Pretraining Large Language Models

Summer

Introduction

This document provides an overview of pretraining large language models (LLMs), discussing scaling laws, dataset preparation, and distributed training strategies. It also highlights trends in LLM development.

State of LLMs

- **Closed Models**: APIs only; no access to model weights or data.
- **Open Models**: Fully open access to model weights, code, and data.
- **Trends**:
  - Longer training durations.
  - Larger model sizes (e.g., GPT-4 has 1,800 billion parameters).
  - Increased context windows.
  - Higher compute budgets.

Scaling Laws

- **Predictable Returns**: Performance scales predictably with data, compute, and model size.
- **Compute-Optimal Models**: Models that minimize loss for a given compute budget.
- **Chinchilla Fix**: Focus on training with more data instead of just increasing model size.

Datasets

- **Goals**: Train general-purpose models with diverse, high-quality text.
- **Challenges**:
  - Maximizing data diversity and quality.
  - Filtering noisy or irrelevant data.

- **Common Sources**:
    - Common Crawl.
    - Curated datasets (e.g., Wikipedia, Arxiv).
    - Synthetic data generation.
- **Filtering Pipelines**:
    - Use heuristics (e.g., perplexity-based filtering).
    - Classifier-based quality evaluation.

Distributed Training

- **Parallelism Strategies**:
    - Data Parallelism: Distribute microbatches across GPUs.
    - Tensor Parallelism: Split matrix computations.
    - Pipeline Parallelism: Share layers across GPUs.
    - Sequence/Context Parallelism: Process sequences in parallel.
- **Mixed Precision Training**:
    - Use FP16/BF16 for faster computation.
    - Experimental approaches include FP8.
- **Optimization Techniques**:
    - ZeRO (Zero Redundancy Optimizer): Reduces memory overhead.
    - Flash Attention: Optimizes attention computation.

Advantages of Pretrained LLMs

- High scalability and transferability.
- State-of-the-art performance across multiple domains.
- Efficient handling of diverse and large-scale datasets.

References

- Leandro von Werra, *Pretraining Large Language Models.*
- Chinchilla Scaling Laws: `https://arxiv.org/abs/2203.15556`.
- Flash Attention: `https://arxiv.org/pdf/2205.14135`.