项目设计报告

姓 名： 王琮皓

班 级： 计算机试验班61

学 号： 2160512008

2020年5月5日

1. 研究背景

本项目设计的选题为“面向知识森林的示意图装配算法研究及其应用”。西安交通大学智能网络与网络安全实验室提出了知识森林的概念[1]，其组织结构为学科-课程-主题-分面树-碎片化知识，如图1所示。

地图的截图

描述已自动生成

图 1 知识森林概览[1]

在知识森林的多级结构中，最上层是计算机科学、数学等学科，第二层是学科类别下的不同课程，图1中的是计算机科学中数据结构课程的知识森林体系。森林中的每一棵树都是一个主题，可以理解为课程中的主要知识点，它们相互关联，又有着独立的分面树。

地图上有字

描述已自动生成

图 2 分面树概览[1]

分面树的结构如图2所示，其中“特点”、“操作”、“定义”等是一级分面，“出栈”、“入栈”等是一级分面下的二级分面。分面树一定存在一级分面，但不是每个一级分面都有二级分面，更多级分面在本项目中不予考虑。最末端的叶子部分是碎片化知识，一般分为文本、图像、视频、音频等不同介质。本项目的目的是将示意图碎片化知识装配正确装配至主题-一（二）级分面下。但示意图一般伴随着文本碎片化知识的出现而出现，所以除了针对示意图的图像处理，文本处理也是项目中不可或缺的一环。

1. 相关研究

2.1 图像处理

当系统获得示意图碎片化知识时，首先要尝试理解图像语义，才能分析其所属的主题-分面类别，所以针对示意图的图像处理是整个算法过程中最为重要的一步。由于后续流程将结合图像与其他介质中包含的信息，跨模态地决定碎片化知识类别，这里我采用提取图像特征的方式处理收集到的示意图，方便后期处理。目前用于提取图像特征的方法有很多，传统方法有颜色、纹理、形状特征提取。这些方法可以处理自然图像，但对于教育领域的示意图来说，它们的处理能力较为有限。深度学习方法快速发展以来，基于卷积神经网络（CNN）的图像深度特征提取方法成功在精确程度上超过了传统特征工程方法，因此本研究采用提取示意图深度特征的方法处理图像。

常用的图像处理CNN模型有VGG[2], Inception [3], ResNet[4]等。早期研究工作中，由于认为CNN网络越深，结构越复杂，表达能力就越强，VGG、Inception等模型的规模越来越大，参数也越来越多。但事实证明，模型的复杂度高到了一定程度后，一味地得加卷积层数并不能不断提高模型的分析能力，反而会增加训练成本，大幅延长网络收敛所需时间。ResNet创新性地将残差学习模块（图3）与CNN结合，解决了复杂网络模型训练难度大的问题的同时，进一步提高了模型的分析处理能力。因此，本项目将采用ResNet预训练模型提取图像深度特征。

手机屏幕截图

描述已自动生成

图 3 残差学习模块[4]

2.2 文本处理

由于示意图数据通常伴随着起到说明作用的文本数据出现，并且教育领域的示意图并不属于自然图像，其深度特征的语义表征能力十分有限，在本项目中，也将进行相关文本内容的处理。

在文本特征提取工作中，我重点使用了Google提出的语言模型BERT[5]，BERT模型站在巨人的肩膀上，融合了Attention机制与Transformer模型，是当前自然语言处理领域较为前沿的算法。该模型目前已支持多种自然语言处理，并可调用预训练模型实现不同任务，如Masked Language Model, Next Sentence Prediction，也可以通过微调模型定制化地实现其他目的，在各个领域有着创新性的应用。

手机屏幕截图

描述已自动生成

图 4 基于BERT的预训练与微调应用[5]

1. 研究内容

3.1 数据集的构建

项目最初尝试过自行组建数据集：借助爬虫API从谷歌/百度图片下载数据，但这样就需要自行标注示意图所属的主题-分面信息，相当于从头构建一个知识森林。工作量大且人工标注准确率无法保证，所以考虑借助实验室现有的YOTTA数据库。

YOTTA数据库提供了丰富的数据和功能完善的API，借助API可以查询课程下的主题，下载主题下包含所有碎片信息的json格式数据文件。如图5所示，这些文件中记录了碎片化知识来源的URL，示意图URL以HTML格式存在img标签中，也记录了碎片ID，文本内容，所属分面信息。借助Python正则表达式提取URL和assembleText，保存png图片和txt文件，实现示意图的下载和相关文本内容的提取。

手机屏幕截图

描述已自动生成

图 5 数据文件示例

3.2 图像特征提取

Pytorch中提供了Resnet50模型，这里直接加载在ImageNet上训练过的模型。其结构如图6所示，Stage4的输出维度是(2048, 7, 7)，经过AveragePooling2D层变为(2048, 1, 1)，经过Flatten层变为(2048)，在该层后面接一个线性连接层，相当于对每个神经节点做f(x)=x的变换，就可以导出长度为2048的一维特征数组。这样就将模型原本全连接层前的数据作为示意图深度特征导出，保存在txt文件中。

