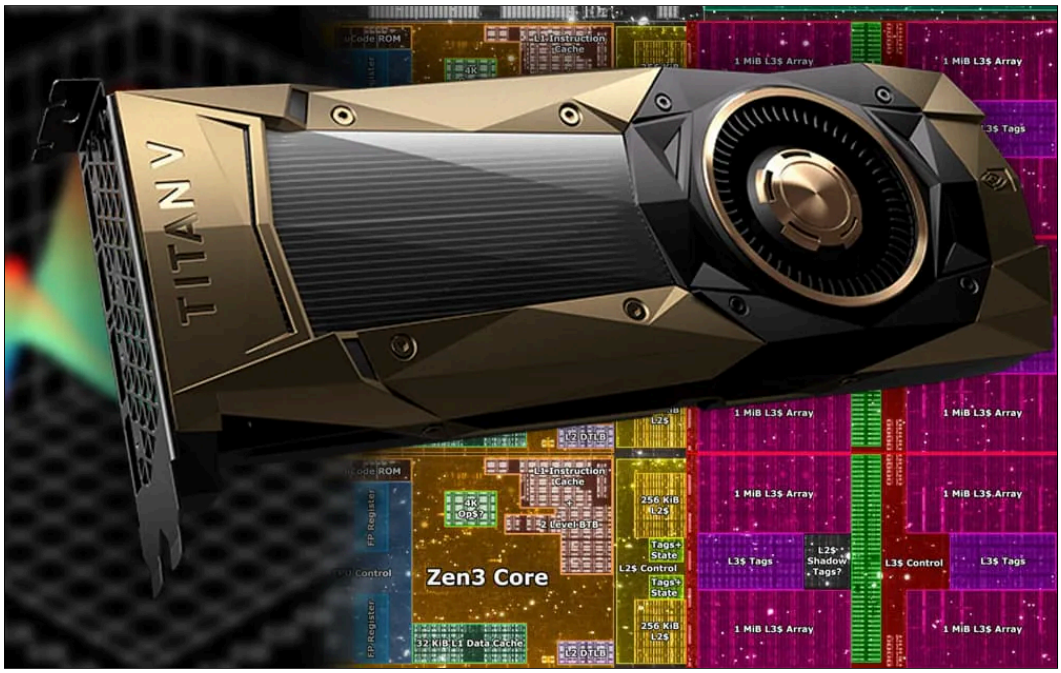
GPUs May Be Better, Not Just Faster, at Training Deep Neural Networks

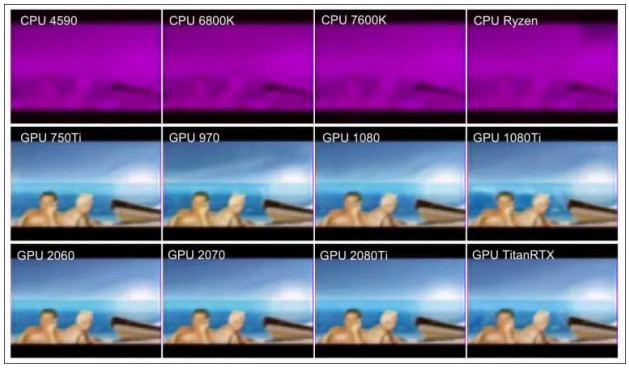
Originally published at https://www.unite.ai/gpus-may-be-better-not-just-faster-at-training-deep-neural-networks/



Researchers from Poland and Japan, working with Sony, have found evidence that machine learning systems trained on GPUs rather than CPUs may contain fewer errors during the training process, and produce superior results, contradicting the common understanding that GPUs simply perform such operations faster, rather than any better.

The research, titled *Impact of GPU Uncertainty on the Training of Predictive Deep Neural Networks*, comes from the Faculty of Psychology and Cognitive Sciences at Adam Mickiewicz University and two Japanese universities, together with SONY Computer Science Laboratories.

