Random Layer Aggregation Effects on Federated Learning Model Alignment and Robustness

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the impact of random layer aggregation on model alignment in federated learning, as measured by cross-domain performance (e.g., transfer learning accuracy on ImageNet and CIFAR-10) and. The advent of federated learning has facilitated large-scale data exchange amongst machine learning models while maintaining privacy. Despite its brief history, federated learning is rapidly evolving to make wider use more practical. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Federated and Transfer Learning: A Survey on Adversaries and Defense Mechanisms. Research question: What is the impact of random layer aggregation on model alignment in federated learning, as measured by cross-domain performance (e.g., transfer learning accuracy on ImageNet and CIFAR-10) and out-of-distribution robustness (e.g., performance on DomainBed benchmarks)?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.1/10.

3 Results

10 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.1/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Outsider Adversaries include attacks by eavesdroppers on the communication line between clients and the FL server.	×	0.06
Outsider Adversaries include attacks by clients of the FL model once it is provided as a service.	×	0.05
Insider Adversaries involve attacks initiated from the server or the edge of the network.	×	0.05
Byzantine and Sybil attacks are identified as two of the most important insider attacks.	×	0.02
Semi-Honest Adversaries are non-aggressive adversaries that attempt to discover the hidden states of other users while a	×	0.02
Semi-Honest Adversaries can only access received information, such as the global model's parameters.	×	0.05
Training Manipulation is defined as the process of learning, affecting, or distorting the FL model itself.	×	0.04
Attackers can damage the integrity of the learning process by attacking the training data or the model during the traini	×	0.06
Inference Manipulation mainly consists of evasion or inference attacks.	×	0.01
Inference Manipulation attacks usually deceive the model into making incorrect decisions or gather information.	×	0.01

References

- http://arxiv.org/abs/2207.02337v1
- http://arxiv.org/abs/2209.05395v1
- http://arxiv.org/abs/2308.13515v5