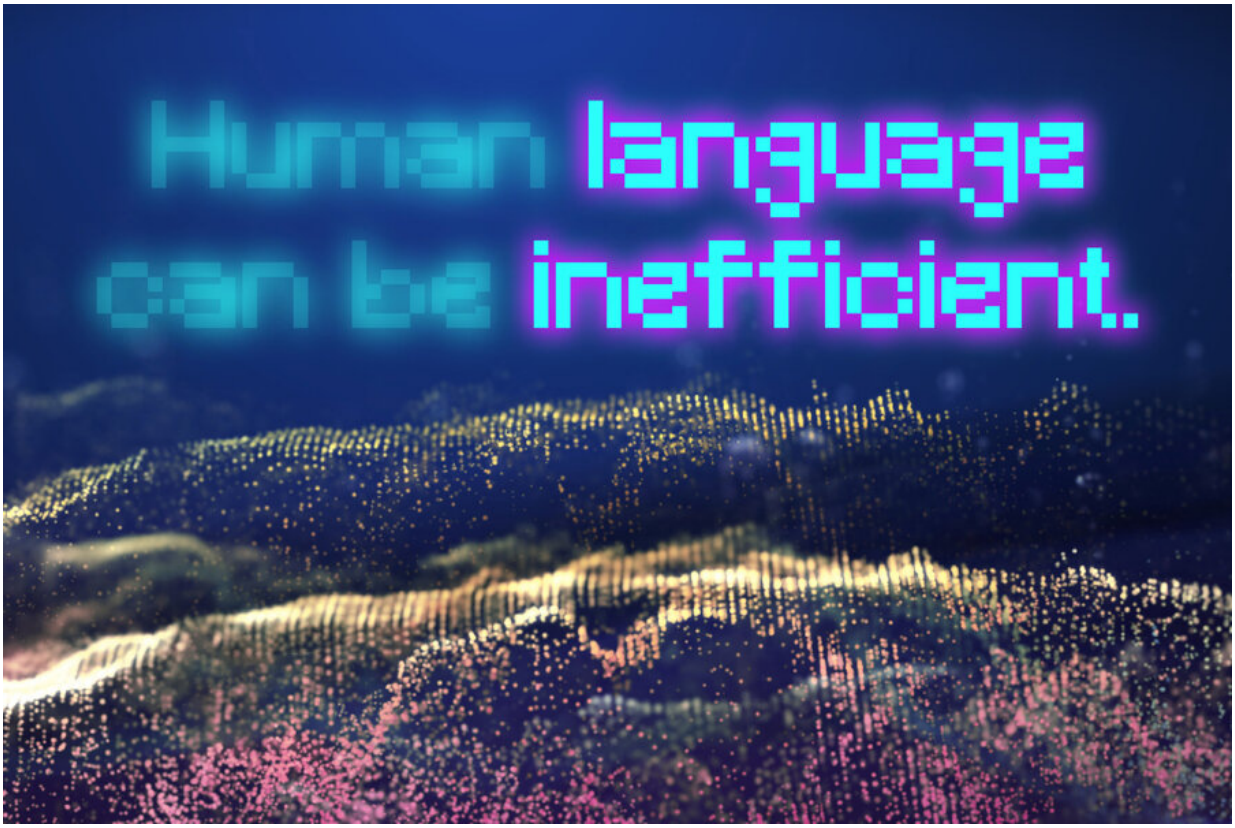
A language learning system that pays attention more efficiently than ever before

February 10 2021, by Daniel Ackerman



MIT researchers developed a hardware and software system that could reduce the computing power, energy, and time required for text analysis and generation. Credit: Jose-Luis Olivares, MIT

Human language can be inefficient. Some words are vital. Others,

expendable.

Reread the first sentence of this story. Just two words, "language" and "inefficient," convey almost the entire meaning of the sentence. The importance of key words underlies a popular new tool for natural language processing (NLP) by computers: the attention mechanism. When coded into a broader NLP algorithm, the attention mechanism homes in on key words rather than treating every word with equal importance. That yields better results in NLP tasks like detecting positive or negative sentiment or predicting which words should come next in a sentence.

The attention mechanism's accuracy often comes at the expense of speed and computing power, however. It runs slowly on general-purpose processors like you might find in consumer-grade computers. So, MIT researchers have designed a combined software-hardware system, dubbed SpAtten, specialized to run the attention mechanism. SpAtten enables more streamlined NLP with less computing power.

"Our system is similar to how the human brain processes language," says Hanrui Wang. "We read very fast and just focus on key words. That's the idea with SpAtten."

