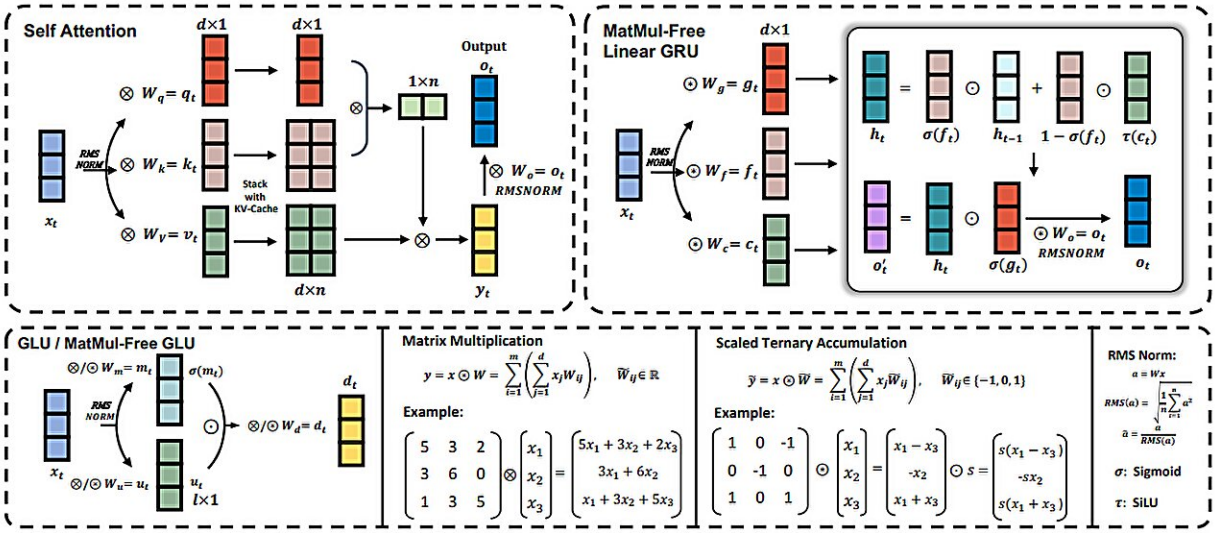
Software engineers develop a way to run AI language models without matrix multiplication

June 26 2024, by Bob Yirka



Overview of the MatMul-free LM. The sequence of operations are shown for vanilla self-attention (top-left), the MatMul-free token mixer (top-right), and Ternary Accumulations. The MatMul-free LM employs a MatMul-free token mixer (MLGRU) and a MatMul-free channel mixer (MatMul-free GLU) to maintain the transformer-like architecture while reducing compute cost. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2406.02528

A team of software engineers at the University of California, working with one colleague from Soochow University and another from LuxiTec, has developed a way to run AI language models without using matrix

multiplication. The team has published a paper on the *arXiv* preprint server describing their new approach and how well it has worked during testing.

As the power of LLMs such as ChatGPT has grown, so too have the computing resources they require. Part of the process of running LLMs involves performing matrix multiplication (MatMul), where data is combined with weights in neural networks to provide likely best answers to queries.

Early on, AI researchers discovered that graphics processing units (GPUs) were ideally suited to neural network applications because they can run multiple processes simultaneously—in this case, multiple MatMuls. But now, even with huge clusters of GPUs, MatMuls have become bottlenecks as the power of LLMs grows along with the number of people using them.

In this new study, the research team claims to have developed a way to run AI language models without the need to carry out MatMuls—and to do it just as efficiently.

To achieve this feat, the research team took a new approach to how data is weighted—they replaced the current method that relies on 16-bit floating points with one that uses just three: {-1, 0, 1} along with new functions that carry out the same types of operations as the prior method.

They also developed new quantization techniques that helped boost performance. With fewer weights, less processing is needed, resulting in the need for less computing power. But they also radically changed the way LLMs are processed by using what they describe as a MatMul-free linear gated recurrent unit (MLGRU) in the place of traditional transformer blocks.

In testing their new ideas, the researchers found that a system using their new approach achieved a performance that was on par with state-of-the-art systems currently in use. At the same time, they found that their system used far less computing power and electricity than is generally the case with traditional systems.

More information: Rui-Jie Zhu et al, Scalable MatMul-free Language Modeling, *arXiv* (2024). DOI: 10.48550/arxiv.2406.02528

Citation: Software engineers develop a way to run AI language models without matrix multiplication (2024, June 26) retrieved 16 August 2024 from https://techxplore.com/news/2024-06-software-ai-language-matrix-multiplication.html