**内容概述：**

1. 拓扑神经网络代码
   1. 添加了数据预处理模块
   2. 写好了基于注意力机制的网络gat\_v2
   3. 添加了部分加速训练的trick
   4. 添加了归一化操作。

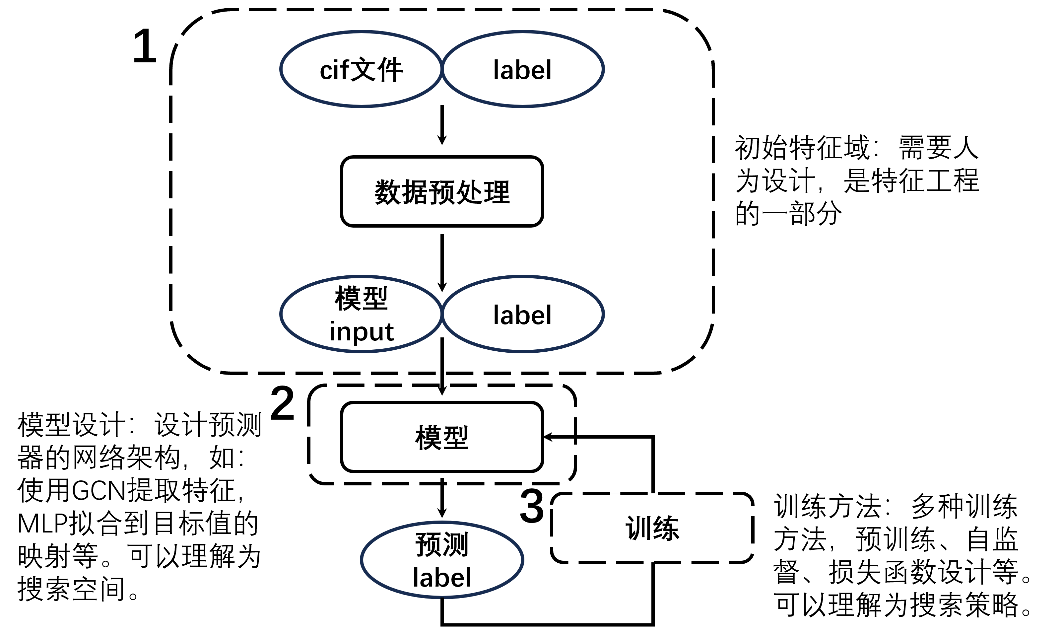
**具体内容：**

1. **添加数据预处理模块**

任务概述：在训练COF-甲烷存储数据时发现数据太大了，**内存中没有办法一次性放下**，**依次读取预处理（读取cif文件构建晶体图）的话又非常耗时**，因为每个epoch都需要进行预处理。所以把预处理操作分解出来，单独写了一个模块，利用进程池并行的处理数据。

模块说明（**留个mark**）：

整个预测任务可以分解成几个子任务，如图所示：



根据图中所示，实验的实现需要完成四个部分：

1. 数据预处理模块：提取cif文件中的信息，转化为模型所需要的输入形式，例如图神经网络所需要的图信息，也可以转化为正弦库伦矩阵后得到特征值向量再使用全连接网络。在这个项目里，我重点考虑基于图、超图和组合复形的包含拓扑信息的数学结构，作为图神经网络、超图神经网络、组合复形神经网络的输入。
2. 数据集实现：因为整个项目是基于pytorch的，所以需要实现一个dataset用于后续数据加载的需要。目前只实现基于map的dataset方式，因为数据集基本上是固定的，后续再实现iterate的dataset，用于在线学习。
3. 模型设计：在这个项目中，我重点考虑基于图、超图和组合复形的数据结构，所以模型可以分为两个部分：用于自动提取特征的图神经网络、超图神经网络、组合复形神经网络；用于拟合到目标值的映射，是一种最一般的深度学习模型架构。**根据训练方法不同，会对模型进行一定的扩充和修改，利用基于重构的预训练方法需要增加一个decoder。但是最终预测器的架构还是如前面所说的两个部分。**
4. 训练方法的选择是多种多样的，如基于设计代理任务的预训练模型，基于重构的预训练，在预测力时增加的力损失函数等，还包括主动机器学习方法等。
5. **写好了基于注意力机制的网络gat\_v2**

**说明：**基于前文所述，是模型上的改进。由于输入的晶体图是根据距离8连边的，这导致了每个原子周围连了许多条边，这在无机晶体中或许合理，正如嵌入原子定理所使用的那样。但是对于有机晶体，原子往往只会和周边少量原子连成边，所以可以将距离缩减为2（这正是形成共价键的距离阈值）。但是这样做的话，原子的观察野就被缩小的，只会考虑距离2内的原子对它的影响，而有机晶体中普遍存在的氢键往往大于2甚至更大，并且对原子的作用力也很大，有时候也需要考虑。**所以可以采用注意力机制，让网络去自己学习哪些原子对自己的影响比较大。**

