Student Research Analyst Final Presentation

Chris Back

How I got started

Background

- Information Systems & Political Science at the University of Texas at Dallas
- Interest: Role that bridges technology and government

Archer Fellowship

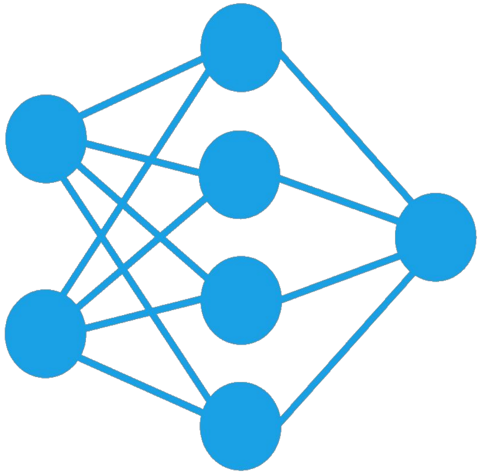
- Semester-long interdisciplinary public policy program from UT system
- Introduced through Archer staff

I hoped to gain exposure to

- High-demand technical skills
- Research writing
- Understanding our government and relevant policy issues

Main Assignments

2 Projects



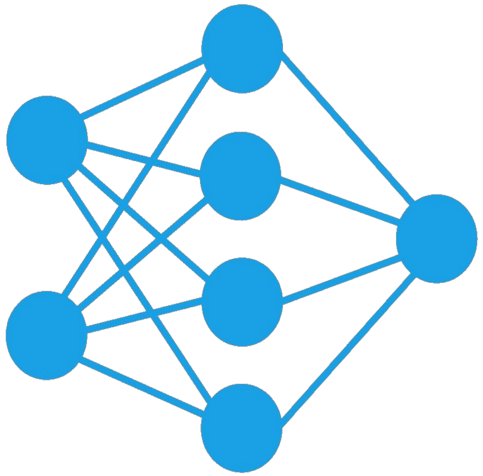
Data and Adversarial Examples

under Drew

High School Cyber Competitions

under Ali and Kayla

Data and Adversarial Examples



Problem Statement

Now we know adversarial attacks can be transferred, how do the difference between two models affect transferability performance?

Task

Assist Drew with designing experiment, then setup and test experiment with AI models and adversarial attacks

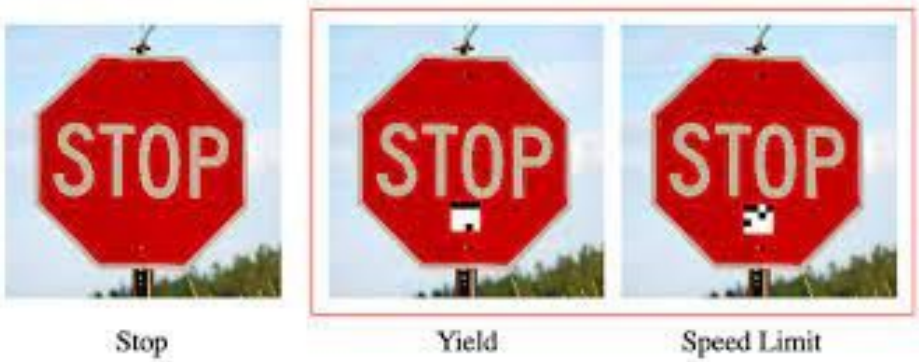
Background

Adversarial Example

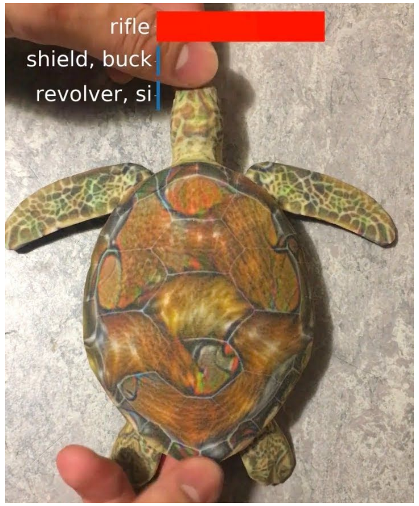
Input to ML model designed to fool a model despite resembling a valid input to a human

Adversarial Attack

Method to generate Adversarial examples



(a) Normal (b) Attack



Attack Method: Fast Gradient Sign Method (FGSM)

Calculates **direction** gradient that maximizes loss, and adds a small **perturbation** to that direction

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(x, y)),$$

x: unperturbed data (regular turtle)
epsilon: small decimal constant to adjust intensity of attack
epsilon*sign: perturbation
x': perturbed data (turtle that fools model it's a gun)

