人工智能导论实验报告

——基于 ArXiv 数据集进行分析和模型训练

彭培烜 2018202201

一、 概述

1. 数据背景

ArXiv 是用于搜索，浏览和下载学术论文的重要工具，为公众和研究社区提供了开放获取学术论文的服务，涉足物理学和计算机科学的众多领域。为了让 ArXiv 可用度更高，康奈尔大学在 Kaggle 上创建了一个免费、开放的数据集，包含了 170 多万学术论文。

2. 实验成果

本次实验基于 ArXiv 数据集，完成和实现了以下任务与功能：

（1） 数据处理，预训练了词向量；

（2） 实现了自然语言问答；

（3） 实现了标题归纳。

二、 基于 word2vec 训练词向量

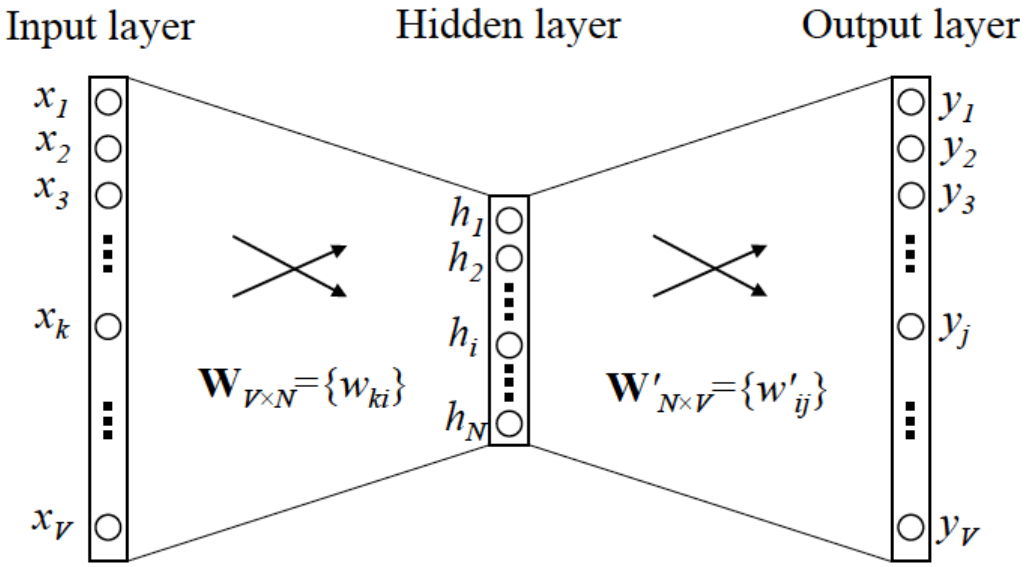
1. Word2vec

Word2vec 是一群用来产生词向量的相关模型，用来预测文本中单词的关联度大小。其工作步骤如下：

（1） 在输入层，一个词被转换为 One-Hot 向量；

（2） 在第一个隐层，输入一个$W \times x + b$，做一个线性模型，此时是一个简单的映射，相当于一个线性回归函数；

（3） 第三层可以看成一个分类器，用 Softmax 进行回归，最后输出每个词对应的概率。



2. 具体实现

此次实验的数据对象为 ArXiv 论文集中论文的摘要(Abstract)，首先需要去除句子中不是字母或数字的字符，简写以及停用词。

```
def clean_sentence(val):
    "remove chars that are not letters or numbers, downcase, then remove stop words"
    regex = re.compile('([^\s\w]|_|\r|\n)+')
    sentence = regex.sub('', val).lower()
    sentence = sentence.split(" ")

    for word in list(sentence):
        if word in STOP_WORDS:
            sentence.remove(word)

    sentence = " ".join(sentence)
    return sentence
```

