A Single Usage Is All You Need: Generating an Entire Text by Using a Causal Language Model Once

Kremer, Yoni (School: Ironi Dalet)

Text-generation AI services are unaffordable for most people at $10 to $20 a month. I developed and published an open-source efficient text-generation algorithm called grouped sampling to enable cheap and accessible AI text-generation services for everyone. Causal language models such as GPT4 are state-of-the-art text generation models that power many popular products like chat-GPT. The naive text generation algorithm requires x usages of a causal language model to generate x words, which makes it inefficient. Grouped sampling is an alternative algorithm that manipulates the input text before passing it to the model, forcing the model to predict the entire output. Grouped sampling only requires one use of a causal language model to generate text of any length, making it much more efficient. I compared grouped sampling and the naive algorithm in translating TED talks. The naive algorithm required 33.049 GPU hours and $17.87. Grouped sampling required 0.028 GPU hours and $0.015. Grouped sampling translated more accurately by 5%-24%, measured using BERT scores. In conclusion, grouped sampling is an accurate and efficient text-generation technique. It is 1180 times faster and cheaper to run than the naive algorithm.