Adversarial Attack and Training on Deep Learning-based Gaze estimation

Author: Clio Feng
Supervisor(s): Dr. G. (Guohao) Lan, Lingyu Du.
Thesis committee: Dr. Guohao Lan, Lingyu Du, Dr. Xucong Zhang

Contact: h.feng-5@student.tudelft.nl

Introduction /Motivation

While gaze estimation has improvement by using deep learning models, research had shown that neural networks are weak against adversarial attacks. Despite researchers has been done numerous on adversarial training, there are little to no studies on adversarial training in gaze estimation.

Research Questions

- What is the different effects of PGD attacks with different experimental settings?
- How adversarial training elevates the adversarial attack on gaze estimation?

Dataset: MPIIFaceGaze, 15 subjects, 3000 images each subjects, each RGB image (448,448)
Baseline Model: Epoch=20, Learning Rate=0.0001, Adam optimizer. **LetNet:** 8 Angular Error

Before Experiment: Attack Visibility

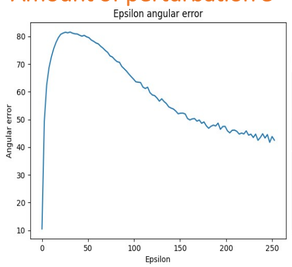
Since one of the key factor in adversarial attack is the visibility of the difference of the adversarial image and the original image, which could expose the present of attacker. Despite some research ignore this factor, from experiments in resent research, it reveals they matters. One way to represent it is through mean squared error (MSE) of image difference. After experiment , we set the human perceptual threshold as image difference is 22.

Projected Gradient Descent Attacks

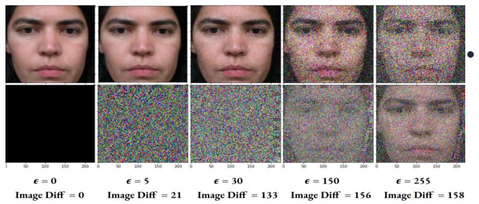
The PGD attack is white-box attack which means the attacker has access to the model gradients. Goal to find the noise that maximises the loss while keeping the size of the noise smaller than a specified amount. For Modification, I change cross-entropy loss for classification tasks to L1 loss.

Default setting: Epsilon = 5 , Step size = 0.6, Steps number = 10 , Random Start

Amount of perturbation ε



Introduction: The bound of the random noise (Epsilon ε). It prevent the noise exceed so visible to human.

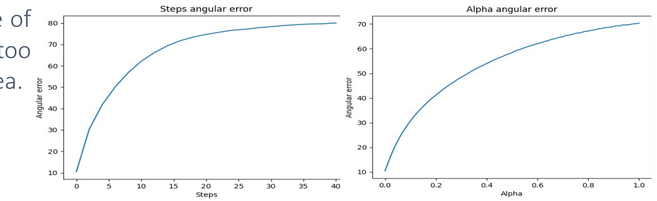


Result:

- As ε increases, the human perceptual level increase
- There is a certain loss, that after it there is not much attack left to explore.

Step Size introduction

Learning rate of the loss function. if we take a too large step, we can explore more area.



Number of Steps Introduction

Stop criteria: Number of iteration

Result: While step size and the number of steps do not impact the human perceptual level as much, they all converge within the bound set by ε when they increase.

Random Start Introduction

Starting at a uniformly random point.

Experiment: Angular Error: TRUE:78.70,FALSE:73.05

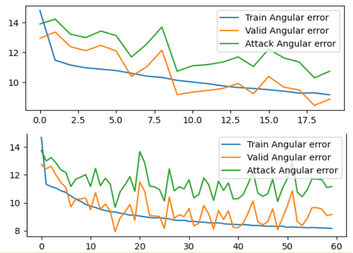
Result: Random start gives out better attack results than no random.

PGD Adversarial training

PGD adversarial training is the most effective adversarial training on the classifier yet. The implementation is followed [1]: Replace the training set with its PGD adversarial counterpart, with modification to L1 Loss for gaze estimation.

