How CLEVER is your neural network?
Robustness evaluation against adversarial examples

Pin-Yu Chen
IBM Research AI

O’Reilly AI Conference @ London 2018
Label it!
Label it! **AI model says:**

**ostrich**
How about this one?
Surprisingly, AI model says:

shoe shop
What is wrong with this AI model?

- This model is one of the BEST image classifier using neural networks


IBM Research AI
Adversarial examples: the evil doublegangers

source: Google Images

IBM Research AI
Why do adversarial examples matter?

- Adversarial attacks on an AI model deployed at test time (aka evasion attacks)
Adversarial examples in different domains

- Images
- Videos
- Texts
- Speech/Audio
- Data analysis
- Electronic health records
- Malware
- Online social network
- and many others

IBM Research AI
Adversarial examples in image captioning

Input: image

Output: caption

Original Top-3 inferred captions:
1. A red stop sign sitting on the side of a road.
2. A stop sign on the corner of a street.
3. A red stop sign sitting on the side of a street.

Adversarial Top-3 captions:
1. A brown teddy bear laying on top of a bed.
2. A brown teddy bear sitting on top of a bed.
3. A large brown teddy bear laying on top of a bed.

Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge, Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, T-PAMI 2017
IBM Research AI
Adversarial examples in speech recognition

What did your hear?

without the dataset the article is useless

Audio Adversarial Examples: Targeted Attacks on Speech-to-Text, Nicholas Carlini and David Wagner, Deep Learning and Security Workshop 2018
IBM Research AI
Adversarial examples in speech recognition

without the dataset the article is useless

What did your hear?

okay google browse to evil.com
Adversarial examples in data regression

Is Ordered Weighted $\| \ell_1 \|$ Regularized Regression Robust to Adversarial Perturbation? A Case Study on OSCAR, Pin-Yu Chen*, Bhanukiran Vinzamuri*, and Sijia Liu, GlobalSIP 2018

IBM Research AI
Adversarial examples in physical world

- Real-time traffic sign detector
- 3D-printed adversarial turtle
- Adversarial eye glasses
Adversarial examples in physical world (1)

• Real-time traffic sign detector
Adversarial examples in physical world (2)

• 3D-printed adversarial turtle

Synthesizing Robust Adversarial Examples

Anish Athalye*12 Logan Engstrom*12 Andrew Ilyas*12 Kevin Kwok2
Adversarial examples in physical world (3)

- Adversarial eye glasses that fool face detector
- Adversarial sticker

---

**Adversarial Patch**

Tom B. Brown, Dandelion Mančić, Aurko Roy, Martín Abadi, Justin Gilmer
(tombrown,dandelion,aurkor,abadi,gilmer)@google.com
Adversarial examples in black-box models

- **White-box setting**: adversary knows everything about your model
- **Black-box setting**: craft adversarial examples with limited knowledge about the target model
  - Unknown training procedure/data/model
  - Unknown output classes
  - Unknown model confidence

---

**Black-box attack via iterative model query (ZOO)**

**Targeted black-box attack on Google Cloud Vision**

---

Black-box Adversarial Attacks with Limited Queries and Information, Andrew Ilyas*, Logan Engstrom*, Anish Athalye*, and Jessy Lin*, ICML 2018
Source: https://www.labsix.org/partial-information-adversarial-examples/
IBM Research AI
Growing concerns about safety-critical settings with AI

Autonomous cars that deploy AI model for traffic signs recognition

Source: Paishun Ting
But with adversarial examples...
Where do adversarial examples come from?

- What is the common theme of adversarial examples in different domains?
Neural Networks: The Engine for Deep Learning

• Applications of neural networks
  - Image processing and understanding
  - Object detection/classification
  - Chatbot, Q&A
  - Machine translation
  - Speech recognition
  - Game playing
  - Robotics
  - Bioinformatics
  - Creativity
  - Drug discovery
  - Reasoning
  - And still a long list...

Source: Paishun Ting
The ImageNet Accuracy Revolution and Arms Race

IBM Research AI

What’s Next?
Beyond human performance

In 2012, the team to first use deep learning was the only team to get their error rate below 25%.
In 2017, 29 of 38 teams got less than 5% wrong.
The following year nearly every team got 25% or fewer wrong.
In the competition’s first year teams had varying success. Every team got at least 25% wrong.
Accuracy ≠ Adversarial Robustness

• Solely pursuing for high-accuracy AI model may get us in trouble...

Our benchmark on 18 ImageNet models reveals a tradeoff in accuracy and robustness
How can we measure and improve adversarial robustness of my AI/ML model?

An explanation of origins of adversarial examples

The CLRVER score for robustness evaluation
Learning to classify is all about drawing a line

- Decision boundary w/ 100% accuracy
- Decision boundary w/ <100% accuracy

Source: Paishun Ting
Connecting adversarial examples to model robustness

Robustness evaluation: how close a reference input is to the (closest) decision boundary

Source: Paishun-Ting, Tsui-Wei Weng
Robustness evaluation is NOT easy

- We still don’t fully understand how neural nets learn to predict
  - calling for interpretable AI
- Training data could be noisy and biased
  - calling for robust and fair AI
- Neural network architecture could be redundant and leading to vulnerable spots
  - calling for efficient and secure AI model
- Need for human-like machine perception and understanding
  - calling for bio-inspired AI model
- Attacks can also benefit and improve upon the progress in AI
  - calling for attack-independent evaluation

Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

- Labeled datasets

IBM Research AI
How do we evaluate adversarial robustness?

