Quantifying Realistic Threats for Deep Learning Models
Zhenyu Zhong, Zhisheng Hu, Xiaowei Chen, Baidu Research Institute
{edwardzhong, zhishenghu, xiaoweichen01}@baidu.com

Goal
- Define safety properties observed from the real-world such that any violations lead to misprediction
- Design standardize pipeline to evaluate threat severity & quantify DNN model robustness
- Shed a light on model robustness for pretrained models from different learning tasks

Motivation
- Intentional adversarial example attacks are less likely to happen due to the lack of practical monetizing scheme by the attackers.
- Real world threats against DNN don’t cease to exist even if there is no attacker for safety-critical scenarios.
- AI industries are in great need of real world threat quantification for the DNN model robustness.

Safety Violation to Resnet101
Ground Truth
A. Luminance
Brightness magpie Contrast african grey
B. Geometric Transformation
Rotation cabbage butterfly Vertical humming bird Horizontal lycanid
C. Blur
Motion Blur madagascar cat Gaussian Blur indri
D. Corruption
Blended Noise bulbul Salt & Pepper fountain Gaussian Noise ptarmigan
E. Weather
Fog african grey Frost fountain Snow cabbage butterfly

Threat Quantification Framework

Fig 1. Pretrained Model Robustness Comparison across 13 DNN architectures on randomly sampled 1k images from ImageNet. The || Perturbation L2 ||2 introduces Misclassification.

Fig 2. Fooling Success Rate: The median minimal L2 distance is the threshold T for each property across all the models. A success is defined as an input image that needs less than T perturbation to achieve model misbehavior.


Preliminary Results

<table>
<thead>
<tr>
<th>Category</th>
<th>Safety Properties Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luminance</td>
<td>Brightness, Contrast Reduction</td>
</tr>
<tr>
<td>Geometric Transformation</td>
<td>Horizontal (Vertical) Translation, Rotation, Spatial</td>
</tr>
<tr>
<td>Blur</td>
<td>Motion Blur, Gaussian Blur</td>
</tr>
<tr>
<td>Corruption</td>
<td>Uniform Noise, Gaussian Noise, Blended Noise, Salt And Pepper Noise</td>
</tr>
<tr>
<td>Weather</td>
<td>Fog, Frost, Snow</td>
</tr>
</tbody>
</table>

Table 1.

Criteria Description
- Misclassification
  $ C(x + δ) \neq G(x) $
- Confidence Misclassification
  $ P_{\text{Baidu}}(x) > \text{threshold}_L, L \neq G(x) $
- Topk Misclassification
  $ G(x) \notin \text{top}_k(x + δ) $
- Original Class Probability
  $ P_e(x + δ) < \text{threshold}_L $ 
- Target Class Probability
  $ P_e(x + δ) > \text{threshold}_L $ 

$x$ is the original input, $x + δ$ is the perturbed input. $C$ is the function returns the class label, $G$ is the ground truth of the input. $P$ is the probability of the input prediction, $e = G(x)$, $L$ is the multiclass label collection, $T$ is the target class.

REFERENCES