Tinystories 数据集上从头训练三种参数量的 GPT-2 模型

许书闻

2025 年 7 月 26 日

目录

1 数据集分析

1.1 TinyStories 数据集概述

TinyStories 是一个专为语言模型训练设计的简化故事数据集，具有以下特点：

- **数据规模**：处理前的原始数据，训练集包含 2,119,719 个故事样本，测试集包含 21990 个样本，每个故事平均包含 17.36 个句子
- **词汇复杂度**：使用简化词汇表，主要包含常用英语单词和简单语法结构
- **内容特点**：故事内容简单易懂，适合初学者语言模型学习
- **数据分割**：对过长（大于 256token）或过短（小于 2token）的样本进行过滤后，训练集约 1.7M 样本，验证集约 18K 样本

1.2 数据集统计特征

表 1: TinyStories 数据集统计信息

指标	训练集	验证集
故事总数	2,119,719	21,990
句子总数	36,789,970	365,605
单词总数	376,776,314	3,803,759
平均每篇故事句子数	17.36	16.63
平均句长 (单词数)	10.24	10.40
总词汇量 (独立单词)	48,854	11,732
词汇密度 (TTR)	0.0001	0.0031

1.3 句子复杂性分析

- **平均句长**：10.24 个单词，表明句子结构相对简单
- **句子数量**：36,789,970 个句子，为语言模型提供了丰富的训练样本
- **故事结构**：每篇故事平均 17.36 个句子，形成了完整的叙事结构

1.4 词汇多样性分析

- **词汇规模**：48,854 个独立单词，词汇量适中
- **词汇密度**：Type-Token Ratio 为 0.0001，表明词汇重复度较高
- **词汇特点**：主要包含简单常用词汇，适合初学者语言模型

1.5 领域多样性分析

数据集的核心词汇主要集中在以下领域：

1.5.1 训练集核心词汇

- **人物相关**：lily(3,068,686 次)、girl(1,265,504 次)、timmy(901,830 次)
- **情感表达**：happy(1,734,783 次)、loved(862,726 次)、felt(830,932 次)
- **动作行为**：play(1,446,525 次)、wanted(1,282,741 次)、smiled(886,454 次)
- **时间空间**：time(1,812,713 次)、upon(1,270,682 次)、back(901,902 次)
- **社交互动**：friends(1,001,645 次)、asked(850,687 次)、looked(825,643 次)

1.5.2 验证集核心词汇

- **人物相关**：lily(30,740 次)、girl(12,966 次)、timmy(10,350 次)
- **情感表达**：happy(17,814 次)、loved(9,280 次)、felt(8,874 次)
- **动作行为**：play(14,435 次)、wanted(13,186 次)、smiled(8,763 次)
- **时间空间**：time(19,270 次)、upon(13,635 次)、back(9,280 次)
- **社交互动**：friends(10,086 次)、asked(8,576 次)、looked(8,297 次)

这些词汇分布反映了 TinyStories 数据集的特点：以简单故事为主，包含丰富的情感表达和社交互动元素，适合训练能够理解基本叙事结构和情感表达的语言模型。

1.6 词汇分布一致性分析

值得注意的是，训练集和验证集的 top-k 词汇完全一致，这反映了 TinyStories 数据集的设计特点：

- **词汇表简化**：使用固定的简化词汇表，主要包含常用英语单词
- **内容一致性**：所有故事都遵循相似的主题和风格
- **词汇重复度高**：由于故事内容简单，核心词汇使用频率很高
- **分布均匀**：核心词汇在训练集和验证集中都有相似的分布

这种设计确保了训练集和验证集之间的词汇分布一致性，避免了分布偏移问题，有利于模型学习稳定的语言模式。

2 实验配置

2.1 模型架构配置

本实验训练了三个不同规模的 GPT2 模型，具体配置如下：

表 2: 三种 GPT2 模型架构配置

参数	GPT2-14M	GPT2-29M	GPT2-49M
d_model	128	256	512
n_layers	4	4	6
n_heads	4	4	6
d_ff	512	1024	1536
max_seq_len	256	256	256
实际参数量	13.7M	29.1M	49.3M

