CASE STUDY

High Performance Computing
Intel® Xeon® Scalable Processors
Intel® Omni-Path Architecture

Switch to CPUs Improves Chest X-Ray Analysis with Deep Learning, Accelerates Neural Machine Translation Models

SURFsara research shows significant advances in deep learning by rethinking algorithms and shifting to CPUs, avoiding memory constraints and improving accuracy and performance.

At a Glance:

- SURF is a cooperative association of more than one hundred Dutch educational and research institutions.
- SURFsara, a part of SURF, is the Dutch national center for High Performance Computing & Data Services.
- Significant break-through resulted by taking advantage of superior memory capacity and simpler programmability.
- The ground-breaking CPU-based approaches are well supported by Intel® Xeon® Scalable processors, which offer outstanding compute performance, memory bandwidth, and compute density.

Executive Summary

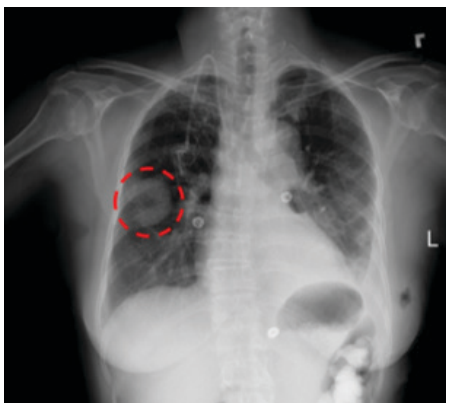
SURF is a cooperative association, of 109 Dutch educational and research institutions, spanning universities, medical centers, vocational schools, and other important Dutch educational and knowledge organizations. Its members have joined forces to create more flexible and improved education and research.

Within SURF, SURFsara assists scientists who are domain experts, but not necessarily HPC or AI experts. SURFsara enjoys a significant concentration of expertise and services for HPC, Networking, Data Services, Visualization, e Science Support and Cloud Services.

SURFsara has been helping realize a transformative impact on the general HPC community using AI-based computing methods.

Challenge

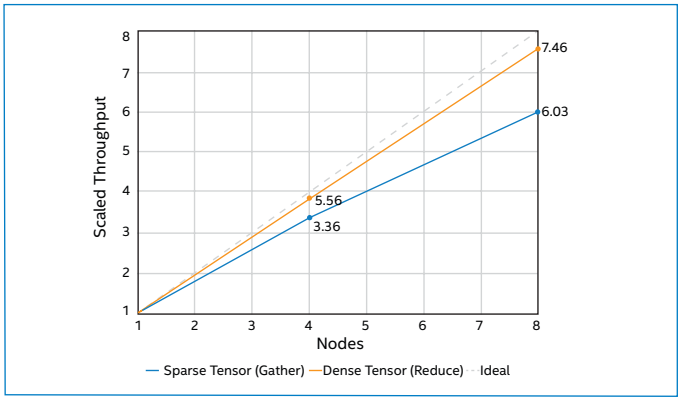
New research shows how some AI algorithms are significantly hampered due to designs limited by the highly constrained local memory available to a GPU. Significant new results for AI algorithms came about when SURFsara researchers, with the memory capacities of Intel CPU-based systems, found ways to apply AI algorithms to new problems that were previously considered too complex to be targeted by a data-intensive approach.



A typical Chest X-Ray image, ready for analysis. X-Ray chest imaging facilities are cheap and widely available, but their images are traditionally more difficult to interpret than the less widely available and much more expensive than those from CAT scanners.

"We are able to improve accuracy and performance when not constrained by the limited local memory of GPU systems. CPU systems offer a much more versatile solution for real-world application of deep learning, with applications found on the forefront of AI research working to solve really hard problems, including those in medicine."

– Valeriu Codreanu, Group Lead, High Performance Machine Learning at SURFsara



Even at only 8 nodes, the rapid decline in scaling of the original (sparse) translation approach dooms any high degree of scale-out—runs at higher levels would be overly cost- and compute-intensive. The new approach (dense) scales well enough to show exceptional scaling results above 256 nodes.

Solution

Researchers have proven that rethinking deep learning and other AI algorithms, without the constraints of GPU local memory, can offer significantly superior results. This rethinking, to harness the superior capabilities available with Intel® processors, led to jumps in the accuracy and performance of chest X-ray analysis, and much improved machine translation capabilities. Researchers also emphasized that they found "programming a CPU is significantly more straightforward than for a GPU."

Fast and Accurate Training of an AI Radiologist

Chest X-ray exams are one of the most frequent and cost-effective medical imaging examinations available—far cheaper and more widely available than chest CT imaging. However, diagnosis of chest X-rays are generally more challenging, more difficult, and less reliable than diagnosis via the more expensive, less available, and more detailed chest CT imaging.

