Efficient LLMs via Switchable and Dynamic Quantization

Background and goal: Large language models (LLMs) have garnered increasing attention due to their remarkable emergent abilities, albeit at the expense of their substantial model size. However, this large size also poses challenges in terms of efficiency. Among various efforts to enhance LLM efficiency, quantization has emerged as a promising approach, offering a favorable accuracy-efficiency trade-off, ease of implementation, and compatibility with hardware. In this project, our goal is to further investigate this method by developing a switchable and dynamic quantization scheme that aims to improve the accuracy-efficiency trade-off of LLMs.

Problem breakdown: To achieve the goal of switchable and dynamic quantization, you need to finish the following steps:

- Step 1: Integrate quantization into the pretrained GPT-2 model. Given the potentially diverse redundancy in each layer of the model, implement the quantization function to support using different bit-width per layer based on the input config.
- Step 2: Add multiple LoRA modules to all linear layers in GPT-2 and implement the function to adaptively activate different LoRA modules in each layer during inference based on the input config.
- Step 3: Tune the GPT-2 model for 1000 iterations on the SQuAD dataset with different per-layer quantization bit-width configurations at the same time (i.e., enabling switchable precision), where different LoRA modules are activated for different per-layer quantization bit-width configurations.
- Step 4: Evaluate the performance of your tuned model under different quantization bit-width configurations on SQuAD dataset. You may decide your own quantization configuration, i.e., the optimal per-layer bit-width configuration to push the accuracy-efficiency trade-off. Try to draw some insights based on your observations.
- Step 5: Instead of jointly training with all bit-widths in Step 3, use the cyclic precision training to dynamically change the training bit-widths throughout the tuning process and investigate whether this quantization strategy, which effectively enhances the accuracy of CNNs, can also improve the downstream accuracy of LLMs.
- Step 6: On top of a pretrained GPT-2, examine whether random precision switch, i.e., dynamic quantization at inference time, can improve GPT-2's adversarial robustness. You can do a literature survey and pick arbitrary widely used adversarial attacks within 3 years to evaluate GPT-2's adversarial robustness.

Implementation Details

- *Codebase references*
  - Model: GPT-2
  - Quantization scheme: QAT-LLM
  - Downstream dataset: SQuAD
- *Deliverables*: Submit your code with a brief ReadMe and a PDF report to answer the following questions:
  - [Step 4] What is the task accuracy achieved after applying various quantization bit-width configurations to the SQuAD dataset?
  - [Step 4] How did you determine the optimal quantization bit-width configurations? Have you gleaned any insights from your observations that could guide future work to further enhance performance?
  - [Step 4] A motivation behind switchable quantization is to support diverse layer-wise quantization configurations simultaneously, accommodating different resource allocation needs. Could you suggest additional training objectives that could more effectively facilitate the mechanism for switching quantization bit-widths?
  - [Step 5] Does this phenomenon align with the observations in CPT (ICLR'21)? If not, what could be the potential reasons?
  - [Step 6] Does this phenomenon align with the observations in Double-Win Quant (ICML'21)? If not, what could be the potential reasons?
  - Based on your explorations of switchable and dynamic quantization, could you propose some promising research directions or questions for further integrating them with LLMs?