昇思 MindSpore

# 通过FFN近似计算实现Transformer网络训练加速

# 目录》
# CONTENTS


# 项目需求

# 项目需求

【背景描述】

Transformer是功能强大的神经网络模型。训练Transformer模型通常需要耗费大量的时间和资源。Transformer的计算主要来自attention层和FFN层，目前已经有不少文献专注于对这两部分进行优化或近似计算来提升模型的计算效率，从而降低训练成本。现在需要在GPU和昇腾上实现FFN的近似计算。

【需求描述】

完成FFN近似计算方案梳理；

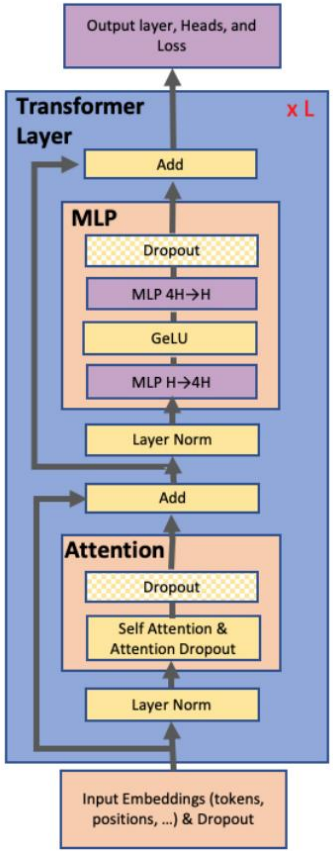
完成GPU/昇腾下代码开发，实现近似计算；

给出API接口，方便调用。

02
PART

# 项目实现

# Transformer结构



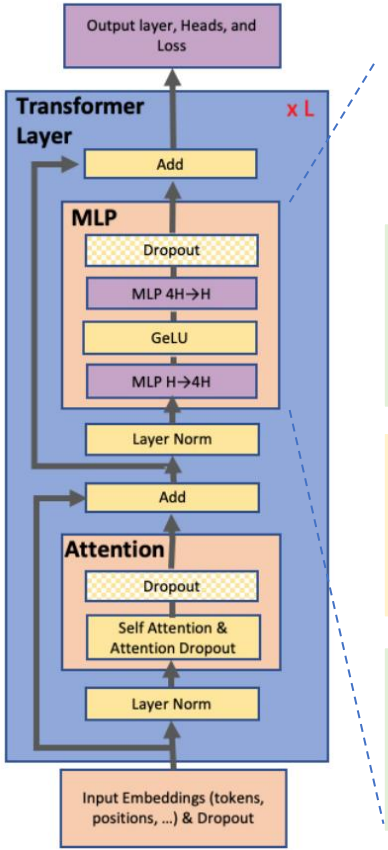

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2,$$

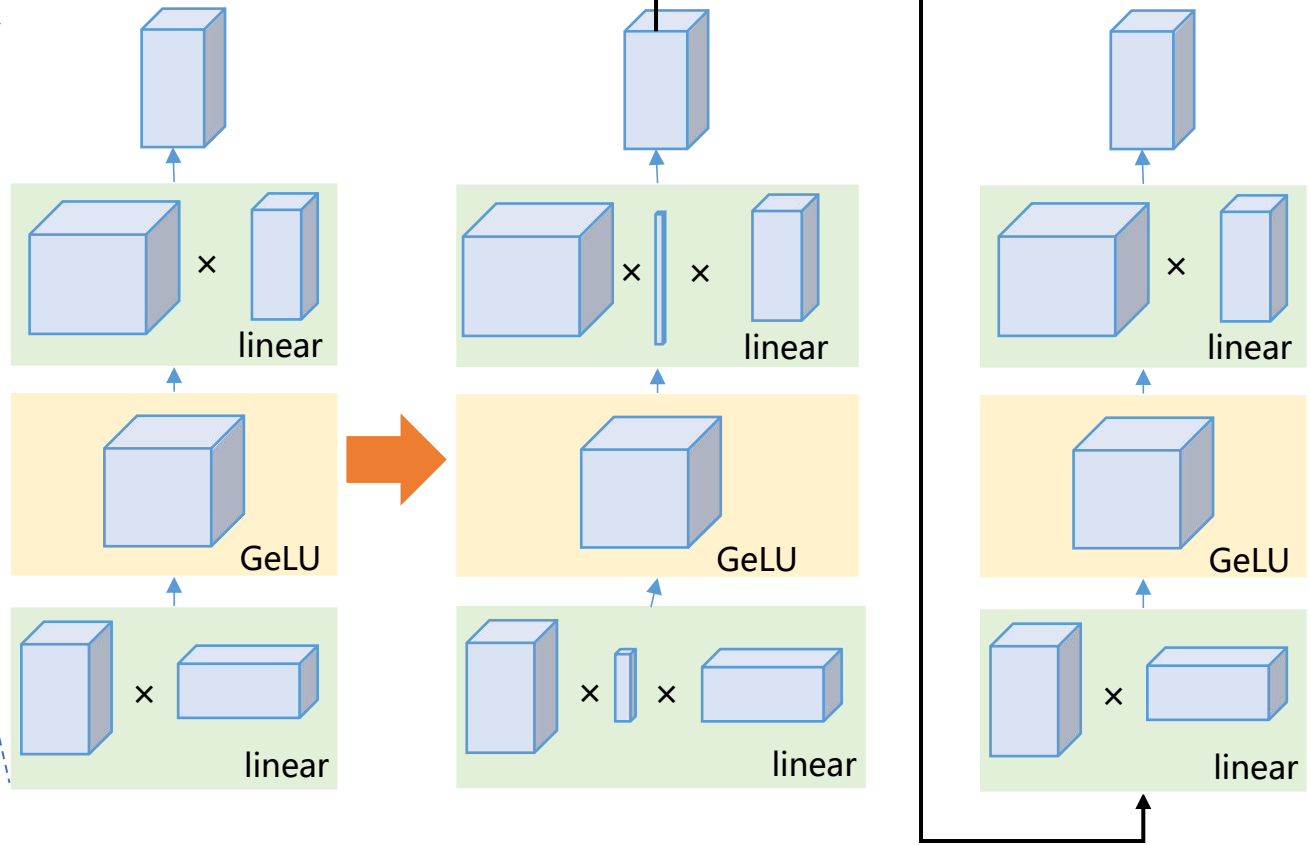
where $W_1 \in \mathrm{R}^{d \times d_f}, W_2 \in \mathrm{R}^{d_f \times d}.$

$$Att(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V,$$

where $Q \in \mathrm{R}^{n \times d_k}, K \in \mathrm{R}^{n \times d_k}, V \in \mathrm{R}^{n \times d_v}$

# FFN近似化加速




普通FFN(MLP)

快速FFN(MLP) 1阶段

快速FFN(MLP) 2阶段

# 软件架构

```
.
├─ research
   └─ FastFFN
      ├─ examples
         ├─ pretrain
