# **任务一：online\_shopping\_10\_cats数据集的文本分类**

1. 数据预处理

本次任务预处理流程大概如下：

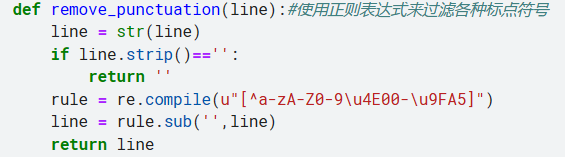
1. 输入数据



1. 处理空值



1. 过滤标点



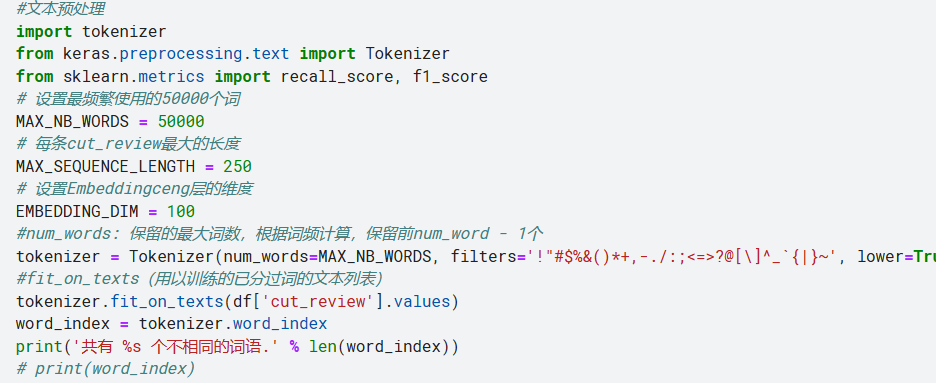
1. 停用词分割



得到结果如下：

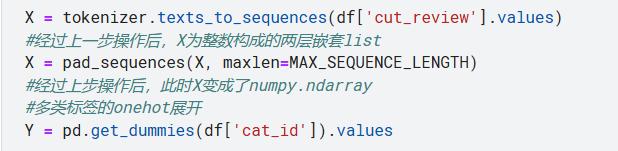
1. 建立词表

按照词语出现次数多少进行排序，取最多的50000个词建立词表，为每个词赋予一个编号。同时通过画出每句话的含词数，选定250为每句话的最大含词数，多余250个词的句子进行截去，少于250个词的句子自动在最前面补0。



