网络表示学习大作业报告

甘云冲
2101213081
信息科学技术学院

1 问题分析

本次作业需要针对三个分子数据集，构建图表示模型，从而预测分子性质。

之所以可以在分子结构上应用图神经网络，是因为在分子当中，原子通过化学键连接，形成一个天然的图结构表示形式。同时，分子也具备一定的空间几何结构，图结构只是部分表现它的空间表示，所以在以往的工作当中也有往模型当中嵌入一些几何结构表示的尝试。

像图神经网络比较常见的运用场景，如社交网络和推荐系统，通常而言用户的交互只会和临近节点相关。但是在化学分子当中，这个性质会出现一定的差异，在拓扑上距离较远的原子对也可能会有较强的相互作用，从而影响整个分子的特性。所以如果仅仅考虑临近原子之间的相互作用，很难构造出一个具有较强泛化能力的分子表征。另一个理想的分子图神经网络，应当能够捕捉到分子中遥远原子之间所包含的信息，类似于分子内氢键，已形成一个泛化性良好的网络，可以在样本外数据上也取得比较好的表现。

事实上，这种性质在蛋白质网络当中也出现，在之前针对于 PPI 数据集的研究当中也曾经有研究人员提出，在蛋白质中，具有相同邻居的两个节点，往往会表现出相同的性质，这个先验假设是不存在的。

可以考虑对于数据集的基本统计信息做一个直观感受，具体如表1当中所示，涉及到的三个数据集中，两个是图回归任务，一个是图分类任务。ESOL 和 Lipop 分别是对于分子水溶性和脂溶性进行预测，输入均为分子的 Smiles 表示，输出为单目标。数据集当中并没有缺失数据。

但是对于 sars 数据集，预测目标为药物分子对于新冠病毒受体人类纤维蛋白受体的结合性和毒性相关的预测，其中 13 个预测目标当中，存在很大的缺失部分，在训练集当中，总共有 11652 个分子，但是最多的预测目标也只包含九千多个有效标签，部分预测目标实际上只有三千个分子是有效的，剩下的都是未知结果，所以数据集当中有这非常大量的数据缺失。具体的基本统计数据如表2当中所展示。并且从统计结果中可以发现，大部分的分子都是无药效的，药效

Dataset	Task	Num of Targets	Metric
ESOL	Regression	1	RMSE
Lipop	Regression	1	RMSE
sars	Classification	13	ROC-AUC

表 1: 涉及到任务相关描述

Target	0	1	2	3	Valid
target_0	8812	277	67	6	9162
target_1	2463	120	27	10	2620
target_2	4181	2032	854	445	7512
target_3	2949	232	148	84	3413
target_4	1585	322	117	25	2049
target_5	1877	134	31	7	2049
target_6	1242	536	206	65	2049
target_7	1898	96	44	11	2049
target_8	6933	293	244	42	7512
target_9	6220	661	423	208	7512
target_10	1843	493	273	11	2620
target_11	1826	444	314	36	2620
target_12	3808	91	54	5	3958

表 2: sars 数据集基本统计情况

越强，对应的分子集合就越少，所以对于每一个预测目标而言，处理的都是一个不平衡数据集上面的分类问题。

通过对于 sars 数据集进行进一步的分析可以发现，事实上在训练集当中存在重复出现的分子表示，并且其非空的目标集合存在重叠。这会导致以下两个问题：

- 对于重复出现的预测分子，会给予高于只出现一次的样本的权重，带来来自训练集分布上的偏见。
- 划分训练集、验证集和测试集的时候，如果采用随机划分，可能会导致相同的分子出现在不同的集合当中，导致信息泄漏，所得到的验证集结果不能很好的展现真实的模型情况

综上，这里考虑对于 sars 数据集进行预处理，对于相同分子进行合并，将其转化为一个训练数据。进行清理之后的数据集统计信息如表3所示，在清理之后，总共一共有 9900 个互不相同的分子表示。

2 模型设计

在模型设计方面，这里采用与特征相匹配的 AttentiveFP 作为基准模型，同时尝试了更新提出的 TrimNet 模型，并且尝试自己对于已有的模型进行部分的修改以达到更好的效果。

