01

基于Transformers的NLP解决方案

（1）基础组件知识回顾

（2）基于Transformers的NLP解决方案

（3）显存优化策略，4G显存跑BERT-Large

基础组件知识回顾

截止目前讲的基础组件

- **Pipeline**
  - 流水线，用于模型推理，封装了完整的推理逻辑，包括数据预处理、模型预测及后处理
- **Tokenizer**
  - 分词器，用于数据预处理，将原始文本输入转换为模型的输入，包括input_ids、attention_mask等
- **Model**
  - 模型，用于加载、创建、保存模型，对Pytorch中的模型进行了封装，同时更好的支持预训练模型
- **Datasets**
  - 数据集，用于数据集加载与预处理，支持加载在线与本地的数据集，提供了数据集层面的处理方法
- **Evaluate**
  - 评估函数，用于对模型的结果进行评估，支持多种任务的评估函数
- **Trainer**
  - 训练器，用于模型训练、评估，支持丰富的配置选项，快速启动模型训练流程

基于Transformers的NLP解决方案

- **以文本分类为例**
  - Step1 导入相关包 > General
  - Step2 加载数据集 > Datasets
  - Step3 数据集划分 > Datastes
  - Step4 数据集预处理 > Tokenizer + Datasets
  - Step5 创建模型 > Model
  - Step6 设置评估函数 > Evaluate
  - Step7 配置训练参数 > TrainingArguments
  - Step8 创建训练器 > Trainer + Data Collator
  - Step9 模型训练、评估、预测（数据集） > Trainer
  - Step10 模型预测（单条） > Pipeline

Transformers显存优化

显存优化策略，4G显存也能跑BERT-Large

- **显存占用简单分析**
  - 模型权重
    - 4Bytes * 模型参数量
  - 优化器状态
    - 8Bytes * 模型参数量，对于常用的AdamW优化器而言
  - 梯度
    - 4Bytes * 模型参数量
  - 前向激活值
    - 取决于序列长度、隐层维度、Batch大小等多个因素

Transformers显存优化

显存优化策略，4G显存也能跑BERT-Large

- 显存优化策略
  - hfl/chinese-macbert-large，330M

优化策略	优化对象	显存占用	训练时间
Baseline (BS 32, MaxLength 128)	-	15.2G	64s
+ Gradient Accumulation (BS 1, GA 32)	前向激活值	7.4G	259s
+ Gradient Checkpoints (BS 1, GA 32)	前向激活值	7.2G	422s
+ Adafactor Optiomizer (BS 1, GA 32)	优化器状态	5.0G	406s
+ Freeze Model (BS 1, GA 32)	前向激活值 / 梯度	3.5G	178s
+ Data Length (BS 1, GA 32, MaxLength 32)	前向激活值	3.4G	126s

关于参数高效微调（如Lora）、cpu offload、flash attention等技巧将在LLM章节进行讲解