**武汉大学国家网络安全学院**

**实验报告**

**课程名称 自然语言处理**

**专业年级 2019级网络空间安全**

**姓 名 庞远心**

**学 号 2019302050244**

**协 作 者 无**

**实验学期 2021-2022　 学年 第一 学期**

**课堂时数 课外时数**

**填写时间 2021 年 10 月 24 日**

|  |
| --- |
| **实验介绍** |
| **【实验名称】：**自然语言第二次实验 |
| **【实验目的】：**   1. 了解神经网络中的基础模型：神经网络是指一系列受生物学和神经科学启发的数学模型。主要是通过对人脑的神经元网络进行抽象，构建人工神经元，并按照一定拓扑结构来建立神经元之间的连接。前馈神经网络（FNN）是最早发明的简单人工神经网络；卷积神经网络（CNN）是一种具有局部连接、权重共享等特性的深层前馈神经网络；循环神经网络（RNN）是一类具有短期记忆能力的神经网络。 2. 了解深度学习框架Pytorch的使用：PyTorch使用python作为开发语言，近年来和TensorFlow, keras, caffe等热门框架一起，成为深度学习开发的主流平台之一。PyTorch的基本元素包含张量(Tensor)、变量(Variable)、神经网络模块(nn.Module)等。 3. 了解使用深度学习解决文本分类任务基本流程：以PyTorch为例，一个常规的文本分类任务代码开发流程是：安装并导入相关的深度学习库、数据获取和预处理、定义神经网络、定义损失函数(loss function)和优化器(optimizer)、训练网络和测试网络。   **【实验环境】：**   1. python 3 2. jieba 3. PyTorch 4. NumPy 5. Sklearn 6. Tensor/boardX 7. tqdmPytorch |
| **实验内容** |
| **【实验步骤】**   * + - 1. 阅读代码，运行并得到结果。       2. 根据已有的TextCNN模型格式，写出BiLSTM模型，代码实现可参考[Chinese-Text-Classification-Pytorch](https://github.com/649453932/Chinese-Text-Classification-Pytorch)。TexcCNN模型相关超参数配置在\Chinese-Text-Classification-Pytorch \models\TextCNN.py     * + - 1. 调整batch size参数，取值分别为[8, 16, 32, 64] ，画出TextCNN训练集和验证集的loss折线图。   TextCNN的loss训练集折线图：    TextCNN的loss验证集折线图：     * + - 1. 调整embedding\_size参数，取值分别为[64, 128, 256] ，画出TextCNN训练集和验证集的loss折线图。   图例：       * + - 1. 调整hidden\_size参数，取值分别为[64, 128, 256] ，画出TextCNN训练集和验证集的loss折线图。 * hidden\_size = 64 * hidden\_size = 128   hidden\_size = 256     * + - 1. 调整learning\_rate参数，取值分别为[1e-2, 5e-3, 1e-3, 5e-4, 1e-4] ，画出TextCNN   训练集和验证集的loss折线图。       * + - 1. 调整dropout参数，取值分别为[0.1, 0.2, 0.3, 0.4, 0.5] ，画出TextCNN训练集和验证集的loss折线图。        * + - 1. 选出最好的参数组合（不局限于以上参数组合），列出参数并将实验结果写入表格  |  |  | | --- | --- | | 模型 | 准确率 | | TextCNN | 91.75% |   根据前几问画图分析和参数（kernel\_size等）的调整，只能通过控制变量法在一定范围内实验，当模型的超参数为下图的值时，拥有最高的准确率，91.75%。   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | Dropout | Num\_pochs | Batch\_size | Pad\_size |  |  |  |  | | 0.5 | 20 | 64 | 32 |  |  |  |  | | Learning\_rate | Embedding\_size | Num\_filters | 准确率 |  |  |  |  | | 1e-3 | 300 | 256 | 91.75% |  |  |  |  |   运行结果：    **【实验分析】：**  插值处理  有些epoch可能并没有计算的结果，此时的loss为零，如果画出图的话可能导致一段有值一段持续低，效果不佳。    于是选择判断，如果某点loss为零，则插值，其值为上一poch值和下一poch值的中值，如上图，epoch8、9、10处此时的loss为零，则epoch8的loss为loss（epoch7+epoch9）/2=0.48,一定程度上让折线看起来更平滑,这样处理后，当图中存在连续下降且值下降越来越慢，可考虑成这一片其实真实值为0，只用考虑看起来稍微尖锐的点，如下图绿色区域内的epoch，loss就为真实值。    （左图为处理后的尖锐点，右图为我们所认为的实际变化曲线）  对于自己的图没有意料之中的平滑的分析猜想：   1. 笔者实验每次都只改变一个超参数，其他参数仍然保持最初的状态。而咨询了其他同学，他们的模型随着超参数变量的变化，会将模型改成历史最佳的超参数组合，有记忆功能，比如在做embedding\_size的参数分析时，batch\_size已经选用了损失最少的组合，导致模型受某一超参数影响变小，其导致的loss减小，所以看起来更平滑。   **【实验总结】：**   * + - 1. 实验感想：   此次实验发现了很多知识上的盲区， 对于不同类型的分类模型，之前大致了解了理论基础，但也没有具体想过怎么实现，本次实验通过调试各个超参数，帮助笔者进一步了解了超参数和神经网络各层的作用，收获了很多关于自然语言处理的新知识，但美中不足在于自己调试的时候走了很多弯路，花了大量的时间，希望这些时间最后都能内化成有用的东西！ |
| **评语及评分（指导教师）** |
| **【评语】：**      **评分：**  **日期：** |

**附件2：**

**实验报告说明**

**1．实验名称：**要用最简练的语言反映实验的内容。

**2．实验目的**：目的要明确，要抓住重点。

**3．实验环境**：实验用的软硬件环境（配置）。

**4．实验方案设计（思路、**步骤和方法等**）**：这是实验报告极其重要的内容。包括概要设计、详细设计和核心算法说明及分析，系统开发工具等。应同时提交程序或设计电子版。

对于**设计型和综合型实验**，在上述内容基础上还应该画出流程图、设计思路和设计方法，再配以相应的文字说明。

对于**创新型实验**，还应注明其创新点、特色。

**5．实验结果分析：**即根据实验过程中所见到的现象和测得的数据，进行对比分析并做出结论（可以将部分测试结果进行截屏）。

**6．实验总结：**对本次实验的心得体会，所遇到的问题及解决方法，其他思考和建议。

**7．评语及评分：**指导教师依据学生的实际报告内容，用简练语言给出本次实验报告的评价和价值。