Communication-Efficient Learning of Deep Networks from Decentralized Data

Summary

Ruofan Liao, 50079732

Federated Learning is a decentralized approach designed to train machine learning models on distributed data while safeguarding data privacy. By keeping data localized on client devices such as smartphones or laptops, it avoids the privacy risks associated with traditional centralized data storage. This paper explores the core architecture of federated learning, its advantages in ensuring data privacy, and the challenges posed by non-independent and identically distributed (non-iid) data.

Client-Server Architecture

Federated learning operates on a client-server architecture, where a central server coordinates the training process across numerous client devices. Each client utilizes its local dataset to perform computations, such as stochastic gradient descent (SGD), generating model updates. These updates are sent to the server, which aggregates them to improve the global model. The updated global model is then distributed back to the clients for the next training iteration. This iterative process ensures that the server only accesses aggregated model updates, not raw data, thereby preserving privacy.

Data Privacy

Enhancing data privacy is a primary motivation for federated learning. By retaining data on client devices and sharing only model updates, federated learning significantly reduces the risks of privacy breaches associated with centralized storage. This approach aligns with principles of data minimization and frameworks such as the 2012 White House Consumer Data Privacy framework. The exchanged updates contain only model improvement information and no raw data, providing an additional layer of privacy protection.

Challenges of Non-iid Data

In practice, client data is often non-independent and identically distributed (non-iid). Each client's dataset may vary significantly in size and distribution, reflecting individual usage patterns rather than a global distribution. This heterogeneity can lead to model bias favoring dominant client data distributions and hinder the generalization ability of the global model. To address this, algorithms such as Federated Averaging (FedAvg) combine local SGD with server-side model averaging,

effectively reducing the number of communication rounds while being robust to imbalanced and non-iid data distributions.

Conclusion

Federated learning offers a promising framework for building shared models on decentralized data, respecting user privacy and tackling the challenges of non-iid data. With appropriate algorithms and sufficient client computational resources, it demonstrates new potential for advancing distributed deep learning.