Core Fundamentals of Federated Learning and Machine Unlearning - Week 5

Phoenix Asange-Harper

Performance Comparison

Model	Dataset Samples	Test Set Accuracy	Unlearn Set Accuracy	Utility Preservation	Forgetting Effectiveness
Baseline Model	60000	0.96	N/A	1	0
Unlearned Model	6000	0.90	0.89	0.93	0.89

This table shows that the retrain-from-scratch method while unlearning on the MNIST data is very effective in terms of removing data from the training set and retaining accuracy. Although the unlearned model didn't reach a low accuracy as expected, the 54,000 samples were in fact removed before retraining from scratch. The high performance on the unlearned set was not due to a failure in deleting the data, it was indicative of the homogeneity of the MNIST dataset. Moreover, the effectiveness of this method is shown by the model accuracy on the test set dropping by only 6% while the size of the dataset dropped by 90%. The performance of this may vary on tasks that are more complex than hand-written digits. The retrain from scratch method effectively forgot a large set of the data and retained good accuracy, showing that this method works well on homogeneous datasets.