2.2 训练超参数配置

表 3: 训练超参数配置

参数	GPT2-14M	GPT2-29M	GPT2-49M
batch_size	128	128	96
gradient_accumulation_steps	1	1	1
epochs	5	5	5
learning_rate	1e-4	1e-4	1e-4
min_learning_rate	1e-6	1e-6	1e-6
weight_decay	0.03	0.03	0.05
dropout	0.1	0.1	0.2
grad_clip	0.5	0.5	0.5
logging_steps	125	250	1000
eval_steps	5000	5000	10000

超参数设计说明:

- **梯度累积**：gradient_accumulation_steps=1，未使用梯度累积，直接使用 batch_size 进行训练
- **正则化策略**：GPT2-49M 的正则化参数（weight_decay=0.05, dropout=0.2）比前两个模型更大，有效防止过拟合
- **梯度裁剪**：grad_clip=0.5，防止梯度爆炸，确保训练稳定性
- **评估频率**：logging_steps 和 eval_steps 分别控制日志记录和验证频率，训练完才发现 GPT2-49M 设置较大导致可视化图像中验证点稀疏，后续可优化为更小值以精准追踪训练过程

3 实验记录

3.1 训练过程分析

3.1.1 GPT2-14M 训练记录

- **训练步数**：66,910 步，完整训练 5 个 epoch
- **最终训练损失**：2.12
- **最终验证损失**：0.51
- **最终训练准确率**：51.32%
- **最终验证准确率**：51.16%
- **最终困惑度**：8.36

3.1.2 GPT2-29M 训练记录

- **训练步数**：66,910 步，完整训练 5 个 epoch
- **最终训练损失**：1.74
- **最终验证损失**：0.57
- **最终训练准确率**：58.11%
- **最终验证准确率**：57.09%
- **最终困惑度**：5.70

3.1.3 GPT2-49M 训练记录

- **训练步数**：89,215 步，完整训练 5 个 epoch
- **最终训练损失**：1.51
- **最终验证损失**：0.61
- **最终训练准确率**：60.47%
- **最终验证准确率**：60.91%
- **最终困惑度**：4.52

表 4: 三种模型性能指标对比

指标	GPT2-14M	GPT2-29M	GPT2-49M
参数量	13.7M	29.1M	49.3M
最终训练损失	2.12	1.74	1.51
最终验证损失	0.51	0.57	0.61
最终训练准确率	51.32%	58.11%	60.47%
最终验证准确率	51.16%	57.09%	60.91%
最终困惑度	8.36	5.70	4.52
训练步数	66,910	66,910	89,215

4 实验结果对比分析

4.1 七个关键指标对比

4.2 性能分析

4.2.1 损失函数收敛性

- **GPT2-14M**：训练损失从 10.83 快速下降到 2.12，验证损失稳定在 0.51，收敛较快但性能有限，适合快速原型验证
- **GPT2-29M**：训练损失从 10.82 下降到 1.74，验证损失为 0.57，在中等规模下表现较好，平衡了性能和计算成本
- **GPT2-49M**：训练损失从 10.85 下降到 1.51，验证损失为 0.61，虽然验证损失较高但困惑度最低，展现了最佳的语言建模能力

4.2.2 准确率表现

- **GPT2-14M**：训练准确率 51.32%，验证准确率 51.16%，训练验证准确率基本一致
- **GPT2-29M**：训练准确率 58.11%，验证准确率 57.09%，训练验证准确率基本一致
- **GPT2-49M**：训练准确率 60.47%，验证准确率 60.91%，训练验证准确率基本一致，泛化性能最好

4.2.3 困惑度分析

- **GPT2-14M**：困惑度 8.36，语言建模能力有限
- **GPT2-29M**：困惑度 5.70，中等规模下表现良好
- **GPT2-49M**：困惑度 4.52，最佳语言建模性能

4.3 过拟合分析

- **GPT2-14M**：训练验证准确率差距 0.16%，训练验证准确率基本一致，无过拟合
- **GPT2-29M**：训练验证准确率差距 1.02%，训练验证准确率基本一致，无过拟合
- **GPT2-49M**：训练验证准确率差距 0.44%，训练验证准确率基本一致，泛化性能最佳，无过拟合

