AbbVie Accelerates Data-Driven Biopharmaceutical Research using Cerebras' Wafer-Scale AI Accelerator

AbbVie explores using a Cerebras Systems AI accelerator to reduce development and training time for natural language processing models.

Jon Stevens, Brian Martin, Mehmed Sariyildiz, AbbVie Inc.
Aarti Ghatkesar, Noah Arthurs, Evren Tumer, Natalia Vassilieva, Andy Hock, Cerebras Systems

Summary

Deep learning and natural language processing (NLP) are revolutionizing biomedical research, but the computational and engineering demands of model training are formidable. One major challenge is the long model training time using conventional computing clusters featuring GPUs – large NLP models such as those used for this type of work often take many days, weeks, or sometimes longer to train on traditional GPU clusters. Additionally, scaling sophisticated models across multi-GPU server clusters requires a substantial amount of programming expertise. These challenges impede implementation, increase time to solution and thereby delay innovation.

"A common challenge we experience with programming and training $BERT_{LARGE}$ models is providing sufficient GPU cluster resources for sufficient periods of time.

The CS-2 system will provide wall-clock improvements that alleviate much of this challenge, while providing a simpler programming model that accelerates our delivery by enabling our teams to iterate more quickly and test more ideas."

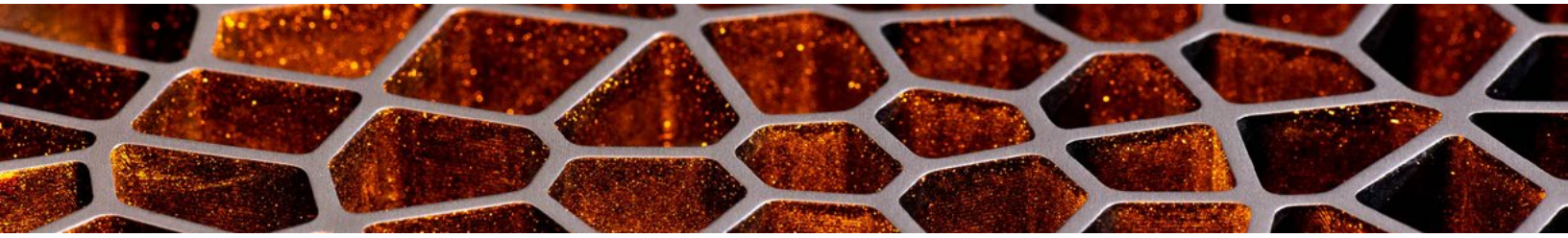
- Brian Martin,
Research Fellow, AbbVie Inc.

AbbVie recently collaborated with Cerebras Systems to compare the performance of a state-of-the-art NLP model on a Cerebras system with a conventional GPU cluster. The results show that the Cerebras CS-2 system delivered the wall-clock compute equivalent of more than 128 GPUs. The CS-2 accelerated model training time by 8x or more, consumed approximately 2/3 less energy, and delivered a 5x reduction in overall time to solution when compared to work done with a conventional 16-GPU cluster. In addition, the simple programming model required fewer machine learning engineering resources to develop and deploy novel models.

Contents

Key Results

- CS-2 delivered the wall-clock compute equivalent of more than 128 GPUs
- CS-2 made training the complex $PubMedBERT_{LARGE}$ model tractable for the first time
- End-to-end development time reduced by almost 5x
- Energy consumption reduced by approximately 3x



Accelerating data-driven research

AbbVie is a research-driven biopharmaceutical company with over 47,000 employees around the globe. The company focuses on cutting-edge R&D and emphasizes a core set of research areas, including immunology, neuroscience, oncology, and virology.

One of the challenges of biopharmaceutical research is keeping abreast of new information as it is published by scientists and doctors around the world. AbbVie's NLP tools include BERT (Bidirectional Encoder Representations from Transformers), a relatively new language processing model that uses bidirectional training and offers a significant improvement in accuracy compared to previous techniques. AbbVie refines and develops their solutions by fine-tuning generic pre-trained models with custom datasets, including internal and external data sets.