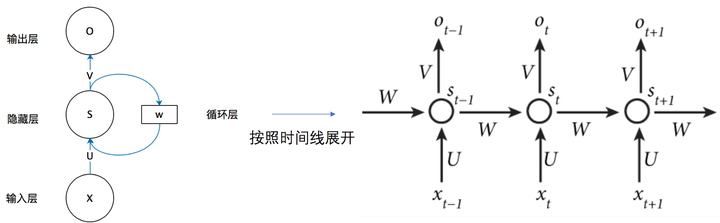
1. 修改句子形式

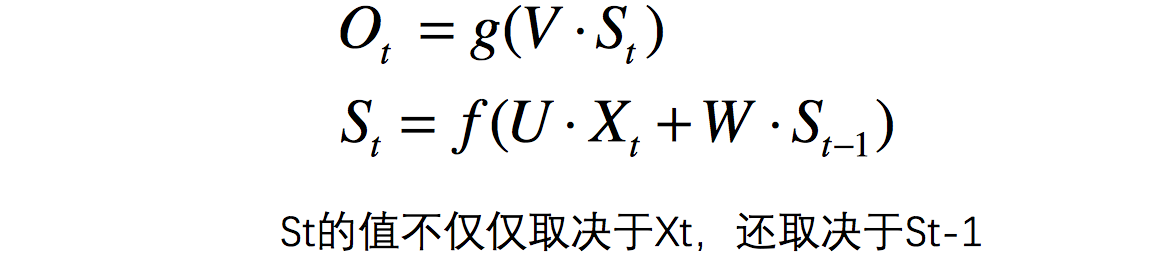
将句子修改为词编号组成的形式，同时对标签进行onehot编码。



1. RNN结构实现

已知RNN原理如图：

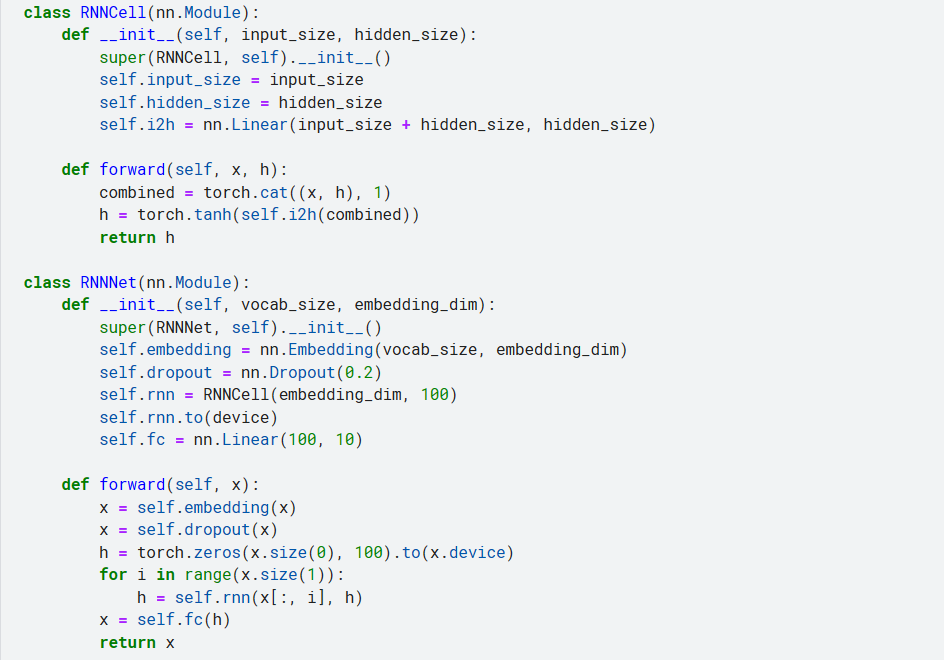




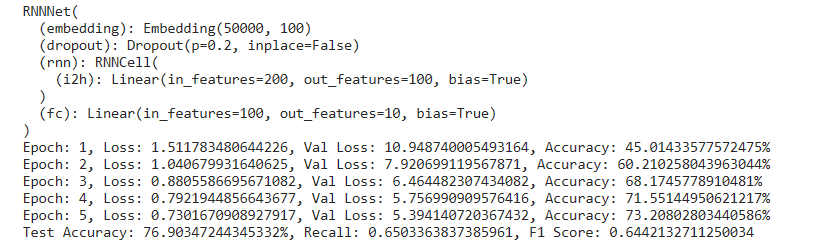
故可实现结构如下，同时在rnn的基础上补充了embedding层dropout层以及全连接层的实现。

其中Embedding层通常用于处理具有大量类别的离散特征，最常见的应用是在自然语言处理中将单词转换为向量。这种转换可以捕捉到单词之间的语义关系，例如相似的单词会被映射到向量空间中的相近位置。Embedding层可以看作是一个查找表，表中的每一行对应一个类别的向量表示。

而Dropout层，用于防止神经网络过拟合，是一种正则化技术。在训练过程中，Dropout层会随机地将输入单元的一部分设置为0。这可以防止网络过度依赖任何一个特征，从而提高模型的泛化能力。：

