INTRODUCTION

Federated learning is a potential solution to constructing a machine learning model from several local data sources that cannot be exchanged or aggregated. These restrictions are essential in areas where data privacy or security is critical such as healthcare and defense. However, federated learning is also valuable to companies that can shift much of the computing workload to local devices instead. Furthermore, the local datasets are not required to be independent and identically distributed. Hence, a shared robust global model is desirable and, in many cases, cannot be produced without some form of collaborative learning.

At a high level, federated learning is an iterative procedure of rounds until it meets some termination criteria. These rounds consist of sending the global model to users and selecting a subset of users to update the global model. Those chosen users train their local copy of the model, and their resulting models are communicated back and aggregated in some manner to create a new global model. Typically, the final local model's gradients or weights are transmitted back to ensure privacy or keep the necessary resources minimal.

Concerns have arisen that the lack of control or knowledge regarding the local training procedure could allow a user, with malicious intent, to create an update that compromises the global model for all future users. An example of such harm is a backdoor attack, where training attempt to get a model to associate a given manipulation of the input data, known as a trigger, with a particular outcome. Recent work, such as by Li et al., creates backdoor triggers that are not detectable in the data by human or computer vision. However, in federated learning, only the resulting model gradients or weights are communicated, so there is potentially no need to hide the trigger in the data. Furthermore, without access to leverage user input data, there is less information available to help detect and prevent such malicious intent.

Our contribution to federated learning is establishing defense criteria for federated learning, which is effective against multiple attackers, model scaling, and attacks before global model convergence. Furthermore, our threshold still allows many regular users to update the global model, resulting in little to no performance degradation and requiring only a minimal increase in computational resources or time.

METHODOLOGY

We start by giving further specifications regarding the federated learning environment. Our interest is training a global model over M communication rounds with N users. Each iteration randomly selects K users, using a specified proportion of the total users, to participate in the model update. After local training, the next global model is the average of returning model weights by the FedAvg procedure, detailed in McMahan et al.

To produce local datasets that do not have to be independent and identically distributed, we sample from any training data set using a Dirichlet distribution with specified parameter alpha. The Dirichlet sample determines the proportion of each class included in that user's dataset. Larger values of alpha produce more balanced class distributions.

We assume that all users, including malicious, have complete control over all aspects of local training, such as learning rate, the number of epochs, and the model weights they return. For simplicity, we will be picking two sets of hyperparameters for benign and malicious users, respectively, to stay consistent across all communication rounds. The malicious users will poison a given proportion of their local data by adding their backdoor trigger to the input and changing the training label to that of the target class. They intend for the model to associate the trigger with the target class and hence have the future global model identify any input with the trigger as belonging to the target class. For global model evaluation, we split the test set, added the backdoor trigger to half, and removed any target class observations from that half's data. We use classification accuracy from the first half to measure model success, and the proportion of the poisoned half predicted as the target class, known as attack success rate, to measure the backdoor's extent.

We must assume that malicious users cannot compromise the aggregation method used to update the global model other than modify their returned local model weights. However, we make the additional assumption that there exists one user who we can be confident is trustworthy to place in charge of gatekeeping the global model for updates. That user will evaluate incoming model weights and determine whether each contribution is allowed to participate in the FedAvg procedure. Their decision will be made independently of other users and thus will not violate any data privacy or security concerns.

Our Contribution

We intend to show that our proposed method of defense holds up against attack settings much more potent than what is realistic.

EXPERIMENTS

Balanced Data

Unbalanced Data

Many Attackers

----------------------------------------------------------

RETIRED

Next, Xie et al. show that the multiple-user nature of federated learning is exploitable to make more potent and lasting backdoor attacks. By geometrically distributing the backdoor trigger across a few malicious users, they could make the global model exhibit the desired behavior at higher rates and for many iterations after the attack had concluded.

Consider our updated versions of the settings presented by Xie et al. (2020). Then, we will show our proposed method protects either context.

1. Repeated Selection - In this setting, the federated learning algorithm selects the malicious user(s) to participate in consecutive updates to the global model. Note that under the assumptions above, if the global entity is indeed using random selection to select a subset of participants each round, that repeated consecutive selection of malicious participants is unlikely by chance alone. Hence, we can view repeated selection as a worst-case scenario given our federated learning environment.
2. Single-Selection - In contrast to the previous setting, the malicious user(s) are present for only a single update to the global model. Given sufficient communications rounds, we expect a malicious user to be re-selected to affect the global model. Therefore, most backdoor attacks will fall somewhere between these two situations.