2.1 AttentiveFP

AttentiveFP 是 19 年上海科技大学提出的用来预测分子药理性和毒性的图注意力模型。其总体架构如图1所示，主要可以分成四个部分来得到最终的结果。

对于最开始的模型输入，原子特征的向量和邻近的原子特征的向量不具有相同的长度。因此，首先通过线性变换和非线性激活来统一向量长度，为每个原子及其相邻原子生成初始状态向量。

Target	0	1	2	3	Valid
target_0	8137	271	67	6	8481
target_1	2410	117	27	10	2564
target_2	3976	1898	797	416	7087
target_3	2882	232	147	84	3344
target_4	1585	322	117	25	2049
target_5	1877	134	31	7	2049
target_6	1242	536	206	65	2049
target_7	1898	96	44	11	2049
target_8	6708	287	242	40	7277
target_9	6055	633	401	188	7277
target_10	1805	479	269	11	2564
target_11	1789	434	305	36	2564
target_12	3692	86	54	5	3837

表 3: sars 数据集

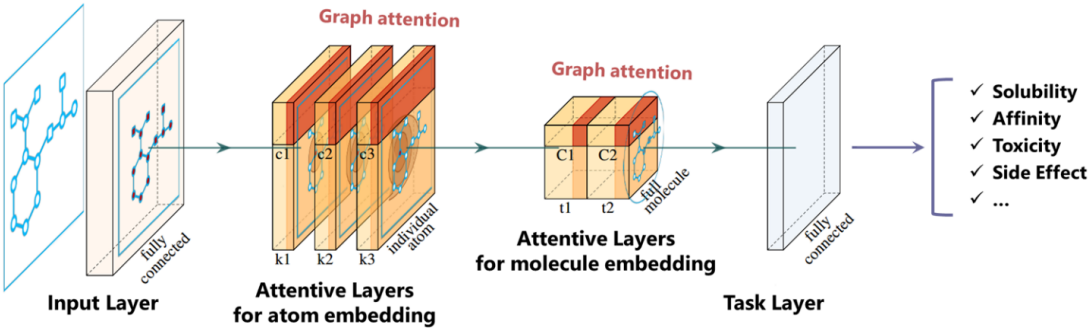

图 1: AttentiveFP 结构

之后为了能够从局部信息当中挖掘更多的信息，采用图注意力机制来对于邻居节点进行信息抽取。在前一步当中得到的初始态向量被送入堆叠的图注意层来进行节点嵌入表示的学习，这使得原子可以使用注意力机制对来自其相邻原子的信息进行聚集和筛选，从其相邻节点中抽取最相关的信息。在每个嵌入注意层的节点中，该过程为每个原子生成一个新的状态向量。在经过几个堆叠的注意层之后，状态向量包含了更多的邻域信息。

其中注意力机制的数学公式可以表示为：

$$e_{vu} = \text{LeakyReLU}(W[h_v \| h_u]) \tag{1}$$

$$\alpha_{vu} = \text{Softmax}(e_{uv}) = \frac{\exp(e_{uv})}{\sum_{u \in N(v)} \exp(e_{uv})} \tag{2}$$

$$C_v = \text{ELU}\left(\sum_{u \in N(v)} \alpha_{vu} \cdot W h_u\right) \tag{3}$$

由于我们所做的是图回归和图分类任务，所以需要将单个原子状态向量组合成一个分子状态向量。在 AttentiveFP 当中采用的方法是，构造一个超级虚拟节点，在这个超级节点与所有的原子都建立连边，采用这个超级虚拟节点的表征作为整个分子的向量表示。

最终得到的状态向量，就是对分子图的结构信息编码的学习表示，随后是用于预测的任务依赖层，根据任务的需要来产生相关的输出。整个网络以端到端的方式进行训练，获得针对特定任务或同时针对多个任务的一组特定网络权值参数。

对应的，AttentiveFP 也可以如大多数的 GNN 一样表示为 Messaging 和 Readout 两个步骤，形式化的可以表示如下，对应于图中的第二个部分与第三个部分。

Messaging:

$$C_v^{k-1} = \sum_{u \in N(v)} M^{k-1}(h_u^{k-1}, h_v^{k-1}) \tag{4}$$

Readout:

$$h_v^k = \text{GRU}^{k-1}(C_v^{k-1}, h_v^k) \tag{5}$$

2.2 TrimNet

针对于 AttentiveFP 的参数量过多的情况，2020 年有学者提出了 TrimNet，大大压缩了 AttentiveFP 的参数量，并且能够在部分数据集上超越 AttentiveFP 的表现。

TrimNet 考虑同样是采用注意力机制来处理分子图，但是它采用的注意力为一个三元组注意力机制，除了两个节点的表示之外，同时将边的表示加入了注意力机制当中。通过多头注意力机制进行信息聚合，相较于 AttentiveFP 当中的注意力机制，能够降低参数量来提升运算效率。

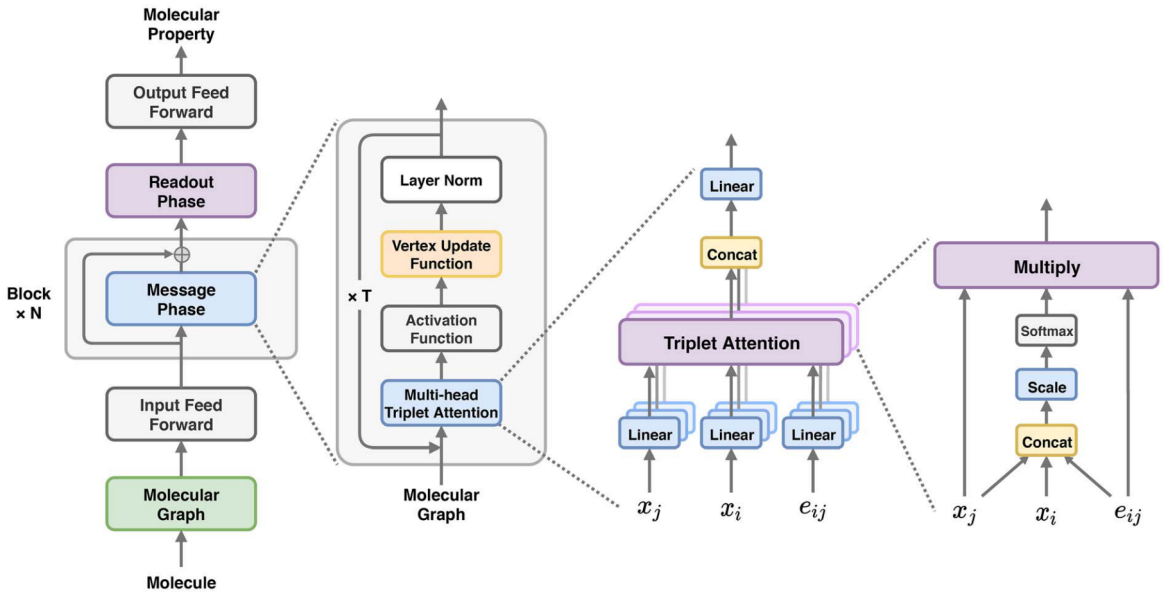

图 2: TrimNet 结构

$$\tau_{vu} = \text{LeakyReLU}(u^T[W_h h_v \| W_e e_{vu} \| W_h h_u]) \tag{6}$$

$$\alpha_{vu} = \text{Softmax}(\tau_{uv}) = \frac{\exp(\tau_{uv})}{\sum_{u \in N(v)} \exp(\tau_{uv})} \tag{7}$$

$$m_v = \sum_{u \in N(v)} \alpha_{vu} \odot W_h h_v \odot W_e e_v u \tag{8}$$

$$h_v^{t+1} = \text{LN}(\text{GRU}(h_v^t, m_v^{t+1})) \tag{9}$$

最终 TrimNet 的总体结构可以表示为在最初的输入阶段应用一个全连接网络来将输入的特征向量转换为特定的向量长度，通过多个堆叠的多头三元组注意力块，来进行原子中的消息传递，最后通过一个 Set2Set 层来和输出的全连接层组合来输出最终分子级别特性的预测结果。