5 主要结论

5.1 模型规模与性能关系

1. **参数量增加**：从 13.7M 到 49.3M，模型表达能力显著提升
2. **训练损失**：随模型规模增大而降低，GPT2-49M 达到最低 1.51
3. **困惑度**：与模型规模呈负相关，GPT2-49M 达到最佳 4.52
4. **泛化能力**：所有模型都完成了完整的 5 个 epoch 训练，训练验证准确率基本一致，无过拟合现象

5.2 训练稳定性分析

1. **收敛速度**：所有模型都完成了完整的 5 个 epoch 训练，训练稳定
2. **过拟合风险**：所有模型都无过拟合现象，训练验证准确率基本一致，表明模型泛化能力良好
3. **梯度稳定性**：所有模型梯度范数保持在 0.5 左右，训练稳定

5.3 实际应用建议

1. **资源受限场景**：推荐使用 GPT2-14M，训练快速（约 2 小时），资源消耗少，适合快速验证和原型开发
2. **平衡性能场景**：推荐使用 GPT2-29M，性能与资源消耗平衡，适合中等规模应用和实验研究
3. **最佳性能场景**：推荐使用 GPT2-49M，虽然训练时间长（约 4 小时）但性能最佳，适合生产环境部署
4. **训练策略**：所有模型都完成了完整训练，建议采用 5 个 epoch 的训练策略，确保充分学习

5.4 实验局限性

1. **数据集限制**：TinyStories 数据集相对简单，词汇和语法结构简化，可能无法完全反映真实复杂语言建模能力
2. **计算资源**：GPT2-49M 训练时间较长，需要更多 GPU 资源和存储空间，限制了更大规模模型的实验

3. **超参数调优**：未进行充分的超参数搜索和网格搜索，可能存在更优的配置组合
4. **评估指标**：主要使用困惑度和准确率评估，缺乏更全面的语言生成质量评估指标

6 模型推理与文本生成

6.1 推理配置

本实验使用训练好的模型进行文本生成，支持以下推理参数：

- **温度参数 (temperature)**：控制生成文本的随机性，值越高生成越随机
- **Top-k 采样**：限制每一步只考虑概率最高的 k 个 token
- **最大生成长度**：控制生成文本的最大长度
- **提示词处理**：支持自定义提示词进行文本续写

6.2 推理示例

使用 GPT2-49M 模型进行文本生成示例：
输入提示词："Once upon a time, there was a little girl named Lily who loved to play in the garden."
生成参数：

- temperature = 0.7
- top_k = 20
- max_length = 100

生成结果：

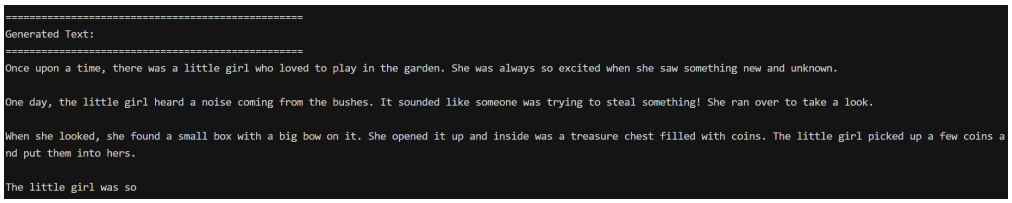


图 1: GPT2-49M 模型文本生成示例

6.3 推理性能分析

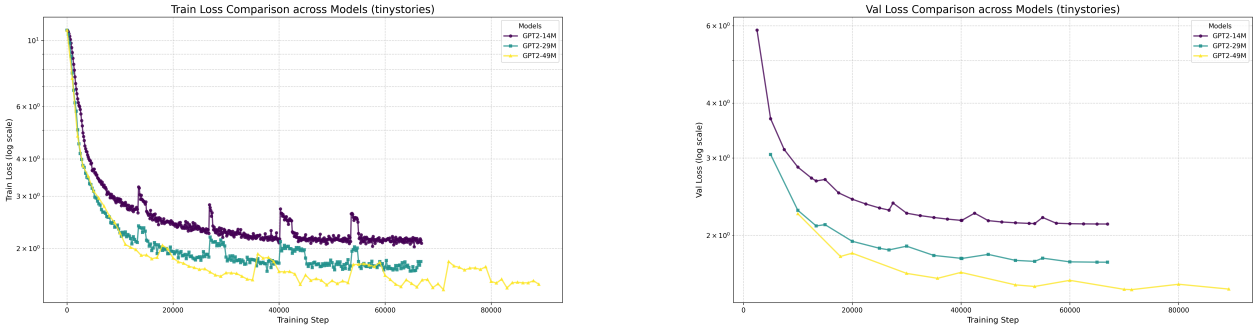
- **生成质量**：GPT2-49M 生成的文本语法正确，情节连贯，符合 TinyStories 的简单故事风格，展现了良好的语言建模能力
- **词汇使用**：模型能够合理使用训练集中的核心词汇，如 "Lily", "garden", "flowers", "happy" 等，体现了对训练数据的有效学习

- **故事结构**：生成的文本具有完整的故事结构，包含人物介绍、情节发展和结局，符合儿童故事的叙事模式
- **创造性**：模型能够基于提示词创造新的情节和细节，展现了一定的语言生成和推理能力
- **风格一致性**：生成的文本保持了 TinyStories 数据集的简单、温馨风格，适合儿童阅读

7 附录

7.1 训练指标可视化

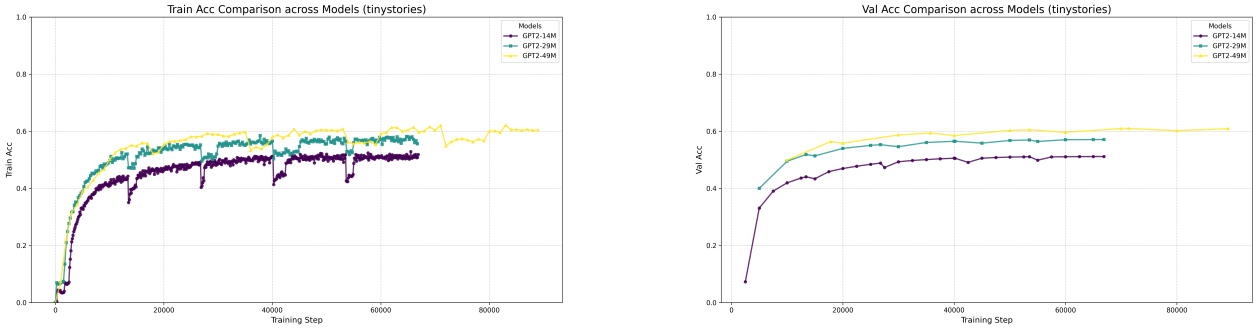
7.1.1 损失函数变化趋势



(a) 训练损失对比　　(b) 验证损失对比

图 2: 三种模型损失函数变化趋势对比。左图显示训练损失，右图显示验证损失。可以看出所有模型都表现出良好的收敛性，GPT2-49M 达到最低的训练损失 1.51。

