Attention??????

RNN容易梯度消失：梯度消失的本质问题是，网络太深了，这里的深代表信息表达的层级而不是拓扑，resnet就是解决这个问题

Transformer 的强大之处同时也是它的弱点：Transformer 中固有的自注意力机制（attention）带来了挑战，主要是由于其二次复杂度造成的，这种复杂度使得该架构在涉及长输入序列或资源受限情况下计算成本高昂且占用内存。

?Transformer ????

1. 以 **RWKV**、**Mamba** 和 **S4** 为代表，它们完全用 **recurrent**（循环）结构去替代 **attention**。

   这种思路是用一个固定的内存记住前面的信息，但目前看来虽然可以记住一定长度，但要达到更长的长度是有难度的。
2. 把 **full attention** 这种密集结构变得稀疏，例如 Meta 的 Mega，在之后的计算中不再需要算所有 attention 矩阵中的每一个元素，模型效率也随之变高。

DeepMind 团队提出的 Hawk 和 Griffin 同样认为没有 attention 是不行的，属于 gated linear RNN，跟 Mega 一样属于混合模型。

现阶段来看，基于现有硬件的算力基础，用 Transformer 去做端侧大模型的难度很高，还是需要在云上完成计算推理等工作，而且应答速度不如人意，终端用户很难接受。

上述投资人评价 RWKV“麻雀虽小，五脏俱全”，总体体验感能达到 GPT-3.5 的 60 分，但并不知道最后能否达到 GPT 的 80 分、90 分。这也是非 Transformer 的问题所在，即如果舍弃了框架的复杂度、可能会牺牲上限的天花板。

Transformer 日益坚固的生态护城河，无论是硬件、系统、应用，都是围绕 **Transformer** 做适配、优化，使得开发其他架构的性价比降低，导致想要开发新的架构越来越难。

Attention???TopK Attention???

众所周知，注意力机制本质上具有稀疏性，因此动态稀疏注意力和基于 TopK的近似方法得到了广泛研究。

然而，这些方法往往伴随着显著的质量下降问题。

目前已有的 KV缓存压缩技术，如 Quest、H2O和 Loki，主要通过筛选出 KV 缓存中注意力得分最高的子集来提高效率。然而，尽管这些方法在实践中表现出一定的效果，基于 TopK 的注意力依然是一种存在偏差的近似方法，且缺乏理论上的严格保障。

这种不足限制了其在高精度场景中的广泛应用。

Revision #1
Created 2025-01-11 09:44:04 UTC by Colin
Updated 2025-01-11 09:44:04 UTC by Colin