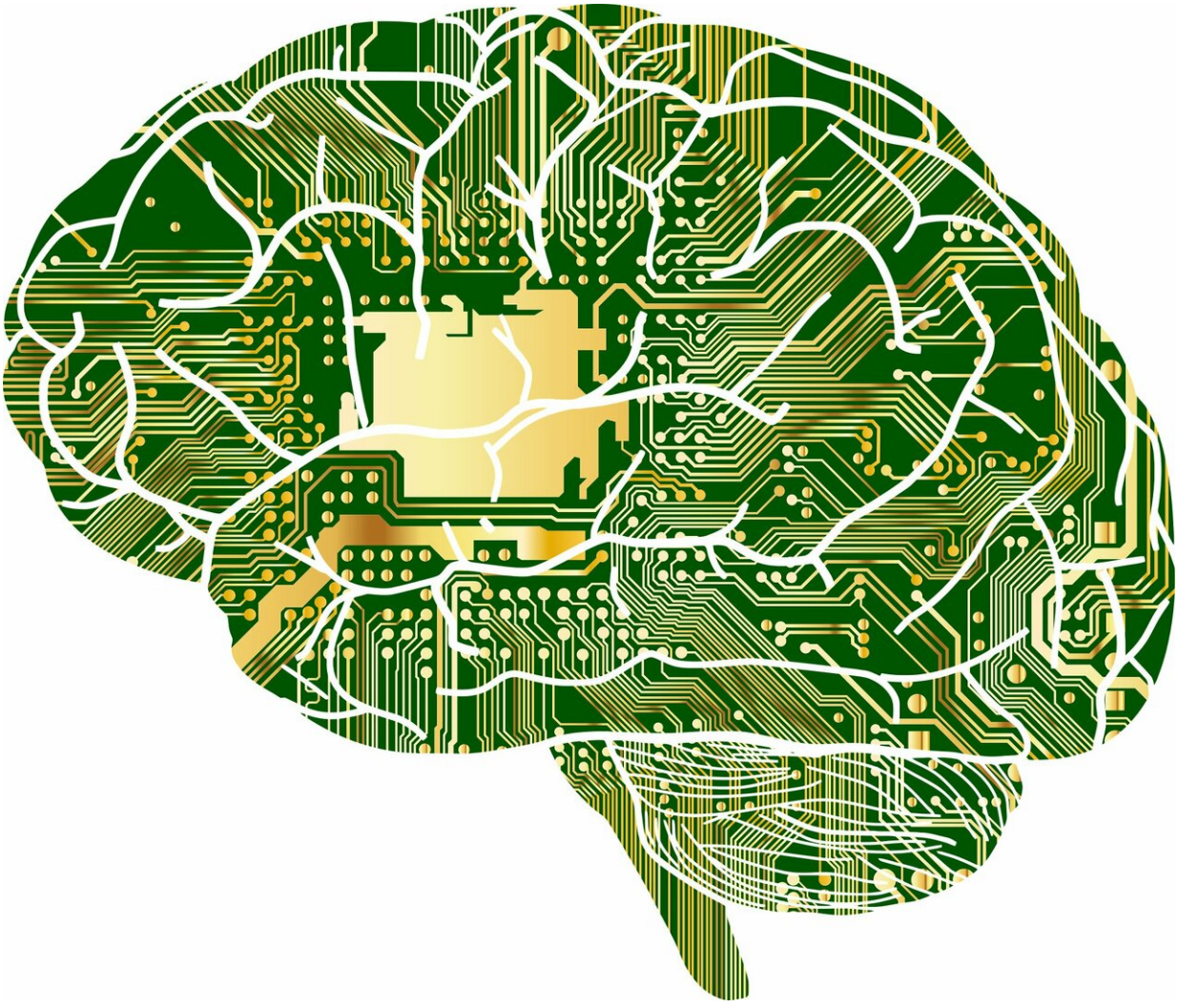
More chat, less duh, on the way thanks to Nvidia AI leaps with BERT

August 14 2019, by Nancy Cohen



Credit: CC0 Public Domain

In the future are chatbots that are even more chatty and less dim-witted. Yes, the day will come when you can easily reflect on how far AI's language skills have come. And upon that reflection, do not ignore Nvidia's contributions in their work with BERT.