2.3 模型改进

在 AttentiveFP 的 Readout 阶段，采用的均为堆叠图注意力层最终的输出，与超级节点的表示进行注意力与 GRU 计算。事实上，由于最终的分子表示为一个超级节点，那么每个原始分子与超级节点有且只有一条邻边，对应原子的嵌入表示实际上可以当成是对应的边特征。这样可以考虑将之前的原子嵌入表示也应用起来，实际上等价于在 readout 阶段添加了跳跃链接，也能够使得整个网络更好学习。

注意到在 AttentiveFP 当中的图注意力层数 l 和最后的步数 t 可能是不一致的，这里采用最为简单的处理方法，当 $t > l$ 的时候，最后的输入均为图注意力层最后一层的输出，否则则表示为跳跃链接：

- $t <= l$: $\hat{c}_t = \frac{m_t + m_l}{2}$
- $t > l$: $\hat{c}_t = m_l$

这一改进仅仅增添了跳跃链接，并没有增添额外的参数。

Model	ESOL	Lipop	sars
TrimNet	0.6848	0.6551	0.6314
AttentiveFP	0.6479	0.5892	0.6745
Ours	**0.6290**	**0.5734**	**0.6834**

表 4: 对比实验结果

3 实验结果

3.1 训练过程

对于 ESOL 与 Lipop 数据集，采用最开始 10% 作为测试集，10% 到 20% 作为验证集，后面的 80% 作为训练集合进行固定的划分结果，所有的模型都在相同的数据集划分下进行以避免划分方法上的差异让模型的比较产生偏差。对于 sars 数据集，在报告的问题分析部分已经描述了相关的数据清理过程，由于对于数据清理的过程会打乱数据原本的次序，所以这里再次对于数据进行打乱，并在所有实验当中固定打乱后的样本顺序。然后同样以类似于之前两个数据集的划分方式进行训练集、验证集和测试集的划分。

所有模型都通过 Adam 优化器进行参数优化，以来验证集上的指标来进行模型筛选，在结果报告部分，报告的为测试集上的结果。

针对于超参数方面，这里没有做自动化的参数搜索，大部分参考了原始论文当中的超参数设置，并且结合数据集规模以及个人对于模型的结果进行了部分的超参数调整。

关于评测指标，对于回归任务仍然适用 RMSE 作为评测指标，但是对于分类任务，在评分脚本当中的 ROC-AUC 计算方法采用最终的分类标签作为输入，丢失了部分信息。这里采用以概率分布作为输入的 ROC-AUC 计算方法来作为评价指标，可以更连续的表示模型的学习效果，不过在这里报告的数值可能会与最终评分脚本的结果存在一定的出入。

3.2 结果报告

结果如表4中所示，其中 ESOL 和 Lipop 的结果都是越低越好，sars 的结果是越高越好。总体来看 AttentiveFP 模型的表现比 TrimNet 要更好，在通过加入对于 Readout 阶段的改进之后，在划分的测试集上可以取得更好的表现。

4 感想与收获

针对大作业从我个人的角度出发，实际上在之前既没有接触过分子相关任务，对于化学并没有什么了解。对于图神经网络也是这个学期开始通过课程开始学习，对于如何处理与建模图结构数据也没有许多经验储备。通过对于这样一个任务，从文献调研到代码实现完成了一个图结构数据建模任务从零到一的过程，在这过程当中收获是非常大的。通过对于开源社区代码的学习，以及对于数据集自身性质的思考，很大地锻炼了解决问题的能力。

对于课程，课程当中涉及了很多近年的论文，内容紧跟着图神经网络最前沿的发展。但是同时也兼顾了图神经网络早期的发展历程，作为一个在之前从来没有接触过图神经网络的零基

础学生，在通过这个学期的学习，对于图神经网络所涉及到的组件以及总体的框架机制都有了一定的了解。并且在各次作业当中，循序渐进地熟悉如何搭建一个图神经网络的基本框架，以及对于常用的基本组件有了一个比较直观的了解。三次作业加上大作业做下来感觉自己已经可以针对于图数据构造网络进行训练和分析来编写代码，收获非常大。