Billions Served: NVIDIA Merlin Helps Fuel Clicks for Online Giants

Deep learning-based systems enable improved recommendation relevance for billions of online consumers.

Author: Scott Martin

Online commerce has rocketed to trillions of dollars worldwide in the past decade, serving billions of consumers. Behind the scenes of this explosive growth in online sales is personalization driven by recommender engines.

Recommenders make shopping deeply personalized. While searching for products on e-commerce sites, they find you. Or suggestions can just appear. This wildly delightful corner of the internet is driven by ever more massive datasets and models.

NVIDIA Merlin is the rocket fuel of recommenders. Boosting training and inference, it enables businesses of all types to better harness data to build recommenders accelerated by NVIDIA GPUs.

The stakes are higher than ever for online businesses. Online sales in 2021 were expected to reach nearly $5 trillion worldwide, according to eMarketer, up nearly 17 percent from the prior year.

On some of the world's largest online sites, even a 1 percent gain in relevance accuracy of recommendations can yield billions more sales.

Investment in recommender systems has become one of the biggest competitive advantages of internet giants today.

The market for recommenders is expected to reach $15.13 billion by 2026, up from $2.12 billion in 2020, according to Mordor Intelligence. The largest and fastest growing segment of the market for recommender engines is in the Asia Pacific region, according to the research firm.

But an industry challenge is that improved relevance requires more data and processing. This data consists of trillions of user-product interactions — clicks, views, — on billions of products and consumer profiles.

Data of this scale can take days to train models. Yet the faster you can spin out new models informed by more data, the better your relevance.

The Merlin collection of models, methods, and libraries, includes tools for building deep learning-based systems capable of handling terabytes of data that can provide better predictions and increase clicks.

U.S. digital advertising is expected to reach $191.1 billion in 2021, up 25.5 percent from the year before, according to eMarketer.

Snap, parent company to social media app Snapchat, is based in Santa Monica, Calif., and has more than 300 million daily active users. It creates ad revenue from its social photo and video messaging service.

"We will continue to focus on delivering strong results for our advertising partners and innovating to expand the capabilities of our platform and better serve our community," said Snap CEO Evan Spiegel in its third-quarter earnings statement.

The technical hurdle for Snap is that it seeks to continue to develop its workload's higher-cost ranking models and expand into more complex models while reducing costs.

The company used NVIDIA GPUs and Merlin to boost its content ranking capabilities.

“Snap used NVIDIA GPUs and Merlin software to improve machine learning inference cost efficiency by 50 percent and decrease serving latency by 2x, providing the compute headroom to experiment and deploy heavier, more accurate ad and content ranking models,” said Nima Khajehnouri, VP of engineering at Snap.

Entertainment giant Tencent, which operates the enormously popular messaging service WeChat and payments platform WeChat Pay, is China’s largest company by market capitalization.

Its engineers need to rapidly iterate on models for its advertising recommendation system, putting increasing demands on its training performance.

“The advertising business is a relatively important business inside Tencent and the recommendation system is used to increase the overall advertising revenue,” said Xiangting Kong, expert engineer at Tencent.

The problem is that accuracy of advertising recommendation can only be improved by training more sample data, including more sample features, but this leads to longer training times that affect model update frequency.

“HugeCTR, as a recommendation training framework, is integrated into the advertising recommendation training system to make the update frequency of model training faster, and more samples can be trained to improve online effects,” he said.

After the training model performance is improved, more data can be trained to improve the accuracy of the model, increasing advertising revenue, he added.

Meituan’s business is at a crowded intersection of food, entertainment and on-demand services, among its 200 service categories. The Chinese internet giant has more than 667 million active users and 8.3 million active merchants.

Jun Huang, a senior technical expert at Meituan, said that if his team can greatly improve performance, it usually prefers to train more samples and more complex models.

The problem for Meituan was that as its models became more and more complex, it became difficult to optimize the training framework deeply, said Huang.

“We are working on integrating NVIDIA HugeCTR into our training system based on A100 GPUs. The cost is also greatly reduced. This is a preliminary optimization result, and there is still much room to optimize in the future,” he said.

Meituan recently reported its average number of transactions per transacting users increased to 32.8 for the trailing 12 months of the second quarter of 2021, compared with 25.7 for the trailing 12 months of the second quarter of 2020.

Learn more about NVIDIA Merlin . Learn more about NVIDIA Triton .

Original URL: https://blogs.nvidia.com/blog/2022/01/18/nvidia-merlin-helps-fuel-clicks-for-online-giants/