The research will be presented this month at the IEEE International Symposium on High-Performance Computer Architecture. Wang is the paper's lead author and a Ph.D. student in the Department of Electrical Engineering and Computer Science. Co-authors include Zhekai Zhang and their advisor, Assistant Professor Song Han.

Since its introduction in 2015, the attention mechanism has been a boon for NLP. It's built into state-of-the-art NLP models like Google's BERT and OpenAI's GPT-3. The attention mechanism's key innovation is selectivity—it can infer which words or phrases in a sentence are most

important, based on comparisons with word patterns the algorithm has previously encountered in a training phase. Despite the attention mechanism's rapid adoption into NLP models, it's not without cost.

NLP models require a hefty load of computer power, thanks in part to the high memory demands of the attention mechanism. "This part is actually the bottleneck for NLP models," says Wang. One challenge he points to is the lack of specialized hardware to run NLP models with the attention mechanism. General-purpose processors, like CPUs and GPUs, have trouble with the attention mechanism's complicated sequence of data movement and arithmetic. And the problem will get worse as NLP models grow more complex, especially for long sentences. "We need algorithmic optimizations and dedicated hardware to process the ever-increasing computational demand," says Wang.

The researchers developed a system called SpAtten to run the attention mechanism more efficiently. Their design encompasses both specialized software and hardware. One key software advance is SpAtten's use of "cascade pruning," or eliminating unnecessary data from the calculations. Once the attention mechanism helps pick a sentence's key words (called tokens), SpAtten prunes away unimportant tokens and eliminates the corresponding computations and data movements. The attention mechanism also includes multiple computation branches (called heads). Similar to tokens, the unimportant heads are identified and pruned away. Once dispatched, the extraneous tokens and heads don't factor into the algorithm's downstream calculations, reducing both computational load and memory access.

To further trim memory use, the researchers also developed a technique called "progressive quantization." The method allows the algorithm to wield data in smaller bitwidth chunks and fetch as few as possible from memory. Lower data precision, corresponding to smaller bitwidth, is used for simple sentences, and higher precision is used for complicated

ones. Intuitively it's like fetching the phrase "cmptr progm" as the low-precision version of "computer program."

Alongside these software advances, the researchers also developed a hardware architecture specialized to run SpAtten and the attention mechanism while minimizing memory access. Their architecture design employs a high degree of "parallelism," meaning multiple operations are processed simultaneously on multiple processing elements, which is useful because the attention mechanism analyzes every word of a sentence at once. The design enables SpAtten to rank the importance of tokens and heads (for potential pruning) in a small number of computer clock cycles. Overall, the software and hardware components of SpAtten combine to eliminate unnecessary or inefficient data manipulation, focusing only on the tasks needed to complete the user's goal.

The philosophy behind the system is captured in its name. SpAtten is a portmanteau of "sparse attention," and the researchers note in the paper that SpAtten is "homophonic with 'spartan," meaning simple and frugal." Wang says, "that's just like our technique here: making the sentence more concise." That concision was borne out in testing.

The researchers coded a simulation of SpAtten's hardware design—they haven't fabricated a physical chip yet—and tested it against competing general-purposes processors. SpAtten ran more than 100 times faster than the next best competitor (a TITAN Xp GPU). Further, SpAtten was more than 1,000 times more energy efficient than competitors, indicating that SpAtten could help trim NLP's substantial electricity demands.

The researchers also integrated SpAtten into their previous work, to help validate their philosophy that hardware and software are best designed in tandem. They built a specialized NLP model architecture for SpAtten, using their Hardware-Aware Transformer (HAT) framework, and

achieved a roughly two times speedup over a more general model.

The researchers think SpAtten could be useful to companies that employ NLP models for the majority of their artificial intelligence workloads. "Our vision for the future is that new algorithms and hardware that remove the redundancy in languages will reduce cost and save on the power budget for data center NLP workloads" says Wang.

On the opposite end of the spectrum, SpAtten could bring NLP to smaller, personal devices. "We can improve the battery life for mobile phone or IoT devices," says Wang, referring to internet-connected "things"—televisions, smart speakers, and the like. "That's especially important because in the future, numerous IoT devices will interact with humans by voice and natural language, so NLP will be the first application we want to employ."

Han says SpAtten's focus on efficiency and redundancy removal is the way forward in NLP research. "Human brains are sparsely activated [by key words]. NLP models that are sparsely activated will be promising in the future," he says. "Not all words are equal—pay attention only to the important ones."

More information: SpAtten: Efficient Sparse Attention Architecture with Cascade Token and Head Pruning. arXiv:2012.09852v2 [cs.AR] arxiv.org/abs/2012.09852

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology