scRNA-seq pretrained model 进展

动机

sly的数据包括多物种的脑区单细胞多组学数据，当前的分析手段选择使用同源基因，因此丢失大量信息。如果能够学习基因embedding，并且验证这些基因embedding是有生物学意义的，就可以通过基因embedding的差异分析基因层面的回声非回声差异。

同时，小鼠数据有详细的参考数据集，allen_2021和allen_2023,通过在这些参考数据上pretrain后再迁移到自有数据上，可以提供更详尽的细胞注释。先前的实验中,allen_2023不能得到有效注释，或许transformer-based模型会提供不一样的结果。

研究问题

1. 能否通过transformer-based 模型学习到有意义的基因embedding? transformer-based模型能否有效解决参考基因组迁移问题?
2. 如果学习到的gene_embedding有意义，可以给出一般化的跨物种学习方法，无需同源基因比对，类似于LLM中词表扩增的做法，对于新的未训练过的物种，在现有模型下扩增基因词表后微调即可解决。
3. 学习到的gene_embedding是否具有组织特异性？同一物种下不同组织的gene_embedding存在怎样的差异。
4. 尝试全新的分词和训练任务，验证自己之前的想法。直接进行 基因名+raw_count 的编码方式。同时掩码预测任务做成raw_count的分类任务。
5. 引入对比学习的策略，改善细胞的表示效果。通过对基因进行抽样，使得全部基因都可以得到训练。
6. 考虑后续整合ATAC-seq数据，构建单细胞多模态模型。

当前进展

完成初步的模型搭建，在一个10K大小的数据集上跑通基本流程，loss能够有效下降，验证模型整体没有问题。当前训练的小模型参数量为57M，仅使用单张4090训练。粗略估计，可以在1h完成100K样本的训练，单卡训练1M数据1个epoch的时间估计在10h。在当前序列长度和数据量(小鼠脑区atalas数据，不超过5M)条件下，计算压力不大。

下面是训练loss:

```
epochs: 0
begin:================================================================================
====
Step 1, Loss: 6.2305
Step 2, Loss: 4.8934
Step 3, Loss: 4.1637
Step 4, Loss: 3.7007
Step 5, Loss: 3.2197
...
...
epochs: 9
begin:================================================================================
====
Step 1, Loss: 0.0060
Step 2, Loss: 0.0147
```

```
Step 3, Loss: 0.0075
Step 4, Loss: 0.0147
Step 5, Loss: 0.0314
```

后续工作

模型部分

- [ ] dataloader 部分，基因采样部分在每个batch内进行。
- [ ] 训练器中，加入参数初始化、优化率scheduler等策略。
- [ ] 整体训练脚本封装，通过config传入参数，支持快速迭代。
- [ ] 训练器中补充验证集部分。
- [ ] 解决seq-length问题，当前训练的序列长度仅有1k，更长的seq-length或许会捕捉到更多的基因互作关系。参考llm中的长文本方案以及使用更高级的计算显卡。
- [ ] 考察混合精度训练、flash-attention等训练方法，合适的话加入脚本提高训练效率。
- [ ] 考察多卡并行训练的方案。
- [ ] 筹备加入多物种数据后扩词表的训练方案。

实验分析部分

- [ ] 先使用现有方案，尽快拿到allen_2021数据的结果和自有小鼠脑区数据的结果并分析。
- [ ] 如果现有方案可行，在当前基础上进行跨物种脑区数据分析。
- [ ] 考虑模型解释性，gene_embedding, count_embedding, attention_score是否与某些现象相关。
- [ ] 准备其他预训练模型做对比评估。

推进部分

先完成可行性部分的实验。模型部分先解决前四条和最后一条，实验分析部分先完成第一条和第三条。确定基本结果可用后再考虑更复杂的模型设计和应用。