# 会议纪要

**会议主题：长序列处理中的注意力机制研究进展**

**会议时间：** 2024年10月28日19：00至20：00

**会议地点：** 西安电子科技大学北校区会议中心303-1会议室

**线下人员：**承楠孙瑞锦 尹志胜 贺靖超 王葳 沈京龙 王秀程 马龙飞 周新阳 权赟昊 张玉洁 黄蕾 胡陆莹 齐阁 邓川 孙兴栋 方忠盛 贾宏刚 郑雯馨 祝馨平 贾昊燏 张智杰 杨杰 邱子仪 刘永红 万佳林 朱煜朋 韩松明 刘苏 尚佳瑶 陈梦豪 郑佩林 孙路路 王天宇 于凡迪 尉家豪 杨双宇

**线上人员：**赵璇 傅连浩 郭译凡 侯毓真 李青壮 许嘉洁 张岳

**记 录 人：**郑雯馨

**会议内容：**解思舀介绍了Attention在处理长序列时的问题和当前解决这些问题的主要研究。

* **承楠：**HBM是什么？

**解思舀：**显存，GPU的一部分

* **承楠：**IO-Awareness是什么意思？

**解思舀：**IO就是Input和Output。

**王秀程：** 因为正常的Attention的计算需要频繁的从HBM和SRAM之间进行访问，HBM的效率没有SRAM的高，Flash Attention就想减少从HBM和和SRAM之间的读写操作，尽可能地在高速内存上(也就是SRAM)进行操作，而不是在HBM（也就是正常的显存）上进行操作。

* **承楠：** SRAM只有20MB?

**解思舀：**对，很小，所以才要分块处理。

**承楠：**那为什么不多放点SRAM？

**王秀程：**太贵了。

* **承楠：**Approximate Attention就是不那么精确吗？

**解思舀：**是的，相当于改变注意力本身的形式。

* **承楠：**Attention模型会不会像普通的神经网络一样有输入序列宽度的限制？

**王秀程：**理论上可以大于限制，但是越大的话，softmax平均之后的值就会越来越小，y值就会特别小，相当于把分布给改了，实际效果不好。

* **承楠：**长序列要解决的问题是你说的这个还是两个问题都有?

**王秀程：**现在这个问题都能克服，主要还是计算量和显存的问题比较难做。

* **承楠：**如果需要星上处理，硬件应该是什么样的？

**解思舀：**我觉得跟卫星本身有关系，考虑它本身的存储和计算框架。

* **承楠：**100万token的上下文怎么理解？

**解思舀：**ring attention其实并没有真实的实现100万，但是理论上可以扩展到100万。

* **承楠：**那这个100万token是一次性输入的吗？

**解思舀：**它是最长的能获取到的前文的长度。

* **承楠：**那这个RWKV效果怎么样，会不会有遗忘的问题？ 为什么它能像你说的这样训练和推理。

**解思舀：**性能后面会提到，会有遗忘的问题，它的主要技术是一个attention改进的公式，和一个token shift（token向前位置的偏移），这个公式可以写成Transformer和RNN的形式。

* **王秀程：** Attention的计算方式在多卡中怎么实现的？

**解思舀：**基于ring attention，可以再加入flash attention这些技术来提高计算速度。

* **王秀程：**那么现在研究AI是把一个完整的模型加载到一张卡里，然后用剩余的内存跑数据，还是把模型加载到不同的卡里，就是模型并行。

**解思舀：**这个完整模型的本质有点像ring attention，每一张卡都放了所有的模型。可能因为模型并行对Attention这样一个很适合用数据并行加速的模型来说研究意义不大。

* **承楠：**如何结合和星网的研究？attention和卫星算力网结合；通信路由和attention结合。

**解思舀：**我回去继续调研。