Adversarial Attack Via Model Substitution

Adversarial Ttransferability Under Different Data Distributions

1. INTRODUCTION

- Adversarial examples cause ML models to miscalssify
  - Adds human-impreceptible “noise” to image
- *Model substitution* allows to create adversaries under “black-box” scenario

2. RESEARCH QUESTION

- What degree of adversarial transferability can be achieved by substitute models trained on a different data distribution than the target

  - Effectiveness under different attacks
- Do models pick same “non-robust” features (Ilyas et. al., 2019)?

3. Method

MODELS

- GoogLeNet
- ResNet-50
- DenseNet-121

ATTACKS

- FGSM
- PGD

D_{NR} Data

- Via Adversarial Ensemble

DATA DISTRIBUTIONS
(Splits based on animal breed)

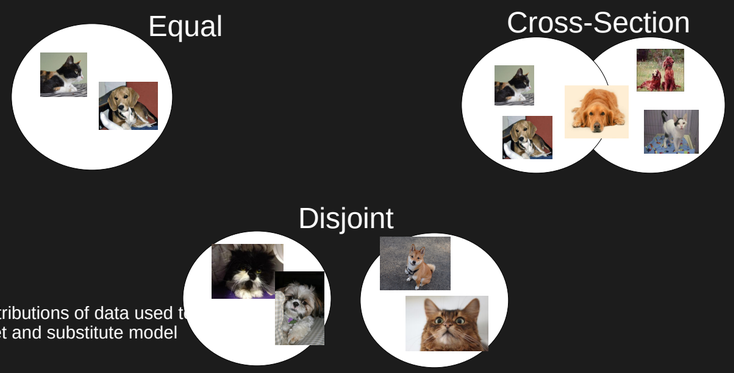


Fig 1. Distributions of data used t
train target and substitute model

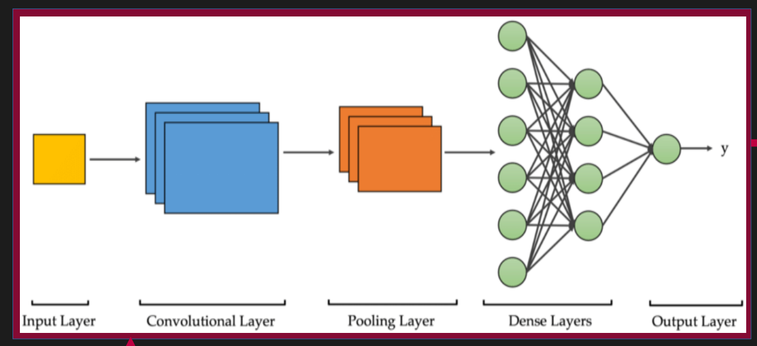


Fig 3. Structure of a CCN with 1 output (A.C, 2012)

Feed to Target Model

Adversarial Perturbation

+

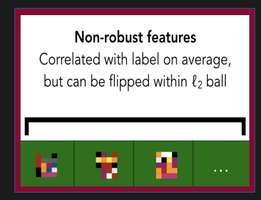
Clasified “Dog” on Substitute Model

Transferable Adversary?

Robust Features: “Dog”

Output: “Cat”

Similarity on non-robust features?



Non-Robust Features: “Cat”

4. RESULTS

- All models perfrom above 95% in normal conditions
- Transferability of adversarial samples between similar substitute models is comparable under the three different data distributions
- Some attacks are more effective at creating transferable samples
- Complex models are more susceptible to adversarial samples

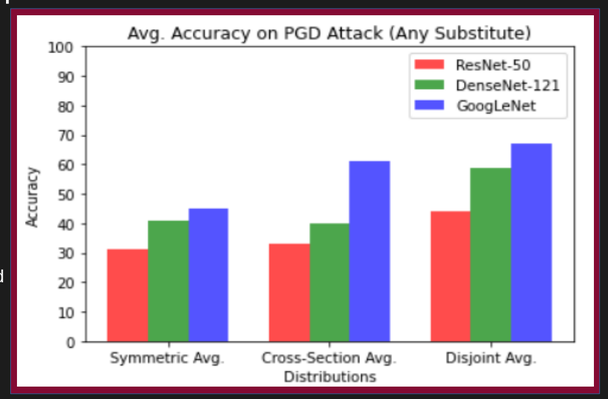


Fig 2. Intra-model and cross-model model accuracy given adversarial input created via PGD

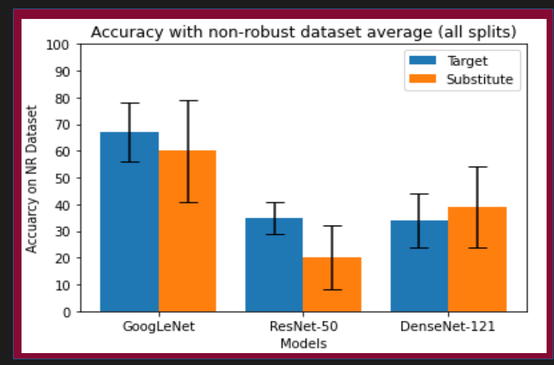


Fig 4. Comparison in evaluation accuracy with non-robust data D_{NR} (Liu et. Al, 2017l)

5. CONCLUSION

- Attacker can use a semantically similar dataset to achieve good adversarial transferability
- Inconclusive evidence to suggest models picking up same non-robust features

Attacker only needs a dataset similar enough to effectively conduct a substitute attack

REFERENCES

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. Advances in neural information processing systems, 32.

C, A. (2021, December 16). An Overview on Convolutional Neural Networks - The Startup. Medium. Retrieved May 12, 2022, from https://medium.com/swlh/an-overview-on-convolutional-neural-networks-ea48e76fb186

Liu, Y., Chen, X., Liu, C., & Song, D. (2016). Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv:1611.02770.

Pietro Vigilanza | p.m.vigilanzalorenzo@student.tudelft.nl

Supervised By: Stefanie Roos, Chi Hong, Jiyue Huang

CSE 3000 | 11.05.2022

TU Delft