组会汇报

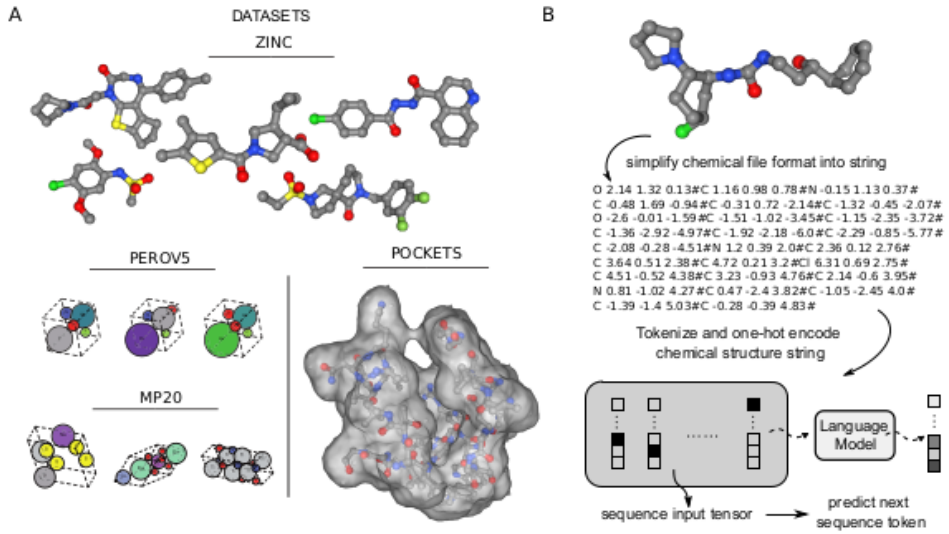
陈钶杰
专业:计算数学

August 22, 2023

目录

文章思想方法

这篇论文介绍了一种新的方法，即使用语言模型进行分子结构的生成。通过特定方法将化学分子的特征提取出来变成序列，然后对分子的所有操作都可以通过对序列变换来完成，文章中要做的分子合成，对序列操作就是进行序列预测。本文通过用语言模型的预测和一些专门用于化学分子序列预测的模型进行比对，来得到结论，模型的训练任务是预测由处理化学文件格式（XYZ、CIF或PDB）生成的序列中的下一个标记。



具体建模

- 化学分子经过特征提取后的序列是包含它的三维空间结构信息的，因此序列的各个元素之间都有一定的关联性，而使用LSTM的化学语言模型不可避免地会遇到学习重要的远距离依赖性的问题。然而，像Transformer这样的其他架构一次处理整个序列，并且不会出现这个问题，因此它们是最有可能成功完成这个任务的模型。
- 模型的训练任务是预测由处理化学文件格式（XYZ、CIF或PDB）生成的序列中的下一个标记。本文用了两种不同的标记化策略，一种是字符级别标记化(LHCH)，，原子+坐标级别标记化(LH-AC)例如CH4两种标记方式分别是C H H H 和C: (0.0, 0.0, 0.0) H: (0.0, 0.0, 1.0) H: (0.0, 1.0, 0.0) H: (1.0, 0.0, 0.0) H: (0.0, -1.0, 0.0) 。

TABLE I. Generation performance for ZINC.

3D	Model	Basic Metrics (%) ↑			WA Metrics ↓		
		Valid	Unique	Novel	MW	SA	QED
Not 3D	Train	100.0	100.0	100.0	0.816	0.013	0.002
	SMLM	98.35	100.0	100.0	3.640	0.049	0.005
	SFLM	100.0	100.0	100.0	3.772	0.085	0.006
	DGMG	79.63	100.0	99.38	88.94	3.163	0.095
	JTVAE	100.0	98.56	100.0	22.63	0.126	0.023
	CGVAE	100.0	100.0	100.0	45.61	0.426	0.038
3D	ENF	1.05	96.37	99.72	168.5	1.886	0.160
	GSchNet	1.20	55.96	98.33	152.7	1.126	0.185
	EDM	77.51	96.40	95.30	101.2	0.939	0.093
	LM-CH	90.13	100.0	100.0	3.912	2.608	0.077
	LM-AC	**98.51**	**100.0**	**100.0**	**1.811**	**0.026**	**0.004**

