Decentralized Deep Learning: How to Create Machine Learning Models While Keeping Data Private

Eric Ihre-Thomason & Tom Hagander

May 2024

Imagine a world where your data helps improve technology without ever leaving your device. Our research in decentralized deep learning aims to make this a reality, by creating better models through collaboration without compromising privacy.

In the rapidly evolving field of machine learning, privacy concerns are becoming more and more important. Traditional methods of training machine learning models often require large amounts of data which often leads to entities collecting data that people would rather keep private. Our thesis explores decentralized deep learning strategies, where individual users or organizations can collaboratively train models without exposing their private data.

Decentralized deep learning works by letting all participating parties calculate their similarity to other participating parties. Then participants choose to collaborate with each other based on their similarity, dictated by a communication strategy. To preserve privacy all collaborators only share their model parameters when they calculate their similarity and when they share their insights. This lets all collaborators that do not have enough data to be able train a well-performing model do so without exposing their private data, even if the data differs significantly between collaborators.

We investigated two main communication strategies: Decentralized Adaptive Clustering (DAC) and Personalized Adaptive Neighbor Matching (PANM). Additionally, we explored various similarity metrics together with these communication strategies.

Our experiments were conducted using various datasets, including CIFAR-10 and Fashion-MNIST, simulating real-world scenarios where data is distributed unevenly among multiple participants. The results highlight the potential of DAC and PANM strategies in improving model performance while maintaining data privacy in various settings.

Our research demonstrates the feasibility of decentralized deep learning and presents some unsolved problems, providing a foundation for future work in this area. By improving communication strategies and the use of similarity metrics, we can make collaborative learning systems more reliable and effective. This has far-reaching implications, from healthcare to finance, where data privacy is highly important, and collaborative learning can drive significant advancements.