

Mini-BioBERT 项目总结报告

作者: Claire Ye CAO

日期: [2025-08-25]

1. 项目背景

- 目标：从零实现一个简化版的 **BioBERT**，掌握 BERT 的内部机制（编码器、注意力机制、预训练任务）。
 - 方法：
 - 手写 `bert_model.py`（包含 `BertConfig`，`MiniBertModel`，`BertForPretraining`）。
 - 使用医学语料（PubMed abstracts）继续预训练（MLM + NSP）。
 - 未来可在下游任务（NER、QA、分类）上验证效果。
-

2. 模型实现

- 架构：Mini-BioBERT
 - Transformer Encoder：4 层
 - Hidden size：256
 - Attention heads：4
 - Feed-forward size：1024
 - Max sequence length：128
 - 预训练任务：
 - **Masked Language Modeling (MLM)**：随机遮盖 15% 的 token
 - **Next Sentence Prediction (NSP)**：预测句子对是否相邻
-

3. 数据准备

- 语料来源：PubMed 文献摘要（[ccdv/pubmed-summarization · Datasets at Hugging Face](#)）
- 规模：采样约 **10 万条摘要**（写入 `pubmed_corpus.txt`）

- 预处理：
 - 切分成句子对，构造 NSP 正负样本
 - 使用 `BertTokenizer` 分词并生成 MLM mask
-

4. 预训练实验

实验配置

- 训练脚本： `pretrain_pubmed.py`
- 语料： `pubmed_corpus.txt`，约 100k 摘要
- Epochs: 5
- Batch size: 32
- Optimizer: AdamW (lr=5e-5)
- 学习率调度: Warmup + Linear decay

日志与结果

- 初始总损失: ≈ 10
 - 训练过程：
 - 在 10 万条摘要上训练 5 个 epoch
 - 总损失逐渐下降至 ≈ 6.5
 - 解读：
 - 模型确实在学习，但当前困惑度仍然偏高
 - 由于语料规模和模型规模有限，尚未达到收敛
-

5. 遇到的问题与限制

- HuggingFace `ccdv/pubmed-summarization` 数据集在 `datasets` 新版本中不兼容，只能通过脚本生成语料。
 - 训练在 CPU 上运行，速度慢，loss 曲线收敛缓慢。
 - 当前模型为“Mini-BERT”(层数小)，收敛速度有限，效果不如完整 BioBERT。
 - 训练步数不足，5 epoch 尚处于 early stage。
-

6. 下一步工作建议

1. 继续大规模预训练

- 使用 `pretrain_pubmed.py` 在 GPU/多卡上训练更长时间。

2. 调参优化

- 尝试更合适的学习率（目前 $5e-5$ 可改成 $1e-4$ / $3e-4$ ）。
- 增加训练步数和 batch size。
- 使用学习率调度（linear warmup）。







3. 下游验证

- 选择一个医学任务（NER: BC5CDR, QA: BioASQ）微调并测试效果。
- 对比：随机初始化 vs 预训练模型。

4. 手写代码

- 当前代码主体由 AI 生成，后续重新手动编写关键模块，以加深对模型底层实现的理解与掌握。

7. 当前成果总结

-  从零实现了一个 Mini-BERT 架构。
-  完成了 MLM+NSP 预训练任务实现。
-  在小规模语料上验证 loss 可以下降（10 → 7）。
-  大规模预训练尚未完成，需要更多算力和时间。
-  下游任务验证尚未开始。
-  熟悉了模型搭建和训练流程，但是代码功底不扎实，需要手写锻炼。