The challenge

AbbVie employs complex NLP transformer models to build services like their Abbelfish Machine Translation and various question answering and search tools. Abbelfish is a language translation service specializing in translation of unique biomedical terminology. Customized question-and-answer search tools such as AbbVie's "Agent-Q" system are based on the BioBERT and other BERT-like transformer models.

Formidable training demands

These models are revolutionizing biomedical research, but the computational and engineering demands of model training are formidable. As AI-powered tools such as Abbelfish become more sophisticated, they become increasingly difficult to improve, modify, or train. Abbelfish covers 180 languages with 32,400 language directions and is based on a sparse mesh Transformer architecture with 128 mixture-of-experts amounting to a 6 billion parameters large network. As we reported previously, an earlier version of Abbelfish with "only" 500 million parameters took more than four months to train.[1] Clearly, this length of time does not allow rapid experimentation.

In addition, prevailing assumptions about the best way to achieve accurate models are being challenged. The conventional method is to take a model pre-trained using ordinary "general-domain" texts and then fine-tune that model with additional training runs using domain-specific texts, such as biomedical literature. A foundational paper by Gu et al, from Microsoft Research[2], demonstrates that more accurate models can be achieved by training from scratch using domain-specific data. However, because training such complex language models has historically been so time-intensive and challenging to program on legacy cluster-based computing infrastructure, this approach has been practically inaccessible for many organizations and users.

The high cost of experimentation

The problem isn't limited to training time. Scaling sophisticated models across multi-GPU server clusters requires a substantial amount of programming expertise. Depending on the scale of the model, familiarity with HPC-style placement algorithms and interconnects may be required, as will changes to the machine learning model itself.

It can take weeks to train large models on GPU clusters. Programming a model to perform and converge on a GPU cluster requires a painstaking process of hyper-parameter tuning. Experiments risk adding weeks to development time and do not guarantee improvements. When using GPU clusters, the programmer must spend significant time and effort to optimize training time to accuracy by tuning batch size, learning rate, number of devices, and layer sizes, while dealing with constraints like per-device memory and available memory bandwidth.

The high cost of experimentation effectively punishes developers for attempting to optimize their models. In some cases, the more capable a model becomes, the higher the cost of improving it.

The solution

In AI compute, large chips process information more quickly. The Cerebras Wafer-Scale Engine (WSE) is the world's largest, most powerful chip (Table 1).

The Cerebras system, powered by the WSE, presents a solution to this scaling problem. The simplified Cerebras development model and high performance allowed Cerebras and AbbVie to execute multiple, full, end-to-end pre-training and fine-tuning runs with various BERT models. PubMed abstracts and full texts were used to evaluate compute performance and time-to-solution, following the architecture and model laid out by Gu et al.

Microsoft has released several PubMedBERT models as part of the Biomedical Language Understanding and Reasoning Benchmark, or BLURB. The first is a $BERT_{BASE}$ model that has been pre-trained on a collection of PubMed abstracts. The second is a $BERT_{BASE}$ model pre-trained using a larger dataset consisting of both PubMed abstracts and full articles from PubMed Central.

We report CS-2 wall clock training time, including from-scratch pre-training for both models, and compare it against the wall clock training times reported in the same Microsoft Research paper for an NVIDIA DGX-2 system equipped with 16 GPUs.

We also report results for a more complex model $BERT_{LARGE}$ trained using the same PubMed abstracts and full text articles. The September 2021 revision of the Microsoft paper states that "$BERT_{LARGE}$ appears to yield improved performance in some preliminary experiments." But it also takes much longer to train and is more painful to experiment with. Therefore, despite the promise of improved performance, $BERT_{LARGE}$ models are not that widely used for pre-training from scratch on domain-specific datasets.

Cerebras' software platform is designed to allow AI/ML researchers to take advantage of its performance without overhauling their code. TensorFlow and PyTorch are all supported through a flexible graph compiler that automatically converts and optimizes the code for execution. In addition to supporting industry standard tools like TensorBoard, Cerebras provides debug and profiling tools to speed the development and deployment process.

