# 实验记录

## 1时间：20190610，16:00

## 2实验目的：

测试网络是否可以运行，测试新整理得到的数据是否可以进入网络，测试2个网络，测试2种新特征。

## 3实验代码存放文件夹：

20190610文件夹中。

## 4实验运行时间：

1天及以上。

## 5实验使用数据信息：

药物作用位点预测，可能存在的药物作用位点标志为1，正样本。其余其他位点均为负样本。

根据6月6日之后整理得到的数据，正样本选取所有正样本点，去同（去掉所有一样的）

负样本来自0.3去冗余后的蛋白质中，去掉处于正样本中的。

## 6.1，数据集正负样本设置，比例设置，

正负样本比大概1：12，按照8:1:1划分训练集、验证集、独立测试集。随机打乱后划分。

训练中随机选取正样本中80%的数量进入训练。取相同数量的负样本。

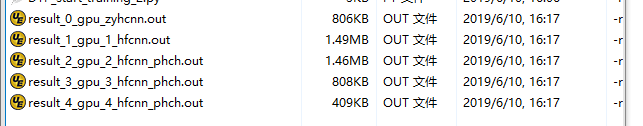
## 6.2，网络设置：

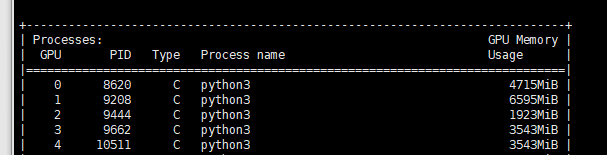
网络有2个，自己的和何老师网络改造后得到的。

使用的特征有2种，one hot编码，蛋白质理化属性。

特征输入部分存在问题：【row】【col】输入的是否需要和样本相同？若不相同会有什么后果？（修改输入数据中的维度，得到了val验证集中的数据反馈。）

## 8实验运行记录



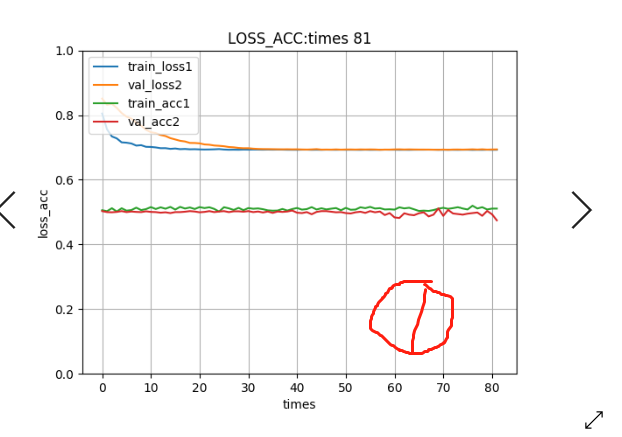


|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 开始时间 | 输出文件 | 结束时间 | GPU编号（数量） | 运行id | 长度 | 文件夹 | 使用网络 | 使用特征 | 其他 | 描述 |
| 1 | 16：00 | 如图 |  | 0 | 如图 | 31 | 0 | Zyh\_cnn | One\_hot | 0.25 |  |
| 2 | 16：00 | 如图 |  | 0 | 如图 | 31 | 1 | Hf\_cnn | One\_hot |  |  |
| 3 | 16：00 | 如图 |  | 0 | 如图 | 31 | 2 | Hf\_cnn | phy\_che |  |  |
| 4 | 16：00 | 如图 |  | 0 | 如图 | 31 | 3 | Zyh\_cnn | phy\_che |  |  |
| 5 | 16：00 | 如图 |  | 0 | 如图 | 31 | 4 | Zyh\_cnn | phy\_che | 输入维度有变换 |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |

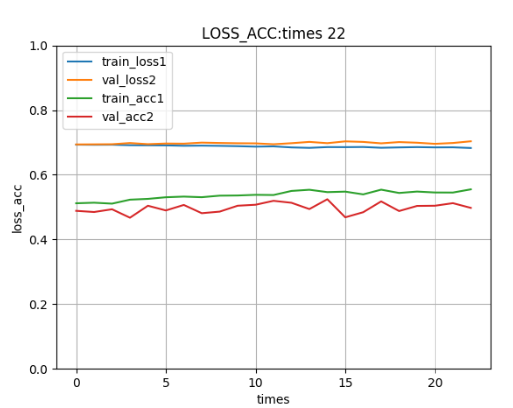
以后根据预期主要变量由小到大设置数值，GPU编号，名字，文件夹都要由小到大设置。

