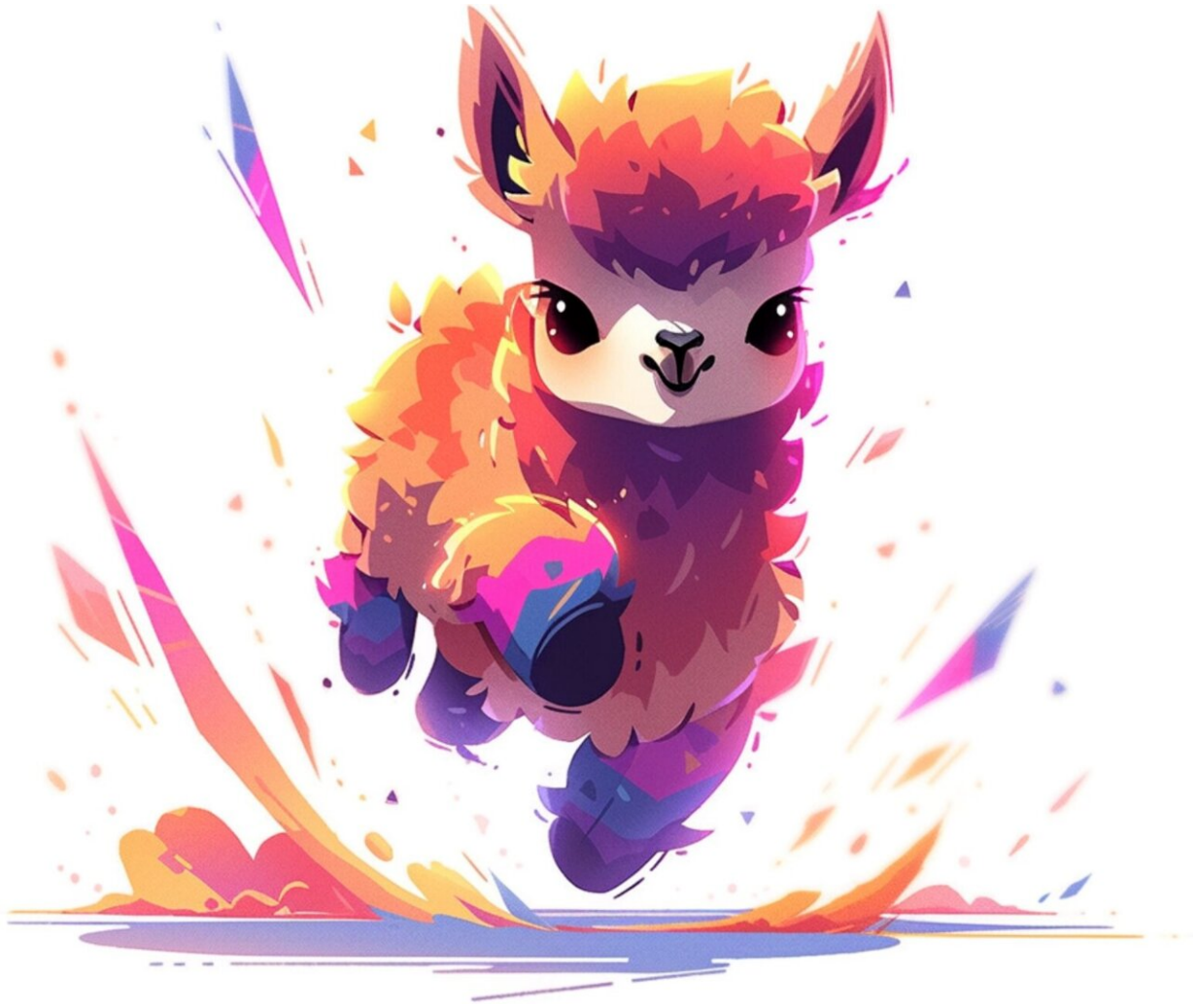
Research team launches first-of-its-kind mini AI model with three trillion-token punch

January 31 2024



TinyLlama--the mini AI model with three trillion-token punch. Credit: SUTD

It's called TinyLlama and it's taken the research world by storm because of how much power it packs.