Transferability

Property of adversarial attacks where attacks generated to specifically fool **model A** can also be used to fool **model B**

ex. Russia wants to fool US satellite to misclassify airplane as a harbor using attack trained from only Russian data

Model A: Attacker Model

ML inside Russia’s image-classifying satellite trained from Russian images of airplanes and harbors **(attacker dataset)**

Model B: Victim Model

ML inside US’ image-classifying satellite trained from US images of airplanes and harbors **(victim dataset)**

Attack Method: Fast Gradient Sign Method (FGSM)

$$x' = x + \epsilon \cdot \mathrm{sign}(\nabla_x \mathcal{L}(x, y)),$$

Attack (Russian Satellite Model)
x: unperturbed data (victim or attacker datasets)
epsilon: small decimal constant to adjust intensity of attack
sign: direction of loss using **attacker model**
epsilon*sign: perturbation
x’: perturbed data (victim or attacker datasets)

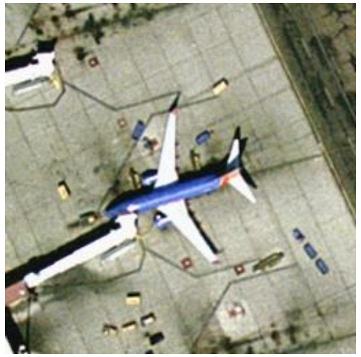
Victim (US Satellite Model)
Fed as **input** to **victim model**

Approach

Problem Statement: Now we know adversarial attacks can be transferred, how do the difference between two models affect transferability performance?

Approach: Blur datasets to create versions that are increasingly dissimilar to the original dataset. Test how effective attacks are when **difference between attacker model and victim model increase**

1. Gather dataset and create multiple manipulated versions of it (1 clean + 5 blurred)
2. Train 6 ML models for each of the 6 datasets
3. Generate FGSM attacks using each of 6 models and try to fool all 6 models (including itself)



Effectiveness of FGSM Attacks on Models Trained on Various Datasets

Datasets used to train **victim model**

Datasets used to train **attacker model**

Dataset	Clean	Blur 1	Blur 2	Blur 3	Blur 4	Blur 5
Clean	100%	?	?	?	?	?
Blur 1	?	100%	?	?	?	?
Blur 2	?	?	100%	?	?	?
Blur 3	?	?	?	100%	?	?
Blur 4	?	?	?	?	100%	?
Blur 5	?	?	?	?	?	100%
Blur 6	?	?	?	?	?	?

Findings

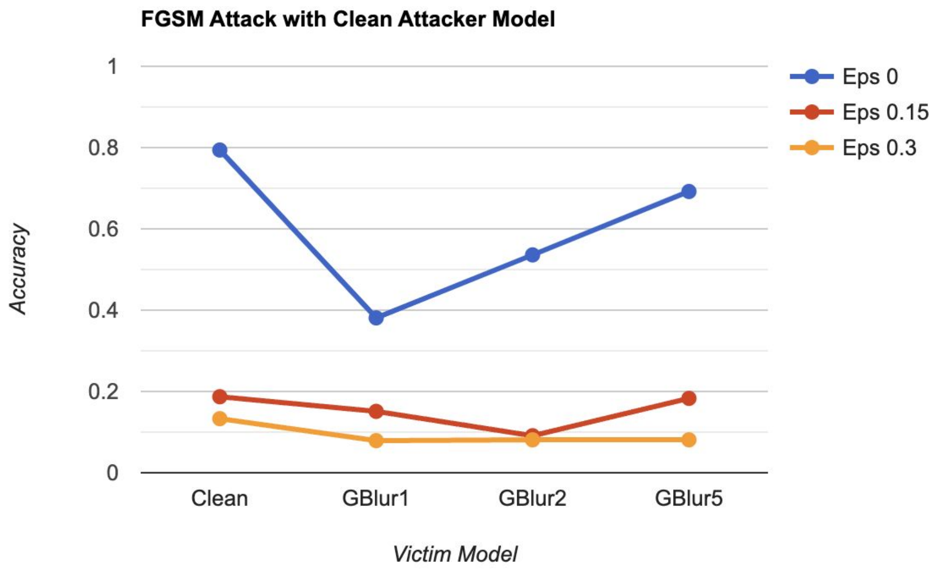
Attacker Model as Dataset (pre-image capture attack)