图片包含 文字, 游戏机

描述已自动生成

图 6 Resnet50模型网络结构[6]

3.3 文本特征提取

使用Google的BERT自然语言处理模型，完整的模型包含BertTokenizer类（文本的编码、预处理），BertModel类（存放模型），和BertConfig类（存放预训练参数）。该模型根据需要处理的语言种类不同，有很多分支。因为本项目文本内容有英文也有中文，所以选择支持多种语言的bert-base-multilingual-cased分支。输入的文本借助BertTokenizer预处理成图7所示形式，通过BertModel和BertConfig分别获取模型和预训练参数，构建完整的预训练模型，即可提取文本特征。

图片包含 游戏机, 键盘

描述已自动生成

图 7 BERT输入转换形式

1. 部分结果展示

4.1 数据集

由于大量示意图URL已失效，或者为国外网址，不能正确下载。同步下载示意图和文本内容会导致只有文本文件没有配套示意图，对于研究示意图装配算法没有意义，还要手动对比删除。这里先只下载示意图，文件命名为assembleId。然后借助os.listdir(dir\_name)将成功下载的图片名称集合成assemble\_list，只捕捉列表中碎片的assembleText项，保存为txt文件。

最终数据的组织形式是课程文件夹（图8）包含所有的主题文件夹（图9），主题文件夹内是原始数据json文件，assemble list，以及示意图、文本文件夹。

电脑萤幕画面

描述已自动生成

图 8 课程文件夹

电脑萤幕的截图

描述已自动生成

图 9 主题文件夹

4.2 提取到的特征

电脑萤幕的截图

描述已自动生成

图 10 示意图深度特征

手机屏幕截图

描述已自动生成

图 11 文本特征示例

1. 问题总结

5.1 批量提取特征

数据库中一门课程的主题类别很多，按照上面介绍的数据组织形式，需要不停地更改路径才能完成全部的特征提取工作。为了方便，使用os.walk和glob文件名模式匹配，只需输入课程文件夹（如，数据结构人工）的路径，即可遍历获取所有图片和文本文件路径，装进files\_list列表。

手机屏幕截图

描述已自动生成

图 12 部分代码

5.2 提取文本特征时Token超长

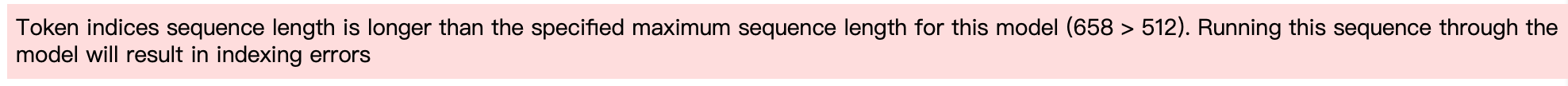


图 13 报错截图

这个问题是由于BERT模型支持的最大句子长度为512，而数据中的文本有些超长了。出现这样问题的一般是一些代码组成的文本，自然语言处理模型可能无法正确处理它们的语义，并且这样的文件不多，我的处理方法是简单的跳过这些碎片。

5.3 Tensor类型出错

问题形式：Expected tensor for argument #1 ‘indices’ to have scalar type Long; but got CUDAFloat

尝试按照报错提示，用torch.LongTensor()定义变量，没能解决这个问题。接下来本项目借助了TorchSnooper这个工具[]，实时监控每个tensor的状态。发现有些碎片只有示意图，无相关说明文本，空的txt文件提取文本特征也没有意义。但在尝试忽略空文本文件时，使用空格识别，而空文档的内容实际是‘\xa0’，就会出现这个错误。使用‘\xa0’判定后，成功解决了这个问题。

手机屏幕截图

描述已自动生成

图 14 TorchSnooper示例

1. 后续计划
   1. 训练跨模态示意图装配模型；
   2. 酌情应用OCR技术捕捉示意图中的文本内容，丰富数据的特征表示；
   3. 调整特征提取的相关参数优化装配效果；
   4. 撰写毕业论文。
2. 参考文献

[1] 西安交通大学智能网络与网络安全教育部重点实验室. 知识森林数据库[OL]. http://yotta.xjtushilei.com:888/yotta/module/construct/pages/kg/index.html. 2017.

[2] Simonyan, Karen, Zisserman, Andrew. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer ence. 2014.

[3] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.

[4] He K , Zhang X , Ren S , et al. Deep Residual Learning for Image Recognition[J]. 2015.

[5] Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

[6] aptx9688. ResNet50网络架构图[OL]. https://www.bilibili.com/read/cv2051292. 2019.

[7] Xiang Gao. TorchSnooper[OL]. https://github.com/zasdfgbnm/TorchSnooper. 2019.

附录 源代码

<https://github.com/nomoresomethingwentwrong/Diploma-Project/tree/master/code>