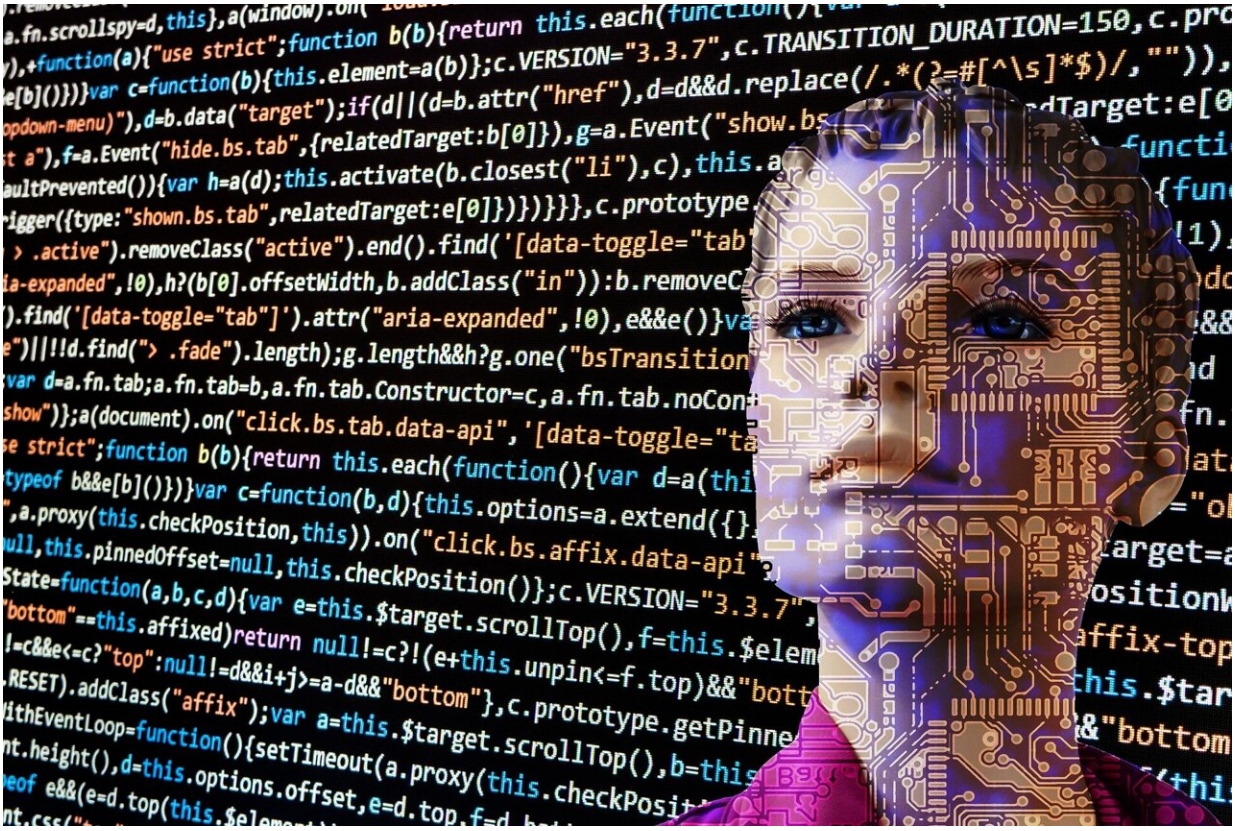
AI researchers tackle longstanding 'data heterogeneity' problem for federated learning

July 11 2022, by Matt Shipman



Datasets used to train AI algorithms may underrepresent older people. Credit: Pixabay/CC0 Public Domain

Researchers from North Carolina State University have developed a new approach to federated learning that allows them to develop accurate

artificial intelligence (AI) models more quickly and accurately. The work focuses on a longstanding problem in federated learning that occurs when there is significant heterogeneity in the various datasets being used to train the AI.

Federated learning is an AI training technique that allows AI systems to improve their performance by drawing on multiple sets of data without compromising the privacy of that data. For example, federated learning could be used to draw on privileged patient data from multiple hospitals in order to improve diagnostic AI tools, without the hospitals having access to data on each other's patients.

Federated learning is a form of machine learning involving multiple devices, called clients. The clients and a centralized server all start with a basic model designed to solve a specific problem. From that starting point, each of the clients then trains its local model using its own data, modifying the model to improve its performance. The clients then send these "updates" to the centralized server. The centralized server draws on these updates to create a hybrid model, with the goal of having the hybrid model perform better than any of the clients on their own. The central server then sends this hybrid model back to each of the clients. This process is repeated until the system's performance has been optimized or reaches an agreed-upon level of accuracy.

"However, sometimes the nature of a client's personal data results in changes to the local model that work well only for the client's own data, but don't work well when applied to other data sets," says Chau-Wai Wong, corresponding author of a paper on the new technique and an assistant professor of electrical and computer engineering at NC State. "In other words, if there is enough heterogeneity in the data of the clients, sometimes a client modifies its local model in a way that actually hurts the performance of the hybrid model."

"Our new approach allows us to resolve the heterogeneity problem more efficiently than previous techniques, while still preserving privacy," says Kai Yue, first author of the paper and a Ph.D. student at NC State. "In addition, if there is enough heterogeneity in the client data, it can be effectively impossible to develop an accurate model using traditional federated learning approaches. But our new approach allows us to develop an accurate model regardless of how heterogeneous the data are."

In the new approach, the updates clients send to the centralized server are reformatted in a way that preserves data privacy, but gives the central server more information about the data characteristics that are relevant to model performance. Specifically, the client sends information to the server in the form of Jacobian matrices. The central server then plugs these matrices into an algorithm that produces an improved model. The central server then distributes the new model to the clients. This process is then repeated, with each iteration leading to model updates that improve system performance.

"One of the central ideas is to avoid iteratively training the local model at each client, instead letting the server directly produce an improved hybrid model based on clients' Jacobian matrices," says Ryan Pilgrim, a co-author of the paper and former graduate student at NC State. "In doing so, the algorithm not only sidesteps multiple communication rounds, but also keeps divergent local updates from degrading the model."

The researchers tested their new approach against industry-standard data sets used to assess federated learning performance, and found the new technique was able to match or surpass the accuracy of federated averaging—which is the benchmark for federated learning. What's more, the new approach was able to match that standard while reducing the number of communication rounds between the server and clients by an

order of magnitude.

"For example, it takes federated averaging 284 rounds of communication to reach an accuracy of 85% in one of the test data sets," Yue says. "We were able to reach 85% accuracy in 26 rounds."

"This is a new, alternative approach to federated learning, making this exploratory work," Wong says. "We're effectively repurposing analytical tools for practical problem-solving. We look forward to getting feedback from the private sector and from the broader federated learning research community about its potential."

The paper, "Neural Tangent Kernel Empowered Federated Learning," will be presented at the 39th International Conference on Machine Learning (ICML), which is being held in Baltimore, Md., July 17-23.

More information: Kai Yue et al, Neural Tangent Kernel Empowered Federated Learning, *arXiv* (2022). arXiv:2110.03681 [cs.LG] arxiv.org/abs/2110.03681

Conference: icml.cc/

Provided by North Carolina State University