The study suggests that 'uncertainties' which deep neural networks exhibit in the face of various hardware and software configurations favor more expensive (and increasingly scarce) graphics processing units, and found in tests that a deep neural network trained exclusively on CPU produced higher error rates over the same number of epochs (the number of times that the system reprocesses the training data over the course of a session).



In this supplemental example from the paper, we see (bottom two rows), similar result quality obtained from a variety of GPUs, and (first row), the inferior results obtained from a range of otherwise very capable CPUs.
Source: https://arxiv.org/pdf/2109.01451.pdf

Strange Phenomena

These preliminary findings do not apply uniformly across popular machine learning algorithms, and in the case of simple autoencoder architectures, the phenomenon does not appear.

Nonetheless the work hints at a possible ‘escape velocity’ for efficacy of training in complex neural networks, where covering the same operations at lower speed and greater training times does not obtain the parity of performance one would expect of mathematical iteration routines.

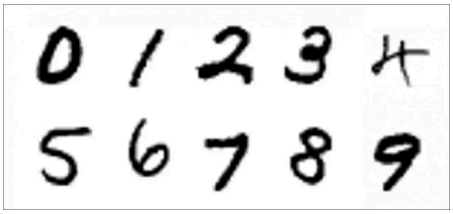
The researchers suggest that this performance disparity could be particular to certain types of neural networks, and that the indeterminate aspects of GPU-specific processing, frequently seen as an obstacle to eventually be overcome, may not only provide notable benefits, but could eventually be intentionally incorporated into later systems. The paper also suggests that the findings could offer deeper insights into brain-related computational processing.

Identifying the peculiarities that increase efficiency and quality of results in this way on GPUs holds the potential for obtaining a deeper insight into ‘black box’ AI architectures, and even for improving CPU performance – though currently, the underlying causes are elusive.

Autoencoder Vs. PredNet

In studying the anomalies, the researchers used a basic autoencoder and also Harvard University’s Predictive Neural Network PredNet, research from 2016 which was designed to explore and attempt to replicate the behavior of the human cerebral cortex.

Both systems are deep neural networks designed to synthesize apposite images through unsupervised learning (with data from which labels were omitted), though the autoencoder deals linearly with one image per batch, which would then produce an output as the next image in a recurring pipeline. The autoencoder was trained on the MNIST handwriting database.



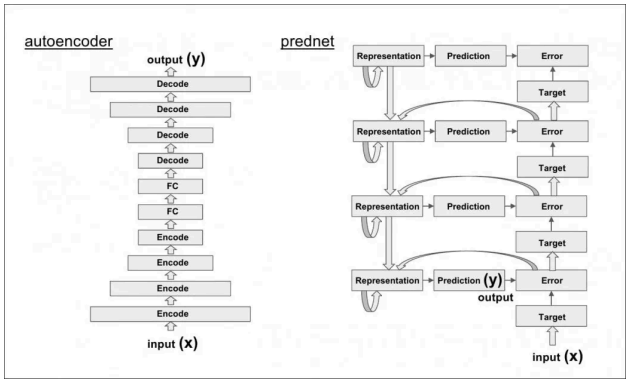
The autoencoder in the researchers’ tests was trained on the MNIST database, which comprises 60,000 training images at 28×28 pixels, anti-aliased for grey-scale induction, as well as 10,000 test images.

By contrast, PredNet evaluates complex video input, and in the case of this research, was trained on the FPSI dataset, which features extensive body-worn video footage of a day in Disney World at Orlando, Florida (Disney was one of the research associates on the 2012 paper).

Image sequences from FPSI, showing first-person views on a day at Disney World.

The two architectures are very different in terms of complexity. The autoencoder is designed to reconstruct images rather than predict target values. By contrast, PredNet features four layers, each of which consists of representation neurons using convolutional long short-term memory (LSTM).

The layers output contextual predictions which are then compared to a target in order to produce an error term that propagates throughout the network. Each of the two models utilize unsupervised learning.