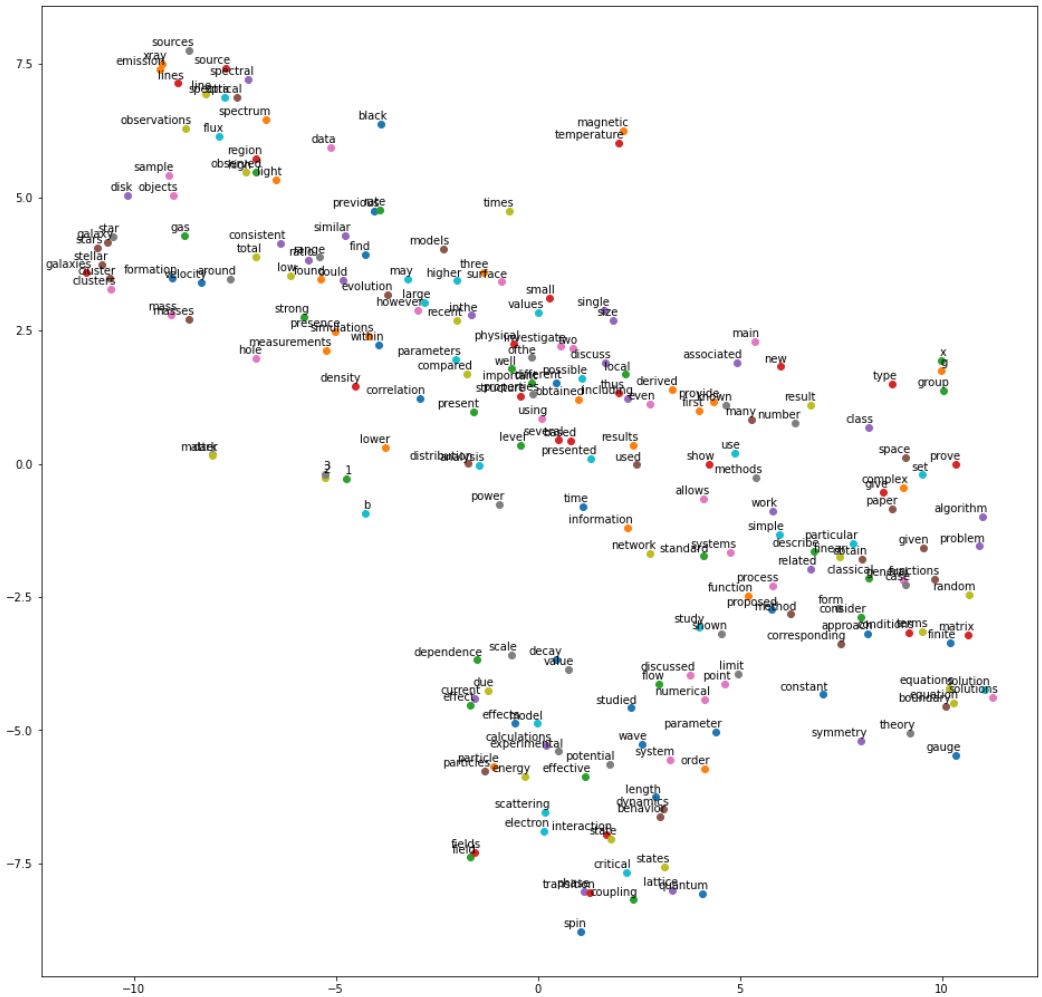
然后通过 word2vec 进行训练

```
model = word2vec.Word2Vec(abstracts, size=100, window=20, min_count=200, workers=4)
```

最后将结果进行可视化处理，即放入坐标系。

3. 运行结果

此次训练处理一百多个单词，进行可视化后如下图



三、 自然语言问答

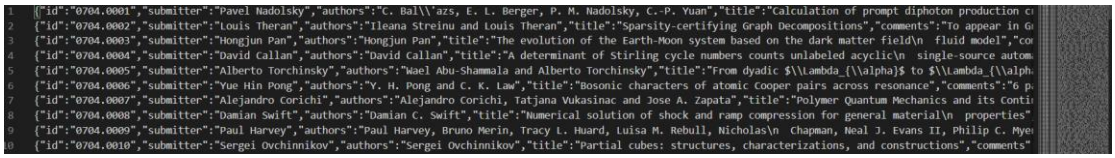
1. 数据属性

论文集有以下属性

ID	ArXiv ID
Submitter	提供者
Authors	作者
Title	标题
Comments	其他信息（如页数和数字）
Journal-Ref	期刊
Abstract	摘要
Categories	分类标签
Version	版本

该模块的实现主要使用了以上的 Title 和 Abstract 属性。

数据截图：



2. Haystack 以及模型训练

自然语言问答，即以人们平常交流的提问题方式对系统进行提问，比如

"*What do we know about Bourin and Uchiyama?*"

"*How is structure of event horizon linked with Morse theory?*"

然后系统通过搜索整合文档库中的数据，给出回答。

为实现该功能，我使用了 Haystack 框架，这是 GitHub 上的一个开源库 (https://github.com/deepset-ai/haystack)，该框架主要由三个部分组成：

Document Store：用于存储搜索文档的数据库，其主要推荐的是 ElasticSearch。

Retriever：从大量文档中识别候选段落的快速，简单的算法，包括 TF-IDF 和 BM25 等，有助于将 Reader 的范围缩小到可以回答给定问题的较小文本单元。

Reader：神经模型，用于从文本中寻找答案。

其中最关键的部分在于 Reader 的训练，在这里我使用了 deepset/roberta-base-squad2 模型进行训练。

```
reader = TransformersReader(model_name_or_path='deepset/roberta-base-squad2',
                            tokenizer='deepset/roberta-base-squad2',
                            context_window_size=500, use_gpu=-1)
```

经过训练后，就能给出答案了。

```
qes = input('Question: ')
# print(qes)
prediction = finder.get_answers(question=qes, top_k_retriever=2, top_k_reader=3)
print_answers(prediction, details='minimal')
```

