td>65.47</td>
<td>67.05</td>
<td>70.69</td>
<td>76.05</td>
</tr>
</tbody>
</table>
Black-Box Attack: Results

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>58.70</td>
<td>61.07</td>
<td>69.14</td>
<td>72.07</td>
<td>76.40</td>
<td>78.06</td>
</tr>
<tr>
<td>FR-ZF-07 (r₁)</td>
<td>3.61</td>
<td>22.15</td>
<td>66.01</td>
<td>69.47</td>
<td>74.01</td>
<td>75.87</td>
</tr>
<tr>
<td>FR-ZF-0712 (r₂)</td>
<td>13.14</td>
<td>1.95</td>
<td>64.61</td>
<td>68.17</td>
<td>72.29</td>
<td>74.68</td>
</tr>
<tr>
<td>FR-VGG-07 (r₃)</td>
<td>56.41</td>
<td>59.31</td>
<td>5.92</td>
<td>48.05</td>
<td>72.84</td>
<td>74.79</td>
</tr>
<tr>
<td>FR-VGG-0712 (r₄)</td>
<td>56.09</td>
<td>58.58</td>
<td>31.84</td>
<td>3.36</td>
<td>70.55</td>
<td>72.78</td>
</tr>
<tr>
<td>r₁ + r₃</td>
<td>3.98</td>
<td>21.63</td>
<td>7.00</td>
<td>44.14</td>
<td>68.89</td>
<td>71.56</td>
</tr>
<tr>
<td>r₁ + r₃ (permuted)</td>
<td>58.30</td>
<td>61.08</td>
<td>68.63</td>
<td>71.82</td>
<td>76.34</td>
<td>77.71</td>
</tr>
<tr>
<td>r₂ + r₄</td>
<td>13.15</td>
<td>2.13</td>
<td>28.92</td>
<td>4.28</td>
<td>63.93</td>
<td>67.25</td>
</tr>
<tr>
<td>r₂ + r₄ (permuted)</td>
<td>58.51</td>
<td>61.09</td>
<td>68.68</td>
<td>71.78</td>
<td>76.23</td>
<td>77.71</td>
</tr>
</tbody>
</table>
Black-Box Attack: Results

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

<table>
<thead>
<tr>
<th>Adversarial Perturbations from</th>
<th>FR-ZF-07</th>
<th>FR-VGG-07</th>
<th>FCN-Alex</th>
<th>FCN-VGG</th>
<th>R-FCN-RN101</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>56.83</td>
<td>68.88</td>
<td>35.73</td>
<td>54.87</td>
<td>80.20</td>
</tr>
<tr>
<td>FR-ZF-07 (r₁)</td>
<td>5.14</td>
<td>66.63</td>
<td>31.74</td>
<td>51.94</td>
<td>76.00</td>
</tr>
<tr>
<td>FR-VGG-07 (r₃)</td>
<td>54.96</td>
<td>7.17</td>
<td>34.53</td>
<td>43.06</td>
<td>74.50</td>
</tr>
<tr>
<td>FCN-Alex (r₅)</td>
<td>55.61</td>
<td>68.62</td>
<td>4.04</td>
<td>54.08</td>
<td>77.09</td>
</tr>
<tr>
<td>FCN-VGG (r₇)</td>
<td>55.24</td>
<td>56.33</td>
<td>33.99</td>
<td>4.10</td>
<td>73.86</td>
</tr>
<tr>
<td>r₁ + r₃ + r₅</td>
<td>5.02</td>
<td>8.75</td>
<td>4.32</td>
<td>37.90</td>
<td>69.07</td>
</tr>
<tr>
<td>r₁ + r₃ + r₇</td>
<td>5.15</td>
<td>5.63</td>
<td>28.48</td>
<td>4.81</td>
<td>65.23</td>
</tr>
<tr>
<td>r₁ + r₅ + r₇</td>
<td>5.14</td>
<td>47.52</td>
<td>4.37</td>
<td>5.20</td>
<td>68.51</td>
</tr>
<tr>
<td>r₃ + r₅ + r₇</td>
<td>53.34</td>
<td>5.94</td>
<td>4.41</td>
<td>4.68</td>
<td>67.57</td>
</tr>
<tr>
<td>r₁ + r₃ + r₅ + r₇</td>
<td>5.05</td>
<td>5.89</td>
<td>4.51</td>
<td>5.09</td>
<td>64.52</td>
</tr>
</tbody>
</table>
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
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

<table>
<thead>
<tr>
<th>Original Image</th>
<th>Adversarial Perturbations</th>
<th>Adversarial Image</th>
<th>Adversarial Result</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Original Image" /></td>
<td><img src="image2.png" alt="Adversarial Perturbations" /></td>
<td><img src="image3.png" alt="Adversarial Image" /></td>
<td><img src="image4.png" alt="Adversarial Result" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Original Image" /></td>
<td><img src="image6.png" alt="Adversarial Perturbations" /></td>
<td><img src="image7.png" alt="Adversarial Image" /></td>
<td><img src="image8.png" alt="Adversarial Result" /></td>
</tr>
<tr>
<td><img src="image9.png" alt="Original Image" /></td>
<td><img src="image10.png" alt="Adversarial Perturbations" /></td>
<td><img src="image11.png" alt="Adversarial Image" /></td>
<td><img src="image12.png" alt="Adversarial Result" /></td>
</tr>
</tbody>
</table>

### Objects
- Background
- Airplane
- Bicycle
- Bird
- Boat
- Bottle
- Bus
- Car
- Cat
- Chair
- Cow
- Cow Dining Table
- Dog
- Dog Train
- 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.
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?