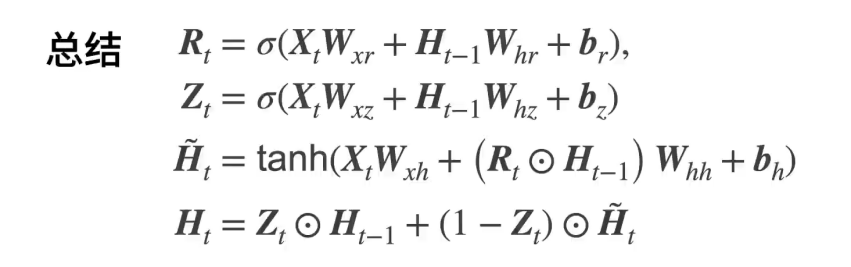


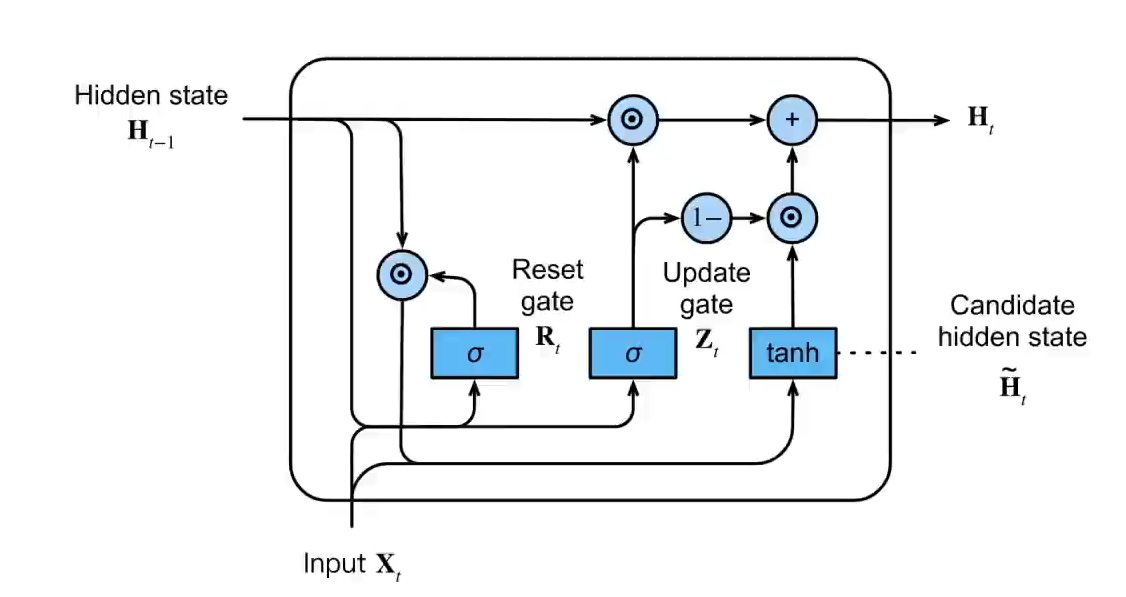
RNN网络实验结构如下：



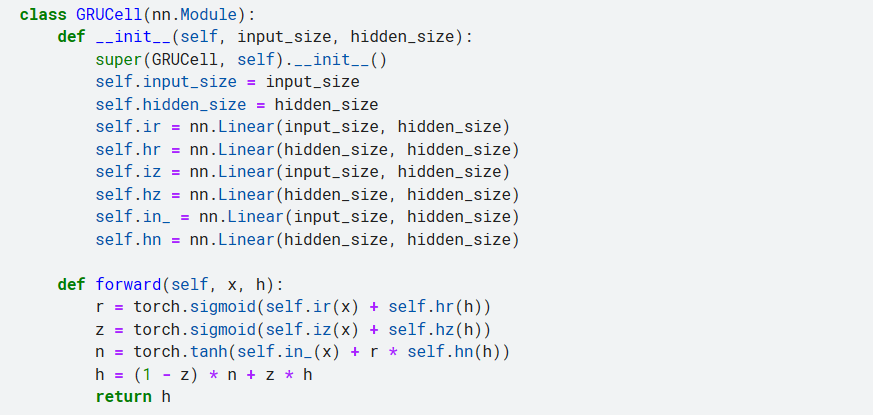
1. GRU结构实现

已知GRU原理如图：

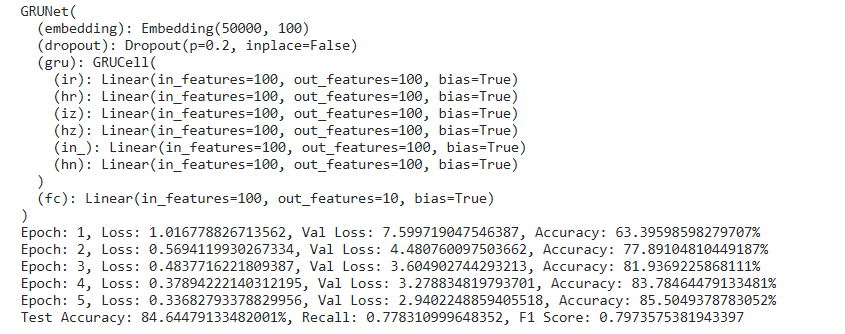




故可实现结构如下：

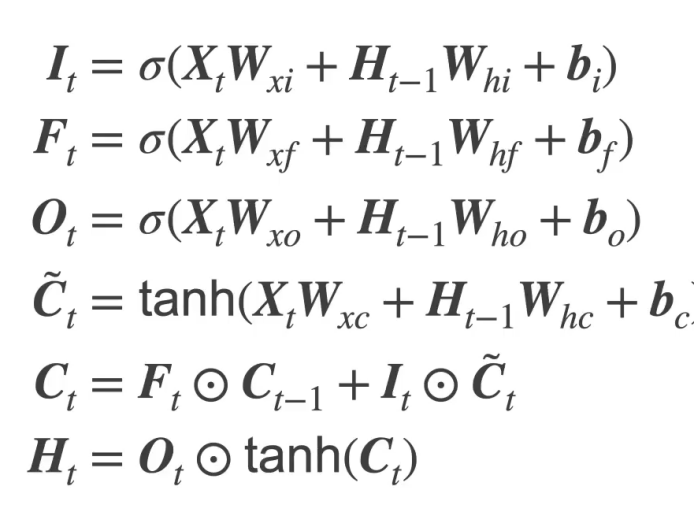


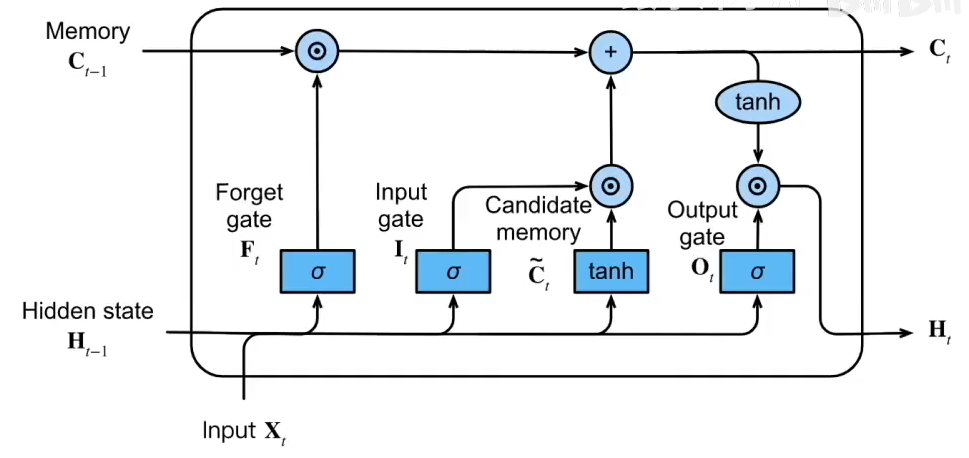