Result:

- PGD adversarial training has some defense against the PGD adversarial attack.
- It is not as effective as the classification task.
- It don't converge around baseline (8 degree)
- Different PGD experimental setting also impact the Performance



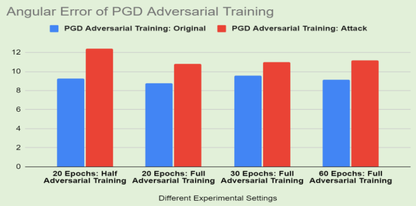
Data argumentation

Result:

- PGD adversarial training require large training time.
- Including the original samples in the training set does not improve the original sample's performance.
- As the number of adversarial samples increases, the performance improved.

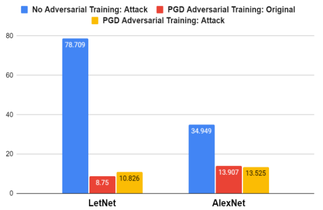
Half/Full adversarial Training:

Replace half/all the training set with its adversarial counterpart



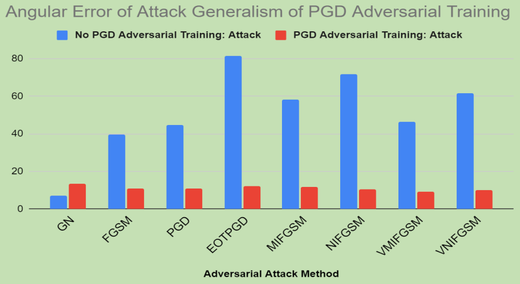
Model Capacity

- As the capacity increase, it becomes more resistant to the adversarial attack when no adversarial Training
- The performance of the adversarial training is not improved as the model capacity increase.



Attack Generalism

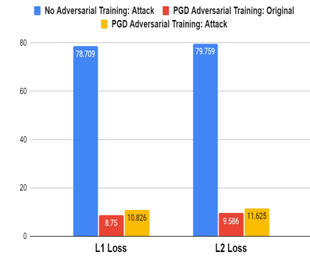
- PGD adversarial training can better defense against other attacks
- Most effective is VNIFGSM attack. Least effective on EOTPGD attack. Not able to against the GN.



Conclusion: PGD adversarial training can defend against other adversarial attacks as well and have a certain level of attack generalism.

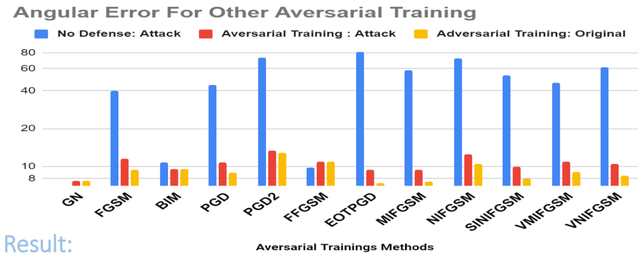
Loss Function

- L2Loss is more susceptible to adversarial samples than L1Loss as L2 is more sensitive to outliners.
- Therefore, L2 Loss perform little worse in adversarial training than L1



Other Adversarial Training

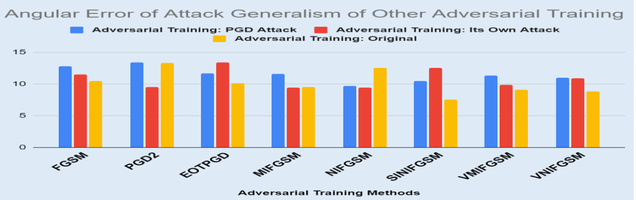
To further explore the effect of other adversarial training, this project is experimented on the other adversarial training that also exploit the gradient from simple to complex. Under the human perceptual threshold (image difference = 22), the experimental setting is applied.



Result:

- For GN, merely randomizing the noise does not affect attacking the model or defense against any attacks
- EOTPGD is the most effective attack, improved more than PGD in defense
- FFGSM (Single step) is not as effective in defense and attack

Attack Generalism



- Result: Depending on the different adversarial attacks, the performance of the adversarial training changes
- Hypothesis: Sometimes, simple adversarial training is not as able against more complex adversarial attacks, as the ratio of simple adversarial training failing is more than complex one.

Reference [1] " Towards deep learning models resistant to adversarial attacks" Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu., 2019