Big data privacy for machine learning just got 100 times cheaper

November 16 2021, by Jade Boyd

Rice University computer scientist Ashumali Shrivastava (left) and graduate student Ben Coleman discovered an inexpensive way to implement rigorous personal data privacy when using or sharing large databases for machine learning. Credit: Jeff Fitlow/Rice University

Rice University computer scientists have discovered an inexpensive way

for tech companies to implement a rigorous form of personal data privacy when using or sharing large databases for machine learning.

"There are many cases where machine learning could benefit society if data privacy could be ensured," said Anshumali Shrivastava, an associate professor of computer science at Rice. "There's huge potential for improving medical treatments or finding patterns of discrimination, for example, if we could train machine learning systems to search for patterns in large databases of medical or financial records. Today, that's essentially impossible because data privacy methods do not scale."

Shrivastava and Rice graduate student Ben Coleman hope to change that with a new method they'll present this week at CCS 2021, the Association for Computing Machinery's annual flagship conference on computer and communications security. Using a technique called locality sensitive hashing, Shirvastava and Coleman found they could create a small summary of an enormous database of sensitive records. Dubbed RACE, their method draws its name from these summaries, or "repeated array of count estimators" sketches.

Coleman said RACE sketches are both safe to make publicly available and useful for algorithms that use kernel sums, one of the basic building blocks of machine learning, and for machine-learning programs that perform common tasks like classification, ranking and regression analysis. He said RACE could allow companies to both reap the benefits of large-scale, distributed machine learning and uphold a rigorous form of data privacy called differential privacy.

Differential privacy, which is used by more than one tech giant, is based on the idea of adding random noise to obscure individual information.

"There are elegant and powerful techniques to meet differential privacy standards today, but none of them scale," Coleman said. "The

computational overhead and the memory requirements grow exponentially as data becomes more dimensional."

Data is increasingly high-dimensional, meaning it contains both many observations and many individual features about each observation.

RACE sketching scales for high-dimensional data, he said. The sketches are small and the computational and memory requirements for constructing them are also easy to distribute.

"Engineers today must either sacrifice their budget or the privacy of their users if they wish to use kernel sums," Shrivastava said. "RACE changes the economics of releasing high-dimensional information with differential privacy. It's simple, fast and 100 times less expensive to run than existing methods."

This is the latest innovation from Shrivasta and his students, who have developed numerous algorithmic strategies to make machine learning and data science faster and more scalable. They and their collaborators have: found a more efficient way for social media companies to keep misinformation from spreading online, discovered how to train large-scale deep learning systems up to 10 times faster for "extreme classification" problems, found a way to more accurately and efficiently estimate the number of identified victims killed in the Syrian civil war, showed it's possible to train deep neural networks as much as 15 times faster on general purpose CPUs (central processing units) than GPUs (graphics processing units), and slashed the amount of time required for searching large metagenomic databases.

More information: Benjamin Coleman et al, A One-Pass Private Sketch for Most Machine Learning Tasks, arXiv:2006.09352 [cs.DS], arxiv.org/abs/2006.09352

Provided by Rice University