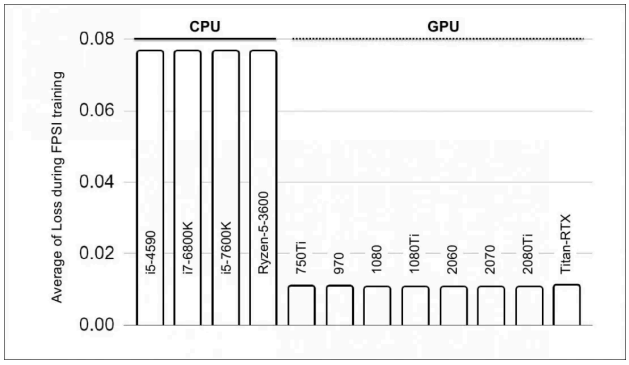
The simple, linear architecture of the autoencoder, and the more labyrinthine and recursive network of PredNet.

Both systems were tested on an array of hardware and software configurations, including CPUs without GPUs (Intel i5-4590, i7-6800K, i5-7600K, or AMD Ryzen-5-3600) and CPUs with GPUs (Intel i5-7600K + NVIDIA NVDA -4.57% GTX-750Ti, i5-7600K + GTX-970, i7-6700K + GTX-1080, i7-7700K + GTX-1080Ti, i7-9700 + RTX-2080Ti, i5-7600K + RTX-2060 super, AMD Ryzen-5-3600 + RTX-2070 super, or i5-9400 + Titan-RTX).

The interactive process viewer htop was used to ensure that all training occurred either on a single thread (on an Intel i7-6800K), on four threads (on an Intel i5-4590 and i5-7600K), or six threads (on an AMD Ryzen-5-3600).

Saddle Points

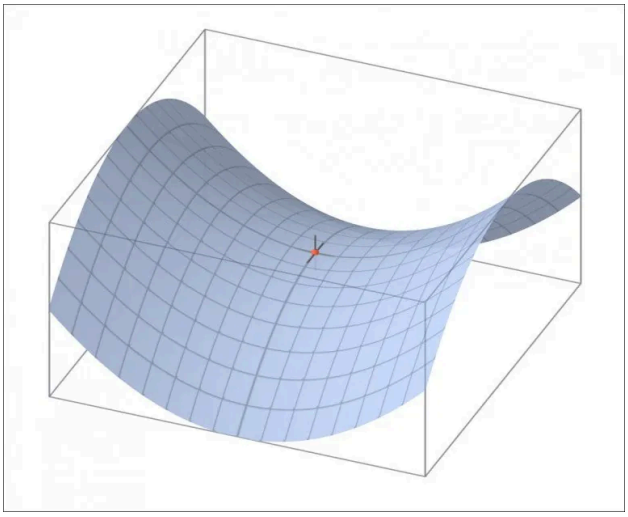
On the autoencoder, the mean difference across all configurations, with and without cuDNN, was not significant. For PredNet, the results were more startling, with notable differences in loss evaluation and quality between CPU and GPU training.



The average loss results for PredNet training across four CPUs and eight GPUs, with the network trained on 5000 video frames in 250 batches, with average loss for the last 1000 frames (50 batches) depicted. cuDNN was turned off.

The researchers conclude that *'Although the mechanism is unclear, the GPU hardware seems to have the ability to advance the training of DNNs.'*

The results indicate that GPUs may be better at avoiding saddle points – the areas in a gradient descent that describe the bottom of a slope.



The nadir of the slopes in a gradient descent is the 'saddle point', named for obvious reasons. Source: https://www.pinterest.com.au/pin/436849232581124086/

Saddle points, though an impediment, have been largely dismissed as easily worked around in recent thought on optimization of stochastic gradient descent (SGD), but the new paper suggests not only that GPUs may be uniquely outfitted to avoid them, but that the influence of saddle points should perhaps be revisited.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More