ECE209AS - Winter 2022
UCLA

“Hariram, Veeramani”
Review for Paper Titled:

“Honest-but-Curious Nets: Sensitive attributes of Private inputs can be secretly coded into the Classifiers’ Outputs”

1- Summary

This paper attacks the problem where a semi-trusted server can train a classifier such that the outputs could carry secret information about a sensitive attribute of the user’s data that is unrelated to the Target Attribute. It demonstrates that ML services can attack their users’ privacy even in a highly restricted setting where they have access to the agreed target computation results only. This shows how Honest-but-Curious (HBC) classifiers can encode a sensitive attribute of their private inputs into classifier’s output by exploiting Output’s Entropy as a side channel and further shows that there are efficient mechanisms to make the HBC classifiers seem like a Standard classifier, making them even more difficult to distinguish. This also discusses Knowledge Distillation as a serious threat of this vulnerability wherein a HBC classifier can transfer its curious capability to student classifiers.

2- Novelty/Contribution

This paper proposes novel solution for the inherent Classifier capacity problem with linear curious classifiers. It is known that any effort in increasing the Curiosity aspect of a linear classifier will have a direct hit on the Honesty hit on the Classifier. Hence, when the secret and target attributes are independent, we need Classifiers with more capacity to accomplish the attack. The paper shows how two logisitc regression classifers each trained separately for a corresponding attribute could be combined by the hard mixture method such that the output is a mixture of the predicted values of the target attribute and the sensitive attribute. The resultant classifier has twice the capacity of the logistic regression while not being a linear classifier anymore. Due to this increased capacity, such method could efficiently operate on independent target and sensitive attributes and still make efficient attacks. The paper consider the worst case White Box knowledge attacks and proposes Regularized Attacks wherein a classifier is enforced to explicitly encode both sensitive and target attributes by regularizing the loss function on the classifier through Argmax and Entropy properties and Parameterized Attacks which can deal with Muticlass in addition to Binary classes by modifying the loss function of the classifier and also utilize an additional model to estimate the sensitive attribute.

3- Evaluation

The above work performs evaluation on two real world datasets such as CelebA dataset and UTKFace Dataset and 4 Convolutional layers plus two fully connected layers with 250K trainable parameters for Classifier function and 2 to 3 Fully connected classifier functions for decoding secret attributes. The corresponding target and sensitive attributes were investigated: Gender vs Race, Smile Detection vs Mouth Open, Hair Colour vs Other (Male/Smiling/Attractive) attributes and Race, Age, Gender Attributes and the results show how effectively the attacks could unveil the sensitive attributes apart from the target attributes under consideration.

4- Unique Strength/Weakness

I like the exhaustive real world evaluation performed in this work along with the novel contributions, the paper suggests how these attacks could be made practically on real world datasets and capture sensitive attributes necessitating the need for defensive mechanisms and related open questions.