Effectiveness of FGSM Attacks on Models Trained on Various Datasets

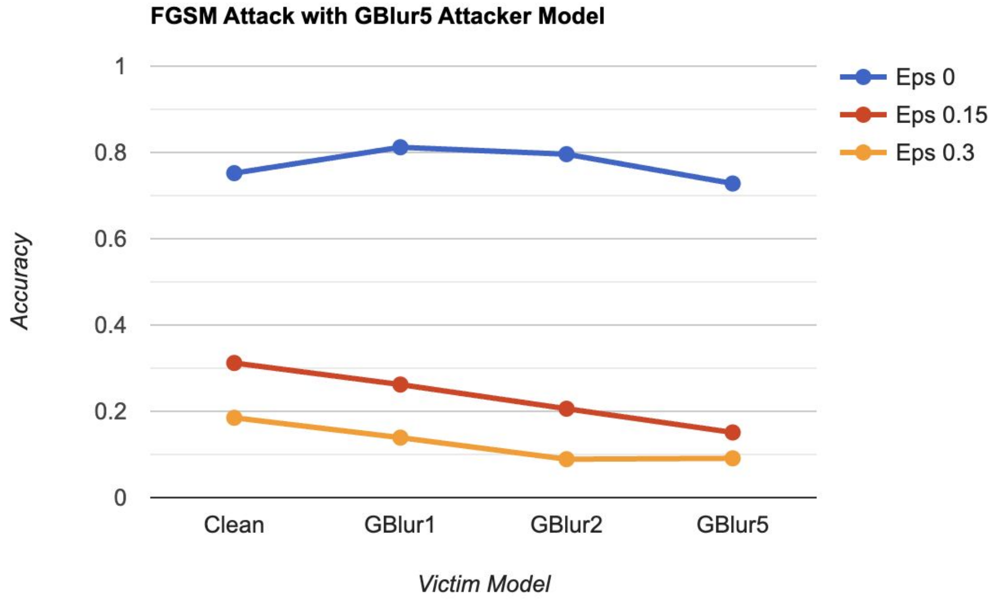
Datasets used to train **victim model** (columns); Datasets used to train **attacker model** (rows)

Dataset	Clean	Blur 1	Blur 2	Blur 3	Blur 4	Blur 5
Clean	100%	?	?	?	?	?
Blur 1	?	100%	?	?	?	?
Blur 2	?	?	100%	?	?	?
Blur 3	?	?	?	100%	?	?
Blur 4	?	?	?	?	100%	?
Blur 5	?	?	?	?	?	100%
Blur 6	?	?	?	?	?	?

Attacker Model: Clean | Victim Model: Varied
Input dataset: Varied



Attacker Model: GBlur5 | Victim Model: Varied
Input dataset: Varied



Theoretically: Accuracy should incr. as victim moves away from clean
Result: Not really shown.

Theoretically: Accuracy should decr. as victim gets closer to GBlur5.
Result: Clearly shown

Findings

Victim Model as Dataset (Post-input capture attack)

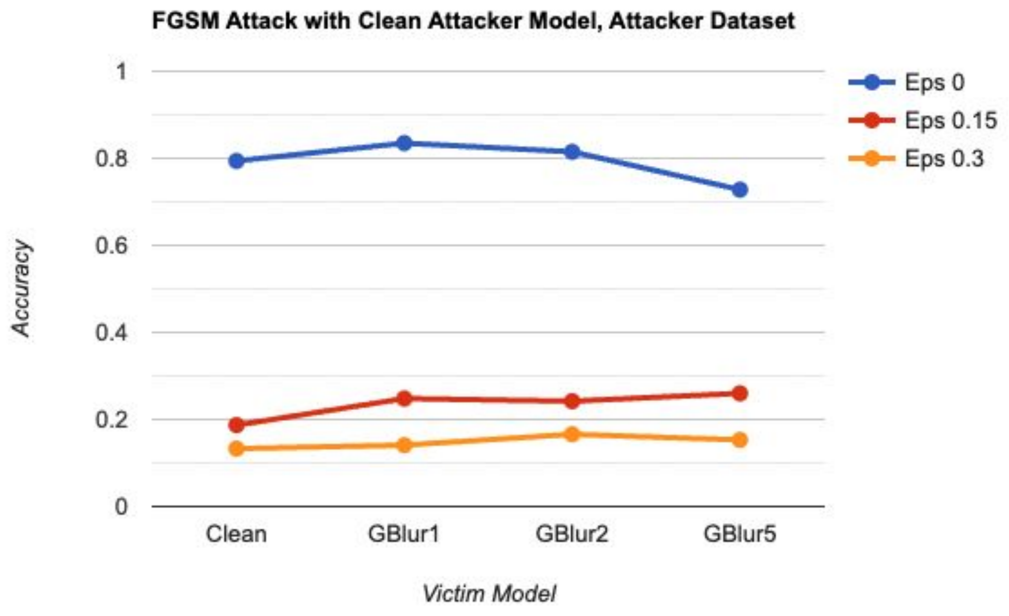
Effectiveness of FGSM Attacks on Models Trained on Various Datasets