Cerebras WSE-2

850,000 cores

46,225 mm2 silicon area

2.6 trillion transistors

40 GB on-chip memory

20 PB/s memory bandwidth

220 Pb/s fabric bandwidth

7nm process technology

Table 1 Characteristics of the Cerebras WSE-2 chip which powers the CS-2 system.

Results

Training time

The CS-2 trained $BERT_{BASE}$ model on the PubMed abstracts and full texts in about a day, which is 8x faster than the 16 GPU DGX-2 server reported by Microsoft Research (Figure 1). A researcher working with Cerebras' CS-2 could pre-train this model over two dozen times in a month, in order to find the best model configuration and training hyperparameters. A DGX-2 would limit the researcher to fewer than four pre-training cycles per month.

Cerebras and AbbVie also trained a custom $BERT_{LARGE}$ model on the same dataset. This represents new work that had not been practical to implement on the team's existing, traditional infrastructure, and delivers higher accuracy downstream performance. Training this larger model took less than two days on a CS-2 system.

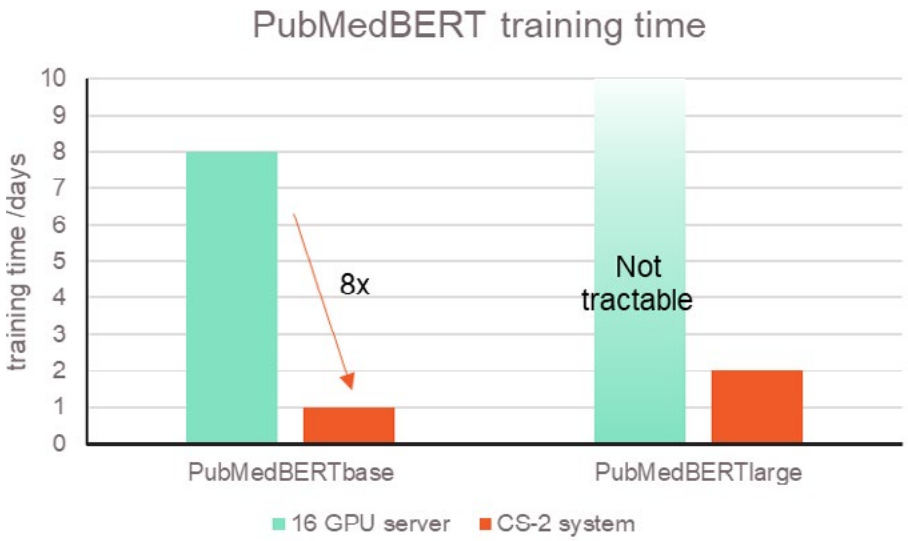


Figure 1 Compared to previously reported results, the CS-2 used for this work delivered 8x faster wall clock training time than a 16-GPU DGX-2 system.

In separate experiments performed on a Cerebras CS-1 system, we verified the Microsoft result and observed improved performance on downstream tasks compared to a smaller $BERT_{BASE}$ model. Note that in this context, "improved performance" means the ability to fine-tune to higher accuracy, which translates to better search results and an improved end-user experience, rather than referring to at metric such as training time.

This result demonstrates that the Cerebras system makes it possible for the first time to take advantage of higher-performing $BERT_{LARGE}$ models in a research environment.

Expected scaling trends

Wall clock training time does not scale linearly using GPU clusters, as illustrated by Figure 2, which charts published results for NVIDIA GPUs for the MLPerf 1.1 benchmark.[3]

MLPerf uses a $BERT_{LARGE}$ model similar to PubMedBERT that is trained in a similar way: a short sequence length is used initially, with a longer sequence length used for the final phase of training. A single DGX-A100 server equipped with 8 A100 GPUs, completed the second phase on the MLPerf benchmarking dataset in 20 minutes, while a massive cluster of 540 DGX-A100 servers took 13.5 seconds. In other words, a 540x increase in the number of GPUs only delivered 89x speed-up.

