Paper Review 3

Title: Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention

Paper Summary:

Linear Transformers address the quadratic complexity of traditional transformers, by improving their speed for processing long sequences. Self-attention is expressed as linear dot-product of kernel feature maps and utilizing the associative property of matrix products, reducing complexity from O(N2) to O(N). Efficiency comes from Linearized Attention, which replaces softmax attention with feature map dot products, improving both computational speed and memory usage. Pre-computing and reusing feature map representations further reduces computational burden, making the model more efficient.

The choice of feature map and kernel function is crucial. While softmax attention uses an infinite-dimensional feature map, practical alternatives like polynomial kernels offer finite-dimensional representations. Experiments are conducted with polynomial kernels and exponential linear unit (ELU) feature maps to ensure positive similarity scores. Causal masking is introduced to linearize the attention mechanism for autoregressive tasks, allowing each position in a sequence to attend only to preceding ones. This approach reduces memory complexity to a constant, even for long sequences.

Linear Transformers can be formulated as RNNs because the linearized causal attention mechanism maintains an internal state, updated recursively with each input, mimicking the behavior of RNNs. In performance evaluations, Linear Transformers demonstrate comparable accuracy to softmax and Reformer models on tasks like image generation (MNIST, CIFAR-10) and speech recognition (WSJ), while being computationally faster. On CIFAR-10, Linear Transformers generate images 4,460 times faster than softmax transformers. In tasks like speech recognition, softmax transformers still show superior convergence and final performance, despite Linear Transformers' computational advantages.

Critique:

Strengths:

- Reduction of complexity from O(N2) to O(N) for handling long sequences efficiently.
- Causal masking enables transformers to behave like RNNs in autoregressive tasks.
- Speed improvements, in image generation tasks, without major performance drops.
- Exploration of alternative kernels, such as polynomial kernels and ELU-based feature maps, adds versatility to the model's application.

Weaknesses:

- Performance trade-offs, especially in speech recognition tasks where softmax transformers demonstrate faster convergence and superior results.
- Did not explore much about the model's effectiveness for shorter sequences.
- Despite computational efficiency, linear transformers may not be universally optimal. How well does linearized attention mechanism generalize to other types of transformers or attention models?
- Choice of kernel function significantly influences model behavior, but did not describe how different kernels impact specific tasks