BlackMamba: Mixture of Experts for State-Space Models

Main Idea: this work looks to combine the Mamba with the MoE architecture. Each of these architectures have unique advantages: Mamba has linear time and memory complexity to increases in context length (is robust to long-range context), while MoE has the advantage of allowing for scaling model's parameters while keeping inference costs fixed at the expense of a larger memory footprint. BlackMamba (Mamba + MoE fusion) is then expected to have the long-range context robustness of Mamba while having the inference efficiency of MoE.

The models experimented with are larger than the previous work done (Mamba-MoE) but could be scaled much more (the models are, in [ACTIVE PARAMETERS/TOTAL PARAMETERS], 340M/1.5B and 630M/2.8B)

Expected Advantages (Synergies) of BlackMamba vs Dense Transformer

- From Mamba
    - Linear computational complexity with respect to input sequence length for both training and inference.
    - Autoregressive generation in constant time and memory.
- From MoE
    - Inference latency and training FLOPs of the equivalent smaller dense base model, while preserving model quality close to an equivalent dense model in terms of total parameters.

MoE Details

- MoE top-k routing is used.
- MoE is compared/evaluated based on:
    - (Forward pass or active parameters) / total parameters ratio
- Similarly to Mixtral8x7B, a relatively small number of experts is used in BlackMamba (even though scaling laws show promise in having many experts) to balance the inference FLOPs and memory cost of MoE (more experts = more memory costs).

Architecture

- BlackMamba consists of replacing a few layers in the Transformer architecture:
    - The MLP/FF layers are replaced by sparse MoE layers.
    - The Attention layers are replaced by Mamba layers.
- BlackMamba was trained on 300B tokens. This is consistent with the scaling done in this paper compared to the previous work trying to combine these architectures (MoE-Mamba was trained on 10B tokens and had significantly smaller model size).
    - 340M/1.5B and 630M/2.8B sized models trained (active parameters/total parameters).
    - 8 experts used per MoE layer.
    - Found a slight advantage in using sequential versus parallel blocks, so prioritized a sequential setup.

        - This is equivalent to depth vs width.
    - Used top-1 routing with the Sinkhorn algorithm to ensure load balancing between experts.
        - Sinkhorn was the same algorithm used in BASE routing. It makes routing more efficient in accelerated hardware (GPUs).
        - A novel version of Sinkhorn was developed, which has faster convergence.
    - Used the Megatron-LM framework for distributed training.
    - Trained using bf16 precision.

Results

- For the same number of active parameters (equal at inference) and the same amount of training FLOPs (equal amount of training), BlackMamba performed significantly better than the Transformer, Transformer-MoE and Mamba equivalents.
- As expected, BlackMamba also showed significant latency improvements over the other architectures. These latency improvements increase with an increase in context length.
    - This indicates that the synergy between Mamba and MoE works.
- In terms of expert balance, most layers show this happens successfully. However, later layers show a clear transition towards expert imbalance.

    - Perhaps this is due to numerical instabilities that show as we get deeper into the network?
    - This pattern of instability in later MoE layers was also shown in the “Faster-MoE” paper.
- BlackMamba leaves room for future work in terms of the Mamba + MoE fusion:
    - Few-shot performance.
    - Quantization and PEFT performance.
    - Fine-tuning, instruction tuning and DPO performance.
    - Are the expert’s specialization dynamics in BlackMamba the same as in Transformer MoEs?

My takeaways:

- The checkpoints of BlackMamba were released, so perhaps some investigation can be done in terms of exploring the expert’s specialization dynamics in the BlackMamba architecture and compare it to regular Transformer MoEs.