Pinterest Boosts Home Feed Engagement 16% With Switch to GPU Acceleration of Recommenders

Image-sharing network taps cuCollections and CUDA Graphs for recommender system, boosting engagement and inference efficiency more than 100x.

Author: Scott Martin

Pinterest has engineered a way to serve its photo-sharing community more of the images they love.

The social-image service, with more than 400 million monthly active users, has trained bigger recommender models for improved accuracy at predicting people's interests.

Pinterest handles hundreds of millions of user requests an hour on any given day. And it must also narrow down relevant images from roughly 300 billion images on the site to roughly 50 for each person.

The last step — ranking the most relevant and engaging content for everyone using Pinterest — required a leap in acceleration to run heftier models, with minimal latency, for better predictions.

Pinterest has improved the accuracy of its recommender models powering people's home feeds and other areas, increasing engagement by as much as 16%.

The leap was enabled by switching from CPUs to NVIDIA GPUs, which could easily be applied next to other areas, including advertising images, according to Pinterest.

"Normally we would be happy with a 2% increase, and 16% is just a beginning for home feeds. We see additional gains — it opens a lot of doors for opportunities," said Pong Eksombatchai, a software engineer at Pinterest.

Transformer models capable of better predictions are shaking up industries from retail to entertainment and advertising. But their leaps in performance gains of the past few years have come with a need to serve models that are some 100x bigger as their number of model parameters and computations skyrockets.

Like many, Pinterest engineers wanted to tap into state-of-the-art recommender models to increase engagement. But serving these massive models on CPUs presented a 100x increase in cost and latency. That wasn't going to maintain its magical user experience — fresh and more appealing images — occurring within a fraction of a second.

"If that latency happened, then obviously our users wouldn't like that very much because they would have to wait forever," said Eksombatchai. "We are pretty close to the limit of what we can do on CPU basically."

The challenge was to serve these hundredfold larger recommender models within the same cost and latency constraints.

Working with NVIDIA, Pinterest engineers began architectural changes to optimize their inference pipeline and recommender models to enable the transition from CPU to GPU cloud instances. The technology transition began late last year and required major changes to how the company manages workloads. The result is a 100x gain in inference efficiency on the same IT budget, meeting their goals.

"We are starting to use really, really big models now. And that is where the GPU comes in — to help make these models possible," Eksombatchai said.

Switching from CPUs to GPUs required rethinking its inference systems architecture. Among other issues, engineers had to change how they send workloads to their inference servers. Fortunately, there are tools to assist in making the transition easier.

The Pinterest inference server built for CPUs had to be altered because it was set up to send smaller batch sizes to its servers. GPUs can handle much larger workloads, so it's necessary to set up larger batch requests to increase efficiency.

One area where this comes into play is with its embedding table lookup module. Embedding tables are used to track interactions between various context-specific features and interests of user profiles. They can track where you navigate, and what people Pin on Pinterest, share or numerous other actions, helping refine predictions on what users might like to click on next.

They are used to incrementally learn user preference based on context in order to make better content recommendations to those using Pinterest. Its embedding table lookup module required two computation steps repeated hundreds of times because of the number of features tracked.

Pinterest engineers greatly reduced this number of operations using a GPU-accelerated concurrent hash table from NVIDIA cuCollections. And they set up a custom consolidated embedding lookup module so they could merge requests into a single lookup. Better results were seen immediately.

"Using cuCollections helped us to remove bottlenecks," said Eksombatchai.

Pinterest relied on CUDA Graphs to eliminate what was remaining of the small batch operations, further optimizing its inference models.

CUDA Graphs helps reduce the CPU interactions when launching on GPUs. They're designed to enable workloads to be defined as graphs rather than single operations. They provide a mechanism to launch multiple GPU operations through a single CPU operation, reducing CPU overheads.

Pinterest enlisted CUDA Graphs to represent the model inference process as a static graph of operation instead of as those individually scheduled. This enabled the computation to be handled as a single unit without any kernel launching overhead.

The company now supports CUDA Graph as a new backend of its model server. When a model is first loaded, the model server runs the model inference once to build the graph instance. This graph can then be run repeatedly in inference to show content on its app or site.

Implementing CUDA Graphs helped Pinterest to significantly reduce inference latency of its recommender models, according to its engineers.

GPUs have enabled Pinterest to do something that was impossible with CPUs on the same budget, and by doing this they can make changes that have a direct impact on various business metrics.

Learn about Pinterest's GPU-driven inference and optimizations at its GTC session, Serving 100x Bigger Recommender Models , and in the Pinterest Engineering blog .

Register for GTC , running Sept. 19-22, for free to attend sessions with NVIDIA and dozens of industry leaders.

Original URL: https://blogs.nvidia.com/blog/2022/08/04/pinterest-gpu-acceleration-recommenders/