大 数 据 分 析HW\_2(2019秋)

@mails.ucas.ac.cn

Homework #2

1. 已知词项文档矩阵

*C*=

利用LSI算法，求词项与文档各自的2维表示。

**解答：**首先对词项文档矩阵C进行SVD分解，令，可得



所以，文档的2维表示为：



词项的2维表示为：



2. 1）与2）选做其中之一，并通过系统截图展示实验结果。鼓励提出并实现自己对算法的改进或创新。有改进或创新的，需在提交作业时明确说明。

1. 实现任意一种基于深度学习的命名实体识别方法，并在公开数据集上进行评估。推荐使用CoNLL-2005数据集。
2. 实现任意一种基于深度学习的有监督关系抽取方法，并在公开数据集上进行评估。推荐使用SemEval 2010 Task 8数据集。

**题目选择：**（1）基于深度学习的命名实体识别(NER)算法实现。

**模型选择：**

本人首先基于End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF论文，实现了其提出的LSTM-CNNs-CRF模型，在Conll-2003数据集上进行了训练和测试，实验结果显示该算法在该数据集上的正确率大概在91%左右。这个正确率不是太好。

在调研的过程中，了解到使用Google-BERT模型进行预处理后可以使文档特征更高效，所以尝试使用该方法进行优化。虽然BERT模型中集成了各种高效的特征提取方法，但是该算法还可以作为一个高效的预处理手段。所以在优化时，使用BERT模型替代原来的word2vec部分，之后仍然使用LSTM+CNN+CRF接在后边，模型的效果提升比较明显，模型效果最好的时候可以达到98%。

**模型说明：**

首先BiLSTM-CNNs-CRF模型的构造流程就像它的名字一样，具有非常明清晰的结构，如图1所示。首先将通过CNN处理字符级别的表示，然后将CNN的输出与词向量拼接，作为BLSTM的输入。需要注意的时BLSTM里边需要有Dropout操作，图中没有表示清楚，之后通过CRF层得出预测结果。

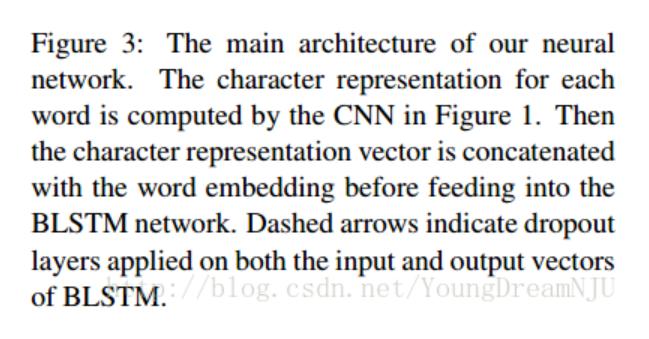
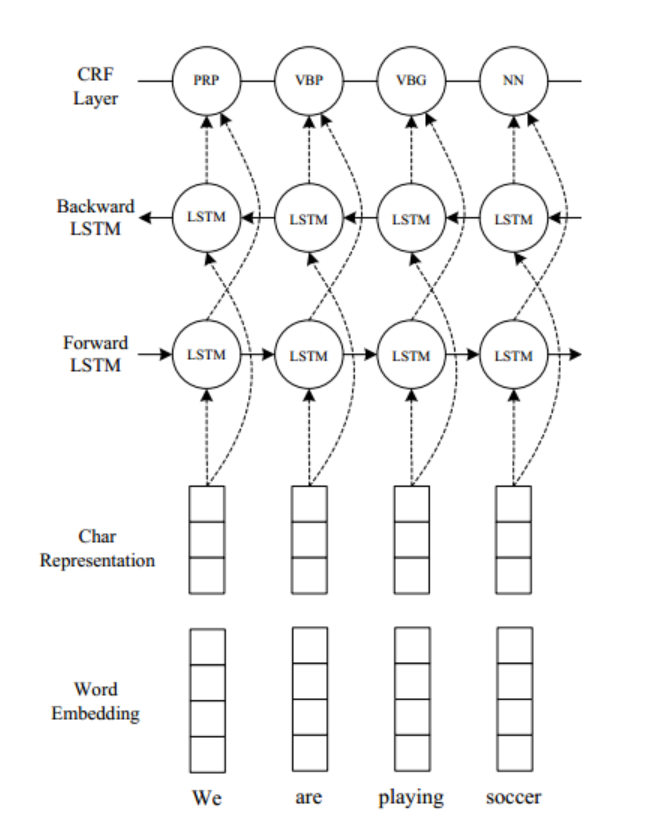