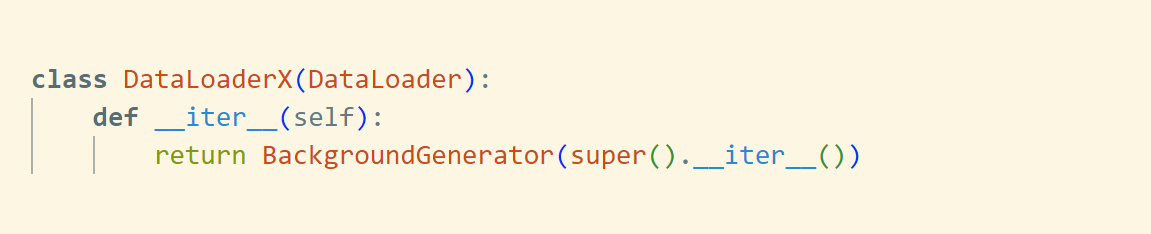
**实现：采用pyg提供的gat\_v2网络。**Gat\_v2是传统的gat的改进，解决了传统gat的静态问题。基于这个，实现了一个预测器模型。

**训练：**目前还在调试代码，添加各种加速trick

1. **添加了加速训练的trick**

任务来源：训练过程包含数据预加载和显卡上训练两部分，并且这两个部分所耗时基本上是一样的，一半一半的样子，所以使用预加载的方式，在前一批数据还在训练的时候就将下一批数据提前加载上去。

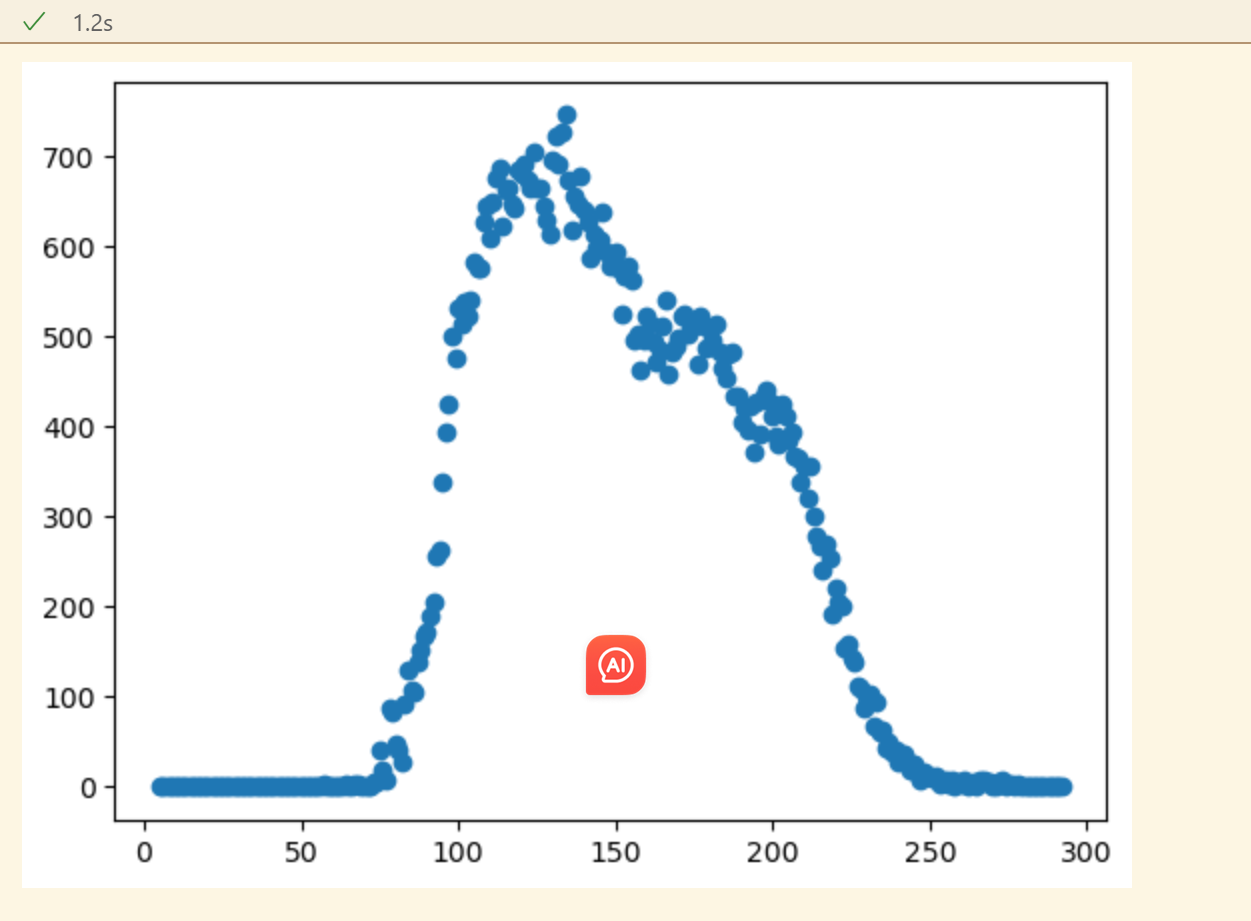
实现方法：



利用prefetch\_generator包，可以实现数据加载不停顿的效果，原始的pytorch利用的dataloader虽然可以实现多进程加载，但是加载完后只有当所有数据都读取完了才会再加载下一批数据，这导致了训练过程被阻塞一段时间，利用这个包，保证训练过程中需要的数据都已经加载好了，不会被阻塞。

