ECE209AS - Winter 2022
UCLA

“Hariram, Veeramani”
Review for Paper Titled:

“High Accuracy and High Fidelity Extraction of Neural Networks”

1- Summary

This work addresses the Model Extraction attack problem by systematically taxonomizing the attack space by focusing around ‘Accuracy’ & ‘Fidelity’ extraction as two key Adversarial Objectives.
Specifically, it incorporates Query based Learning attacks for High Accuracy extraction and proposes techniques to enhance Query Efficiency. Then it contributes to the first of its kind Practical Functionally Equivalent Attack employing Direct Extraction which achieves High Fidelity. Furthermore, it proposes hybrid combination of Learning Extraction and Direct Extraction to improve both the Accuracy and Fidelity Extraction Objectives.
It proposes a mathematical Algorithm for Functional Equivalent Model extraction, only using the standard model prediciton-queries without any other additional information such as the access to Gradient queries or Physical side channels (power channels) while also achieving Fidelity Extraction of the model as a key Solution. It also analyzes the proposed method’s degradation as the number of target model’s parameters increases and contributes to hybrid techniques to counter the problem

2- Novelty/Contribution

This work signifies the importance of Fidelity Extraction while trying to employ Black-Box Adversarial attacks and highlights how attacks such as Membership Inference, Adversarial Overlearning of the target model might leverage this. It also proposes a mathematical framework to systematically craft this attack.

3- Evaluation

Accuracy Extraction using Fully Supervised vs Semi-Supervised Techniques and Fiedlity extraction using Direct Functional Extraction reveal consistent gains with negligible variance while accomplishing this in a Query efficient way using Unsupervised method (25% lesser queries than supervised learning) gives a convincing view of making a more practical attack.

4- Unique Strength/Weakness

I highly like the aspect of this work’s exhaustive analysis of its objectives of Accuracy and Fiedlity Extraction using different Attack Techniques and it’s importance link to Adversarial Attacks.