Noise Robustness of Deep Neural Networks

Yuxin Jiang, Chiaai Lin, Qingyu Zhu

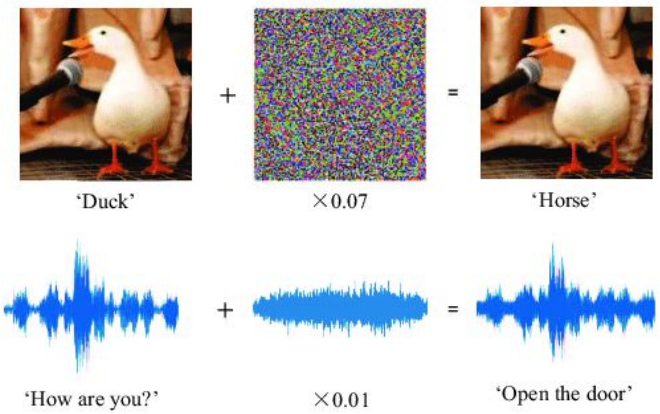
Dec. 9th 2019
18737-SA

Outline

- Background
- Problem Definition
- Experiments
- Analysis
- Conclusion
- Future Work
- References

Background

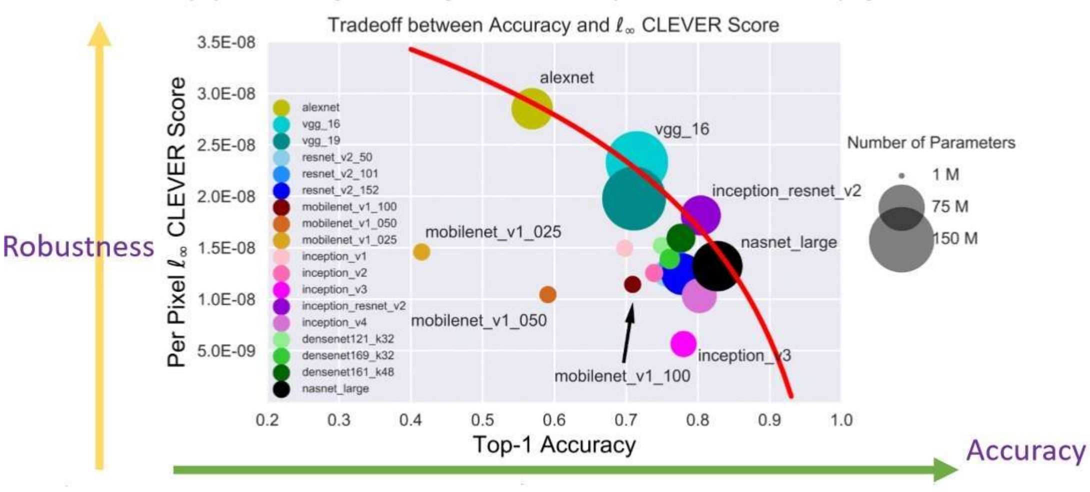
Deep neural networks (DNNs) are vulnerable to maliciously designed adversarial examples, which poses a significant risk in applying DNNs to safety-critical applications. E.g. Captcha, Image/Voice Recognition



Gong, Yuan & Poellabauer, Christian. (2018). Protecting Voice Controlled Systems Using Sound Source Identification Based on Acoustic Cues.

Background

Trade-off between **Robustness** and **Performance**



Junko Yoshida. (2019). “AI Tradeoff: Accuracy or Robustness?” [Available at]https://www.eetimes.com/ai-tradeoff-accuracy-or-robustness/

Problem Definition

Motivation Simple implementation of DNNs are prone to security vulnerabilities, we analyse and quantify the robustness of the DNN used in image recognition.

