NYU, NVIDIA Collaborate on Large Language Model to Predict Patient Readmission

NYUTron, an AI model featured today in Nature, is deployed at NYU Langone Health.

Author: Anthony Costa

Getting discharged from the hospital is a major milestone for patients — but sometimes, it's not the end of their road to recovery. Nearly 15% of hospital patients in the U.S. are readmitted within 30 days of their initial discharge, which is often associated with worse outcomes and higher costs for both patients and hospitals.

Researchers at NYU Langone Health, the academic medical center of New York University, have collaborated with NVIDIA experts to develop a large language model (LLM) that predicts a patient's risk of 30-day readmission, as well as other clinical outcomes.

Deployed in the healthcare system's six inpatient facilities, the NYUTron model — featured today in the scientific journal Nature — provides doctors with AI-driven insights that could help them identify patients in need of a clinical intervention to reduce the likelihood of readmission.

"When you discharge a patient from the hospital, you don't expect them to need to return, or you probably should have kept them in the hospital longer," said Dr. Eric Oermann, assistant professor of radiology and neurosurgery at NYU Grossman School of Medicine and a lead collaborator on NYUTron. "Using analysis from the AI model, we could soon empower clinicians to prevent or fix situations that put patients at a higher risk of readmission."

The model has so far been applied to more than 50,000 patient discharged in NYU's healthcare system, where it shares predictions of readmission risk with physicians via email notifications. Oermann's team is next planning a clinical trial to test whether interventions based on NYUTron's analyses reduce readmission rates.

The U.S. government tracks 30-day readmission rates as an indicator of the quality of care hospitals are providing. Medical institutions with high rates are fined — a level of scrutiny that incentivizes hospitals to improve their discharge process.

There are plenty of reasons why a recently discharged patient may need to be readmitted to the hospital — among them, infection, overprescription of antibiotics, surgical drains that were removed too early. If these risk factors can be spotted earlier, doctors could intervene by adjusting treatment plans or monitoring patients in the hospital for longer.

"While there have been computational models to predict patient readmission since the 1980s, we're treating this as a natural language processing task that requires a health system-scale corpus of clinical text," Oermann said. "We trained our LLM on the unstructured data of electronic health records to see if it could capture insights that people haven't considered before."

NYUTron was pretrained on 10 years of health records from NYU Langone Health: more than 4 billion words of clinical notes representing nearly 400,000 patients. The model achieved an accuracy improvement of more than 10 percent over a state-of-the-art machine learning model to predict readmission.

Once the LLM was trained for the initial use case of 30-day readmission, the team was able to spin out four other predictive algorithms in around a week. These include predicting the length of a patient's hospital stay, the likelihood of in-hospital mortality, and the chances of a patient's insurance claims being denied.

“Running a hospital is in some ways like managing a hotel,” said Oermann. “Insights that help hospitals operate more efficiently means more beds and better care for a greater number of patients.”

NYUTron is an LLM with hundreds of millions of parameters, trained using the NVIDIA NeMo Megatron framework on a large cluster of NVIDIA A100 Tensor Core GPUs .

“Much of the conversation around language models right now is around gargantuan, general-purpose models with billions of parameters, trained on messy datasets using hundreds or thousands of GPUs,” Oermann said. “We’re instead using medium-sized models trained on highly refined data to accomplish healthcare-specific tasks.”

To optimize the model for inference in real-world hospitals, the team developed a modified version of the NVIDIA Triton open-source software for streamlined AI model deployment using the NVIDIA TensorRT software development kit.

“To deploy a model like this in a live healthcare environment, it has to run efficiently,” Oermann said. “Triton delivers everything you want in an inference framework, making our model blazing fast.”

Oermann’s team found that after pretraining their LLM, fine-tuning it onsite with a specific hospital’s data helped to significantly boost accuracy — a trait that could help other healthcare institutions deploy similar models.

“Not all hospitals have the resources to train a large language model from scratch in-house, but they can adopt a pretrained model like NYUTron and then fine-tune it with a small sample of local data using GPUs in the cloud,” he said. “That’s within reach of almost everyone in healthcare.”

To learn more about NYUTron, read the Nature paper and watch this NVIDIA and NYU talk on demand .

Original URL:
https://blogs.nvidia.com/blog/2023/06/07/nyu-large-language-model-patient-readmission-nature/