Qeios

Peer Review

Review of: "ARWKV: Pretraining Is Not What We Need – An RNN-Attention-Based Language Model Born From Transformer"

Tobias Strauß[1]

1. Institute of Mathematics, Universität Rostock, Germany

First of all: Perhaps I am not the right person to review this paper. I have never heard of RWKV before. So sorry if I have misunderstood some points.

The paper investigates the distillation of RNN-based LMs, so-called RWKV models. The authors use Qwen models as teachers and vary different architectural details.

Strengths: I like the idea of fundamentally questioning the Transformer architecture. The disadvantage of the square memory requirement of the Attention modules is addressed by the RNN layers.

Weaknesses: The paper is not understandable without reading several other papers beforehand. Layer-wise distillation and one-step distillation are fundamental to this article but not explained, nor is there a reference. What is direct preference optimization?
For me, the architecture section must be clearer. The authors write that they keep certain things from the Qwen architecture and replace the rest with RWKV-attention. What is the rest? If you keep the SwiGLU activation, where is it used? Is RWKV-attention the same as the time mixing module?

The terms in equations (3) and (4) are not explained. Not even the dimension is clear. In both equations, there are 3 different symbols for multiplication, but nowhere is it explained what each symbol means.

Moreover, several abbreviations are not introduced, e.g., KV, QRWK, QKV, GOA, h800.

doi.org/10.32388/Y331KE

Minor critiques:

It would help me if the loss function was explicitly written down in section 3.2.

Which metric is used in Table 1? Maybe you could highlight the important results?

There are several typos and inconsistent spellings that should be corrected.

Declarations

Potential competing interests: No potential competing interests to declare.