LLaMa

Research Problem/Objective

The paper addresses two primary challenges in Large Language Model (LLM) development:

- The reliance of most state-of-the-art (SOTA) LLMs on proprietary and inaccessible datasets, hindering open research and reproducibility.
- The prevailing focus on scaling model size based on training compute budget (as per scaling laws like Chinchilla's), often neglecting the critical aspect of inference efficiency, which is paramount for deployment.
  The objective is to demonstrate that it is possible to train SOTA or highly competitive LLMs using only publicly available data, and to investigate whether training smaller models on significantly more data than typically prescribed can yield better performance per inference cost.

Core Idea/Hypothesis

The central hypothesis is that training relatively smaller LLMs (7B to 65B parameters) on a substantially larger corpus of tokens (trillions) than suggested by training-compute-optimal scaling laws can achieve performance comparable or superior to much larger models (like GPT-3 175B, PaLM 540B) trained on less data. The underlying idea is that for a target performance level, a smaller model trained longer will ultimately be more efficient (cheaper) at inference time, even if it requires a longer training period. A secondary core idea is that SOTA performance is attainable without resorting to proprietary datasets.

Model Architecture (Conceptual Overview)

LLaMA models are based on the standard Transformer architecture but incorporate several established improvements aimed at enhancing performance and training stability:

- **Pre-normalization (Inspired by GPT-3):** Layer normalization (specifically RMSNorm by Zhang & Sennrich, 2019) is applied before each transformer sub-layer, rather than after, to improve training stability.
- **SwiGLU Activation Function (Inspired by PaLM):** The ReLU activation function is replaced by SwiGLU (Shazeer, 2020), which has shown performance benefits in other LLMs. The feed-forward dimension is adjusted (2/3 * 4d instead of 4d as in PaLM).
- **Rotary Positional Embeddings (RoPE) (Inspired by GPTNeo):** Absolute positional embeddings are discarded in favor of Rotary Positional Embeddings (Su et al., 2021), applied at each layer of the network to incorporate sequence position information.
  The models follow a standard decoder-only, autoregressive structure. The paper presents a family of models ranging in size from 7B to 65B parameters, varying primarily

in dimensions, number of heads, and number of layers. No fundamentally novel architectural components are introduced; the design represents a specific selection and combination of known effective techniques.

Key Innovations & Contributions

- **Training Strategy Validation:** Empirically demonstrating the effectiveness of training smaller models (7B-65B) on significantly larger datasets (1T-1.4T tokens) than prior work or scaling laws suggested, achieving high performance with potentially better inference efficiency.
- **Public Data Sufficiency:** Providing strong evidence that SOTA-level LLM performance can be reached using exclusively publicly available data sources, challenging the perceived need for proprietary data.
- **High-Performance Open Models:** Developing and releasing a family of foundation models (LLaMA 7B, 13B, 33B, 65B) that show competitive or superior performance compared to prominent (often closed) models (e.g., LLaMA-13B outperforming GPT-3 175B; LLaMA-65B competitive with Chinchilla-70B, PaLM-540B). This release aims to democratize research.

Research Methodology & Evaluation (High-Level)

- **Methodology:** Pre-training Transformer models (LLaMA architecture) of varying sizes (7B to 65B parameters) on a large corpus (~1.4T tokens) aggregated from diverse, publicly available text sources.
- **Datasets:** Pre-training data included processed CommonCrawl, C4, GitHub code, Wikipedia, Books (Gutenberg, Books3), ArXiv papers, and Stack Exchange Q&A. Evaluation was performed on a broad suite of 20 standard NLP benchmarks covering Common Sense Reasoning (e.g., BoolQ, HellaSwag, WinoGrande), Closed-book QA (Natural Questions, TriviaQA), Reading Comprehension (RACE), Mathematical Reasoning (MATH, GSM8k), Code Generation (HumanEval, MBPP), and Massive Multitask Language Understanding (MMLU).
- **Evaluation Metrics:** Standard metrics relevant to each task were used, primarily accuracy, exact match (for QA), and pass@k (for code generation), evaluated in zero-shot and few-shot settings. Performance was compared against contemporary LLMs (GPT-3, Gopher, Chinchilla, PaLM, OPT, etc.). Training loss was tracked. Bias and toxicity were measured using benchmarks like RealToxicityPrompts, CrowS-Pairs, WinoGender, and TruthfulQA.

Significant Findings & Results

- LLaMA-13B surpassed the performance of GPT-3 (175B) on most benchmarks, despite being over 10 times smaller in parameter count.
- LLaMA-65B demonstrated performance competitive with Chinchilla-70B and PaLM-540B across a wide array of tasks, confirming the viability of the training strategy and public

data approach.
- Performance generally scaled well with model size and the amount of training data processed.
- The models exhibited strong zero-shot and few-shot learning capabilities across the evaluated tasks.
- Brief instruction fine-tuning significantly improved performance on benchmarks like MMLU, indicating the models are amenable to downstream adaptation.
- Analyses confirmed the presence of social biases (gender, religion etc.) and toxicity in the models, reflecting the nature of large-scale web data, with toxicity tending to increase with model size. Truthfulness, while improved compared to GPT-3, remained imperfect.

Conclusions & Implications

The authors conclude that state-of-the-art foundation LLMs can be developed using solely public datasets, removing a significant barrier to open research. The results validate their hypothesis that training smaller models on vast amounts of data is an effective strategy for achieving high performance, particularly when considering inference efficiency. The release of the LLaMA models is intended to facilitate broader research access and accelerate development in LLM capabilities, robustness, and safety. The findings suggest a potential shift in focus towards optimizing the data-to-model-size ratio for inference efficiency, rather than purely for training compute optimality. Future work includes further investigation into instruction tuning and potentially scaling models further based on the positive trends observed.