****

**本 科 生 实 习 报 告**

实习名称： 综合实习二

学 院： 大数据与智能工程学院

专 业： 计算机科学与技术

年 级： 2023计科专升本（1班）

指导教师： 赵喜 张宏伟

姓 名： 字源彬

学 号： 20231159025

二○二四年六月

西南林业大学教务处

**实习环节情况说明**

1.实习应拟定实习指导书或计划，实习报告应规定基本内容，拟定框架要求。

2.实习期间，学生应每日填写实习日志，内容包括：实习时间、地点、单位、内容、收获和体会，也可摘抄实习实测数据资料。实习结束后按照要求认真撰写实习报告。实习日志和实习报告在实习完成后一起交指导教师，作为实习成绩评定的依据。

3.指导教师按照实习计划的要求，根据学生的实习日志、实习报告、纪律、表现等综合评定实习成绩。

凡有下列情况之一者实习成绩为不合格：

（1）未达到实习计划规定的基本要求；

（2）实习报告混乱，分析有原则性错误；实习日志缺失，或内容大部分未完成；

（3）实习缺课三分之一以上或无故旷课三天以上；

（4）实习中严重违反纪律。

实习天数：　　6天　　实习时间：　　8:20 至 18:00

实习单位及地点：　经管楼211、215

|  |
| --- |
| 一、实习的任务、作用和目的：  **任务：**  1. 以实际动手解决某一具体领域的问题为任务导向，综合利用前导课程中所学习到的理论与技术，如数据结构、程序设计和机器学习基本方法等，并将其应用到解决实际问题的过程中来，规划并设计一个规模适中，具有一定实际应用价值的算法模型。  2. 研究评论文本的情感表达特点，提取情感相关的特征，尝试使用深度学习模型如LSTM、Transformer进行情感分类，设计一个可用于自动化产品评论情感分析的算法系统。探索DNA序列的特征提取方法，如k-mer频率、GC含量等，尝试使用朴素贝叶斯、支持向量机等经典分类算法，设计一个可用于自动识别未知DNA序列物种归属的算法系统。  3. 在实习项目完成的基础上，完成撰写实习论文的任务。  4. 使用Django完成MVT的基本操作，建立一个简单内容管理系统，包括用户注册、登录等基本功能。  5. 学习GitHub项目clone或下载，完成lookie-lookie项目的基本分析和应用。  6. 学习keras框架下进行植物叶片病害识别的基本流程，包括数据集整理、神经网络搭建与训练、训练集、验证集和测试集的划分，完成训练和测试，绘制loss-epoch和accuracy -epoch等图形。  **作用：**  整合一年来学习的专业知识，在实际项目的学习中，围绕信息系统建设、机器学习等具体项目，提升专业应用能力，包括环境配置、框架应用、代码实现和报告撰写等。  **目的：**  学习以Python语言为基础的Web框架和深度学习框架，掌握简单CMS系统搭建的基本概念和方法，熟悉深度学习训练和测试的基本流程，为毕业设计地顺利开展打下扎实的专业基础。 |
| 二、实习主要内容（含工具、方法等）：  1. 使用Python的数据处理库（如Pandas、NumPy），收集并清洗文本数据。  2. 使用scikit-learn等机器学习库提取文本特征，如词袋模型、TF-IDF、Word2Vec等。  3. 使用scikit-learn提供的经典机器学习模型，如逻辑回归、朴素贝叶斯、SVM等，尝试使用深度学习模型,如TensorFlow或PyTorch实现的LSTM、Transformer模型，采用交叉验证、网格搜索等方法优化模型参数。  4. 完成Django下载并配置运行环境，实现多用户注册和登录。  5. 学习lookie-lookie项目，在各种物理环境下进行测试。下载 UnityEyes（windows）程序，使用该程序生成合成数据样本，分析项目自带程序visualize.py，对每一行代码进行注释说明该行代码的作用。  6. 建立虚拟环境，配置tensorflow（keras）环境，获取数据集、数据集分析处理、构建神经网络模型、测试一条数据样本、激活并输出特征图、配置训练器、根据训练结果绘制训练和验证准确性（accuracy）图和损失图、测试模型、输出结果。 |