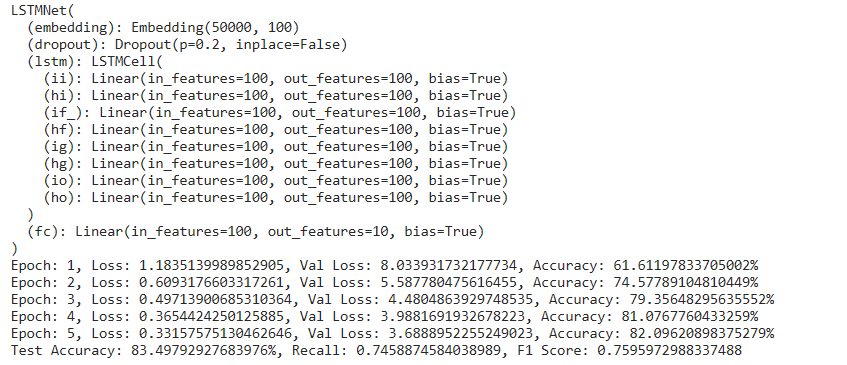
1. LSTM结构实现

已知LSTM原理如图：



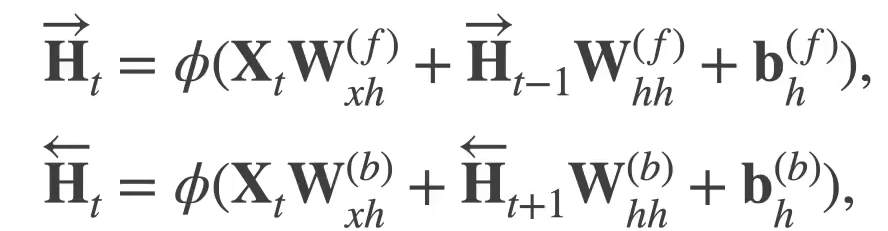


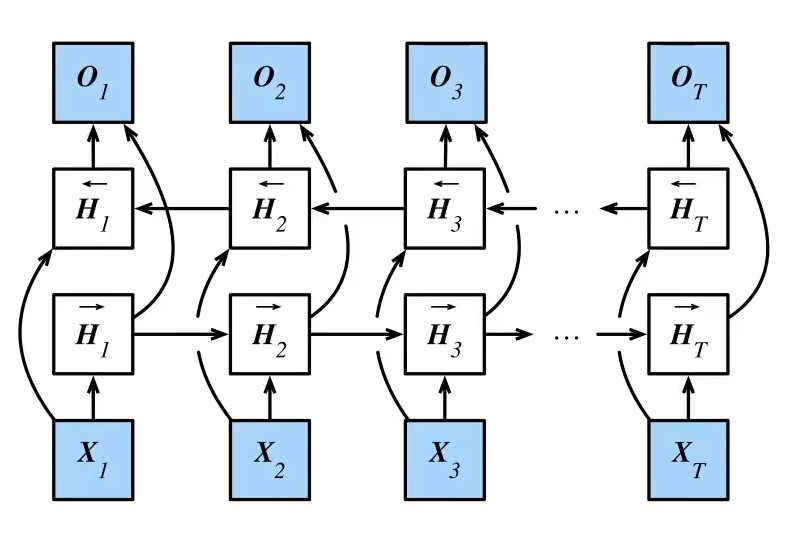
实验结果如下：

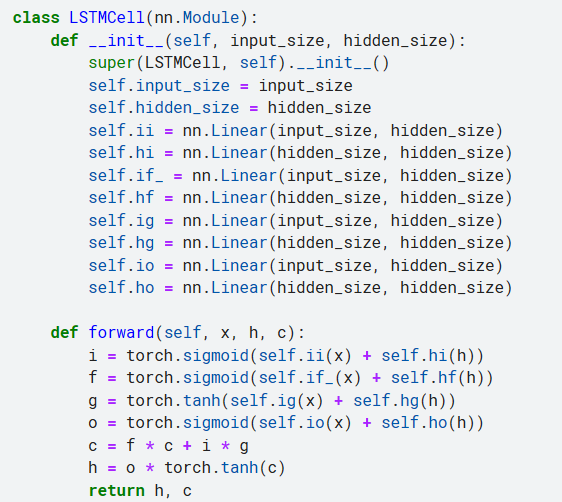


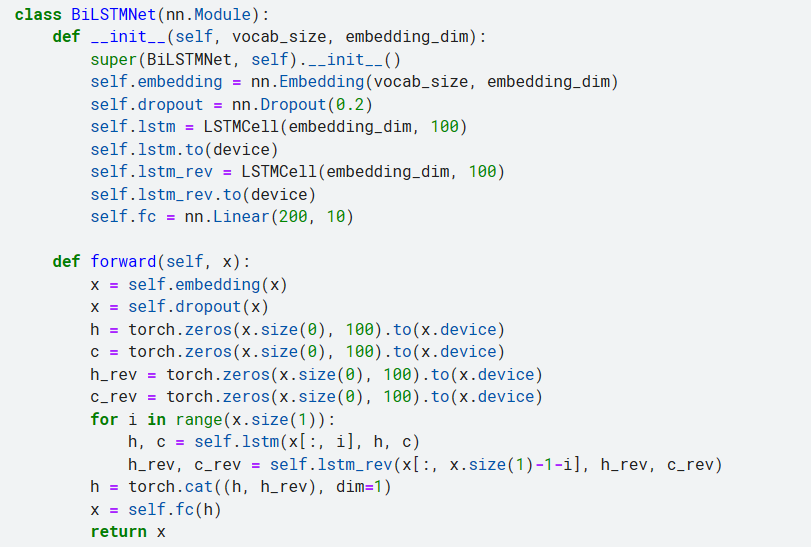
1. Bi-LSTM结构实现

