What is embedding table? Embedding learning is a core technique for modeling categorical features in deep recommendation models. It maps sparse categorical features into dense vectors.

What is the problem? Industrial recommendation models demand an extremely large number of parameters for embedding tables, requiring multi-terabyte memory. We have to partition the tables and put them in multiple GPUs.

Background of Embedding Tables

Challenges

- It is challenging to estimate the costs. The total cost of multiple tables in a shard is not the sum of the single table costs within the shard due to parallelism and operator fusion.
- The partitioning problem is known to be NP-hard.

AutoShard Framework

We present our novel practice in Meta, namely AutoShard, based on cost modeling and deep reinforcement learning (RL).

Results

(a) Degree of balance

(a) Degree of balance

(b) Speedup

(b) Speedup

Figure: AutoShard significantly outperforms all the baselines.