基于 Tensorflow 的 Text Summarizaion 模型自动生成新闻标题

田 江 童薇羽

(景德镇陶瓷大学 江西 景德镇 333000)

【摘 要】随着人工智能的快速兴起，Google 发布的深度学习框架 TensorFlow 在短短两年内，就成为了当前最流行的深度学习项目。在图像处理、音频处理、自然语言处理和推荐系统等场景中，TensorFlow 都有着丰富的应用。虽然开源没多久，但是 TensorFlow 正在快速的参与到我们的工作生活当中。

【关键词】Tensorflow; Text Summarizaion; 自动 生成

一、研究背景

随着互联网的迅速发展，网络中的新闻资源呈指数级增长；在众多的新闻中，如何让用户又快有好的阅读到自己感兴趣的新闻资讯成为了当下的研究热点；本文为某新闻企业通过接入智能推荐系统，在其 APP 端增加智能推荐模块，就能为 APP 用户私人订制感兴趣的新闻。

二、研究方案

文本自动总结的模型一直都是深度学习中的研究热点。有一些诸如 TFIDF 和 TextRank 之类常规算法，其基本原理是直接抽取文本中重要的句子。目前常用的模型是 seq2seq，它是基于 Encoder－Decoder 的一个结构，首先将原始文本中的句子 encode 成一个固定大小的向量，然后通过 decoder 部分一个字符一个字符生成目标句子。

Tensor 意味着数据，Flow 意味着流动、计算和映射，这也体现出数据是有向的流动、计算和映射。TensorFlow 的结构由会话（session），图（graph），节点（operation）和边（tensor）组成，它使用图（graph）来表示计算任务，图在被称之为会话（Session）的上下文（context）中执行，其状态是通过变量（Variable）来维护的，使用 feed 和 fetch 可以为任意的操作（arbitrary operation）赋值或者从其中获取数据。

这篇文章中我们将采用基于 Tensorflow 的 Seq2seq + Attention 模型，训练一个新闻标题自动生成模型。加入 Attention 注意力分配机制，是为了使 Decoder 在生成新的目标句子时，可以得到前面 Encoder 编码阶段每个字符隐藏层的信息向量，提高生成目标序列的准确度。

三、数据处理

样本数据为某企业新闻客户端 2016 年 11 月份的新闻，超过 10M 的语料数据，包含新闻标题和新闻正文信息。由于在 Encoder 编码阶段处理的信息会直接影响到整个模型的效果，所以对新闻数据的预处理工作需要非常细致。对新闻中的特殊字符、日期、英文、数字以及链接都要进行替换处理。

文本预处理后，就是训练样本的准备工作。这里的 Source 序列，就是新闻的正文内容，待预测的 Target 目标序列是新闻标题。为了保证效果，正文部分不能过长，这里设定分词后的正文不超过 100 个词，不足用 PAD 字符补齐，设定标题不超过 20 个词。在生成训练样本的时候，定义了 create_ vocabulary（） 方法来创建词典，data_ to_ id（） 方法把训练样本（train_ data. txt）转化为对应的词 ID。

```
# train_data.txt
故宫 长城 央视 大楼 鸟巢 水立方 有名 地方 不胜枚举 地界 老百姓 生活 相关 市井 本土 北京 胡同 推荐 北京 胡同 游 攻略 瞧瞧 北京 普通百姓 生活
北京 是因为 它作 多年 首都 作 这么久 首都 全 是因为 北京 优秀 全赖 800 年前 千万别 张国字 脸 迷惑 他本 名叫 完颜 迪古 金朝 第四位 皇帝 历史

# train_title.txt
北京 什么 地方 最 出名
不用 再 思考 为什么 来 北京
```

四、算法解析

Seq2Seq 是一个基于输入的 sequence，预测一个未知 sequence 的模型。模型由 Encoder 编码阶段和 Decoder 解码阶段两部分构成。模型编码阶段 Encoder 的 RNN 每次会输入一个字符代表的向量，将输入序列编码成一个固定长度的向量；解码阶段的 RNN 会一个一个字符地解码，如预测为 X。在训练阶段时会强制将上一步解码的输出作为下一步解码的输入，即 X 会作为下一步预测 Y 时的输入。

当编码阶段输入的序列过长时，解码阶段 LSTM 模型将无法针对最早的输入序列解码。Attention 注意力分配机制，在解码阶段每一步解码时，都会有一个输入，对输入序列所有隐含层的信息进行加权求和，能够很好的解决这个问题。