|  |  |
| --- | --- |
| 2024/6/17--2024/6/19  利用情感分析方法，对电影评论进行文字识别，识别结果为正面评论或负面评论，具体步骤如下：  (1)**导入包与数据** 从Keras中导入IMDb数据集。IMDb数据集是从IMDb网站收集的电影评论，一共有50 000条数据，分为训练集和测试集（各25 000条），其中每条影评的文字内容都已标记为正面评  价或负面评价。  (2)**序列填充** 对训练集和测试集的特征值分别进行序列填充。序列填充的目的是将不同长 度的文本序列调整为相同长度，以满足如下要求： 1. 输入要求：某些模型和算法对输入序列的长度有限制。例如，循环神经网络 （RNN）的输入序列长度必须是固定的，而卷积神经网络（CNN）的输入需要 是相同形状的张量。通过进行序列填充，可以满足模型对输入长度的要求。 2. 批处理效率：在进行批处理时，通常要求将相同长度的样本组成一个批次进 行并行计算。如果输入序列长度不同，会导致批次中的样本无法对齐，从而无法 高效地进行并行计算。通过进行序列填充，可以确保批次中的样本长度相同，提 高计算效率。3 3. 数据对齐：在某些任务中，如机器翻译或文本分类，需要对齐不同长度的输 入序列。通过将输入序列进行填充，可以在相同位置对齐词语或字符，使得模型 能够更好地学习序列之间的关系。 4. 特征表示：在使用词嵌入（Word Embedding）等技术将文本转换为向量表示 时，要求输入序列的长度相同。填充可以确保输入序列的长度一致，使得每个样 本都能以相同的方式表示为固定长度的向量。  (3)**构建长短期记忆神经网络模型** 使用Keras的顺序模型构建长短期记忆神经网络模型。影评情感分析网络模型采用长短 期记忆神经网络，它包括一个嵌入层、一个长短期记忆层、两个Dropout层和一个输出层。 主要步骤为： 1. 构建空的顺序模型。 2. 为网络模型添加嵌入层，词汇表的大小为 4 000，转换后输出向量的长度为 32，输入序列的长度为 400。 3. 为网络模型添加 Dropout 层，丢弃概率为 0.3。 4. 为网络模型添加长短期记忆层，输出的维度为 32。 5. 为网络模型添加 Dropout 层，丢弃概率为 0.3。 6. 为网络模型添加隐藏层作为输出层，结点个数为 1，使用 Sigmoid 函数作为 激活函数。 7. 使用 summary()函数，显示模型各层的参数信息。  (4)**训练模型** 编译、训练和评估长短期记忆神经网络模型。该步骤包含以下的主要内容： 1. 编译长短期记忆神经网络模型，其中，优化器使用 RMSprop 优化器，损失 函数使用二元交叉熵损失函数，性能评估函数使用准确率函数。 2. 训练长短期记忆神经网络模型，并将训练结果保存在 history 中。其中，设置 批量的大小为 64，迭代次数为 10，验证集的数据占比为 0.2。 3. 使用测试集数据对长短期记忆神经网络模型进行评估，其中，设置批量的大 小为 64，日志显示模式为 2。 (5) **可视化训练的结果** 绘制训练集和验证集的损失函数值和准确率的变化曲线。该步骤的主要内容如下： 1. 读取 history 的 history 属性，分别将训练集损失函数值赋值给变量 loss、训 练集准确率赋值给变量 acc、验证集损失函数值赋值给变量 val\_loss、验证集准确 率赋值给变量 val\_acc。 2. 创建子图，在子图 1 中绘制损失函数值的折线图。 3. 在子图 2 中绘制准确率的折线图。  (6)**应用模型** 从影评网站复制影评信息，应用网络模型，显示预测结果。主要内容如下： 1. 定义字典变量 dict，键值为 0 表示正面评论，键值为 1 表示负面评论。 2. 定义情感分析预测函数，参数为评论字符串。 a. 定义分词器对象，对评论字符串分词并数字化 b. 将数字化的词汇转换为向量序列 c. 对向量序列分别进行序列填充，填充方式为'post'，剪裁方式为'post'， maxlen 为 400 d. 使用长短期记忆神经网络模型预测文本评论 e. 显示评论和预测结果。5 3. 访问 https://www.imdb.com，复制一段文字影评，将其作为字符串赋值给变 量 test\_text，此处复制一段满分的影评赋值给变量 test\_text。 4. 调用情感分析预测函数，显示评论和预测结果。  (7) 使用其他方法与上述方法进行比较  2024/6/20  1.完成Django下载并配置运行环境    在pychar里面创建好项目。在菜单栏下面的文件菜单选择Settings，再定位到自己的项目，搜索Django安装。如上图所示。  2.完成迁移数据库（Model）建立      在models.py文件中创建model模型，数据库默认使用Django提供的****SQLite数据库。****    利用py manage.py makemigrations和py. mange.py migrate 命令来创建迁移文件和实现迁移。  3.实现简单模板（Template），实现基本业务逻辑（View），比如增加笔记主题和相应条目内容    在项目的settings.py中添加你的应用名到INSTALLED\_APPS列表中    创建模型    创建视图      创建url和模板。最终效果如下所示：    创建的学习日志首页。配置不同的url条目来实现不同页面之间的跳转。    增加学习日志。    在日志中增加新的内容。    修改日志中的内容。  4 4.实现多用户注册和登录    在应用中定义url路由。   在主URL配置中包含应用的URL   使用Django的内置User模型和auth框架来实现注册和登录功能。    为登录和注册应用模板，最后使用默认的数据库**SQLite，数据库的使用在上述第3点已经说明。