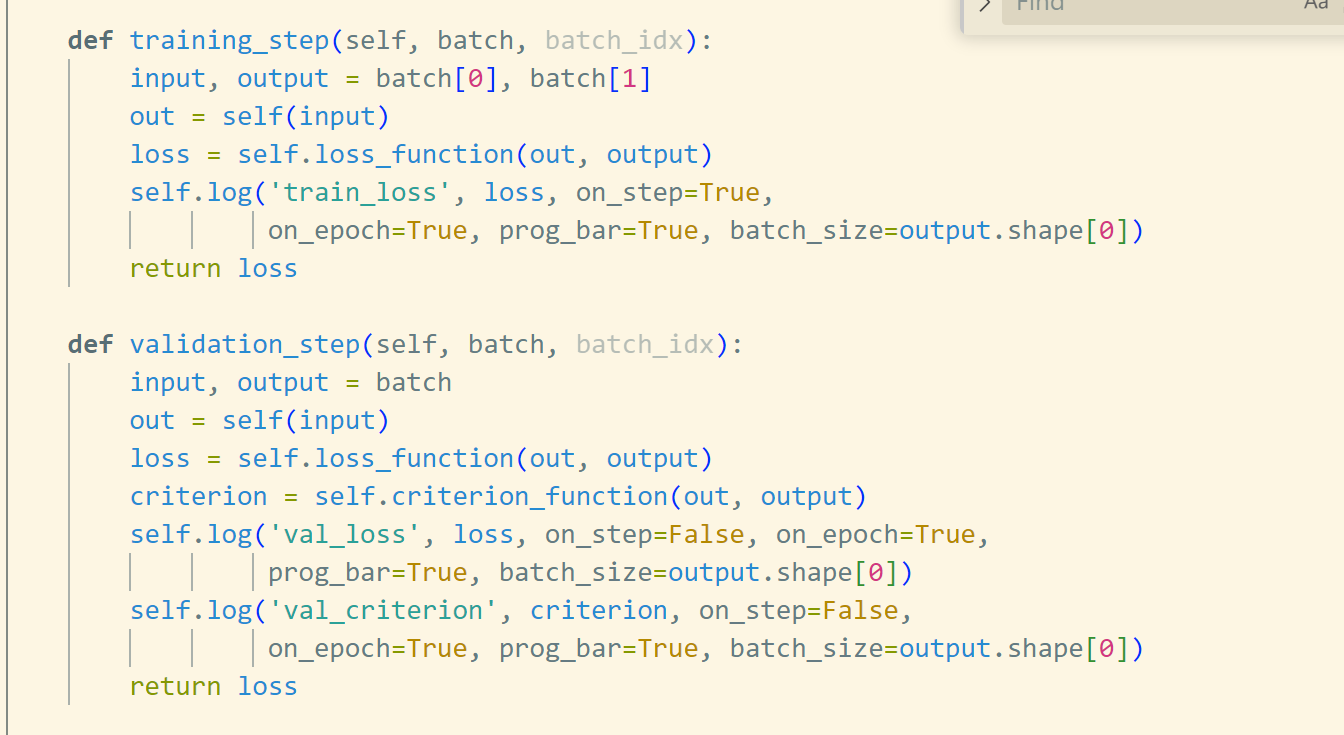
1. **添加归一化操作**

**原因：**目标输出的区间跨度比较大，我对数据根据值的分布画了一张图，



横坐标是属性值，纵坐标是属性值落在区间[x,x+1)上的数量。

**实现：**由于现在的整个项目都是基于pytorch lightning 的，现在归一化操作没有之前那样方便，所以采用的方法是在训练前根据训练集的属性值计算均值、标准差，然后存到DataModule的属性值中，并在batch加载到device之前对目标值进行归一化操作。但是这种方法存在一个问题，我们在评估模型效果的时候需要计算一个评估值，一般采用mae，这个mae计算的是归一化后的mae，所以最终得到的mae还需要再乘以标准差才是真正的mae。



怎样在训练的时候就把归一化得到的均值标准差传进去，从而计算得到的criterion就是实际的评估，仍然需要思考。

下周安排：

1. 调式代码
   1. 上面说的归一化在训练过程中的实现
   2. 验证写好的模型和论文中的吻合：代码分析和结果分析
2. 实验
   1. 不同层对于结果的影响3，5，7（之前做过这个实验，结果式5层效果最好）
   2. 验证截断距离的选择对结果的影响，对比截断距离8和2
   3. 验证注意力机制对结果的影响，对比5层网络，截断距离设置为8，加入注意力机制和一般图网络对结果的影响。