## 9实验结果存放位置

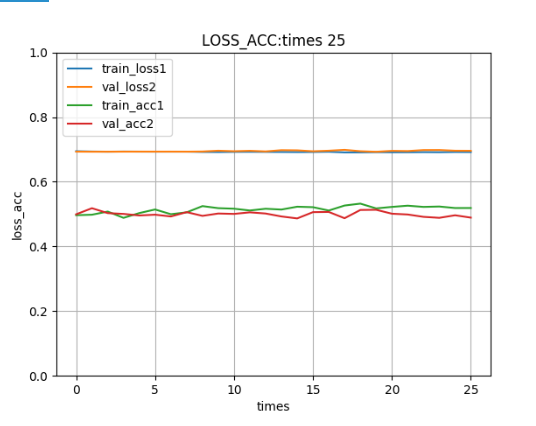
## 10结果分析



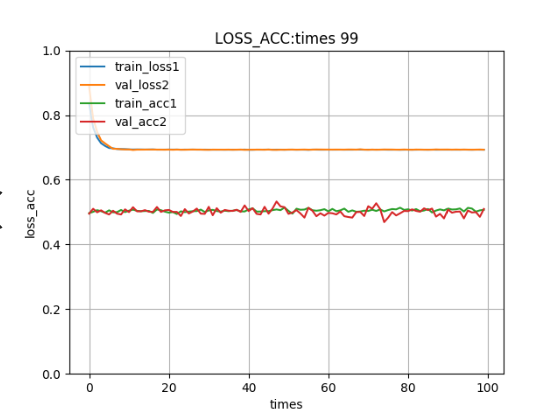
1号实验zyh网络没啥学习效果。



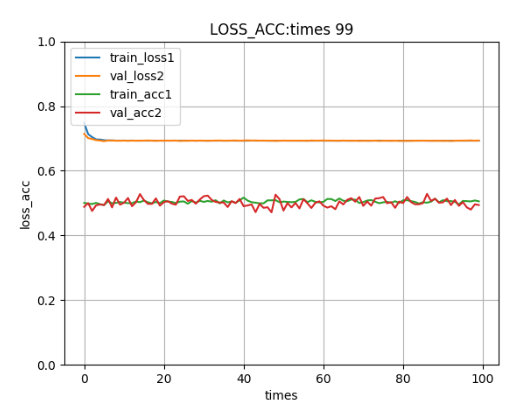
2号实验hf效果略微提升，acc有所升高，但val曲线不好。

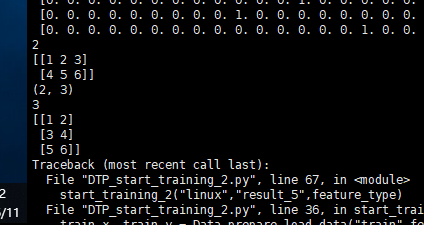


3号实验使用hf的one hot网络跑理化属性没啥成效



4号实验使用zyh网络运行理化属性也没啥收获。



5号实验在4号实验的基础上，调整了输入的shape 。但没有报错。理由是输入的维度进行了reshap

我们使用的是一维卷积，若输入二维向量，将会被自动“reshape”,

之前：我输入的矩阵是（13170，31，20）

原因在于，针对每个样本，我需要依次将存储文件中的氨基酸特征矩阵（one hot）变成numpy格式。

在存储特征过程中，我使用的是list.append模式。

对于长度为31的氨基酸，我得到的list是一个包含31个小list 的大list

就是这样：

【【20个0或1】，【】，【】，。31个。，【】，【】】这时，得到的矩阵是一个（31，20）的矩阵，31row,20col.

若这样的矩阵被reshape,会是这样的效果：第一row 后面接着第二row…总共接31个，形成这样一种一维序列。

（相当于顺次显示了氨基酸总类特征。这个模式下卷积核为64取得有趣的acc提升！！！）

根据冰总的理论：

这个模式是不合理的。

另一种输入更合理。也是应该的：矩阵应该是一个(20\*31)的矩阵，20row,31col. 一行后面接着下一行。这样卷积的话可以提取到序列中的特性。这才是理论上行得通的编码方式。

综上，我修改了读取数据过程中的矩阵，让它变成一种看似更合理的模式。具体的提升将在6月11日的实验结果中展示。

之后将构建2维卷积网络，卷积核设置为长条形。这样将会很大程度上增强序列方向的特征提取模式。

## 

## 11总结

1网络随意尝试效率太低。下次调整网络，计划找一个效果好的。或者根据先验知识调整。

2数据是关键，今天发现数据问题比较严重。也很心疼自己的网络：先告诉它这个是正样本，一会儿又告诉它这个这个是负样本。。。假阴性样本多？