将分词后的新闻文本数据拆分为训练样本和测试样本，共四个文件：train_ data. txt，train_ title. txt，test_ data. txt，test_ title. txt 。新闻正文内容和其对应的新闻标题需要分开存放在两个文件内，一行为一条新闻样本。

五、实证效果

运行脚本，训练好的模型将被保存下来，部分预测好的 Text Summarizaion 如下：

ID	新闻正文	新闻标题	模型生成标题
112882	故宫 长城 央视 大楼 鸟巢 水立方 有名 地方 不胜枚举 地界 老百姓 生活 相关 市井 本土 北京 胡同 推荐 北京 胡同 游 攻略 瞧瞧 北京 普通百姓 生活 每条 胡同 故事 细细的 品味 体会 魅力 pstrong 烟袋 斜街 strongp 烟袋 斜街 位于 地安门 外 大街 鼓楼 前 什刹海 前海 北侧 此街 东西 斜形 走向 全长 232 米 烟袋 斜街 元朝 时期 抄近 道 走出 一条 烟袋 斜街 当年 居住 旗人 嗜好 抽烟 烟叶 装在 烟袋 中 烟袋 需求 与日俱增 斜 街上 一户 一户 开起 烟袋 铺 街道 宛如 一只 烟袋 得名 烟袋 斜街 街道 两侧 建筑 典雅 朴素 颇具 明清 传统 风格 其前 店 居 形式 呈现出 古风 犹存 市井 风情 展现出 浓郁 北京 传统 风貌 烟袋 斜街 北京 北城 有名气 文化街	北 京 什 么 地 方 最 出名	北京 有名 地界 旅游 攻略 推荐
112803	北京 是因为 它作 多年 首都 作 这么久 首都 全 是因为 北京 优秀 全赖 800 年前 千万别 张国字 脸 迷惑 他本 名叫 完颜 迪古 金朝 第四位 皇帝 历史 鼎鼎大名 海陵王 皇帝 先 爷爷 说起 爷爷 完 颜阿骨 打即 金太祖 金朝 第一位 皇帝 公元 岁 完 颜阿骨 东北 白山黑水 间 无数次 厮杀 终于 灭 辽朝 建立 金国 建都 宁府 哈尔滨市 城区 白城 电视剧 中 阿骨 打是 右边 那位 穿 高档 动物 皮草 北京 优秀 全赖 年前 不错 细心 读者 发现 这位 阿骨 丐帮 帮主 乔峰 结拜兄弟 那位 金太祖 六年 皇帝 去世 女真族 兄终弟及 传统 金太祖 弟弟 完 颜晟 即位 金太宗 金太宗 岁 去世 前 不想 皇位 传给 弟弟 想 传给 儿子 太祖 太宗 两派 子孙 夺位 几个 回合 有人 举牌 北京 优秀	不 用 再 思 考 为 什 么 来 北京	北京 优秀 文化 遗产
112337	中国 多地 雾霾 齐发 城市 发布 雾霾 预警 城市 朦胧 模式 华北 黄淮 地迎 本轮 雾霾 最重 时段 京津冀 省份 局部 重度 霾 北京 今夜 污染物 迎来 本次 污染 峰值 今晨 上午 四川 湖南 局地 有强 浓雾 应对 重 污染 天气 京津冀 环保部门 联动 执法 停工 停产 禁行 中小学 停止 室外 活动 京津冀 省份 今日 局部 重度 霾 夜间 空气 中 湿度 增大 污染物 扩散 条件 转差 华北 黄淮 空气质量 下降 华北 黄淮 地雾霾 袭 北京 河北 天津 发布 重 污染 预警 中央气象台 预计 18 白天 夜间 华北 黄淮 本轮 雾霾 过程 北京 南部 天津 西部 河北 中部 河南 中部 陕西 关中 山西 中部 局部地区 重度 霾 另据 交通运输 部 发布 路况	多 地 雾 霾 齐 发 中 国 发 布 雾 霾 预警	中国 发布 多地 雾霾 预警

六、总结

随着互联网的迅速发展，网络中的新闻资源呈指数级增长，通过深度学习自动生成的标题往往能很直观的体现新闻的主题内容，便于读者快速的浏览新闻，准确选择自己感兴趣的内容，节约时间成本，能够给读者带来很好的体验感。

智能推荐已经成为一种势不可挡的趋势，随着人工智能的发展，算法推荐必将成为内容领域的主流之一。如果将基于 Tensorflow 的 LSTM 主题分类的个性化推荐和非个性化推荐相结合，不仅能很好的解决用户冷启动问题，而且可以满足企业的个性化需求和用户的实时智能推荐。

作者简介：田江（1987－），男，汉族，江西上饶人，统计学硕士。