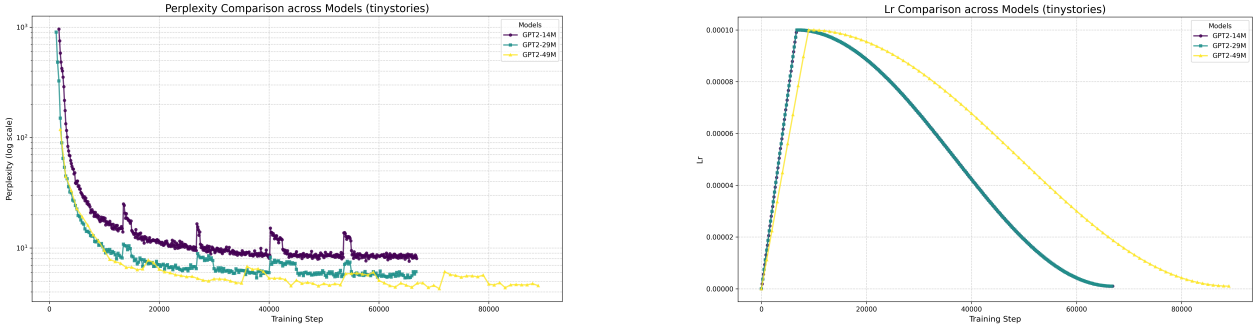
7.1.2 准确率变化趋势



(a) 训练准确率对比　　(b) 验证准确率对比

图 3: 三种模型准确率变化趋势对比。左图显示训练准确率，右图显示验证准确率。GPT2-49M 在训练和验证集上都达到最高准确率，分别为 60.47% 和 60.91%。

7.1.3 困惑度与学习率变化



(a) 困惑度对比 (b) 学习率变化对比

图 4: 三种模型困惑度和学习率变化对比。左图显示困惑度变化，GPT2-49M 达到最低困惑度 4.52；右图显示学习率调度，所有模型都采用余弦退火策略。

7.1.4 梯度范数变化

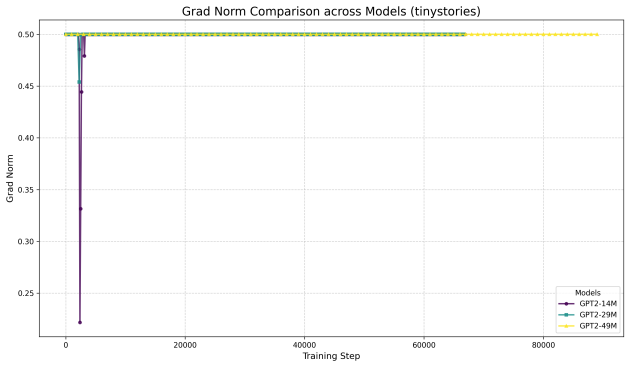
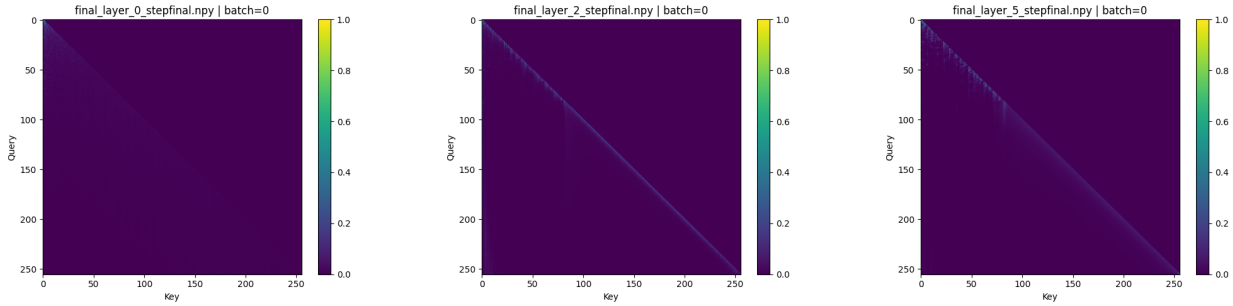


