读文章

https://mp.weixin.qq.com/s/miEziKZNdlnRym0qJlvqLw

1. 内存墙问题既涉及有限的容量、内存传输带宽，也涉及其延迟（比带宽更难改善）
2. 传统趋势表明，新的NN模型架构是根据研究人员在单芯片内的访问权限而开发的，而不是使用复杂的分布式内存方法
3. 事实上，人工智能加速器的重大突破之一就是使用半精度（FP16）算术，这使得硬件计算能力提高了10倍以上。然而，利用当前的优化方法，在不降低精度的情况下进一步降低精度（从半精度降低到INT8）一直是个难题。最近一个很有前途的趋势是混合使用FP8和FP16（甚至最近的FP4）。
4. 在NLP中，训练最新SOTA Trans former模型的计算成本以750×/2yrs的速度递增，模型参数大小以410×/2yrs的速度递增。相比之下，硬件FLOPS峰值的扩展速度为3.0×/2yrs，而DRAM和互连带宽的扩展速度分别为1.6×/2yrs和1.4×/2yrs，越来越落后。

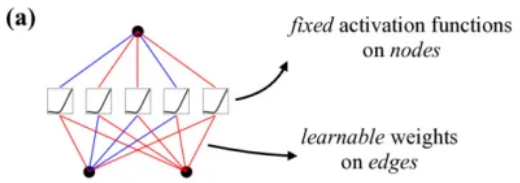
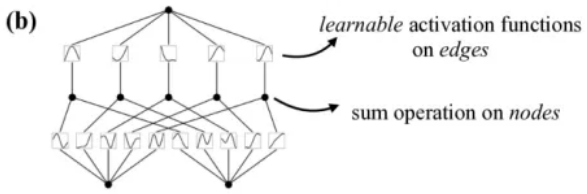
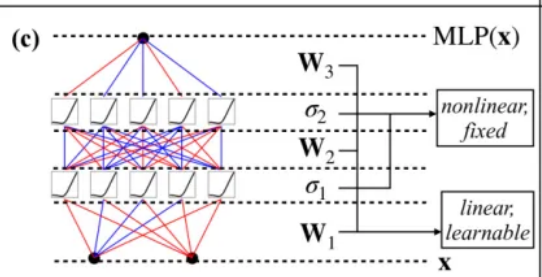
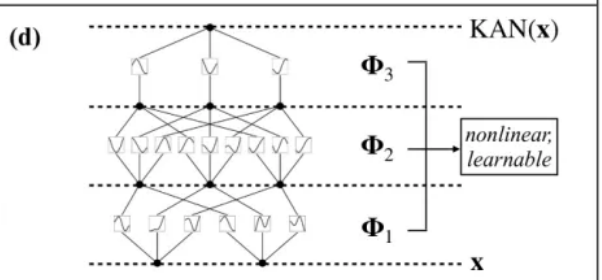
DRAM模拟器

https://people.inf.ethz.ch/omutlu/pub/Ramulator2_arxiv23.pdf

感知机 KAN MLP

https://mp.weixin.qq.com/s/Hrp5v5enYlx3cVwtG63d6w

KAN的核心创新在于，它不是像传统MLP那样在网络的神经元上应用固定的激活函数，而是在权重上应用可学习的激活函数。这些一维激活函数被参数化为样条曲线，从而使得网络能够以一种更灵活、更接近Kolmogorov-Arnold表示定理的方式来处理和学习输入数据的复杂关系。这种设计使得KAN能够以一种理论上更高效、更通用的方式逼近复杂的函数关系，理论上可能在某些任务上超越了MLP的性能。

Model	Multi-Layer Perceptron (MLP)	Kolmogorov-Arnold Network (KAN)
Theorem	Universal Approximation Theorem	Kolmogorov-Arnold Representation Theorem
Formula (Shallow)	$f(\mathbf{x}) \approx \sum_{i=1}^{N(\epsilon)} a_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$	$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right)$
Model (Shallow)		
Formula (Deep)	$\mathrm{MLP}(\mathbf{x}) = (\mathbf{W}_3 \circ \sigma_2 \circ \mathbf{W}_2 \circ \sigma_1 \circ \mathbf{W}_1)(\mathbf{x})$	$\mathrm{KAN}(\mathbf{x}) = (\mathbf{\Phi}_3 \circ \mathbf{\Phi}_2 \circ \mathbf{\Phi}_1)(\mathbf{x})$
Model (Deep)		

Revision #1
Created 11 January 2025 09:44:04 by Colin
Updated 11 January 2025 09:44:05 by Colin