Published on *The National Law Review* https://natlawreview.com

5 Questions with Mike DeCesaris: AI/ML Efficiency Driven by GPUs

Article By:

Cornerstone Research

5 Questions is a periodic feature produced by Cornerstone Research, which asks our professionals, senior advisors, or affiliated experts to answer five questions.

We interview Mike DeCesaris, vice president of Cornerstone Research's Data Science Center, about the benefits of working with GPUs, and how they enhance artificial intelligence (AI) and machine learning (ML) techniques.

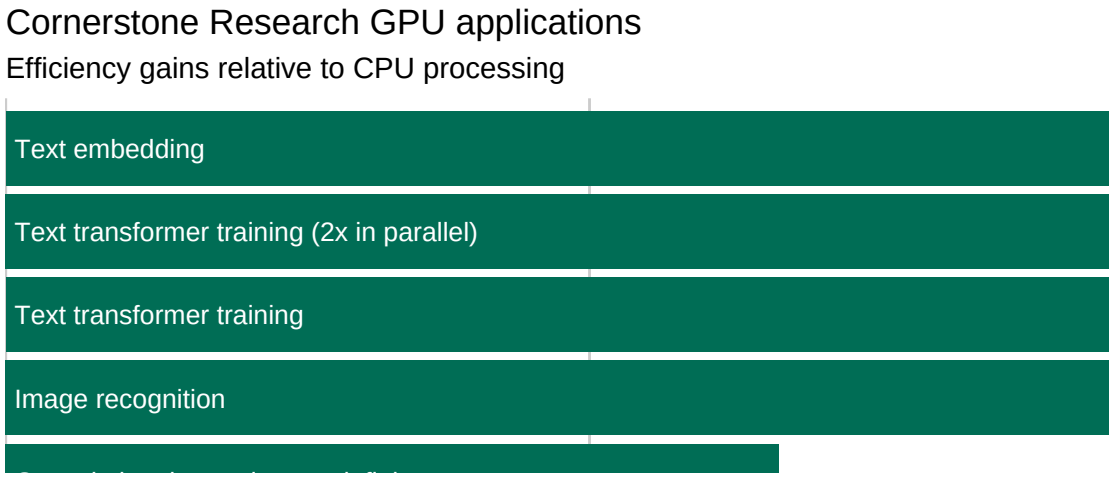
What are GPUs?

Specialized graphics processing units (GPUs), as the name suggests, were originally designed decades ago for the efficient performance of operations common to the processing of images and video. These processes heavily feature matrix-based mathematical calculations. People are generally more familiar with central processing units (CPUs), which are found in laptops, phones, and smart devices, and can perform many different types of operations.

In the early 2000s, researchers realized that, because machine learning algorithms often feature the same type of calculations as graphics processing algorithms, GPUs could provide a more efficient alternative to CPU-based computation for machine learning. Despite availability and cost constraints relative to CPUs in recent years, GPU-based computation has become the de facto standard for machine learning or neural network training.

What are the benefits of using GPUs?

The key benefit is efficiency. The computing efficiency that GPUs provide does more than streamline the analytical process. It facilitates more extensive model training for greater accuracy, expands the scope of the model search process to guard against alternative specifications, makes feasible certain models that previously were infeasible, and allows for additional sensitivities on alternative datasets to ensure robustness.



How do GPUs support expert testimony?

AI-based systems substitute human decisions with data-driven ones. This can reduce subjectivity and error when processing large volumes of complex information. We utilize AI and ML to drive automation of increasingly complex tasks and unlock new approaches for analysis, including using both supervised and unsupervised learning. These techniques are supported by our in-house GPUs.

How does the Data Science Center leverage GPU computing?

We use GPUs at all stages of the case lifecycle, from discovery to economic analysis, and for all types of data, from standard tabular data to text and images. Some of these applications rely on applications where GPU computing has become popular, like neural networks, while others rely on more customized analytical frameworks. Some examples follow.

Matrix arithmetic

GPUs enable us to perform custom matrix arithmetic at rapid speed. For example, in antitrust matters, we often need to calculate the distance between all suppliers and all consumers (coordinate pairs). Migrating this computation from CPUs to GPUs enables us to calculate distances between nearly 100 million coordinate pairs per second.

Deep neural networks

Much of the excitement surrounding GPU-based computation focuses on neural networks. While capable of handling routine classification and regression problems, additional task-specific neural network architectures provide a framework for specialized analyses of text, images, and sound. Given the complexity of these models and the volume of data required to generate reliable results, their use is effectively infeasible without GPU computing resources. When training a popular multi-class image model on a GPU, we experienced a 25,000% speedup compared to running the same process on a single CPU. We leverage this efficiency in content analyses for consumer fraud matters, where we design text and image classifiers to characterize the intended audience of at-issue marketing materials.

Boosted trees

As GPU computing has become more ubiquitous, popular machine learning software packages have increasingly included GPU-based computation options in their offerings. We often use boosted trees

in regression and classification problems. These models sequentially aggregate multiple simple decision trees into a larger, more accurate learner. Compared to deep neural networks, which may feature hundreds of millions of parameters, these models are smaller and thus require less data and training time to produce generalizable inferences. These advantages lead them to be more useful than deep neural networks for many of the types of analyses that we regularly encounter. Switching to GPU-based training processes enables us to train models for these tasks nearly 100 times faster than the corresponding CPU specification.

Language models

Language models, often based on one or more deep learning techniques, can classify, parse, and generate text. We employ large language models to extract specific pieces of information, parse relationships between entities, identify semantic relationships, and supplement traditional term-based features in text classification problems, such as in the quantification of social media sentiment surrounding a public entity in defamation matters.

Unsurprisingly, given all that these models can do, processing documents through these models via CPU can introduce significant delays to the analytical process. With just a single GPU, we can segment documents into individual components and fully process several hundreds of sentences per second.

What developments can we expect in this space in the future?

GPUs and GPU-related software will continue to evolve. New hardware may feature more cores, faster cores, and more memory to accommodate larger models and data batches. New software may make it even easier to share models and data across multiple GPUs.

Other developments may involve different devices altogether. To address some of the inefficiencies still present in GPU computing, machine learning practitioners have increasingly turned to application-specific integrated circuits (ASIC) and field-programmable gate arrays (FPGAs). For example, Google’s tensor processing unit (TPU) is an ASIC designed specifically to perform calculations for its TensorFlow software package for machine learning. FPGAs offer more flexibility and are typically used to deploy machine learning models in production environments that require low latency, high bandwidth, and minimal energy consumption.

We continue to monitor developments in this space to ensure that we continue to provide best-in-class service to our clients and experts.

National Law Review, Volume XII, Number 305