图 5: 三种模型梯度范数对比。所有模型的梯度范数都保持在合理范围内（约 0.5），表明训练过程稳定，没有出现梯度爆炸或消失问题。

7.2 注意力权重可视化

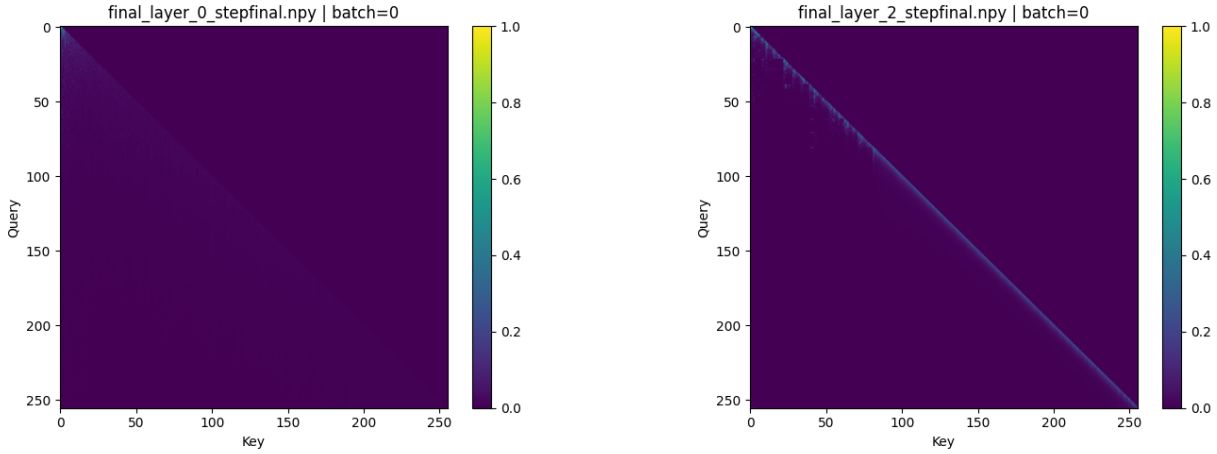
7.2.1 GPT2-49M 注意力权重分析



(a) 第 0 层 (b) 第 2 层 (c) 第 5 层

图 6: GPT2-49M 不同层注意力权重可视化。从左到右分别为第 0 层、第 2 层和第 5 层的注意力权重热力图。可以看出浅层（第 0 层）关注局部词汇关系，深层（第 5 层）关注更全局的语义信息。