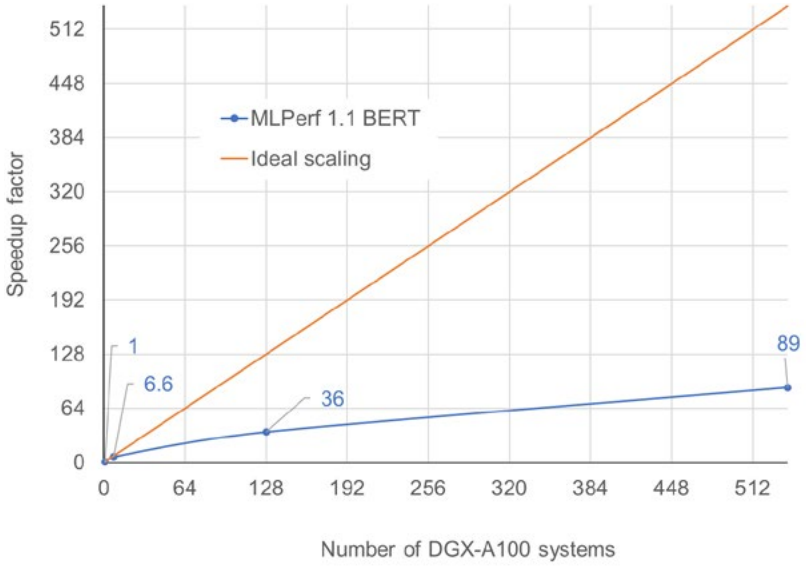


Figure 2 Non-linear scaling performance of GPU clusters (MLPerf 1.1 benchmark results)

It is reasonable to assume that the same scaling trend would obtain for PubMedBERT trained on DGX-2 servers. Thus, while the CS-2 trained the baseline PubMedBERT model about 8x faster than reported results on a DGX-2 server, it would take significantly more than 8x DGX-2 servers to achieve the same training time. In other words, significantly more than 128 GPUs would be needed to achieve the same training time as a single CS-2 system.

End-to-end development time

In addition to increased performance, the CS-2 system provides simplified setup and use, with a nearly 4x improvement in overall time to solution compared to an equivalent deployment on GPUs (Table 2,Figure 3). Improving setup speed generally helps every research project and allows researchers to spend more time fine-tuning parameters and conducting research and less on GPU engineering.

ML research step	Time on GPU server	Time on CS-2
Software setup	~ 1 week	~ 0.2 week
Writing the model	~ 2 weeks	~ 1 week
Debugging training functionality	~ 2 weeks	~ 1 week
Hyperparameter tuning: performance optimization	~ 2 weeks	~ 0.1 week
Initial model training to target accuracy	~ 2 weeks	~ 0.2 weeks
Additional training runs to refine model variant	~ 8 weeks	~ 1 weeks
Total time to solution (TTS)	**17 weeks**	**3.5 weeks**

Table 2 Comparison of the steps in overall time to solution for a typical ML research cycle for NLP model training application such as BERT pre-training on PubMed with a multi-GPU server such as NVIDIA DGX-2 and the Cerebras CS-2 system