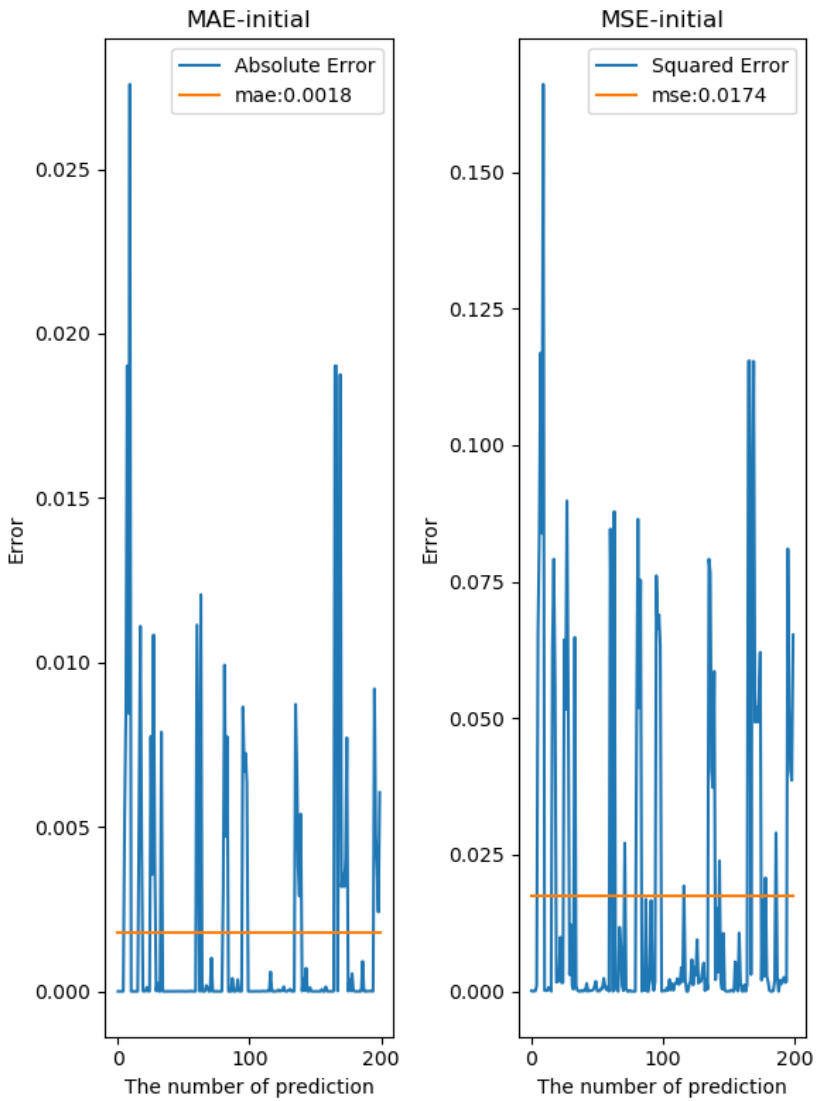
- 实验用到了在化学中使用的一些有三维特征提取能力等长短时记忆递归神经网络模型，以及两种标记化处理的语言模型。
- 结论：通过字符级别和坐标级别标记化的语言模型在性能上与使用

