LG-GAN: Label Guided Adversarial Network for Flexible Targeted Attack of Point Cloud-based Deep Networks

Hang Zhou\textsuperscript{1}  Dongdong Chen\textsuperscript{2}  Jing Liao\textsuperscript{3} \\
Kejiang Chen\textsuperscript{1}  Xiaoyi Dong\textsuperscript{1}  Kunlin Liu\textsuperscript{1} \\
Weiming Zhang\textsuperscript{1}  Gang Hua\textsuperscript{4}  Nenghai Yu\textsuperscript{1}

\textsuperscript{1}University of Science and Technology of China \quad \textsuperscript{2}Microsoft Research \\
\textsuperscript{3}City University of Hong Kong \quad \textsuperscript{4}Wormpex AI Research
Problem

Point shifting/adding/dropping

Neural network

Adversarial example attack

car

house

Threat!
Motivation

Related work

Current attack methods:
• Optimization-based:
  High attack success rate/slow runtime/visible outliers
• Gradient-based:
  Fast runtime/low attack success rate

Motivation

Generation based adversarial examples will avoid creating outliers and be fast in generation with high attack success rates.
Objective loss functions

Generator:
\[
L_g = L_{cls} + \alpha L_{rec} + \beta L_{dis}
\]
\[
L_{cls} = - \left[ t \log H(\hat{P}) + (1 - t) \log (H(1 - \hat{P})) \right]
\]
where \( \hat{P} = G_\theta(P, t) \)

\( L_{rec} \) is \( \ell_2 \) distance

\[
L_{dis}(\hat{P}) = \|1 - D_\theta(\hat{P})\|_2^2
\]

Discriminator:
\[
L_D(P, \hat{P}) = \frac{1}{2} \|D_\theta(\hat{P})\|_2^2 + \frac{1}{2} \|1 - D_\theta(P)\|_2^2
\]
Results

- clean plane
- C&W L2 attack
- C&W chamfer attack
- C&W hausdorff attack
- C&W cluster attack
- C&W object attack
- IFGM attack (to toilet)
- Single-layered LG-GAN attack (to vase)
- LG attack (to sofa)
- LG-GAN attack (to lamp)
## Results

<table>
<thead>
<tr>
<th>Target</th>
<th>Defense (SRS)</th>
<th>Defense (DUP-Net)</th>
<th>$l_2$ dist (meter)</th>
<th>Chamfer dist (meter)</th>
<th>Time (second)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C&amp;W + $l_2$</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.006</td>
</tr>
<tr>
<td>C&amp;W + Hausdorff</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>—</td>
<td>0.005</td>
</tr>
<tr>
<td>C&amp;W + Chamfer</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>—</td>
<td>0.005</td>
</tr>
<tr>
<td>C&amp;W + 3 clusters</td>
<td>94.7</td>
<td>2.7</td>
<td>0</td>
<td>—</td>
<td>0.120</td>
</tr>
<tr>
<td>C&amp;W + 3 objects</td>
<td>97.3</td>
<td>3.1</td>
<td>0</td>
<td>—</td>
<td>0.064</td>
</tr>
<tr>
<td>FGSM</td>
<td>12.2</td>
<td>5.2</td>
<td>2.8</td>
<td>0.15</td>
<td>0.129</td>
</tr>
<tr>
<td>IFGM</td>
<td>73.0</td>
<td>14.5</td>
<td>3.3</td>
<td>0.31</td>
<td>0.132</td>
</tr>
</tbody>
</table>

| LG + Chamfer (ours) | 96.1 | 75.4 | 13.9 | 0.63  | 0.137 | 0.037 |
| single-layered LG-GAN (ours) | 97.6 | 80.2 | 37.8 | 0.27  | 0.032 | 0.053 |
| LG (ours)          | 97.1 | 85.0 | 72.0 | 0.25  | 0.028 | 0.033 |
| LG-GAN (ours)      | **98.3** | **88.8** | **84.8** | 0.35  | 0.038 | 0.040 |

Table: Attack success rate (%, second to fourth column), distance (fifth-sixth column) between original sample and adversarial sample (meter per object) and generating time (second per object) on attacking PointNet. “Target” stands for white-box attacks. The hyper-parameter setting of two gray-box attacks is: for the simple random sampling (SRS) defense model, percentage of random dropped points is 60%~90%; for DUP-Net defense model, $k = 50$ and $\alpha = 0.9$ from [39]. The default LG-GAN (ours) consists of multi-layered label embedding, $l_2$ loss and GAN loss.
Thank You