Datasets used to train **victim model**

Datasets used to train **attacker model**

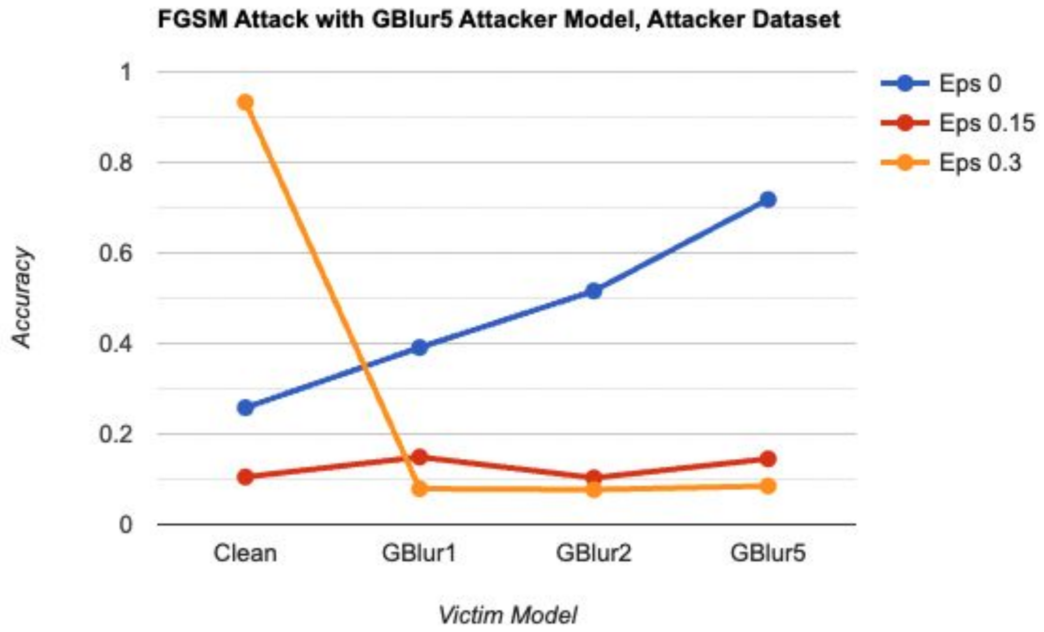
Dataset	Clean	Blur 1	Blur 2	Blur 3	Blur 4	Blur 5
Clean	100%	?	?	?	?	?
Blur 1	?	100%	?	?	?	?
Blur 2	?	?	100%	?	?	?
Blur 3	?	?	?	100%	?	?
Blur 4	?	?	?	?	100%	?
Blur 5	?	?	?	?	?	100%
Blur 6	?	?	?	?	?	?

Attacker Model: Clean | Victim Model: Varied
Input dataset: Clean



Theoretically: Accuracy should incr. as victim moves away from clean.
Result: Slightly shown.

Attacker Model: GBlur5 | Victim Model: Varied
Input dataset: GBlur5



Theoretically: Accuracy should decr. as victim gets closer to GBlur5.
Result: Somewhat shown. Shown strongly when clean is victim, eps 0.3.

Data and Adversarial Examples

Conclusion

- Very preliminary results, but similarity of datasets between attacker and victim seems significant to attack performance

Takeaways

- AI/Ml concepts- gradient descent, deep learning, adversarial attacks
- Programming with PyTorch
- Designing a technical research experiment
- Technology issues relevant to national

High School Cyber Competitions

Objective

Report on the landscape of HS cybersecurity competitions and how they contribute to future workforce.

Inform policymakers on:

- Current scope of competitions and its benefits to industry
- Success factors among top performing schools
- Possible barriers to entry for disadvantaged schools

Task

To assist Kayla and Ali on all parts of the project

- Collect a comprehensive list of past cybersecurity competitions + details
- Identify and collect relevant data points for participating schools
- Interview high school educators
- Write findings into draft

High School Cyber Competitions

Notable Findings

- Nearly all top and bottom performing schools offer CS courses, not cyber
- Extracurricular support is significant dividing factor
- Competitions starting to be treated like a sport

Takeaways

- Relevant socioeconomic, curriculum, performance factors for researching education
- Designing a qualitative research for government audiences
- Writing in the style of policy publications
- Interviewing with results in mind

Big Picture

What I was given

- Perfect blend of technical and writing skills
- Priceless network of professionals in the field
- Became more informed of policy ecosystem and relevant issues

Moving Forward

- Technical roles that are government or policy facing

Thank you!

References

Rey Reza Wiyatno, Anqi Xu, Ousmane Dia, Archy de Berker. “Adversarial Examples in Modern Machine Learning: A Review,” arXiv:1911.05268