Can Differential Privacy Applied to Split Learning Mitigate the Feature Space Hijacking Attack (FSHA)?

Greg Gawron, Phil Stubbings

Background

Split Learning and Differential Privacy

Split Learning trains a part (split) of a neural network directly at every private silo, making sure that no data in raw form leaves the silos.

A Coordinator network (Fig 1. below) combines the output from silo sub-networks to produce a prediction. During training, the gradient for the first layer of the coordinator network is propagated to each sub-network.

- Split Learning (SL) has been shown to be vulnerable to reconstruction privacy attacks
- Differential Privacy (DP) works by adding noise to data to be kept private, aiming at “hiding”any specific individual data points. It seems to be a natural extension of Split Learning
- Collaborative learning (not SL specifically) with DP has been shown to be vulnerable

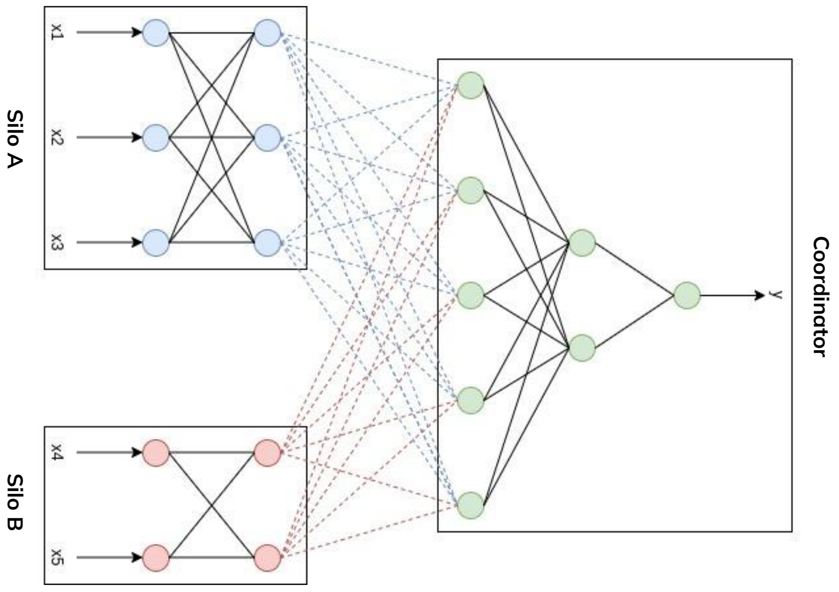


Figure 1: Split Learning architecture with Coordinator network

Experiments

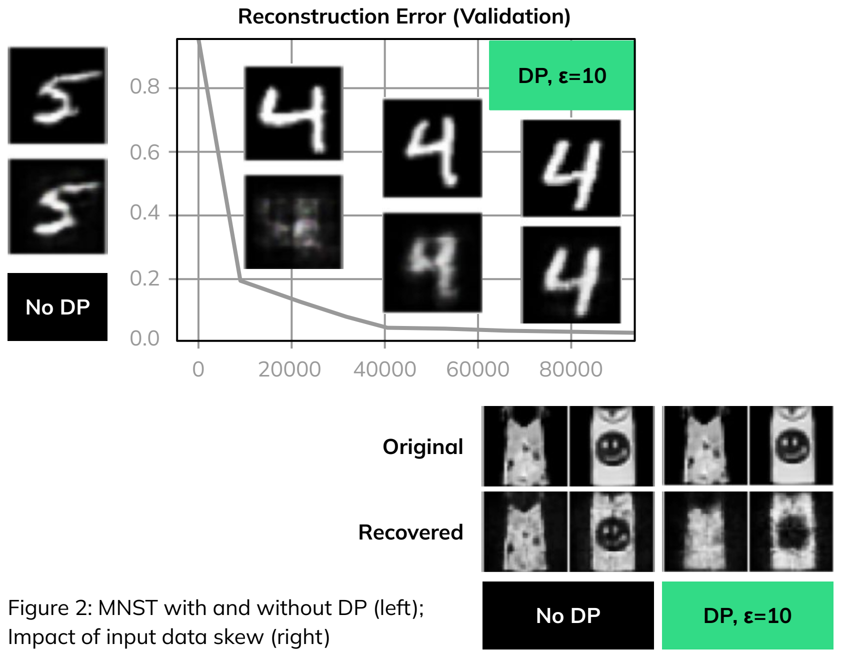


Figure 2: MNST with and without DP (left); Impact of input data skew (right)

FSHA on Split Learning with and without Differential Privacy

We first apply DP-parameterized noise to the gradients during training with the objective to mitigate the FSHA attack. We observe that, while the attack is significantly slowed down, recovery of the underlying data is still possible given enough time.

