https://arxiv.org/abs/2204.02311

PaLM分析

PaLM全称Pathways Language Model，但是这个模型其实与去年Jeaf dean宣传的Pathways模型差异比较大，不是多任务/多模态、没有稀疏激活/动态路由，仍然是SPMD：

模型结构-GPT3的改进版本

1. 纯Decoder，类似GPT3的结构，**稠密模型**，8B/62B/540B；
2. 激活函数用SwiGLU：Swish(xW) xV **此激活函数计算量更大，但是精度收益较大；**
3. Parallel Layers：y = x + MLP(LayerNorm(x + Attention(LayerNorm(x)))->y = x + MLP(LayerNorm(x)) + Attention(LayerNorm(x)) **改变算法，利用MLP+Attention的算子融合，提速15%，会对精度有微小影响；**
4. Multi-Query Attention: key和value单头，query多头 **改变算法，减小计算量，Attention部分计算量减小约2/3；**
    - 标准的多头注意力在自回归解码期间在加速器硬件上的效率很低，因为键/值张量在示例之间不共享。文中模型让key/value映射被每个头共享，而Query相互独立，该方法提升了解码器的自回归时间。
5. RoPE Embedding：旋转式相对位置编码 **改变算法，精度会有收益（长序列更友好）；llama used this**
6. No bias，No dropout **越来越多的大模型开始采用这种方式；**
    - 可以增加大模型的训练稳定性。
7. Adafactor **略微影响精度，减少优化器状态，节省内存。**
8. 优化词表
   使用**SentencePiece**（通过统计方法，将频繁出现的字符串作为词，然后形成词库进行切分），使切分的粒度会更大一些。使用256K的token表，词表以外的文本被切分成utf-8字符。

$$\mathrm{SwiGLU}(x, W, V, b, c, \beta) = \mathrm{Swish}_\beta(xW + b) \otimes (xV + c)$$

$$Swish_\beta(x) = x\sigma(\beta x)$$
$$\sigma(z) \sim \ sigmoid\ function$$

PALM训练

使用 PathWay 方法训练模型，在两个TPU v4 Pods上训练，在每个Pod中包含由3072个TPU v4芯片链接的768个主机。允许在不使用任何pipeline并行的情况下高效的在6144个芯片上训练。
pipeline方式有更多的相互等待时间，而pathway复杂度更高。每个TPU v4 Pod都包含模型参数的完全拷贝。
详见：Pathway原理。

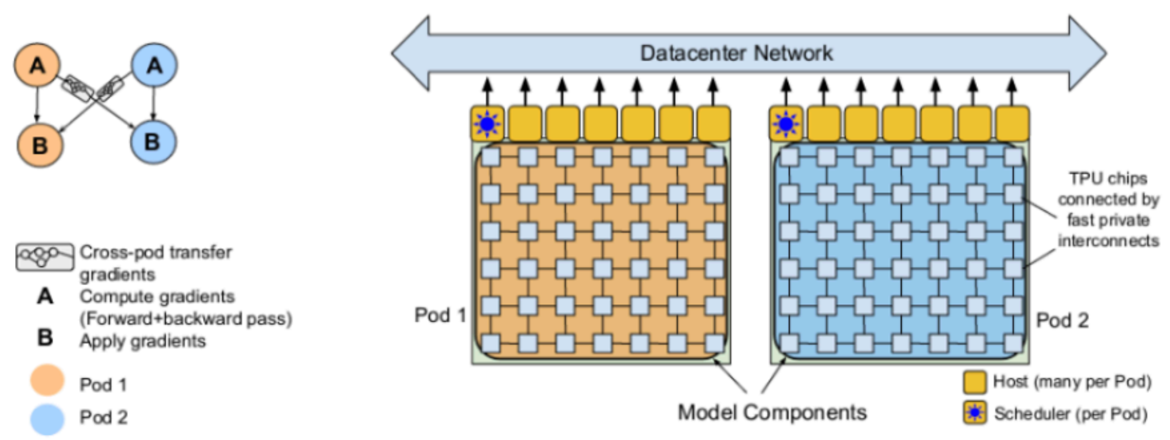

Figure 2: The Pathways system (Barham et al., 2022) scales training across two TPU v4 pods using two-way data parallelism at the pod level.

Model	# of Parameters (in billions)	Accelerator chips	Model FLOPS utilization
GPT-3	175B	V100	21.3%
Gopher	280B	4096 TPU v3	32.5%
Megatron-Turing NLG	530B	2240 A100	30.2%
PaLM	540B	6144 TPU v4	46.2%

Table 3: Model FLOPs utilization of PaLM and prior large models. PaLM achieves a notably high MFU because of several optimizations across the model, compiler, and parallelism strategy. The corresponding hardware FLOPs utilization of PaLM is 57.8%. Details of the calculation are in Appendix B.

相比之前模型，PaLM在由于对模型、编译器和并行策略进行了多项优化，实现了非常高的MFU，对应的硬件FLOPs利用率也更高。

1. **PaLM与Pathways的关系**
2. **Palm 阅读**

PaLM Analysis:

1. PaLM stands for Pathways Language Model, but this model is actually quite different from the Pathways model that Jeff Dean promoted last year. It is not a multitask/multimodal model and does not have sparse activation/dynamic routing. It is still an SPMD model with the following structure:
2. Pure decoder, similar to the structure of GPT-3, a dense model with 8B/62B/540B parameters.
3. Uses the SwiGLU activation function: Swish(xW) xV. This activation function has a higher computational cost, but provides greater precision gains.
4. Parallel Layers: y = x + MLP(LayerNorm(x + Attention(LayerNorm(x)))->y = x + MLP(LayerNorm(x)) + Attention(LayerNorm(x)). The algorithm has been changed to use the operator fusion of MLP+Attention, which speeds up the model by 15% with a small impact on accuracy.
5. Multi-Query Attention: single-headed key and value, multi-headed query. The algorithm has been changed to reduce the computational cost of Attention by approximately 2/3.
    - Standard multi-head attention is not efficient on accelerator hardware during autoregressive decoding because the key/value tensors are not shared between examples. In this model, the key/value mappings are shared by each head, while the queries are independent of each other. This method improves the autoregressive decoding time of the decoder.
6. RoPE Embedding: rotation-based relative position encoding. The algorithm has been changed to improve accuracy (more friendly to long sequences). This was also used in Llama.
7. No bias, no dropout. This approach is increasingly being used by larger models to increase training stability.
    - This can increase the training stability of larger models.
8. Adafactor: slightly affects accuracy and reduces optimizer states to save memory.
9. Optimized vocabulary: SentencePiece is used (a statistical method that takes frequently occurring strings as words and forms a vocabulary for segmentation), resulting in larger segments. A token table of 256K is used and text outside the vocabulary is split into UTF-8 characters.