Adversarial Examples for Robust Detection of Synthetic Media

Nicholas Carlini
Google
or, why you shouldn't trust machine learning

Nicholas Carlini
Google
or, why you shouldn't trust machine learning
... ever

Nicholas Carlini
Google
Andrew Walz
Andrew Walz
Congressional Candidate
Andrew Walz
Congressional Candidate
Verified by Twitter
Andrew Walz

Congressional Candidate

Verified by Twitter

Not a real person
Andrew Walz
Congressional Candidate
Verified by Twitter
Not a real person

FAKE
Andrew Walz
Congressional Candidate

Verified by Twitter

Not a real person
... how?
Four trivial attacks

Joint work with Hany Farid
G(Z, R)  \quad \text{FAKE}

G(Z, R_{\text{adv}})  \quad \text{REAL}
G(Z₁, R₁)  G(Z₂, R₂)  G(Z₃, R₃)

FAKE  FAKE  FAKE
\[ G(Z_1, R_{adv}) \quad G(Z_2, R_{adv}) \quad G(Z_3, R_{adv}) \]
\[ G(Z_1, R_1) \quad G(Z_2, R_2) \quad G(Z_3, R_3) \]
G(Z₁, R₁)   G(Z₂, R₂)   G(Z₃, R₃)

FAKE        FAKE        FAKE
Adversarial Distribution Shifts
or, why you shouldn't trust machine learning ... ever

Nicholas Carlini
Google
Natural Distribution Shifts

Joint work with Rohan Taori, Achal Dave, Vaishaal Shankar, Benjamin Recht, Ludwig Schmidt
What we want

1. Someone wants to know what breed of dog they just saw on the street
2. They take out their phone
3. Open up the camera app
4. Take a picture, and run a ResNet on the image
What we have

1. Someone wants to know what breed of dog they just saw on the street
2. They take out their phone
3. Open up the camera app
Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht*  Rebecca Roelofs  Ludwig Schmidt  Vaishaal Shankar
UC Berkeley  UC Berkeley  UC Berkeley  UC Berkeley

Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% – 15% on CIFAR-10 and 11% – 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models’ inability to generalize to slightly “harder” images than those found in the original test sets.
Now we have a new dataset.

Identical in every way to the original.

How do models do on this new dataset?
Possible explanations

1. It's just a harder dataset
2. Adaptive overfitting
3. Distribution shift
Possible explanations

1. It's just a harder dataset
2. Adaptive overfitting
3. Distribution shift
Possible explanations

1. It's just a harder dataset
2. Adaptive overfitting
3. Distribution shift
Adaptive Overfitting
Possible explanations

1. It's just a harder dataset
2. Adaptive overfitting
3. Distribution shift
Our paper: BIG DATA
Formalization:
Effective Robustness
Hypothetical Robustness Intervention

- ImageNetV2 (top-1, %)
- ImageNet (top-1, %)

- y = x
- Baseline accuracy
- Hypothetical robust model
- Standard models

Effective Robustness
Hypothetical Robustness Intervention

- **ImageNetV2 (top-1, %)**
- **ImageNet (top-1, %)**

- **Baseline accuracy**
- **Hypothetical robust model**
- **Standard models**

- **Effective Robustness**

---

The graph shows a scatter plot comparing ImageNetV2 (top-1) and ImageNet (top-1) accuracies. The data points are aligned along a line, indicating a strong correlation. A star denotes an effective robust model, positioned above the line, suggesting improved performance compared to standard models.
So what helps?
Lessons
If you use machine learning, you're vulnerable
Don't go and try to solve adversarially robust forensic detection
Do consider evaluating the robustness of your classifiers
Machine learning is not robust, in neither adversarial nor natural data settings.