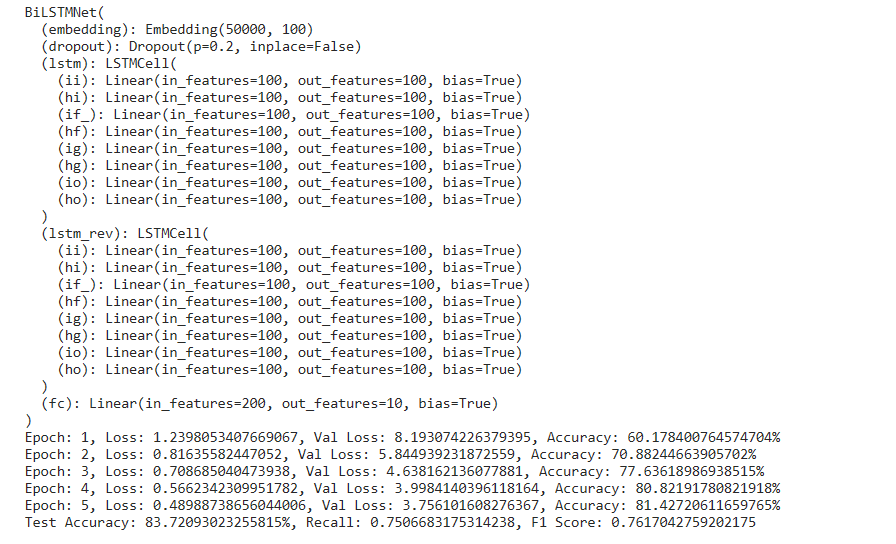
已知Bi-LSTM原理如下：











1. 实验结果对比分析

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | RNN | LSTM | GRU | Bi-LSTM |
| Accuracy | 76.90% | 83.50% | 84.64% | 83.72% |
| Recall | 0.650 | 0.746 | 0.778 | 0.751 |
| F1 Score | 0.644 | 0.760 | 0.791 | 0.762 |