Early and accurate diagnosis of emphysema and pneumonia can save lives. Emphysema, estimated to affect 16 million Americans,[1] is life threatening and early detection is critical to halt its progression. According to the World Health Organization (WHO), 64 million people worldwide have emphysema or some form of pulmonary disease,[2] and WHO predicts this will grow to be the third leading cause of death worldwide by 2030. WHO estimates[3] that pneumonia is the single largest cause of death in children worldwide—killing 808,694 children worldwide in 2017. In the U.S., the American Thoracic Society[4] estimates that one million adults seek hospital care annually due to pneumonia, and the CDC reported 49,157 deaths from pneumonia.

"The general perception might be that the world is digital, and everything is readily available to be analyzed by computers, however that is not true," says Damian Podareanu, HPC & AI Consultant at SURFsara. "Institutions are making strides toward that goal, but we are by no means at that stage today. Our work with CPUs helps take the data we do have, and do more with it. It also highlights a future of enormous possibilities using CPUs as additional high-quality datasets become available over the next few years."

CPU-based deep learning techniques for chest X ray analysis derived superior results compared to previous GPU-based approaches, in part by avoiding the downsampling of 1024x1024 images to 224x224 (necessary when using a GPU-based system). Downsampling the inputs sacrifices much of the rich visual information present in medical data.

Their work on Chest X-Rays can be found in a paper on their techniques, and their most recent work on scaling-out on CPUs including use of collapsed ensembles. A summary of their most recent work can be found on a poster presented at ISC19, including their latest achievements in classification accuracy.

Higher Accuracy than Prior Work

Using CPUs to escape the memory constraints of GPU-based systems, the researchers found techniques that led to much higher accuracy than prior techniques. Scale-out, large-batch training did prove to be an effective way to speed up neural network training using Intel® Xeon® Scalable platforms for chest X-ray analysis. Their experiments led to improving classification accuracy without significantly increasing the total number of passes through the dataset (epochs) required to obtain an effective neural network model. The upscaled version of ResNet-50, called ResNet-59, utilized the full 1024x1024 images to improve classification accuracy even further. Their scale-out work, when training a large AmoebaNet model (168 million parameters), managed to further boost classification (to obtain a mean AUROC of 0.842) outperforming prior work on all 14 different pathologies (Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural Thickening, Cardiomegaly, Nodule Mass, and Hernia).[6]

One of the key techniques for the scale-out experiments was the use of ensembles. These train multiple learners to construct a set of hypotheses in parallel and do a reduction operation to combine them. The researchers found them to be adequate for efficiently harnessing the power of the CPUs, increasing final classification accuracy, and efficiently keeping the total training time under control.

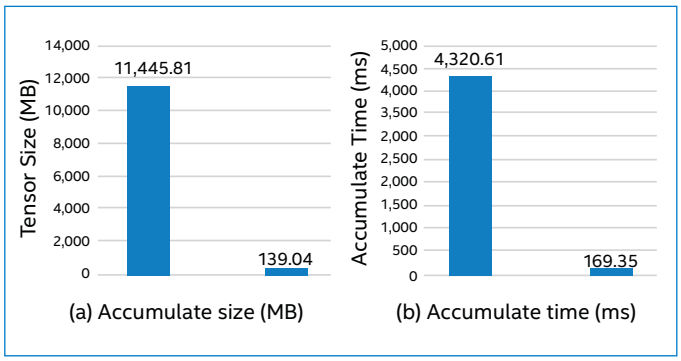
Advancing Neural Machine Translation

Neural machine translation (NMT), such as the Transformer Model based on the Attention Model—using neural networks to translate human language—is an area of active research with the goal of dramatically improving machine translation performance. Current state-of-the-art approaches have hit roadblocks due to excessive memory use. Working with researchers at Uber, Amazon, Dell EMC, and Intel, SURFsara researchers reported modifications made to the Horovod MPI-based distributed training framework to reduce memory usage for transformer models by converting assumed-sparse tensors to dense tensors, and subsequently replacing sparse gradient gather with dense gradient reduction.