- **Game-based approach**
  - Specify a set of players (attacks and defenses)
  - Benchmark the performance against each attacker-defender pair
    - The metric/rank could be exploited;
      - No guarantee on unseen threats and future attacks

- **Verification-based approach**
  - Attack-independent: does not use attacks for evaluation
  - Can provide a robustness certificate for safety-critical or reliability-sensitive applications: e.g., no attacks can alter the decision of the AI model if the attack strength is limited
    - Optimal verification is provably difficult for large neural nets – computationally impractical

CLEVER: a tale of two approaches

- An attack-independent, model-agnostic robustness metric that is efficient to compute
- Derived from theoretical robustness analysis for verification of neural networks: Cross Lipschitz Extreme Value for nEtwork Robustness
- Use of extreme value theory for efficient estimation of minimum distortion
- Scalable to large neural networks
- Open-source codes: https://github.com/IBM/CLEVER-Robustness-Score
How do we use CLEVER?

**Before-After robustness comparison**
- Will my model become more robust if I do/use $X$?

**Other use cases**
- Characterize the behaviors and properties of adversarial examples
- Hyperparameter selection for adversarial attacks and defenses
- Reward-driven model robustness improvement
Examples of CLEVER

• CLEVER enables robustness comparison between different
  ❑ Threat models
  ❑ Datasets
  ❑ Neural network architectures
  ❑ Defense mechanisms
Where to Find CLEVER? It’s ART

Adversarial Robustness Toolbox (ART)

External: https://github.com/IBM/adversarial-robustness-toolbox

- Python library, 7K lines of code
- State-of-the-art attacks, defenses and robustness metrics

```
from keras.datasets import mnist
from keras.models import load_model

from art.attacks import CarliniWagnerAttack
from art.classifier import KerasClassifier
from art.metrics import loss_sensitivity

# Load data
(x_train, y_train), (x_test, y_test) = mnist.load_data()

# Load model and build classifier
model = load_model('my_favorite_keras_model.h5')
classifier = KerasClassifier(model=model)

# Perform attack
attack = CarliniWagnerAttack(classifier)
adv_x_test = attack.generate(x_test)

# Compute metrics on model robustness
print('loss_sensitivity(classifier, x_test)
```

Evasion attacks
- FGSM
- JSMA

Evasion defenses
- Feature squeezing
- Spatial smoothing

Poisoning detection
- Detection based on clustering activations

Robustness metrics
- CLEVER
- Empirical robustness

Open-source release @ RSA 2018:
- 3.5K+ views of IBM blogs
- 100+ news outlets covering release
- 1.3M+ Social Media potential impressions
- 5K+ views of GitHub repo

Also available at https://github.com/IBM/CLEVER-Robustness-Score

IBM Research AI
Take-aways

• Adversarial robustness is a new AI standard
  ❑ Robustness does not come for free: adversarial examples exist in digital space, physical world, and different domains
  ❑ High accuracy ≠ Good robustness
  ❑ Arms race: adversary-aware AI v.s. AI for adversary
• How to evaluate the robustness of my AI model?
  ❑ CLEVER: an attack-independent robustness score
  ❑ Robustness comparison in before-after setting
  ❑ Where to find CLEVER? It’s ART!

IBM Research AI
Beyond Robustness: Trusted AI

IBM Research is building and enabling AI solutions people can trust.

Robustness
We are working to ensure the security and reliability of AI systems by exposing and fixing their vulnerabilities, identifying new attacks and defense, designing new adversarial training methods to strengthen against attack, and developing new metrics to evaluate robustness.

Fairness
To encourage the adoption of AI, we must ensure it does not take on and amplify our biases. We are creating methodologies to detect and mitigate bias through the life cycle of AI applications.

Explainability
Knowing how an AI system arrives at an outcome is key to trust, particularly for enterprise AI. To improve transparency, we are researching local and global interpretability of models and their output, training for interpretable models and visualization of information flow within models, and teaching explanations.

Lineage
Lineage services can infuse trust in AI systems by ensuring all their components and events are traceable. We are developing services like instrumentation and event generation, scalable event ingestion and management, and efficient lineage query services to manage the complete lifecycle of AI systems.
Collaborators: Tsui-Wei Weng(MIT), Luca Daniel(MIT), Honnge Chen(MIT) Huan Zhang(UCLA), Cho-Jui Hsieh(UCLA), Jinfeng Yi(JD AI), Yupeng Gao(IBM), Bhanukiran Vinzamuri(IBM), Sijia Liu(IBM), Yash Sharma, Su Dong, Chun-Chen Tu(UMich), Paishun Ting(Umich)

- MIT-IBM Watson AI Lab: David Cox, Lisa Amini
- IBM Research AI – Learning Group: Payel Das, Saska Mojsilovic
- IBM AI-Security Group: Ian Molloy, Mathieu Sinn, and their teams
- IBM Big Check Demo: Casey Dugan and her team
- IBM DLaaS Group: Evelyn Duesterwald and her team

Personal Website: www.pinyuchen.com
Twitter: pinyuchen.tw