ECE209AS - Winter 2022
UCLA

“Hariram, Veeramani”

Review for Paper Titled:

# “Overlearning Reveals Sensitive Attributes”

## 1- Summary

This paper focuses on the problem on Overlearning where a Model trained for a simplistic objective function learns to recognize sensitive attributes implicitly that are not part of the learning objective. It proposes two significant, alarming results which are intrinsic to some tasks, and which cannot be prevented by Censoring unwanted attributes. This highlights the inadequacy of Censoring as a Privacy Protection Technology First, Inference- time representations of an overlearned model reveal sensitive attributes of the input, breaking defensive privacy protections such as Model Partitioning. Second, it shows how an overlearned model can be ‘repurposed’ for a different privacy-violating task even in the absence of the original data. Also, such unintentionally learned concepts are neither finer nor coarse-grained versions of the model’s labels nor statistically correlated with them.

## 2- Novelty/Contribution

This paper contributes to the De-censoring attacks and shows how it can be used to evade the Censoring Defense mechanisms. It proposes Model Repurposing Attack where it uses a model M to create another model that, when applied directly predicts its sensitive attributes. This also focuses on Inference time attack which applies an input to a given Model ‘M’ and uses M’s representation of that input to predict its sensitive attributes and measure the leakage properties of such attacks. It proposes a mathematical formulation for De-censoring as an optimization problem with a feature space loss comprising a Transformer that the adversary wants to learn where training with a feature-space loss has been proposed for synthesizing more censored representations to arrive at and match with uncensored representations.

## 3- Evaluation

The paper evaluates proposed attacks over various datasets, tasks such as Health, UTKFace, FaceScrub, Place365, Twitter, Yelp, PIPA and reveals alarming results about the accuracy with which the sensitive attribute could be inferred even though some features were not even part of the training dataset.

## 4- Unique Strength/Weakness

The paper emphasizes that the structural complexity of the data as one of the key reasons for overlearning. That is unlike structured, less sophisticated datasets, for data generated from more complex distributions, networks learn more complex solutions leading to the emergence of features that are more general than the learning objective which consequently leads to overlearning.