当前问题解决方案：

两个方向，我还不知道选哪个？

1. 尝试新的网络。（担心是自己输入数据问题、网络卷积方向设置问题、基础参数设置问题、网络是否适用问题、正负样本数量是否合理问题。）需要老师指点一下什么样的网络值得尝试？指点一下分别从哪几个步骤或按照什么标准依次检查网络？使用哪些特征？
2. 重新提取一遍数据，从头开始确保数据源头没有问题。（选取某一理化属性类的小分子药物位点，做对比。）

课题简介：找到膜蛋白序列中的，和药物小分子有相互作用关系的位点。

数据处理流程：

数据来源：

1,PDB数据库中，提取膜蛋白及其相关信息。在每个记录中，找到结合位点相关信息。筛选出配体中属于药物的小分子。找到结合位点，标识为结合位点。

2,药物小分子列表来自于drugbank..

去冗余以及正负样本处理。

全部样本蛋白质条数3232条。0.3cdhit去冗余后446条。

正样本选取：

在全部样本中，按照窗口大小为【31】提取正样本序列，22141条正样本。去掉完全相同的序列，可获得10290个正样本。（长度为31的正样本，有很多完全相同的序列，这部分我将其去掉了。）

有些正样本在另一条蛋白质中，并不是结合位点。因为我们的目标是找到所有“和药物有相互作用的潜在的点”，所以这部分正样本应该予以保留。

负样本选取：

为了避免“冗余的蛋白质序列”进入训练，所以我使用0.3cdhit去冗余后446条来提取负样本。得到141770个负样本序列。其中去掉相同的负样本序列得到138822个负样本。其中去掉其中存在于正样本集合中的样本，得到138807个负样本序列。