F1值是一种用于评估分类模型性能的指标，特别是在数据不平衡的情况下。F1值是精确度（查准率）和召回率（查全率）的调和平均数。精确度是指你预测为正的样本中有多少是真正的正样本，召回率是指所有真正的正样本中有多少被你预测为正。F1值越高，模型的性能越好。

其中RNN结果最差，因为存在梯度消失梯度爆炸等问题，在处理这类长文本时情况不佳。

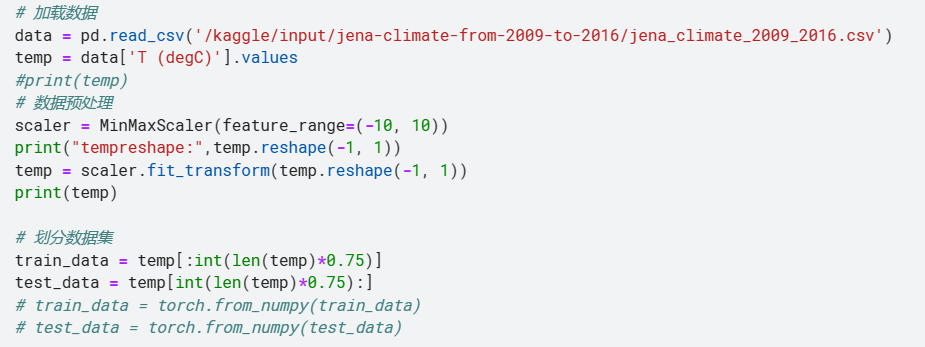
LSTM引入了门机制解决了梯度消失的问题，使得模型能够学习长期依赖。然而，LSTM的参数较多，计算复杂度较高，所以在相同的较小epoch下，收敛速度可能没有简化版的GRU快。