文章的相关启示

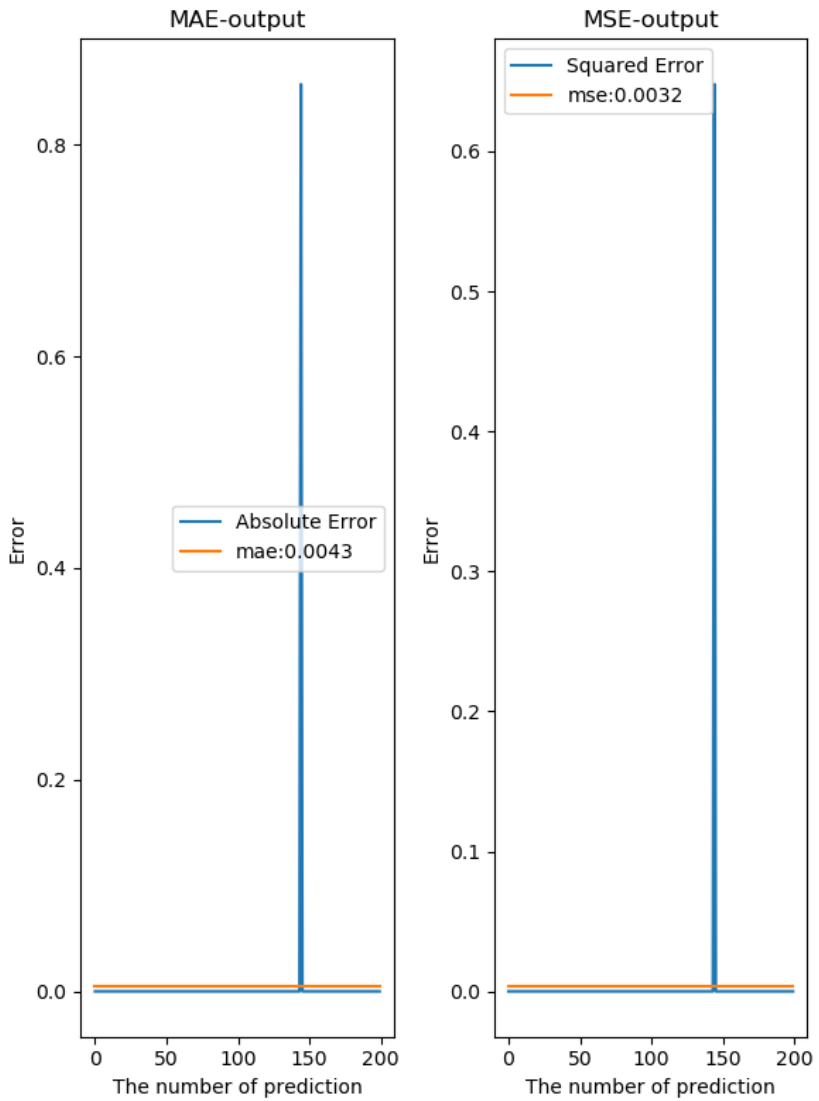
1. 语言模型是基于transformer架构的，所以能够完成序列预测的任务
2. 语言模型相比较LSTM模型，能够更好的捕捉全局信息，所以对序列之间较远元素的相关信息也能捕捉出。
3. 在进行预测的时候，语言模型的原子+坐标级别标记化的预测效果优于字符级别标记化说明语言模型能够提取字符数字混合的信息并进行预测。

RevIn的具体做法

1. 将数据集进行规范化
   将训练和测试的数据集进行规范化
2. 使用语言模型对规范化的数据集进行训练
3. 预测规范化后的数据结果
4. 将预测的规范化数据进行反规范化，来得到最终的结果(还没做完)



(a) 未用RevIN的测试结果，pass:100%



(b) RevIN的测试结果,pass:100%

实验结论分析

- MAE这个指标更加注重误差的绝对值，相对于MAE，MSE更加敏感，对较大误差的惩罚更重，因为误差被平方了。经过RevIN处理以后，绝对误差和之前差不多，但是均方根误差明显减小，说明通过使用RevIN以后，大大提高了预测的稳定性。

auto-cot方法


```
4 {"labels": "0.9982, 0.9991, 1.0000, 1.0009, 1.0018", "predict": "Sure, let's take a look at the
  -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.000
  1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000
  .0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000,
  0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000,
  000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -
  00, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1
  0, -1.0000, -1.0000,"}
5 {"labels": "0.9982, 0.9991, 1.0000, 1.0009, 1.0018", "predict": "Sure, let's start with the firs
  0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000,
  000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -
  00, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1
  0, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.
  , -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0
   -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.0000, -1.00
  -1.0000, -1."}
6 {"labels": "0.9982, 0.9991, 1.0000, 1.0009, 1.0018", "predict": "To make a prediction for the la
  owing steps:\n\n1.  If the current element is -1, then the next element will be -1, as it is alr
  he next element will be -1, as it is already -1.\n3. If the current element is -1, then the next
  ent element is -1, then the next element will be -1, as it is -1.\n5. If the current element is
  .\n\nTherefore, the final prediction for the last column of the time series is -1."}
```


- chatglm模型自动产生的提示效果不太好，可能需要比较强大的语言模型

下一步计划及相关问题

- 将RevIN处理的进行泛化性测试，看看效果如何?
- 解决RevIN在处理序列过程中遇到的一些问题。
- 继续看相关论文

谢谢老师和同学们的聆听!