作业三 序列标注

1. 思路

本次作业要完成 nlp 四大基础任务之一的序列标注任务，也叫做命名实体识别。即是在给定文本中能够对词性、人名地名等特定信息进行标注。

实验主要采用循环神经网络进行搭建，每一条样本输入是一条句子（对应的嵌入向量表示），该样本的标签也是一个等长的句子标签，其中每一个元素对应句子中每一个字的标签。如样本为'我 爱 北 京'，则该样本的标签为'O O B-LOC I-LOC'。然后通过循环神经网络再结合交叉熵损失函数进行训练。但根据课上所学知识，我们知道直接用 LSTM 虽然可以完成该任务，但是往往会预测出现一些不可能真实存在的结果，如连续两个 B-LOC 标签。因此我们可以增加条件随机场 CRF 模型在 LSTM 层之后，使用梯度下降自动去学习 CRF 模型的参数，这样可以获得比只使用 LSTM 好的结果。

2. 模型概况

Input（）
Embedding（input_dim=5000, output_dim=50）
BiLSTM(units=100)
BiLSTM(units=150)
CRF(units=7)

3. 编程实现

为了实现上述模型我们首先需要对输入进行处理。首先使用 Tokenizer 库进行分词，词典大小设置为 5000，之后对训练样本进行 padding（这里选择 maxlength 进行 padding），之后对标签也要做相应的 padding（这里 padding 的内容直接选用 O 标签）。
使用 keras 库进行模型的搭建，堆叠型模型声明如下

```
model = keras.Sequential()
model.add(keras.layers.Embedding(input_dim=5000, output_dim=50, input_length=100))
model.add(keras.layers.Bidirectional(keras.layers.LSTM(100,return_sequences=True)))
model.add(keras.layers.Bidirectional(keras.layers.LSTM(150,return_sequences=True)))
# model.add(keras.layers.Bidirectional(keras.layers.LSTM(150,return_sequences=True)))
model.add(keras_contrib.layers.CRF(units=7,learn_mode='marginal',sparse_target=True))
```

优化器选用 Adam，损失函数使用稀疏交叉熵，batchsize 设为 128，训练 20 轮。之后用训练好的模型，在测试集上测试即可。

4. 实验结果

（1）两层双向 LSTM 结果：

	O	B-LOC	I-LOC	B-PER	I-PER	B-ORG	I-ORG
Precision	0.9940	0.9080	**0.9226**	0.9063	0.9026	0.8131	0.8626
Recall	0.9978	0.8471	0.7927	0.8135	**0.8852**	0.7551	0.7894
F1	0.9959	0.8765	0.8527	0.8574	**0.8938**	0.7830	0.8244

	With 'O'	Without 'O'
Macro-precision	0.9013	0.8859
Macro-recall	0.8401	0.8138
Macro-f1	0.8691	0.8480

```
标签为 0 时的指标： 0.994037340815886 0.997770275149852 0.995900309954588
标签为 1 时的指标： 0.907973174366617 0.8470629127563434 0.8764610681532098
标签为 2 时的指标： 0.9226490066225166 0.7926718252162039 0.8527359529930224
标签为 3 时的指标： 0.9062676453980801 0.8134820070957932 0.8573717948717948
标签为 4 时的指标： 0.9025681758009002 0.8852246169826019 0.8938122705820661
标签为 5 时的指标： 0.8131067961165048 0.7550713749060857 0.7830151928320999
标签为 6 时的指标： 0.8625939487377144 0.7894179894179895 0.8243853025140437
平均指标为： [0.90131373 0.84010014 0.86909741]
去除O后的平均指标： [0.88585979 0.81382179 0.8479636 ]
```

（2）三层双向 LSTM 结果，效果提升不明显

	With 'O'	Without 'O'
Macro-precision	0.8863	0.8682
Macro-recall	0.8610	0.8383
Macro-f1	0.8726	0.8520

```
标签为 0 时的指标： 0.9952056306803588 0.9968901206037409 0.996047163450806
标签为 1 时的指标： 0.9107998525617398 0.8588807785888077 0.8840787119856888
标签为 2 时的指标： 0.9148712559117184 0.7924442421483842 0.8492682926829269
标签为 3 时的指标： 0.8611111111111112 0.8484541307653319 0.8547357671687515
标签为 4 时的指标： 0.8469801862019575 0.9213191378862633 0.8825870646766169
标签为 5 时的指标： 0.812111801242236 0.7858752817430503 0.7987781596029019
标签为 6 时的指标： 0.8631150573436922 0.8229276895943562 0.84254243409173
平均指标为： [0.88631356 0.8609702  0.8725768 ]
去除O后的平均指标： [0.86816488 0.83831688 0.85199841]
```