ECE209AS - Winter 2022
UCLA

“Hariram, Veeramani”
Review for Paper Titled:

“Blind Backdoors in Deep Learning Models”

1- Summary

This work proposes a new Blind attack- injecting Backdoors by compromising the loss computation in the Model-training code.
Precisely, the work extends the target model to perform the main task as well as a Backdoor task and uses Multiple Gradient Descent Algorithm (MGDA) and Franke- Wolfe Optimizer as part of loss computation pipeline to achieve the optimal accuracy balance between the model’s conflicting objectives.
Current work further illustrates the power of Blind attacks by injecting Single pixel, physical backdoors, special backdoors that switch the model to a different privacy-violating functionality, semantic backdoors inherently containing a feature word/name which can switch inputs to be classified as False positive (without even poisoning inputs) at inference time. It evaluates previous defenses against these attack models and proposes more robust defenses.

2- Novelty/Contribution

Blind attack cannot measure the accuracy of models trained using the code nor change the weights/loss coefficients after the ML code has been deployed. Furthermore, fixed coefficients may not achieve the optimal balance between the conflicting objectives of the main task and the backdoor task. This blind attack has been made plausible by conceiving and incorporating MGDA algorithm and Optimizer in a robust to compromise ML model’s pipeline without even knowing about the model architecture, hyperparameters, optimizer used to update the weights, loss criterion, learning rate. By superimposing the main task’s loss along with Backdoor task’s loss and Defense evasion loss, which are computed by operating on Dataset picked from a similar distribution and applying transformations, MGDA runs in a closed loop to compromise the ML model without requiring additional info and can evade existing defenses(Levade)

3- Evaluation

Use of MGDA is evaluated on ImageNet Backdoors, MNIST Dataset(Calculator Backdoors) and Convert Facial Identification Backdoors. The resultant metrics give a convincing superiority of this methodology in terms of Attacker’s performance against previously reported works of setting/fixing loss values after exhaustive experiments and poisoning batches of training data with Backdoored inputs.
Current work also evaluates the overhead caused by MGDA computation. The overhead compromises of one forward pass for each backdoored batch and two backward passes to find the scaling coefficients for multiple losses and offer measures to drastically reducing overhead such as Attacking while model is close to convergence and performing a constrained attack by attacking selective vulnerable batch of Dataset. These measures reduce the overhead required for the attack and makes it more effective.

4- Unique Strength/Weakness

An interesting and robust way to defend against the attack is also proposed which consists of processing the computational graphs associated with the main task and the compromised computational graph associated with the Backdoor task during the blind attack.