            ├─ pretrain_gptfast.sh #可执行脚本文件
            └─ pretrain_gptfast_post.sh #可执行脚本文件
         └─ weight_transform.py #模型转换文件
         ├─ transformer
            ├─ model
               ├─ CoRe_Transformer.py
               ├─ gptfast.py
               ├─ layers.py
               ├─ loss.py
               ├─ moe.py
               ├─ op_parallel_config.py
            ├─ gptfast_traner.py
            └─ gptfast_traner_post.py
```


# 实验

## 实验平台

| 环境 | 依赖 | 配置/版本 |
| --- | --- | --- |
| 硬件环境 | GPU类型 | NVIDIA V100 |
| 软件环境 | 操作系统 | Ubuntu 18.04.6 |
| | Python | 3.7.10 |
| | mindspore-gpu | 1.8.1 |

# 实验配置

| 超参数 | | 数值 |
| --- | --- | --- |
| model | batchsize | 16 |
| | seq_length | 1024 |
| | vocab_size | 50257 |
| | hidden_size | 768 |
| | num_layers | 6 |
| | num_heads | 12 |
| baseline epoch size | | 1 |
| fast ffn第一阶段 epoch size | | 1 |
| fast ffn第二阶段 epoch size | | 1 |
| start_lr | | 5e-6 |
| end_lr | | 1e-7 |
| warmup_step | | 2000 |