图1 BiLSTM-CNNs-CRF模型结构与训练流程

可以知道，这个模型需要的word向量需要提前训练好，这就给了我们选择的空间。了解到Google-BERT模型中有效果很好的预训练机制，所以它可以给我们提供一个很好的词向量表示。

BERT模型有很好的性质，首先训练双向语言模型时以减小的概率把少量的词替成了Mask或者另一个随机的词，其次它增加了预测下一句的loss，这就使得该模型输出的词向量表示更高效。所以这里直接用BERT模型代替前边的CNN模型和word2vec模型，其输出直接给LSTM，此时LSTM后边不需要接CRF，就可以得到想要的效果。为什么这么说呢？本身CRF层的作用就是学习标签之间的依赖关系，这一点在BERT学习过程中就可以得到，BERT的双向语言模型强制性加上了记忆属性，可以记忆标签顺序关系，所以说不用CRF应该也可以。

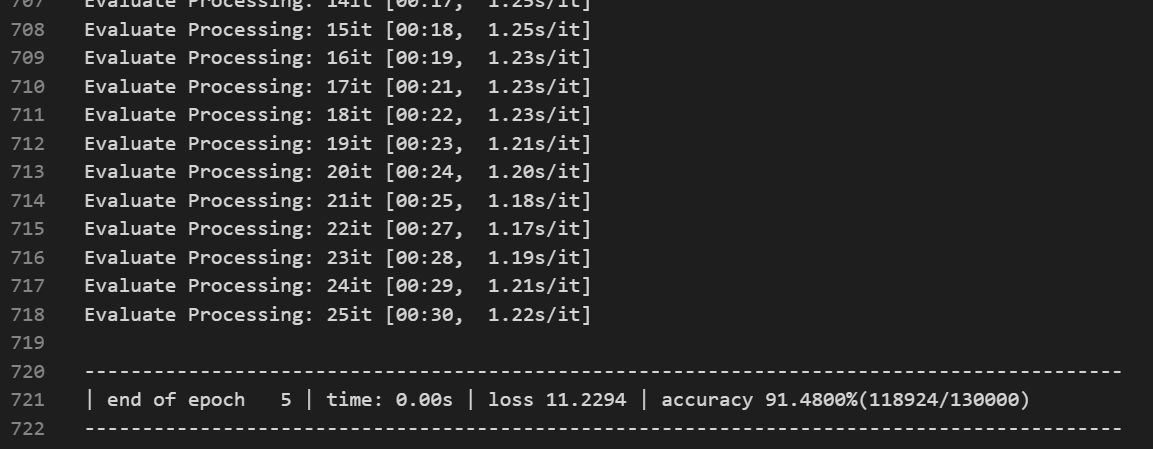
这一点在可以在试验中验证。

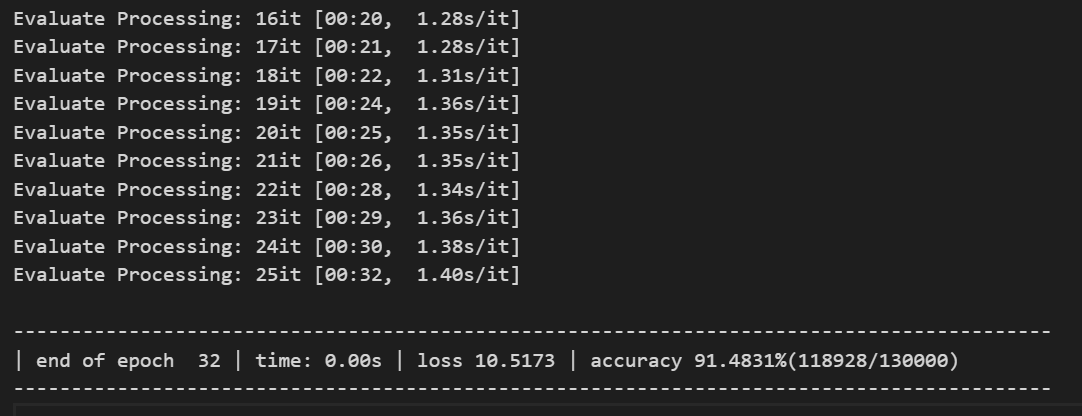
**实验结果：**

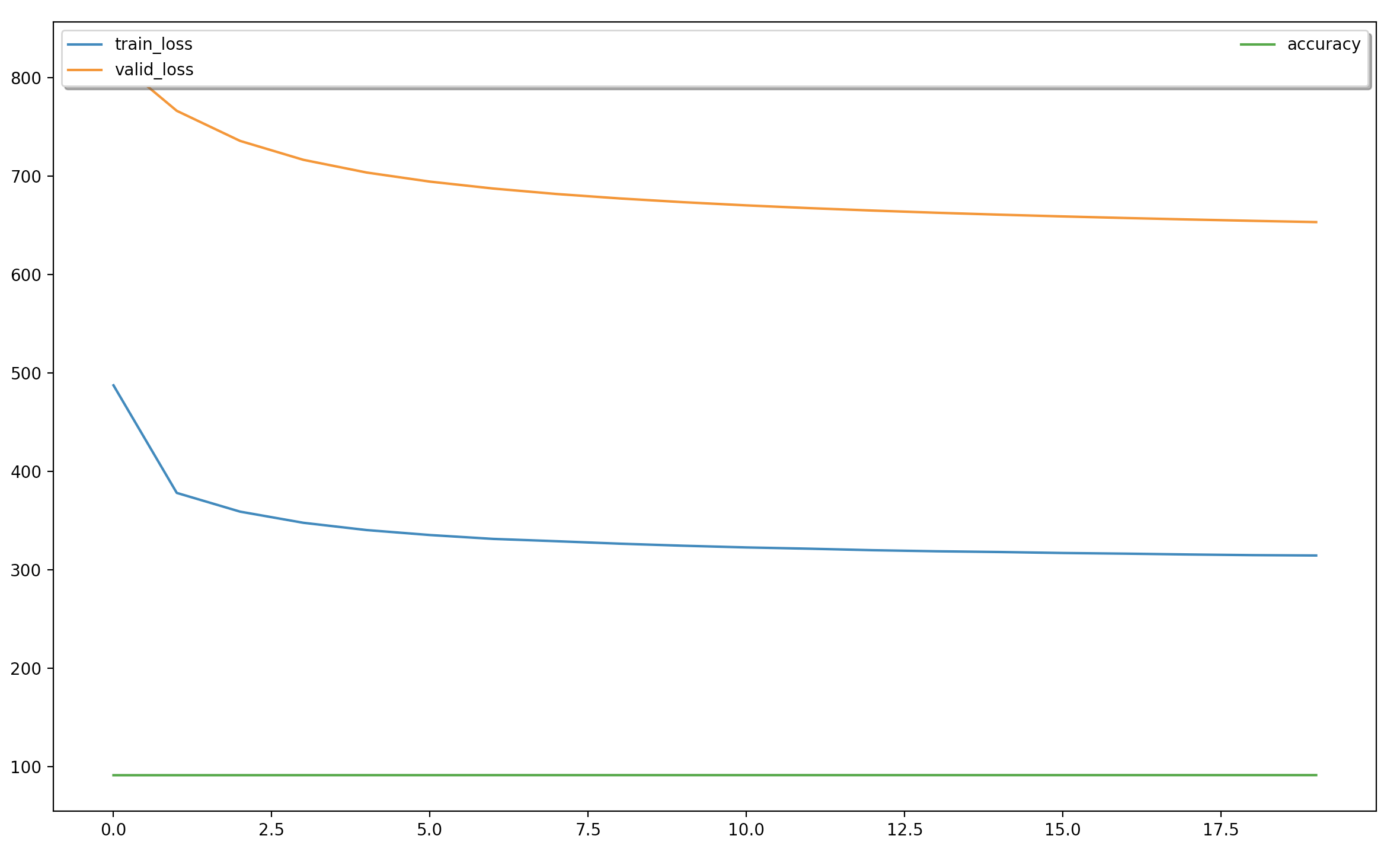
实验系统：Ubuntu16.04

实验环境：python 3.5、pytorch 0.4.0、numpy 1.13.1、tqdm

迭代过程中的几个截图：







**项目代码：**Github地址

<https://github.com/thesouther/LSTM-CNN-CRF-NER>

**主要贡献：**

本项目的主要贡献就是实现了LSTM-CNN-NER的模型代码，用BERT模型进行了改进，并且说明了在加入BERT模型后，原来模型中的CNN和CRF单元可以丢弃。