Paper Title: "WAFFLE: Watermarking in Federated Learning"

Paper Link: https://arxiv.org/abs/2008.07298

1.Summary:

WAFFLE is a watermarking technique for DNN models trained using federated learning. It embeds a resilient watermark into models without requiring access to training data. WAFFLE introduces a retraining step at the server after each aggregation of local models into the global model, resulting in only a negligible degradation in test accuracy. WAFFLE also introduces a novel technique to generate the backdoor used as a watermark, which outperforms prior techniques in terms of communication and computational overhead. WAFFLE satisfies the requirements of demonstrating ownership and model utility in extreme non-IIDness. However, pre-embedded and post-embedded watermarking techniques are not feasible for watermarking federated learning models as they fail to satisfy certain requirements and are not resilient to removal attacks .

1.1Motivation:

The motivation of WAFFLE is to address the issue of model theft in federated learning, where the resulting model is available on every client device, increasing the risk of unauthorized access and theft by malicious clients .The aim of WAFFLE is to provide a means for model owners to demonstrate ownership of their models through watermarking techniques .WAFFLE introduces a retraining step at the server after each aggregation of local models into the global model, efficiently embedding a resilient watermark into models without requiring access to training data .The goal of WAFFLE is to embed a watermark into the DNN models trained using federated learning, ensuring that the

watermark is resistant to removal attacks and maintaining test accuracy with only a negligible degradation (-0.17%)

1.2 Contribution:

In this paper,WAFFLE introduces a solution for addressing the ownership problem in client-server federated learning. It efficiently embeds a resilient watermark into DNN models trained using federated learning, incurring only a negligible degradation in test accuracy (-0.17%

1.3 Methodology:

WAFFLE introduces a watermarking technique for DNN models trained using federated learning.It involves a retraining step at the server after each aggregation of local models into the global model.WAFFLE efficiently embeds a resilient watermark into the models without requiring access to training data.The watermark is generated using a novel technique called WAFFLEPATTERN, which is independent of training data and suitable for federated learning.WAFFLEPATTERN uses the same pattern for each class, helping the model converge and overfit to the watermark pattern.The watermarking process is executed after the aggregation step of federated learning and is independent of the aggregation method used.WAFFLE can be combined with other robust aggregation methods such as Krum, trimmed mean, or median.The methodology ensures that the watermark is resistant to removal attacks and maintains test accuracy with only a negligible degradation.

1.4 Conclusion:

WAFFLE is the first approach to watermark DNN models trained using federated learning.It addresses the problem of demonstrating ownership of models trained via client-server federated learning.WAFFLE efficiently embeds a resilient watermark into models with only a negligible degradation in test accuracy (-0.17%).It introduces the novel technique of WAFFLEPATTERN for generating watermarks, which is independent of training data and suitable for federated learning.WAFFLE ensures that the watermark is resistant to removal attacks and maintains test accuracy.Pre-embedding and post-embedding techniques, which are commonly used for watermarking, are not feasible for watermarking federated learning models.WAFFLE overcomes the limitations of existing

watermarking techniques and provides a reliable method for demonstrating ownership of models trained in a federated learning setting.

2.Limitations:

2.1 First Limitation: Existing watermarking techniques for centralized machine learning cannot be directly implemented in client-server federated learning, as they assume full control over the training process. Pre-embedded watermarks in federated learning models can be easily removed after several aggregation rounds, while post-embedded watermarks are not resilient to removal attacks such as fine-tuning and pruning.
2.2 Second Limitation: WAFFLE does not satisfy the requirements of W1 (demonstration of ownership) and W2 (robustness) as defined for an effective watermarking scheme in federated learning. The computational overhead of WAFFLE is higher in CIFAR10 due to the absence of batch normalization layers in the VGG16 model used.WAFFLE is designed for a single-owner scenario, and extending it to multiple owners may present challenges such as an increase in the size of the watermark set and a potential decrease in utility.

3.Synthesis:

WAFFLE can be applied in scenarios where federated learning is used to train machine learning models on client devices, such as mobile devices or edge devices, while maintaining privacy and efficiency.It can be used by model owners to demonstrate ownership of their models and deter model theft by embedding watermarks into the models.The resilience of the watermark against removal attacks makes WAFFLE suitable for applications where the models need to be protected from unauthorized use or distribution.Future research can explore the extension of WAFFLE to support multiple owners, enabling collaborative federated learning scenarios.Further investigation can be done to evaluate the performance of WAFFLE in different federated learning settings and with various types of models and datasets.The

computational overhead of WAFFLE can be optimized to make it more efficient for resource-constrained devices.