An Empirical Look at Gradient-based Black-box Adversarial Attacks on Deep Neural Networks Using One-point Residual Estimates

Author: Joost Jansen
Supervisors: Stefanie Roos, Jihue Huang, Chi Hong

BACKGROUND

- **Adversarial attacks on Deep Neural Networks (DNN)**: Adding *imperceptible perturbation* to an image results in DNN to *misclassify* the image
- **Black-box:** *model not known* to the attacker, only input-output correspondence (*queries*).
- **Gradient-based Attacks:** use *estimated gradient* to minimise the class probability of the image.

Different gradient estimators		
Estimator	**Number of queries per gradient estimation**	**Accuracy**
Two-point central	2b	**+++**
Two-point Forward/Backward	b + 1	**++**
One-point residual	b	**++**

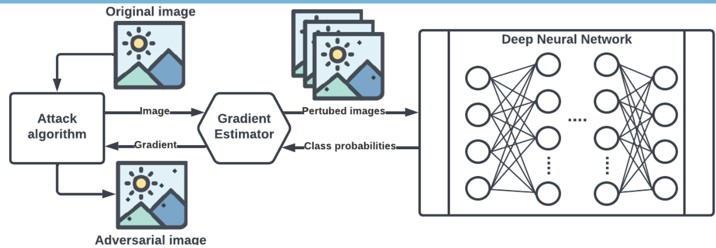
RESEARCH QUESTION

- *Do one-point residual estimates* ***improve*** *untargeted gradient-based adversarial attacks in terms of* ***reducing the number of queries while maintaining accuracy?***

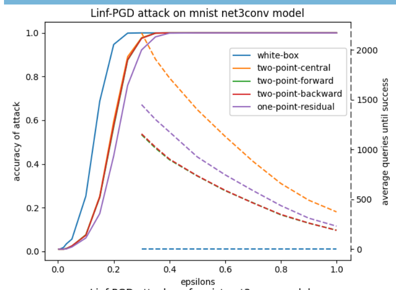
METHODOLOGY

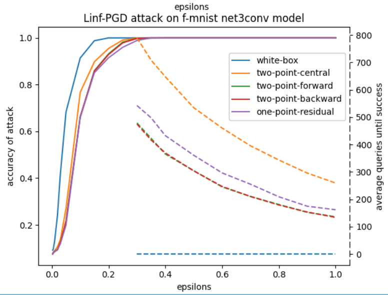
- **Compare different gradient estimators** to the one-point residual estimate:
  - Accuracy of attack
  - Average number of queries until a succesfull adversarial created
- Using **different PGD-attacks and datasets**

A GRADIENT-BASED ADVERSARIAL ATTACK



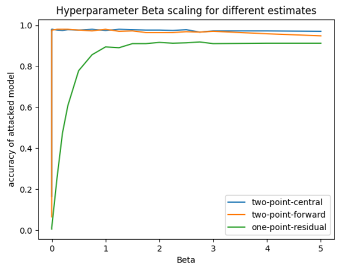
RESULTS





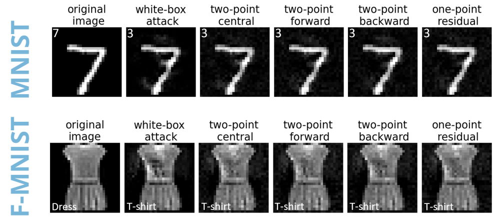
- **MNIST:** One-point residual estimates have a **lower** accuracy compared to the two-point estimates
- **One-point residual:** Less queries per iteration still leads to a **higher** average number of queries until a succesfull adversarial is created.
- **F-MNIST:** One-point residual estimates have a **corresponding** accuracy compared to the two-point estimates

Hyperparameters



- The one-point residual estimate is **sensitive** to some hyperparameters
- Determined optimal Hyperparameters for one-point residual estimates

EXAMPLES



DISCUSSION

- Limited to **only** PGD attacks
- **Bounded** by computational power, estimates were only tested on low dimensional datasets

CONCLUSION

- One-point residual estimates **do not** maintain accuracy for ***strong DNN's***
- One-point residual estimates do maintain accuracy for ***weaker DNN's***
- Although it uses les queries per iteration, one-point residual estimates **do not improve query efficiency**

FUTURE WORK

- Test estimates on more **complex datasets , models** and **other attacks**
- Use **grid search** to find all optimal hyperparameters

REFERENCES

1. Yan Zhang, Yi Zhou, Kaiyi Ji, and Michael M. Zavlanos. A new one-point residual-feedback oracle for black-box learning and control. Automatica, 136, 2, 2022.
2. Pu Zhao, Pin-Yu Chen, Siyue Wang, and Xue Lin. Towards query-efficient black-box adversary with zeroth-order natural gradient descent. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 6909–6916, 2020.
3. Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural net-works. Proceedings of the AAAI Conference on Artificial Intelligence, 33:742–749, 07 2019

Full paper available at:
https://pure.tudelft.nl/admin/files/123281254/Final_paper.pdf