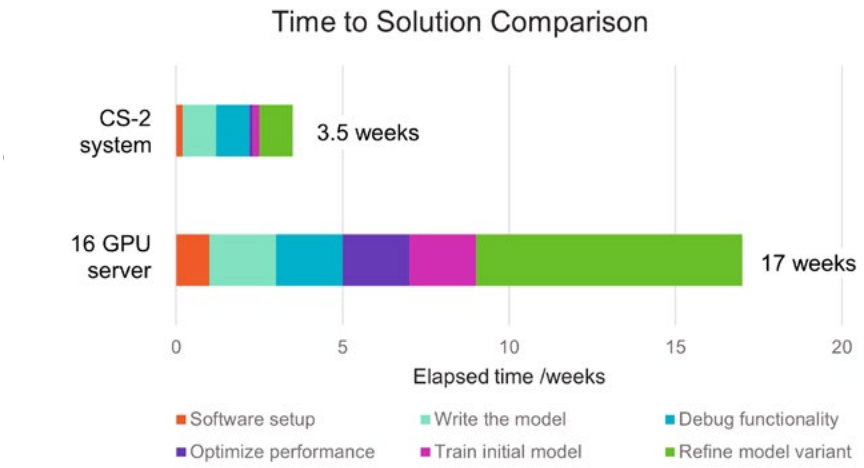


Figure 3 The Cerebras solution demonstrated a significant improvement in overall time-to-solution for a typical NLP research cycle

Energy Consumption

The amount of energy required to train neural networks is an important consideration.[4] We compared the expected power consumed during the initial model training phase. We restricted the analysis to this phase because both the CS-2 and the GPU server would be running at full performance throughout that time. However, since GPUs cannot run this workload at full utilization because of inherent memory bandwidth limitations, we assume a 0.75x reduction from the published peak power consumption.[5] As Table 3 shows, the CS-2 consumed approximately 1/3 the energy of the GPU server.

	16 GPU Server	CS-2	Notes
Training time /days	14	1.4	
Power /kW	7.5	23.1	GPU power reduced from published peak of 10kW to model low utilization
Energy consumed /kWh	2520	776	
Energy fraction		31%	
Energy ratio		3.2x	

Table 3 Energy consumption comparison. The Cerebras solution demonstrated a significant reduction in energy required to complete a single training run.

Further work

AbbVie and Cerebras plan to extend this work using a Cerebras CS-2 system installed at the National Center for Supercomputing Applications (NCSA) at University of Illinois Urbana-Champaign. We will continue to develop complex BERT models for multi-language processing and explore state-of-the-art language models like T5 and GPT-J.

Conclusions

AbbVie increasingly leverages large pre-trained NLP language models such as BERT for biomedical question answering and machine translation. These models have traditionally required a multi-GPU cluster to train and fine-tune. By partnering with Cerebras to train a custom BERT model on the CS-2 Wafer Scale Engine, AbbVie has begun to explore alternatives to GPUs

which can greatly reduce the time it takes to develop, train, fine-tune and deploy one of these models. In addition, this result demonstrates that the horsepower of the Cerebras system makes it possible for the first time to take advantage of higher-performing $BERT_{LARGE}$ models in a research environment.

The Cerebras CS-2 system and software platform demonstrated faster training, easier setup and scaling for a wide variety of models, compared to the GPU-centric approach. Cerebras worked with AbbVie to train a custom BERT model on a CS-2 using PubMed abstracts and full texts, to compare both the total training times and the comparative accuracy. The CS-2 significantly accelerated training times and reduced setup complexity by allowing the entire training model to execute on a single wafer-scale system. The CS-2 also delivered the wall-clock compute equivalent of multiple DGX-2 systems, while simultaneously eliminating the issues of sub-linear scaling.

To explore how Cerebras systems can accelerate your research or to see a demo, please contact us at www.cerebras.net/get-demo.

References

1 G Anthony Reina , Mehmed Sariyildiz, Brian Martin, Andrew Lamkin, Jason Lee, "Accelerating Natural Language Processing Inference Models using Processor Optimized Capabilities," white paper 2021, https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/healthcare-biopharmaceutical-research-white-paper.pdf

2 Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," 2021, https://arxiv.org/pdf/2007.15779.pdf

3 MLPerf training 1.1 results https://mlcommons.org/en/news/mlperf-training-v11/

4 Emma Strubell, Ananya Ganesh, Andrew McCallum, "Energy and Policy Considerations for Deep Learning in NLP", 2019, https://arxiv.org/abs/1906.02243

5 NVIDIA® DGX-2 datasheet https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/dgx-2/dgx-2-print-datasheet-738070-nvidia-a4-web-uk.pdf