**      登录和注册效果如上所示。  2024/6/21  1.尽自己最大努力，训练到最好的精确度（可在各种情况下训练或测试，眼镜、遮挡、光线等），截屏并简要分析说明写入实习报告    该模型通过摄像头，获取到人的信息，然后定位出人的面部信息，从面部信息中截取出人的眼睛，通过截取眼睛的图像，传递给我们的模型，模型主要截取出来眼睛位置的坐标信息，最后实现一个预测。该模型地数据集是我们自己通过客户端自己了收集的，所以它会受到各种因素的影响。戴眼镜和不戴眼镜、光线强弱、人的坐姿等都会影响我们数据的准确率，从而影响最终的预测效果。  2. 回答以下问题：  lookie-lookie项目中每个文件起什么作用？  答：clmtrackr.js文件检测和定位人脸，dataset.js保存我们收集的数据，facetracker.js实现脸部追踪，globals.js放置全局的方法，heat.js绘制热图，main.js处理事件交互，主要是一些事件监听器。mouse.js移动鼠标来收集数据，training.js训练集数据，ui.js更新页面的元素和显示信息。  调用摄像头对应代码是什么？  $(document).ready(function() {  const video = $('#webcam')[0];  const overlay = $('#overlay')[0];  const overlayCC = overlay.getContext('2d');  const ctrack = new clm.tracker();  ctrack.init();  function getEyesRectangle(positions) {  const minX = positions[23][0] - 5;  const maxX = positions[28][0] + 5;  const minY = positions[24][1] - 5;  const maxY = positions[26][1] + 5;  const width = maxX - minX;  const height = maxY - minY;  return [minX, minY, width, height];  }  调用摄像头代码如上所示。  对眼睛部分的截屏图像和对应的（x，y）屏幕坐标存在哪里？  答：我们可以简单地将所有示例存储在一个大**的训练集中**，但在机器学习中，要确保你的模型不仅仅是在学习数据。这就是为什么我们应该将一些示例保存在单独的**验证集中**。然后，我们可以检查模型在看不见的数据上的表现，并确保它没有过度拟合到训练集。为此，我拆分了 20% 进入验证集。  神经网络模型由几层构成，分别起什么作用？  答：这是一个顺序模型，第一个是二维卷积层，从输入的数据中提取有用的特征；接下来是最大池化层，用来降低计算复杂度；再添加一个丢弃层来防止过拟合；最后是一个全连接层，输出x和y坐标。  这个项目属于2D眼动跟踪还是属于3D眼动跟踪，为什么？  答：2D；对于面部识别或者提取出眼睛部分，我们可能更多的是去绘制一些点和线或者是一些二维图形，这些通过2D渲染都可以准确实现。  将回答写入实习报告  3.下载UnityEyes（windows）程序（226MB），使用该程序生成合成数据样本，分析项目自带程序visualize.py，对每一行代码进行注释说明该行代码的作用    2024/6/24   1. 建立虚拟环境，配置tensorflow（keras）环境，从Kaggle下载“植物病害数据集（比赛专用）”，分别是2020版（823.79MB）和2021版（16.1GB）   安装Anaconda,并且新建立一个虚拟环境，如下所示：      在pycharm里面创建新项目，并且环境应用为新建的condan环境，如下所示：     1. 获取数据集、数据集分析处理、构建神经网络模型、测试一条数据样本、激活并输出特征图、配置训练器、根据训练结果绘制训练和验证准确性（accuracy）图和损失图、测试模型、输出结果。可参考 [https://www.cnblogs.com/MuBenYou/p/16996088.html](https://www.cnblogs.com/MuBenYou/p/16996088.html" \t "_blank)   数据集从Kaggle下载“植物病害数据集（比赛专用）”，2020版（823.79MB），代码参考[https://www.cnblogs.com/MuBenYou/p/16996088.html](https://www.cnblogs.com/MuBenYou/p/16996088.html" \t "_blank) 网站，稍微做了一些改动，如下所示：    手动计算迭代器中的测试数。  折线图和训练次数如下所示：        实习总结：通过这次实习，首先，我掌握了利用Django框架开发网站，能够实现登录、注册操作，并且能够对数据库做一些简单的增删查改。其次，对计算机视觉问题有了一定的了解，通过实际使用他人开发的lookie-lookie项目，既通过网络摄像头来判断用户处于网站的哪个位置。最后，通过一个“植物病害”的一个项目，理解了机器学习的实现流程，特别是了解了数据集的好坏对于识别结果的影响。这次实习不局限于一个方面，而是通过多个方面，不同的项目，不同的角度来培养我们的实际操作能力。 | |
|  | 教师评语：  实习态度认真，按照实习计划和实习要求，完成相应实习内容。  成绩：  指导教师（签名）：C:\Users\Administrator\AppData\Local\Microsoft\Windows\INetCache\Content.Word\赵喜-张宏伟-签名.jpg  2024年7月10日 |