Impact of input data skew

If we remove some of the classes from the public data set, the attack achieves lower reconstruction precision, especially with DP applied. This demonstrates that the reconstructions are really generated out of the feature space driven by the access to similar public data.

The closer the public and private data distributions are, the more successful the attacks.

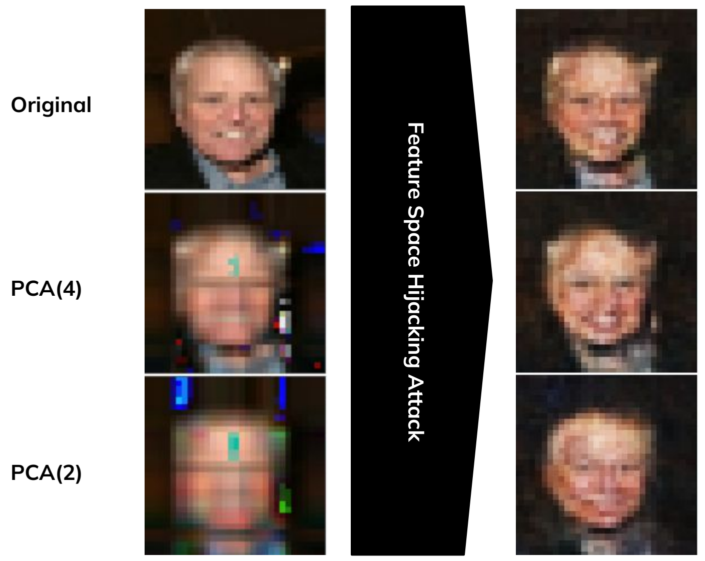


Figure 3: Dimensionality reduction (CelebA)

Dimensionality reduction experiment

Given that training a split neural network with differential privacy can at most delay the successful outcome of a FSHA attack, we find that additional measures are necessary when deploying SL in untrusted environments.

- The inclusion of data preprocessing with dimensionality reduction prior to training is an effective countermeasure to the FSHA attack
- In Fig. 3, input data are compressed by retaining 4 or 2 principal components; the underlying ML task is still possible while limiting the ability of FSHA to recover the original image
- In practice, determining the number of principal components to retain depends on the complexity of the ML task at hand and the required level of privacy; in Fig. 3, FSHA is still able to recover high-level features such as hair colour

Using Differential Privacy with Split Learning only **delays** Feature Space Hijacking Attacks

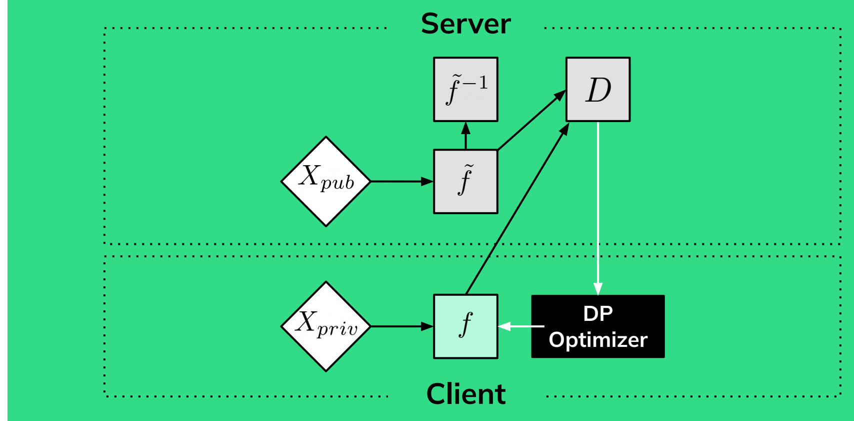
To mitigate:

- Detect and stop learning early
- Make sure it’s difficult to create good public data (not always possible)
- Reduce dimensionality of the input data (utility loss might make it infeasible, and it might still leak private features)
- Secure the compute environment (technologies like Intel SGX; MPC; HE)

Architecture

In FSHA*, an attacker controls the server and trains an auto encoder and Discriminator in a GAN setting.

DP is applied to the gradients coming from the server with the objective to thwart or slow down the attack.



* Pasquini, D.; Ateniese, G.; and Bernaschi, M. 2021. Unleashing the Tiger: Inference Attacks on Split Learning. arXiv:2012.02670.