Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Paper

The Mamba model addresses the inefficiencies of Transformers, especially with long sequences. The paper introduces a new selective SSM where the SSM parameters are input-dependent, allowing the model to focus on relevant information.

Motivation

SSMs use linear time and memory but because the parameters are predefined, it cannot perform content-based reasoning effectively. Transformers are able to perform content-based reasoning but do so in quadratic time and memory. Selective SSMs have the benefits of both with the linear time and memory from the SSM architecture but by modifying the SSM parameters during inference, the model can focus on relevant content.

Architecture

State Space Models (SSMs)

Similar to a recurrent neural network where each output is affected by previous information and the current input. However, each step contains a state matrix, h_t, which is affected by the previous state and the current input, x_t. This state, along with the current input, affects the current output, y_t. These are all done through matrix multiplications.

$$h_t = Ah_{t-1} + Bx_t \quad y_t = Ch_t + Dx_t$$

Selection Mechanism

A variant of the state space model architecture where the SSM parameters are input-dependent, to focus on relevant information and disregard irrelevant data dynamically. This selective SSM architecture does not require attention or MLP blocks. The SSM architecture uses a convolution layer to determine how the input values affect each output value through the hidden state. This convolution layer includes learnable parameters, along with the learned matrices.

Limitations

1. The model is only compared against small model sizes and has not been tested against open source 7B LLMs like Llama.

Mamba was rejected from ICLR 2024 due to missing Long Range Arena (LRA) results, a standard for evaluating long-sequence models, and the use of perplexity as the primary metric was challenged.