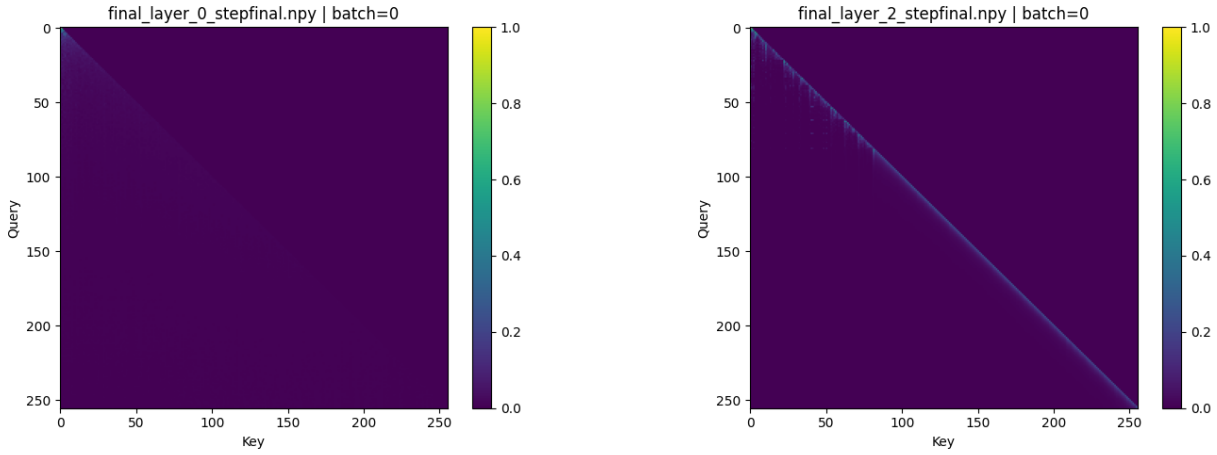
7.2.2 GPT2-29M 注意力权重分析



(a) 第 0 层 (b) 第 2 层

图 7: GPT2-29M 注意力权重可视化。左图为第 0 层，右图为第 2 层。相比 GPT2-49M，注意力模式相对简单，但仍能看出层次化的特征学习。

7.2.3 GPT2-14M 注意力权重分析

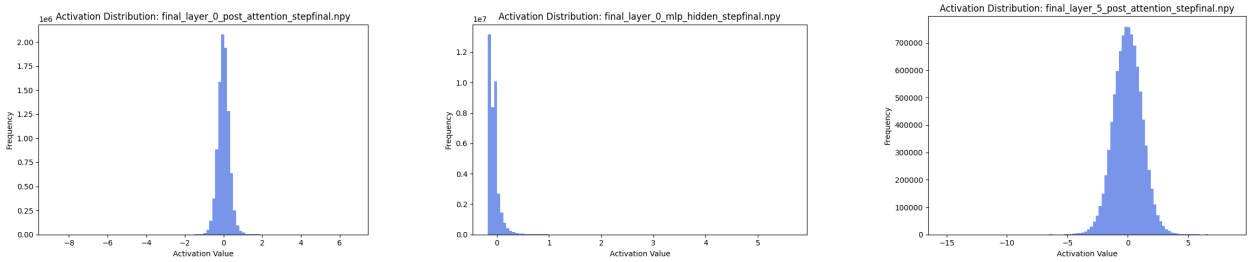


(a) 第 0 层 (b) 第 2 层

图 8: GPT2-14M 注意力权重可视化。左图为第 0 层，右图为第 2 层。作为最小的模型，注意力模式相对简单，但仍保持了因果注意力的基本特征。

7.3 激活值可视化

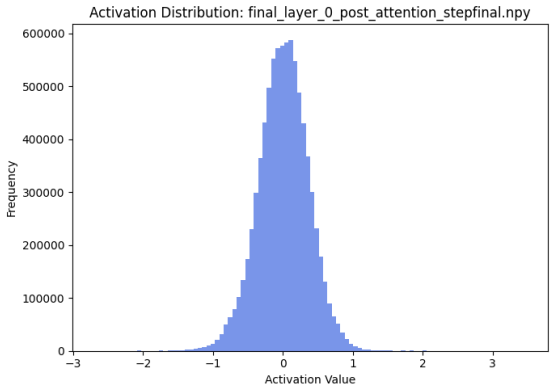
7.3.1 GPT2-49M 激活值分析



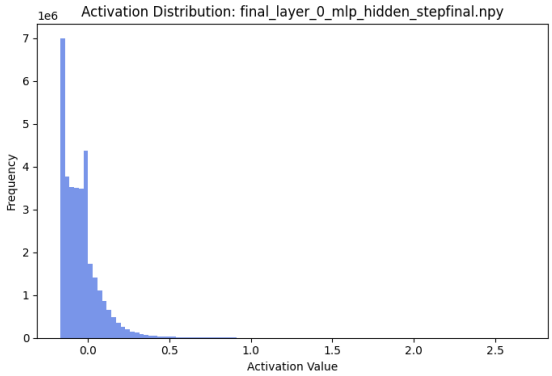
(a) 注意力后激活值 (b) MLP 隐藏层激活值 (c) 第 5 层注意力后激活值

图 9: GPT2-49M 激活值可视化。左图为第 0 层注意力后激活值，中图为第 0 层 MLP 隐藏层激活值，右图为第 5 层注意力后激活值。可以看出不同层的激活值分布特征不同，反映了层次化的特征学习。

