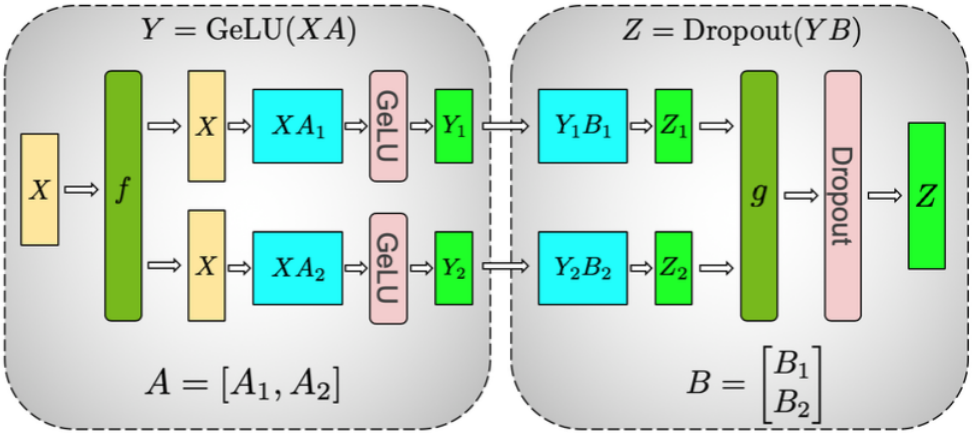
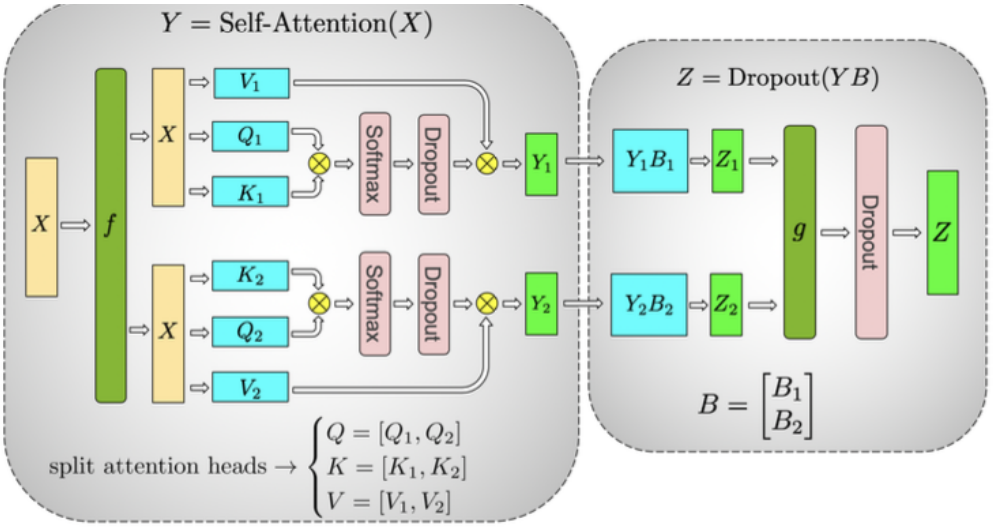
# megatron学习文档




- mlp模块模型并行
  - GeLU激活函数为非线性激活函数，不满足加法结合律，也即f(a)+f(b)!=f(a+b)
  - 因此在前半部分按行切分，后半部分按列切分




- self-attention模块模型并行
  - 按照多头先进行并行化处理，后按照列进行切分
- loss，cross-entropy损失并行：
  - logits和label并行：https://zhuanlan.zhihu.com/p/497672789
  - torch支持：https://github.com/kaiyuyue/torchshard/blob/main/torchshard/nn/functional.py#L40
- 参考资料
  - 知乎讲解：https://zhuanlan.zhihu.com/p/650234985

- scalling low相关知识
  - https://arxiv.org/pdf/2403.06563.pdf
  - 调整bath size的目的
    - 找到训练时间与计算机资源的最佳平衡点

        - 根据loss来挑选一个可以使得时间和资源相对最优的batch size
        - 使用最大的batch size来减少训练步数
    - 参考资料
        - https://zhuanlan.zhihu.com/p/671327709

- llm评估
    - https://github.com/Eladlev/lm-evaluation-harness