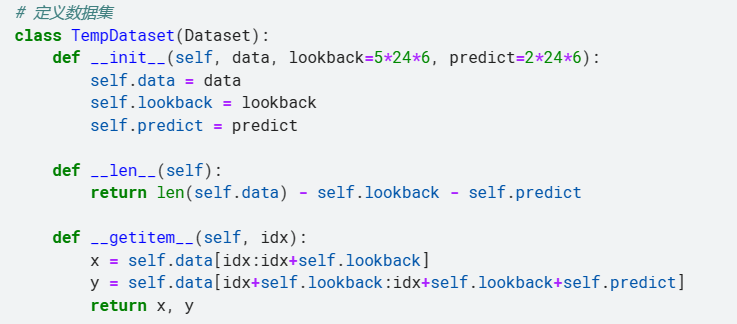
Bi-LSTM引入了两个方向的信息流，计算复杂度更高，在较小epoch下可能收敛也较慢。

GRU是简化版的LSTM，实现的性能却差不多，计算效率更高，所以GRU结果最好。

# **任务二：jena\_climate\_2009\_2016数据集的天气预测**

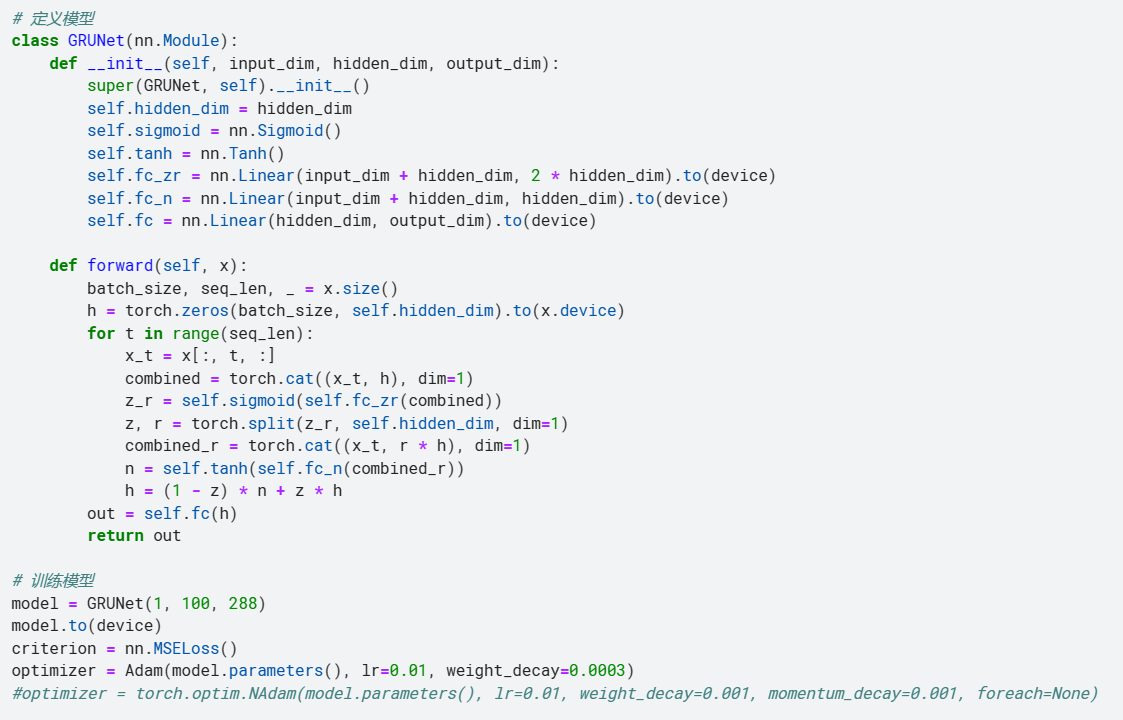
1. 数据预处理

本次实验将温度值映射到[-10，10]的范围内处理。并实现自定义数据集，进行迭代器，初始化等函数自定义。 



1. GRU实现

GRU模型结构如下，与上面的GRU结构略有不同，删去了embedding层和dropout层



1. 实验结果与分析

定义测试集中一个星期为一个batch，每次只输出一个batch的图表；在一个星期中，前5天的温度作为测试输入，输出后2天的温度，并与后2天的真实温度做对比。个人感觉预测结果不太好，分析原因在于：1.GRU能力受限或者是调参没完全调明白 2.预测温度的输入是否不能只是温度，应该也将其他的气温因素一同输入进行预测，可能会有更好的效果。





选取了一些较好的拟合结果如下：