Neural Machine Translation reached new heights by leaning on CPU capabilities including superior memory capacity. Their code using a dense representation resulted in a more than 82x reduction (11446 MB to 139 MB) in the amount of memory required by a 64-node run.[7] It also, saw a more than 25x reduction in time required for the accumulation operation (4321 ms to 169 ms).[8]

Six Hours instead of a Month of Computation

Once the researchers shifted from sparse representations to dense matrix representations for their NMT work, their new implementation opened the door for much-improved



Space/time for tensor accumulated (sparse gather vs. dense reduce)

scaling. What would take one month when using a single node, is now reduced to slightly over six hours when using 200 nodes (121 times faster).[9] This result can significantly increase the productivity for NMT researchers by allowing the use of CPU-based HPC infrastructures. Researchers reported that their ability to maintain very high scaling efficiencies up to the 300-node level suggests that continued scale-out is worthwhile beyond what they have tried thus far. That is certainly far better than the inability to scale beyond eight nodes effectively when they started. CPU-only scaling tests achieved 91% weak scaling efficiency up to 1200 MPI processes (300 nodes), and up to 65% strong scaling efficiency up to 400 MPI processes (200 nodes), using Intel® Xeon® Scalable Platinum 8160 processors interconnected with 100Gbps Intel® Omni-Path fabric, on the Stampede2 supercomputer at TACC.[10]

The software changes which they discuss in their paper have been upstreamed into Horovod 0.15.2 and later.

Their reports of record setting results,[6, 7, 8, 9, 10] stemming from their shift to CPUs, help highlight ways that 2nd Gen Intel® Xeon® Scalable processors with Intel® Deep Learning Boost (Intel® DL Boost) can offer leadership in AI, including on the tough problems found on the frontier of AI research.

Solution Ingredients

Experiments were run on the Zenith cluster in the Dell EMC HPC & AI Innovation Lab, as well as the Stampede2 cluster at the Texas Advanced Computing Center (TACC) in Austin, Texas, both featuring Intel processors and Intel Omni-Path fabric. In both cases, the researchers used Python 2.7, with an optimized version of TensorFlow that utilizes the Intel® Math Kernel Library (MKL), and modifications to Horovod that are available to everyone now in the versions 0.15.2 and later.

Each Zenith node consists of dual Intel® Xeon® Scalable Gold 6148/F processors, 192GB of memory, and an M.2 boot drive to house the operating system that does not provide user-accessible local storage. Nodes are interconnected by a 100 Gbps Intel Omni-Path fabric, and shared storage is provided by a combination of NFS (for HOME directories) and Lustre filesystems.

Work on the Stampede2, used the SKX partition, which consists of 1,736 nodes. Each node is outfitted with dual Intel Xeon Scalable Platinum 8160 processors, 192 GB of memory, and 200 GB internal SSD drive for the operating system and local /tmp. All nodes are interconnected with 100 Gbps Intel Omni-Path fabric and connected to Lustre-based shared filesystems.

Where to Get More Information

For more information on solutions for HPC from Intel, visit intel.com/hpc

Learn more about the work on scaling out on CPUs in this paper.

Find SURFsara's work on Chest X-Rays including success with scale out, with DellEMC and Intel, in their earlier blog, their paper on collapsed ensembles, and the most recent work in this ISC19 poster.

Discover insights on Neural Machine Translation in this ISC19 paper.

Learn more about Dell EMC Power Edge Servers at dellemc.com/hpc.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

[1] https://www.cdc.gov/copd/index.html

[2] https://www.who.int/respiratory/copd/en/

[3] https://www.who.int/news-room/fact-sheets/detail/pneumonia

[4] https://www.thoracic.org/patients/patient-resources/resources/top-pneumonia-facts.pdf

[5] https://www.cdc.gov/nchs/fastats/pneumonia.htm

[6] These results are reported in Valeriu Codreanu, Damian Podăreanu, Lucas A. Wilson, Srinivas Varadharajan, Vikram A. Saletore: High Performance Computing 34th International Conference, ISC High Performance 2019, Frankfurt/Main, Germany, June 16–20, 2019, RP22, Supercomputer-scale training of large AI Radiology models. It is also available directly at https://tinyurl.com/ISC19-RP22.

[7, 8, 9, 10] These results are reported in: Derya Cavdar, Valeriu Codreanu, Can Karakus, John A. Lockman III, Damian Podăreanu, Vikram A. Saletore, Alexander Sergeev, Don D. Smith II, Victor Suthichai, Quy Ta, Srinivas Varadharajan, Lucas A. Wilson, Rengan Xu, Pei Yang: High Performance Computing 34th International Conference, ISC High Performance 2019, Frankfurt/Main, Germany, June 16-20, 2019, Proceedings. Lecture Notes in Computer Science 11501, Springer 2019, ISBN 978-3-030-20655-0, Densifying Assumed-Sparse Tensors—Improving Memory Efficiency and MPI Collective Performance During Tensor Accumulation for Parallelized Training of Neural Machine Translation Models. 23-39; the paper is also available directly at https://arxiv.org/pdf/1905.04035.pdf

0919/RJMJ/SY/PDF Please Recycle 351545-001US