Qeios

Peer Review

Review of: "Low-Bit Quantization Favors Undertrained LLMs: Scaling Laws for Quantized LLMs with 100T Training Tokens"

Cheng Gong[1]

1. Nankai University, China

Summary: This study has conducted numerous experiments to explore and establish the relationships between Quality of Interest (QiD) and various Large Language Model (LLM) training attributes, including the number of parameters, the number of training tokens, and the quantization bits. The research demonstrates that QiD is biased towards undertrained LLMs and attempts to leverage these relationships to predict the required number of training tokens for LLMs with different model sizes by evaluating QiD under specified bitwidth conditions.

Strengths: The paper presents a substantial amount of experimental work across various LLMs and datasets, which is commendable.

Weaknesses:

1. The paper appears to be a comprehensive experimental report with derived conclusions, lacking the necessary theoretical analysis. The motivation behind the paper seems weak, and the application scenarios are not clearly defined. To be precise, it is unclear what problem this paper addresses or in which scenarios it can be applied. If the research is conducted for the sake of research alone, it may seem meaningless.
2. The conclusions drawn in the paper seem to be common knowledge, such as the notion that QiD favors undertrained LLMs. However, the paper fails to provide any groundbreaking explanations, which significantly undermines the contribution of the study.

Questions:

doi.org/10.32388/MYDHM6

In the foundational theories of deep learning, model capacity and dataset size are two crucial concepts. The larger the model capacity and the dataset size, the more likely it is to achieve the ideal generalization error lower bound, known as the Bayes error. In simple terms, larger datasets and model capacities theoretically result in smaller generalization errors, and current scaling laws have experimentally supported this theory. In this paper, the number of parameters corresponds to model capacity, and the number of training tokens corresponds to the training dataset. Based on the theoretical analysis of model capacity and dataset size, the proposal to use QiD to predict the required training tokens for different LLMs seems pointless, as more training tokens are generally better (assuming the variance of noise in the dataset remains constant). Could the authors explain the motivation behind this predictive method?

Another fundamental concept in machine learning is overfitting and underfitting, which refer to the model capacity exceeding or falling short of the data complexity, respectively. This concept is quite similar to the terms undertrained and overtrained used in this paper. Why not use overfitting and underfitting to describe these phenomena? Or, what are the differences between these two sets of concepts?

It is well-known that quantization affects data precision, thereby reducing the model's expressive power, i.e., reducing model capacity. In simple terms, for underfitting models, quantization exacerbates underfitting, leading to a sharp decline in model performance. For overfitting models, quantization can mitigate overfitting, thereby alleviating the decline in model performance or even enhancing it (if fine-tuned based on Quantization Aware Training). The paper suggests that QiD favors undertrained LLMs, which I believe is merely a phenomenon resulting from the quantization of underfitting models. Undertrained LLMs train a small model on a large dataset, where the model capacity is inherently insufficient. Quantization further aggravates the insufficiency of model capacity, severely affecting model performance. This can also explain all the experimental trends in the paper, such as the sharp increase in QiD with the number of training tokens, as illustrated in Figure 1. This is because, with a fixed number of model parameters, more data leads to more severe underfitting, resulting in more significant QiD.

Justification for Score: The paper presents a significant amount of experimental data and explores the relationship between QiD and LLM training properties. However, the lack of theoretical underpinning and the unclear application scenarios limit its impact and contribution to the field. The paper's conclusions do not offer new insights into the well-known phenomenon of QiD favoring undertrained

LLMs. The score reflects the need for a stronger theoretical foundation and clearer practical applications to enhance the paper's value.

Declarations

Potential competing interests: No potential competing interests to declare.