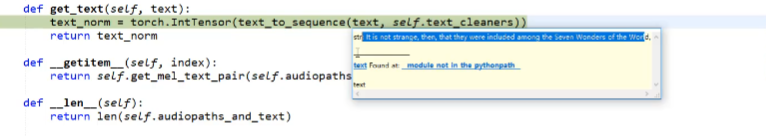
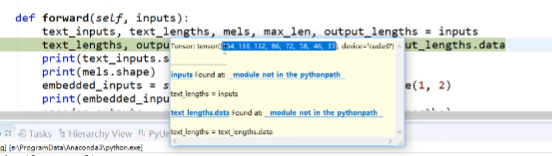
Data Loader:

语音数据，读取，分窗，做短时傅里叶变换，mel滤波器

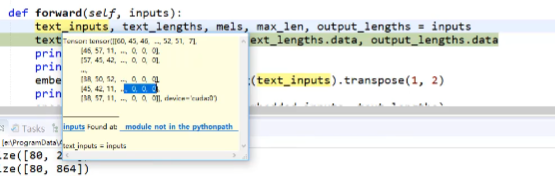
Text Loader:



把每个字符转换成一个id 比如I 46 b 57

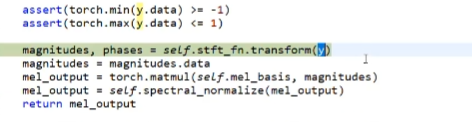


可以看到每一个序列的长度都不一样，所以进行补零操作



这个就是输入数据 维度（8，sequence ）

对于声音信号，先做短时傅里叶变换，再做mel特征（80 filter）



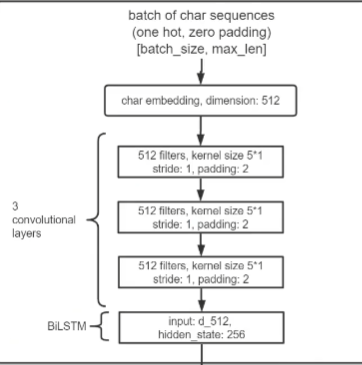
(batch dimention, sequence length, 80 filter) target number

Encoder:

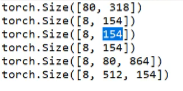
已经拿到数据了

Embedding: 对text数据进行处理 之前的维度是（8，max sequence length）

处理之后的维度是（8,512，max sequence length）相当于将原来简单的数字进行了编码（如何编码的）

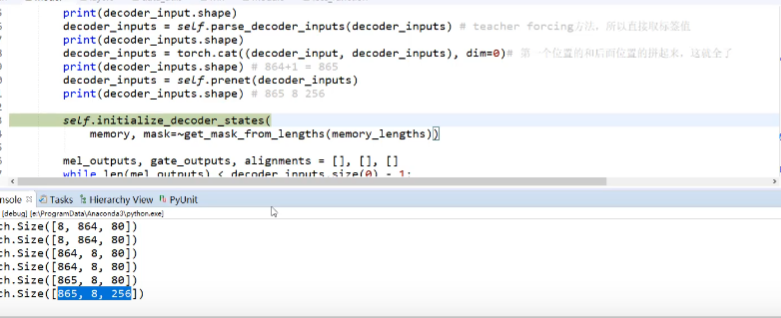


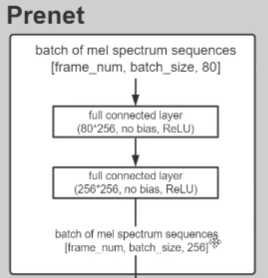
Dimension is 512



Prenet:

做了两个linear function将80维度特征转换成256维度



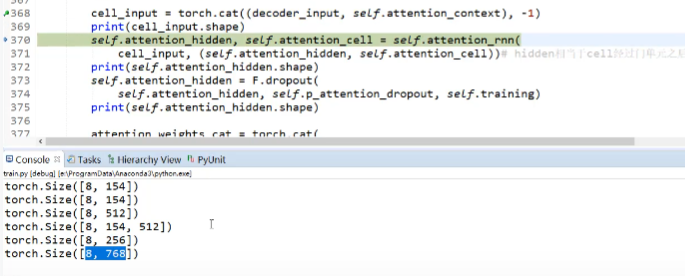
 （但是我不知道为什么要做这一步）

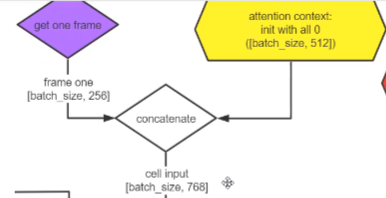
Decoder:

(8, 514, 152) encoder输出 （8,80,864）

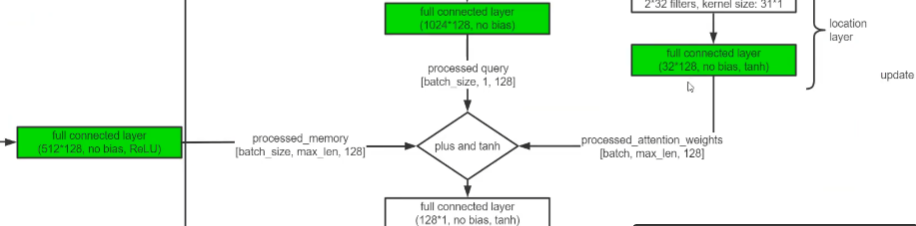
Attention:

之前的decoder input (865,8,256) 提取one frame 维度应该是（8,256）将该维度与attention context维度拼接，得到的维度是（8，768）

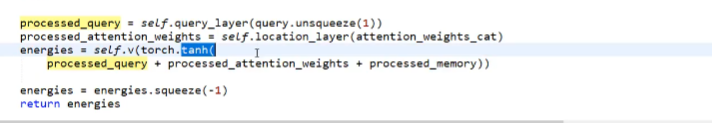




将合成后的特征通过lstm得到隐层特征（8,1024）

累加attention, 提供position信息，decoder提供语音信息，encoder提供文本信息，三个attention汇总

三个隐层特征，经过卷积层做处理，



三个加在一起

权重乘decoder\_output