Developed by Associate Professor Lu Wei of Singapore University of Technology and Design (SUTD), research assistant Mr. Zhang Peiyuan, and Ph.D. students, Mr. Zeng Guangtao, and Mr. Wang Tianduo, TinyLlama is a 1.1 billion parameter open-sourced small language model that has outperformed other open-source models of comparable sizes across several benchmarks. A total of three trillion tokens of datasets were pre-trained on TinyLlama within just four months.

Current large language models (LLMs) such as ChatGPT or Google Bard, developed by large technology firms such as OpenAI or Google, are managed by thousands or even tens of thousands of graphic processing units (GPUs) and require users to connect online to their massive servers. TinyLlama, in contrast, is built on just 16 GPUs and takes up only 550MB of Random Access Memory (RAM). In other words, TinyLlama can readily be deployed on mobile devices, enabling everyone to carry a "mini ChatGPT" in their pocket wherever they go.

According to Marktechpost, a California-based Artificial Intelligence news platform with a community of over 1.5 million AI professionals and developers, TinyLlama's performance in common-sense reasoning and problem-solving tasks highlights the potential of smaller models to achieve high performance when trained with a substantial amount of data. It also opens up new possibilities for research and application in natural language processing, especially in scenarios where computational resources are limited.

Said Prof Lu, also the Director of the StatNLP Research Group, which focuses on natural language processing research, "The importance of small language models cannot be understated, and the reason why TinyLlama was specifically created to be open-sourced was that it will

democratize language models by allowing smaller tech companies and research labs to build and develop their own models for a variety of applications. As researchers, our plan is to lay the foundations for small language models, with the aim of making significant scientific advancements in the field.

"Smaller tech firms as well as individual researchers and developers are increasingly demanding small language models that require less resources to run. These models, such as TinyLlama, are therefore more feasible for them to build and more optimal for edge devices such as mobile phones. The compactness of such models also allows them to cater to a multitude of applications that demand real-time machine translation without an internet connection. This means that users can access the language model offline. They need not send their personal information to the server when using it, and through the technique called 'fine-tuning,' we are able to improve it further," Prof Lu added.

The team behind TinyLlama—from left to right: SUTD Ph.D. students, Zeng Guangtao and Wang Tianduo, Associate Prof Lu Wei and Research Assistant, Zhang Peiyuan. Credit: SUTD

TinyLlama's innovative approach lies in its construction. It is based on the architecture and tokenizer of Llama 2 and incorporates several state-of-the-art technologies. One such technology is FlashAttention, which enhances computational efficiency. Despite its smaller size than some of its predecessors, TinyLlama exhibits exceptional performance in various downstream tasks. It has successfully challenged the notion that larger models are always better, demonstrating that models with fewer parameters can still achieve high levels of effectiveness when trained

with extensive and diverse datasets.

With its compact architecture and exceptional performance, TinyLlama can enable end-user applications on mobile devices and serve as a lightweight platform for language model research.

Firms such as leading global consumer internet company Sea Limited and DSO National Laboratories, a national defense research and development organization, have downloaded the TinyLlama source code from GitHub for research purposes.

Dr. Liu Qian, Research Scientist and Team Lead, Natural Language Processing Group at Sea AI Lab, said, "In our language model research projects, we've utilized the TinyLlama project as a nimble and efficient testbed. Its codebase follows a compact and well-organized structure, which allows easy modifications for diverse purposes. With access to several 1B model checkpoints, we swiftly validate hypotheses, obtaining faster feedback compared to the Llama-7b models.

"Notably, TinyLlama's optimization enhancements significantly boost GPU utilization, outperforming the Hugging Face transformers library. This combination of swift prototyping and efficient training positions TinyLlama as a valuable tool, facilitating accelerated iterations in the research community."

TinyLlama is currently available on GitHub, a platform and cloud-based service for developers to store and manage their code. It was trending as the Number One code on Hugging Face, a platform for hosting AI-related projects, out of over 460,000 models for about a week from 3 January 2024. Plans are underway to further improve TinyLlama.

Provided by Singapore University of Technology and Design