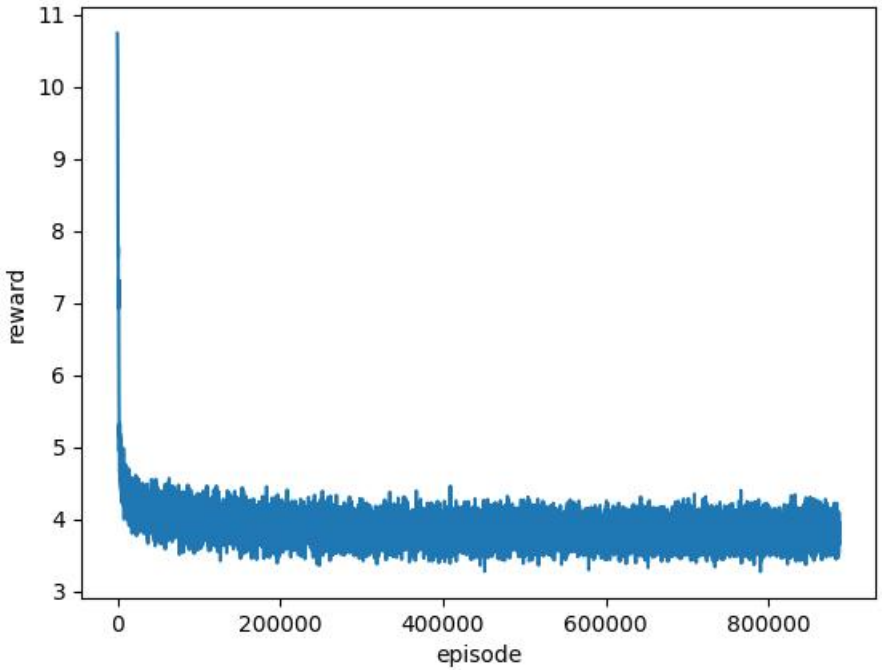
## 数据集

上游数据集：OpenWebText

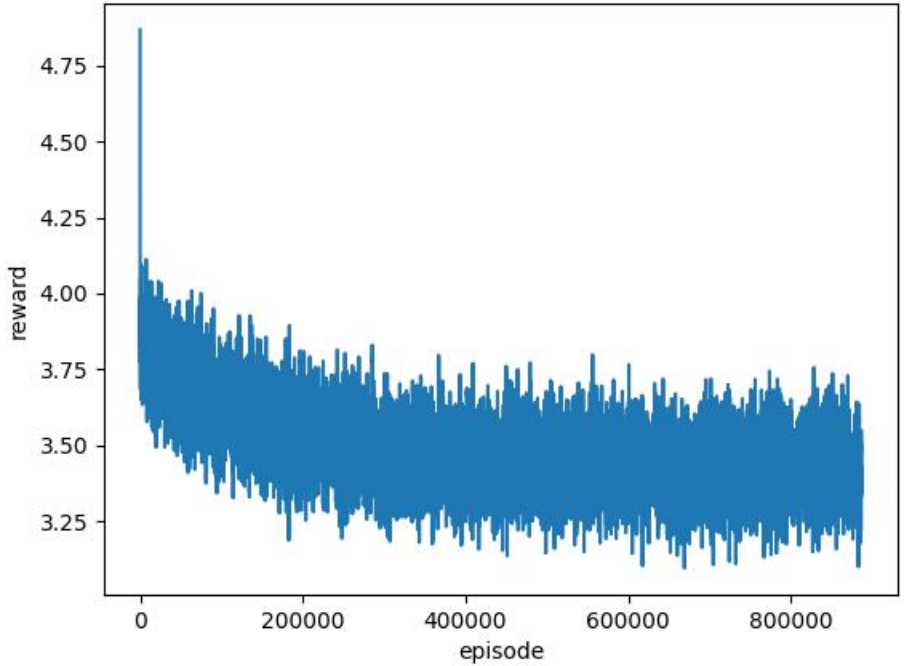
下游数据集：wiki-2

下游任务：language model

# 预训练结果




FFN近似化一阶段




FFN近似化二阶段

# 下游任务实验结果

| 任务 | Average Loss | PPL Average Loss |
| --- | --- | --- |
| 无预训练权重 | 6.657122 | 778.307782 |
| Fast FFN | 4.392275 | 80.824075 |

加速阶段加速：10.7%