3. 结果演示

给出问题如下：

```
Question: How can we enhance sensitivity to the signal?
```

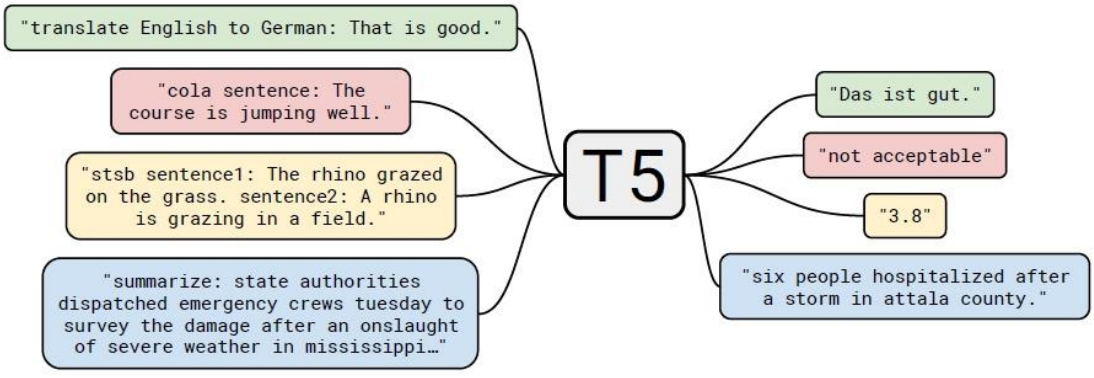
给出回答：

```
{   'answer': 'in frequency regimes',
    'context': '  Circular dichroism (CD) is in widespread use as a means '
               'of determining\n'
               'enantiomeric excess. We show how slow-light phenomena in '
               'dispersive structured\n'
               'media allow for a reduction in the required optical path '
               'length of an order of\n'
               'magnitude. The same ideas may be used to enhance the '
               'sensitivity of CD\n'
               'measurements while maintaining the same optical path '
               'length through the sample.\n'
               'Finally, the sensitivity may be enhanced in frequency '
               'regimes where CD data is\n'
               'typically not accessible due to a modest chiral response '
               'of the enantiomers.\n'},
```

四、 标题归纳

1. T5 模型

这个模块通过对给定摘要(Abstract)进行分析，然后归纳总结出一个契合的标题。是一个“文本到文本”的任务，因此通过 T5 模型进行训练。



T5 主要包含模型结构、注意力掩码机制、Objectives、C4 和训练策略五部分。

其模型结构为 Encoder-Decoder 结构，注意力掩码机制为 Fully-visible， Objectives 包含常用预训练方法、Mask 策略、Mask 比率和 Span 长度四个部分。

此次训练使用了 2017-2020 年的论文数据

```
for paper in metadata:
    paper_dict = json.loads(paper)
    ref = paper_dict.get('journal-ref')
    try:
        year = int(ref[-4:])
        if 2016 < year < 2021:
            years.append(year)
            titles.append(paper_dict.get('title'))
            abstracts.append(paper_dict.get('abstract'))
    except:
        pass
```

然后进行模型参数配置和训练

```
model_args = {
    "reprocess_input_data": True,
    "overwrite_output_dir": True,
    "max_seq_length": 512,
    "train_batch_size": 16,
    "num_train_epochs": 4,
}

# model = Seq2SeqModel(encoder_decoder_type="bart",
                       # encoder_decoder_name="facebook/bart-base",
                       # args=model_args)
model = T5Model("t5-small", args=model_args, use_cuda=True)
model.train_model(train_df)
results = model.eval_model(eval_df)
```

标题归纳系统便完成了

2. 结果演示

```
Actual Title: Computational intelligence for qualitative coaching diagnostics: Automated assessment of tennis swings to improve performance and safety

Predicted Title: ['Personalized qualitative feedback for tennis swing technique using 3D video video']

Actual Abstract: ['summarize:  Coaching technology, wearables and exergames can provide quantitative
feedback based on measured activity, but there is little evidence of
qualitative feedback to aid technique improvement. To achieve personalised
qualitative feedback, we demonstrated a proof-of-concept prototype combining
kinesiology and computational intelligence that could help improving tennis
swing technique utilising three-dimensional tennis motion data acquired from
multi-camera video. Expert data labelling relied on virtual 3D stick figure
replay. Diverse assessment criteria for novice to intermediate skill levels and
configurable coaching scenarios matched with a variety of tennis swings (22
backhands and 21 forehands), included good technique and common errors. A set
of selected coaching rules was transferred to adaptive assessment modules able
to learn from data, evolve their internal structures and produce autonomous
personalised feedback including verbal cues over virtual camera 3D replay and
an end-of-session progress report. The prototype demonstrated autonomous
assessment on future data based on learning from prior examples, aligned with
skill level, flexible coaching scenarios and coaching rules. The generated
intuitive diagnostic feedback consisted of elements of safety and performance
for tennis swing technique, where each swing sample was compared with the
expert. For safety aspects of the relative swing width, the prototype showed
improved assessment ...']
```

五、 总结

为了完成此次项目，用到了许多深度训练模型，在开始阶段，由于不熟悉相关代码框架结构，遇到了许多困难，比如总时报错，无法运行以及内存不足等问题。经过不断地查阅资料，逐一解决。此外，我通过完成此次项目，了解和学习到了许多具体的模型比如 word2vec，RoBERt 等，为以后进一步学习打下了基础。