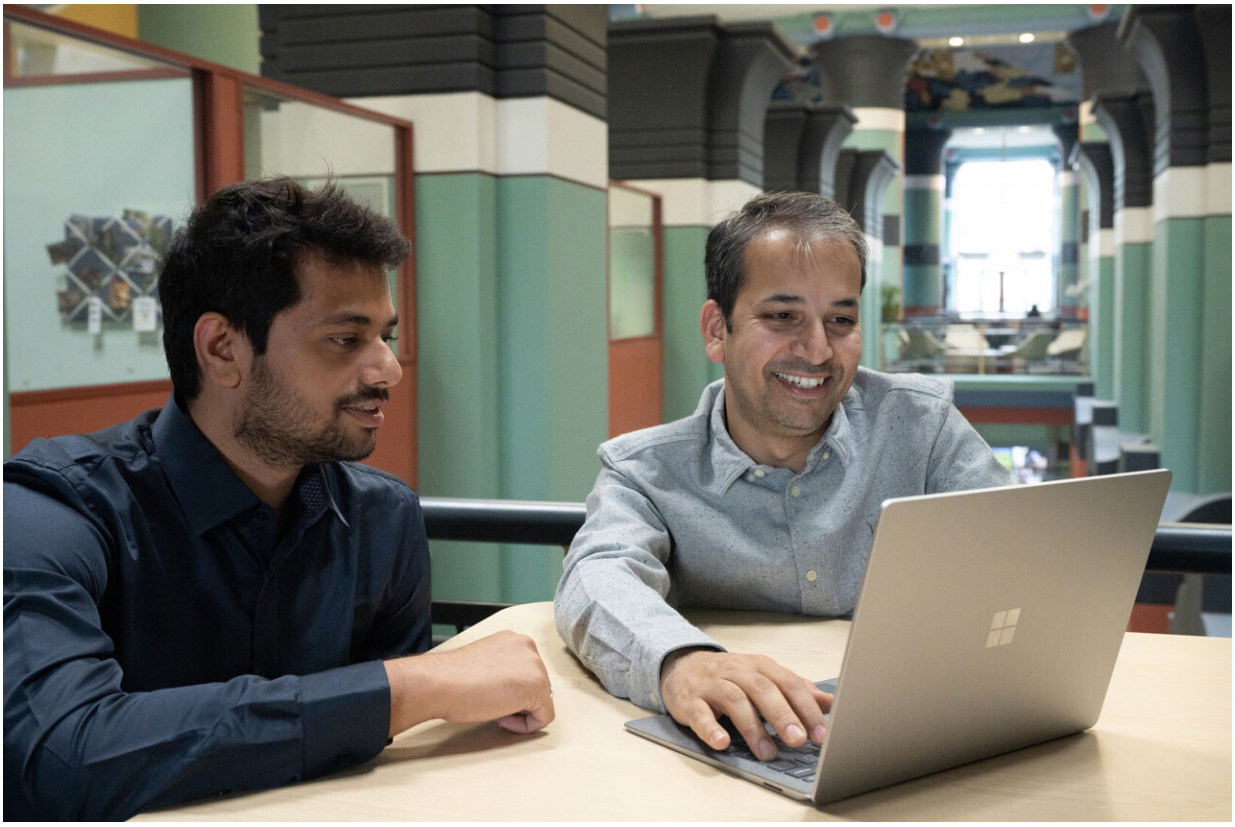
ROBE Array could let small companies access popular form of AI

August 29 2022, by Jade Boyd



Rice University computer scientists Anshumali Shrivastava (right) and Aditya Desai created ROBE Array, a breakthrough low-memory technique for deep-learning recommendation models, a popular form of artificial intelligence that learns to make suggestions users will find relevant. Credit: Jeff Fitlow/Rice University

A breakthrough low-memory technique by Rice University computer scientists could put one of the most resource-intensive forms of artificial intelligence—deep-learning recommendation models (DLRM)—within reach of small companies.

DLRM recommendation systems are a popular form of AI that learns to make suggestions users will find relevant. But with top-of-the-line training models requiring more than a hundred terabytes of memory and supercomputer-scale processing, they've only been available to a short list of technology giants with deep pockets.

Rice's "random offset block embedding array," or ROBE Array, could change that. It's an algorithmic approach for slashing the size of DLRM memory structures called embedding tables, and it will be presented this week at the Conference on Machine Learning and Systems (MLSys 2022) in Santa Clara, California, where it earned Outstanding Paper honors.

"Using just 100 megabytes of memory and a single GPU, we showed we could match the training times and double the inference efficiency of state-of-the-art DLRM training methods that require 100 gigabytes of memory and multiple processors," said Anshumali Shrivastava, an associate professor of computer science at Rice who's presenting the research at MLSys 2022 with ROBE Array co-creators Aditya Desai, a Rice graduate student in Shrivastava's research group, and Li Chou, a former postdoctoral researcher at Rice who is now at West Texas A&M University.

"ROBE Array sets a new baseline for DLRM compression," Shrivastava said. "And it brings DLRM within reach of average users who do not have access to the high-end hardware or the engineering expertise one needs to train models that are hundreds of terabytes in size."

DLRM systems are machine learning algorithms that learn from data.

For example, a recommendation system that suggests products for shoppers would be trained with data from past transactions, including the search terms users provided, which products they were offered and which, if any, they purchased. One way to improve the accuracy of recommendations is to sort training data into more categories. For example, rather than putting all shampoos in a single category, a company could create categories for men's, women's and children's shampoos.

For training, these categorical representations are organized in memory structures called embedding tables, and Desai said the size of those tables "have exploded" due to increased categorization.

"Embedding tables now account for more than 99.9% of the overall memory footprint of DLRM models," Desai said. "This leads to a host of problems. For example, they can't be trained in a purely parallel fashion because the model has to be broken into pieces and distributed across multiple training nodes and GPUs. And after they're trained and in production, looking up information in embedded tables accounts for about 80% of the time required to return a suggestion to a user."

Shrivastava said ROBE Array does away with the need for storing embedding tables by using a data-indexing method called hashing to create "a single array of learned parameters that is a compressed representation of the embedding table." Accessing embedding information from the array can then be performed "using GPU-friendly universal hashing," he said.

Shrivastava, Desai and Chou tested ROBE Array using the sought after DLRM MLPerf benchmark, which measures how fast a system can train models to a target quality metric. Using a number of benchmark data sets, they found ROBE Array could match or beat previously published DLRM techniques in terms of training accuracy even after compressing

the model by three orders of magnitude.

"Our results clearly show that most deep-learning benchmarks can be completely overturned by fundamental algorithms," Shrivastava said. "Given the global chip shortage, this is welcome news for the future of AI."

ROBE Array isn't Shrivastava's first big splash at MLSys. At MLSys 2020, his group unveiled SLIDE, a "sub-linear deep learning engine" that ran on commodity CPUs and could outperform GPU-based trainers. They followed up at MLSys 2021, showing vectorization and memory optimization accelerators could boost SLIDE's performance, allowing it to train deep neural nets up to 15 times faster than top GPU systems.

More information: Random Offset Block Embedding (ROBE) for compressed embedding tables in deep learning recommendation systems

Provided by Rice University