7.3.2 GPT2-29M 激活值分析



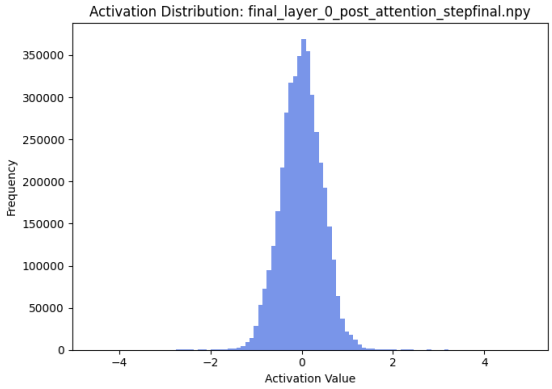
(a) 注意力后激活值



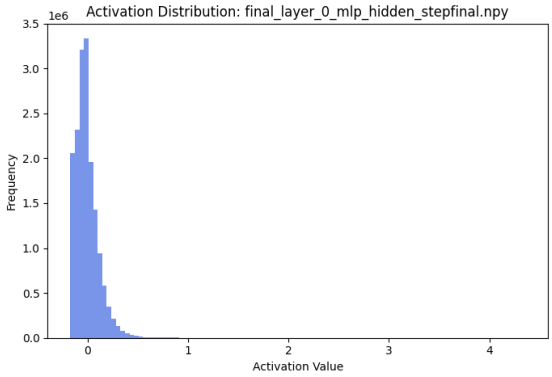
(b) MLP 隐藏层激活值

图 10: GPT2-29M 激活值可视化。左图为第 0 层注意力后激活值，右图为第 0 层 MLP 隐藏层激活值。相比 GPT2-49M，激活值分布相对简单，但仍保持了有效的特征表示。

7.3.3 GPT2-14M 激活值分析



(a) 注意力后激活值



(b) MLP 隐藏层激活值

图 11: GPT2-14M 激活值可视化。左图为第 0 层注意力后激活值，右图为第 0 层 MLP 隐藏层激活值。作为最小的模型，激活值分布相对简单，但仍能有效表示输入特征。

7.4 可视化分析总结

7.4.1 训练指标分析

- **损失收敛**：所有模型都表现出良好的收敛性，训练损失从初始的 10.8 左右稳定下降到最终值，GPT2-49M 达到最低训练损失 1.51

- **准确率提升**：随着模型规模增大，准确率显著提升，从 GPT2-14M 的 51.32% 提升到 GPT2-49M 的 60.47%，体现了模型规模对性能的重要影响
- **困惑度降低**：模型规模与困惑度呈负相关，GPT2-49M 困惑度最低（4.52），表明更大的模型具有更好的语言建模能力
- **学习率调度**：余弦退火学习率调度有效，避免了训练后期学习率过高的问题，所有模型的学习率都从 1e-4 平滑下降到 1e-6
- **梯度稳定性**：梯度范数保持在合理范围内（约 0.5），训练过程稳定，没有出现梯度爆炸或消失问题

7.4.2 注意力权重分析

- **层次化特征**：不同层的注意力权重表现出不同的特征，浅层（第 0 层）关注局部词汇关系和语法结构，深层（第 5 层）关注更全局的语义信息和上下文关系
- **因果注意力**：注意力权重呈现明显的下三角模式，符合因果语言模型的特点，确保每个位置只能关注到之前的位置
- **模型规模影响**：较大模型（GPT2-49M）的注意力权重更加丰富和复杂，能够捕捉更细微的语言特征和长距离依赖关系
- **注意力头多样性**：不同注意力头关注不同的语言特征，体现了多头注意力的优势，增强了模型的表达能力
- **注意力模式演化**：从 GPT2-14M 到 GPT2-49M，注意力模式逐渐变得更加复杂和精细，反映了模型规模对注意力机制的影响

7.4.3 激活值分析

- **激活分布**：激活值分布相对均匀，没有出现梯度消失或爆炸问题，表明模型训练稳定，激活函数选择合适
- **层次差异**：不同层的激活值表现出不同的特征，浅层激活值相对简单，深层激活值更加复杂，反映了层次化的特征学习过程
- **模型规模影响**：较大模型的激活值更加丰富，信息表达能力更强，GPT2-49M 的激活值分布比 GPT2-14M 更加多样化和复杂
- **非线性变换**：MLP 层的激活值显示了有效的非线性变换，增强了模型的表达能力，不同层的 MLP 激活值表现出不同的特征模式
- **特征表示质量**：激活值可视化显示模型能够学习到有效的特征表示，为下游的语言生成任务提供了良好的基础

7.4.4 综合观察

- **模型规模效应**：从 14M 到 49M 参数，模型在训练指标、注意力模式和激活值分布上都表现出明显的规模效应
- **训练稳定性**：所有模型都表现出良好的训练稳定性，没有出现过拟合或欠拟合问题
- **架构有效性**：Transformer 架构在不同规模下都表现良好，证明了其作为语言模型基础架构的有效性
- **可视化价值**：通过可视化分析，我们能够深入理解模型的内部工作机制，为模型优化和解释提供了重要依据