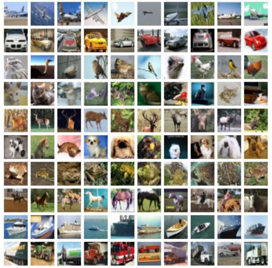
Approach

1. Read literatures in related field
2. Use a public image dataset to train a DNN
3. Compare the performance of self-trained and pre-trained DNN
4. Test the robustness of models against image changes of blur, contrast and brightness

Goals

1. Identify common adversarial examples targeting image recognition DNNs
2. Understand why such attack works
3. Propose suggestions for improving the security level of DNNs

Experiment Setups

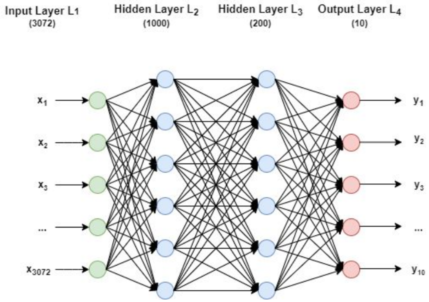


Training Data: **CIFAR-10 dataset**

Libraries: **Pytorch, GLUON, OpenCV**

Self-trained DNN Model (53% AVG Accuracy) vs. **Pre-trained DNN Model (99% AVG Accuracy)**

Feed Forward Network w/ 2 Hidden Layers



GLUON Pre-trained Model on CIFAR-10

Convolutional Neural Network (**cifar_resnet20_v1**)

Experiment 1 - Kernel Convolution (Blurring)

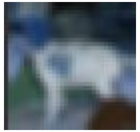
Change ind from 0 to 3

kernel_size = 3 + 2 * ind

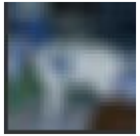
kernel = np.ones((kernel_size, kernel_size), dtype=np.float32)

kernel /= (kernel_size * kernel_size)

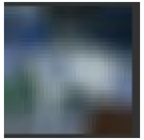
53.02%



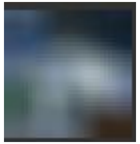
26.17%



24.71%



22.46%



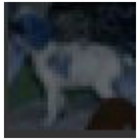
19.86%

Experiment 2 - Contrast and Brightness

$g(x)=\alpha f(x)+\beta$, α = *contrast*, β = *brightness*


22.32%


53.02%


17.11%


19.17%


23.92%


53.02%


24.76%


20.54%

Analysis

Blur:
The average accuracy decreases
More blurry: Classified as “plane” and “ship” increased

Contrast:
The average accuracy would decrease as intuition.
Higher contrast: Classified as “plane” and “ship” increased
Lower contrast: Classified as “car” and “cat” increased

Brightness
The average accuracy would decrease
Brighter: Classified as “plane” and “ship” increased
Darker: Classified as “car” and “cat” increased

Conclusion

Good for Robustness	Bad for Robustness
More Hidden Layers	Less Hidden Layers
Convolution	General Matrix Multiplication
Adversarial Training	Natural Training

Future Work

- Train a more sophisticated DNN with more hidden layers and can learn features at various levels of abstraction.
- Use more completed methods to quantify the accuracy and robustness of models.
- Test DNN’s robustness against more image changes like rotate and saturation.

Project Repository

https://github.com/zqy-nku/Noise-Robustness-of-DNN

References

- Learning Multiple Layers of Features from Tiny Images, Alex Krizhevsky, 2009.
- Gilmer & Hendrycks, "A Discussion of 'Adversarial Examples Are Not Bugs, They Are Features': Adversarial Example Researchers Need to Expand What is Meant by 'Robustness'", Distill, 2019.
- Gong, Yuan & Poellabauer, Christian. (2018). Protecting Voice Controlled Systems Using Sound Source Identification Based on Acoustic Cues.
- Junko Yoshida. (2019). “AI Tradeoff: Accuracy or Robustness?” [Web] https://www.eetimes.com/ai-tradeoff-accuracy-or-robustness/

Work Split

	Chiaai Lin	Qingyu Zhu	Yuxin Jiang
Project Proposal		✓	
Self-DNN Training		✓	
Self-DNN and Pre-trained DNN Comparison			✓
Data Collection	✓		
Robustness Analysis Experiments	✓		
PowerPoint			✓

Q&A