至此，我们得到了10290个正样本和138807个负样本。

其中，正样本选取过程中没有去冗余：理由是虽然序列可能有冗余，但它们都是正样本点，不应该被去掉。

负样本使用的是经过了去冗余处理的蛋白质序列。这样可以合理避免掉冗余造成的神经网络学到的偏差。

数据集划分：根据正负样本数量，各自按照8：1：1的数量划分训练集、测试集、验证集。

（以one\_hot编码方式）

train\_pos\_one\_hot\_data : 8232

val\_pos\_one\_hot\_data : 1029

test\_pos\_one\_hot\_data : 1029

train\_neg\_one\_hot\_data : 111045

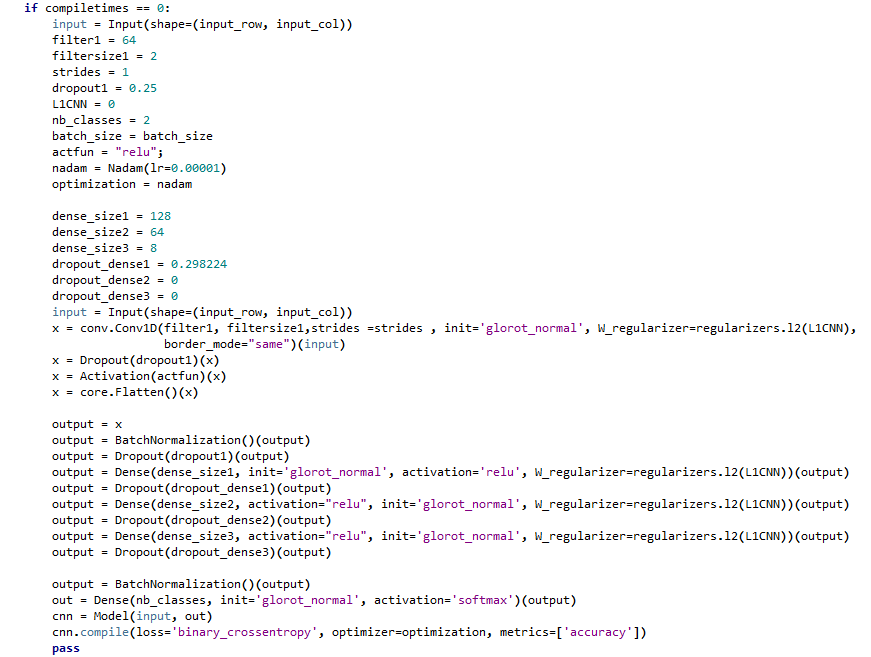
val\_neg\_one\_hot\_data : 13881

test\_neg\_one\_hot\_data : 13881

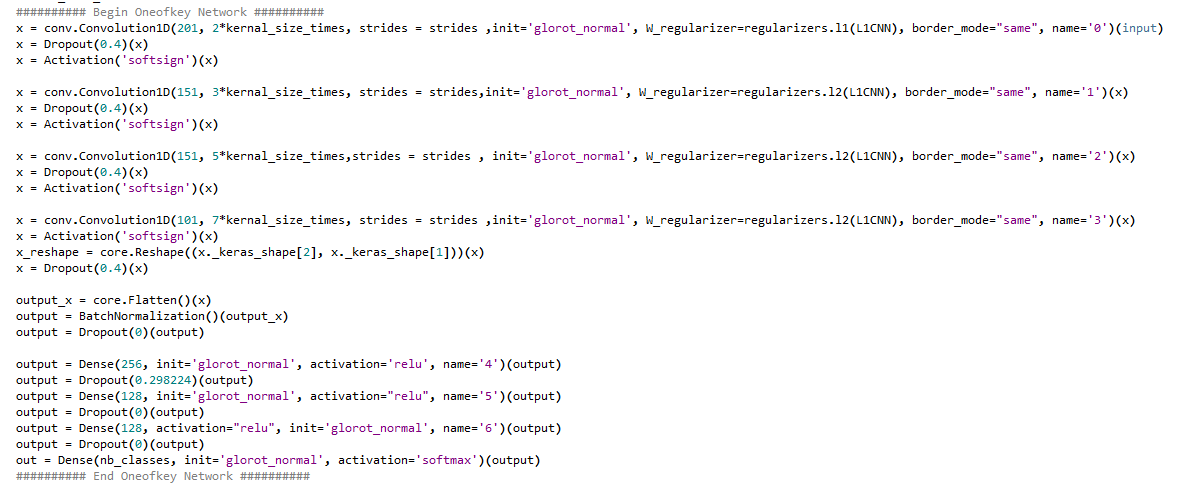
训练时选取正负样本的策略：随机选取正样本集合中的80%的数据（6585条），随机选取负样本中和正样本数量相同的数据（也选取6585条），组成数据集放入模型进行训练。循环多次。

使用过的网络：

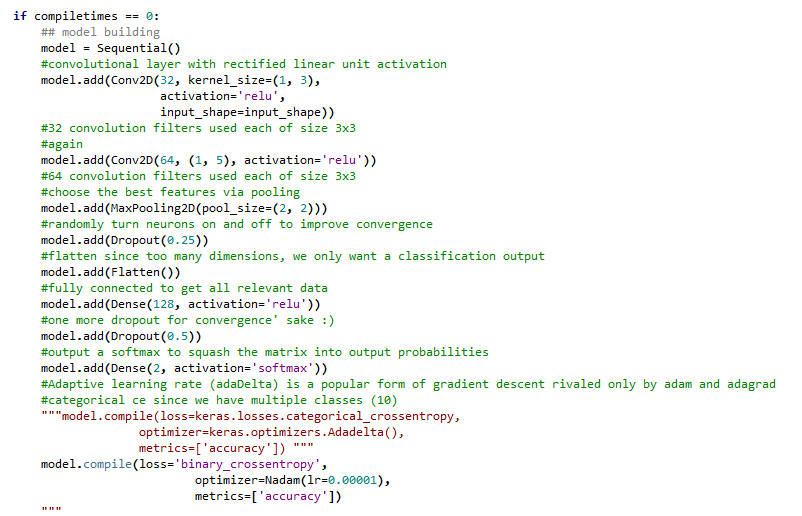
1. 自己构建的简单一维卷积神经网络：1层卷积、3层全连接。



1. 何飞老师泛素化论文中构建的神经网络：3层卷积，3层全连接



1. 自己尝试的简单二维卷积神经网络：



最近三天内，分别对：filter、filtersize、strides、kernal\_size(二维卷积核大小)、卷积方向（横着卷还是竖着卷）、网络层数、单层网络大小、one-hot编码或理化属性编码（泛素化代码中的10个理化属性）等等问题做了尝试。单一变量，每个变量设置2~3组对照。均无显著提升或变化：acc稳定在0.5，loss稳定在0.693.

其他支援：大师姐路畅帮忙调试。将我的数据放入她的机器学习模型中、刘喆的胶囊网络中，inception（？我也没记住是哪个词）网络中，均无学习效果。还计算了one-hot编码和lable之间的皮尔森相关系数，显示没有相关性。

寻求老师的帮助和指点~