https://doi.org/10.1038/s41587-023-01770-3

Data sharing in the age of deep learning

Check for updates

How can we protect personal information and the integrity of artificial intelligence models when sharing data?

High-quality, large datasets are the cornerstone of successful deep-learning algorithms. Although algorithmic advances can sometimes achieve better prediction accuracy without using more data, the size of training data remains the one most important factor for success. Perplexity — which is a measure for how well a model predicts a sample — improves roughly linearly with more data, but recent examples from natural language processing have shown that capabilities of models emerge when the training data are large enough.

There are clear advantages to combining data from different sources, but in the biotechnology sector individual privacy as well as intellectual property concerns often stand in the way of sharing data. This is a detriment to the whole field. Federated learning has been proposed as one solution to this dilemma. The concept of federated learning is that no raw data are shared between the participants; instead, local models are trained for each data silo or repository, followed by multiple iterations of model aggregation into a global model, distribution of the global model to all participants and retraining on the local data silos. Personal privacy or intellectual property is protected, and the artificial intelligence (AI) model can still be trained.

In addition to increasing the size of available training data, training on multiple datasets derived from multiple sources also has potential to reduce biases and lead to models with higher generalizability. Biases in AI have received a lot of media attention in the realm of text and image generation, but the same types of representational biases of race, social class, gender and so on also exist in many datasets that are relevant for the biotechnology sector (such as sequencing data). Although new 'big data' collection projects explicitly aim to sample in a fair and representative way, existing inequalities will continue to persist. Additionally, biases in biological datasets extend far beyond this human-centric view. For example, there are large amounts of detailed data available for a few model organisms, but very sparse data for large numbers of species. There are few cell lines that are very well characterized, and high-throughput screens are biased to particular classes of chemicals. Although the combination of multiple datasets cannot mitigate the problem of bias completely, in most cases the representational bias of the combination will be lower than that of individual datasets. How much of this advantage can be exploited in the federated learning regime is a matter of active debate, but in many cases it seems to be somewhere in between the case of training only local models and the centralized paradigm in which all data are combined.

It is encouraging to see that multiple federated learning projects have successfully been implemented in the past few years on different scales, even though there are organizational challenges to using a distributed learning approach. For example, the melloddy project is a collaboration between ten pharmaceutical companies and seven technology and academic strategic partners that was completed last year; a study (I. Dayan et al. *Nat. Med.* **27**, 1735–1743; 2023) predicted clinical outcomes in patients with COVID-19 with data collected across 20 institutions; and a study (J. Ogier du Terrail et al. *Nat. Med.* **29**, 135–146; 2023) from a collaboration between multiple hospitals that predicted response to neoadjuvant chemotherapy in triple-negative breast cancer was published at the beginning of this year. These projects have demonstrated the potential applications of data sharing and using AI models, while respecting privacy and intellectual property.

Of course, there are still concerns relating to data leakage and security with any analysis that uses sensitive data. Although the raw data never leave the organization that provides the dataset, it has been shown that, in some cases, raw data can be recovered from the model weights and their updates in a so-called gradient inversion attack. In less extreme scenarios, partial information about raw data can be leaked.

These attacks on data privacy can be defended against, using large batches of training data that tend to obscure the effect of individual records, differential privacy, secure multiparty computation or homomorphic encryption (a form of encryption that enables computation on encrypted data but comes at the cost of substantial computational overhead and limitations to the types of computations that can be performed). Although effective defenses against data leakage are possible, concerns remain that, with ever-increasing computing power, algorithms that are considered secure today might become breakable in the future and data could be reconstructed from retrospective datasets.

In addition to privacy, the security of federated learning systems needs to be ensured — a matter that has received far less attention in the biotechnology or healthcare sector. The decentralized nature of the federated learning paradigm lends itself to attacks such as data or model poisoning or the creation of backdoors: if one participant sends carefully manipulated model updates, they can corrupt the performance of the trained global model on specific subtasks.

Although some protections against backdoor attacks exist, they are mostly based on noise injection and negatively affect the benign performance of the model. With large financial incentives at stake in the biotechnology and healthcare sectors, these types of attacks should not be ignored. Even without malicious intent, problems can arise from different data curation and quality control processes that have a detrimental effect on global model performance.

The incentives for data sharing are clear, but although technologies such as federated learning can overcome some of the obstacles related to privacy and intellectual property, their application still is the exception and not the rule. Where intellectual property is concerned, models from game theory might help to set the right incentives such that those parties that contribute the most or highest quality data may also reap larger benefits. As it is challenging to defend against an internal threat, nontechnological strategies — such as the careful selection of partners, tests for data curation compliance, and trusted validation datasets and procedures — will need to be developed and standardized. Most probably, a combination of technological, organizational, regulatory and legislative solutions will be required to enable the shift from competition to data-private, secure and collaborative machine learning for a large number of players.

Published online: 5 April 2023