OK, we will refrain from calling AI language skills dim-witted. Nvidia phrased it more tactfully in its announcement on August 13. "Limited conversational AI services" have existed for several years but it has been extremely difficult for chatbots, intelligent personal assistants and search engines to operate with human-level comprehension due to the inability to deploy extremely large AI models in real time, said the company.

That has changed. Nvidia said key optimizations added to its AI platform helped achieve speed records in AI training and inference. *HotHardware* cut to the chase in assessing the impact of this work. "Nvdia smashed records for conversational AI training which could "turbocharge" mainstream assistants such as Alexa and Siri.

Back to BERT which has already earned a rightful place in natural language processing. A November 2018 announcement from Google appeared on its Google AI blog:

"One of the biggest challenges in natural language processing (NLP) is the shortage of training data...most task-specific datasets contain only a few thousand or a few hundred thousand human-labeled training examples... To help close this gap in data, researchers have developed a variety of techniques for training general purpose language representation models using the enormous amount of unannotated text on the web (known as pre-training). The pre-trained model can then be fine-tuned on small-data NLP tasks like question answering and sentiment analysis, resulting in substantial accuracy improvements compared to training on these datasets from scratch.

"This week, we open sourced a new technique for NLP pre-training called Bidirectional Encoder Representations from Transformers, or BERT."

Well, that was "this week" in 2018 and now it is this week in 2019. Nvidia's developer blog announced Tuesday that Nvidia clocked the world's fastest BERT training time. NVIDIA DGX SuperPOD trained BERT-Large in just 53 minutes.

As Darrell Etherington said in *TechCrunch*, this means "the hour mark" in training BERT was broken (53 minutes). Etherington said, "Nvidia's AI platform was able to train the model in less than an hour, a record-breaking achievement at just 53 minutes."

Nvidia's Shar Narasimhan blogged that a key advantage of BERT was that it doesn't need to be pre-trained with labeled data, so it can learn using any plain text. This advantage opens the door to massive datasets. BERT's numbers: Narasimhan said it was generally "pre-trained on a concatenation of BooksCorpus (800 million words) and the English Wikipedia (2.5 billion words), to form a total dataset of 3.3 billion words."

Nvidia's news release of August 13 said early adopters of the company's performance advances included Microsoft and startups harnessing its platform to develop language-based services for customers. Microsoft Bing is using its Azure AI platform and Nvidia technology to run BERT.

Rangan Majumde, group program manager, Microsoft Bing, said that Bing further optimized the inferencing of BERT. He said they achieved "two times the latency reduction and five times throughput improvement during inference using Azure NVIDIA GPUs compared with a CPU-based platform."

David Cardinal in *ExtremeTech* had more details on what Nvidia brought to the table in advancing BERT: "Nvidia has demonstrated that it can now train BERT (Google's reference language model) in under an hour on a DGX SuperPOD consisting of 1,472 Tesla V100-SXM3-32GB GPUs, 92 DGX-2H servers, and 10 Mellanox Infiniband per node."

Also part of Nvidia's bragging rights on the AI front is a language model based on Transformers, the technology building block used for BERT. Nvidia said "With a focus on developers' ever-increasing need for larger models, NVIDIA Research built and trained the world's largest language model based on Transformers, the technology building block used for BERT and a growing number of other natural language AI models. NVIDIA's custom model, with 8.3 billion parameters, is 24 times the size of BERT-Large."

According to Nvidia, they "built the world's largest transformer based language model on top of existing deep learning hardware, software, and models. In doing so, we successfully surpassed the limitations posed by traditional single GPU training by implementing a simple and efficient model parallel approach with only a few targeted modifications to the existing PyTorch transformer implementations."

More information: devblogs.nvidia.com/training-bert-with-gpus/

nvidianews.nvidia.com/news/nvi … me-conversational-ai