机器学习工程师纳米学位毕业项目

《句子相似度匹配》开题报告

陈宁

2019 年 6 月 25 号

目录

1. 背景介绍

引用

让机器来理解人类语言一直都是人工智能的梦想，最先从词到短语到句子，再到段落到整篇文章。所有的方式都是将字符串转换为向量，最终从数学的角度来理解语义。

2. 问题

引用

本项目中提供了已经配对好的句子对，需要用已有的数据集进行训练，最终可以预测两个句子的相似性。这属于监督学习类型。可以用编辑距离类似的方法来解决。可以把每一个句子看成是一个向量，再求两个向量之间的关系。这种关系在每一个句子对中几乎都存在。把这种关系保存下来，当有新的一对未知关系的句子对需要检测时，那么就可以先根据数据模型算出他们的关系，再和训练数据的句子对的关系对比，可以得到新的句子对的相似性。

3. 数据集

Quora Querstion Pairs 数据集是 Quora 于 2017 年公开的句子匹配数据集，其通过给定两个句子的一致性标签标注，从而来判断句子是否一致。

Quora 数据集训练集共包含 40K 的句子对，且其完全来自于 Quora 网站自身，Quora 在发布数据集的同时，在 Kaggle 平台，发起了 Quora 句子相似度匹配大赛，共有 3307 支队伍参加了本次句子相似度匹配大赛，参赛队伍不仅包括来自麻省理工学院、伦敦大学学院、北京大学、清华大学、中科院计算所等高校研究所，也包括了来自微软、Airbnb、IBM 等工业界的人员。

本项目使用 Kaggle 端的数据集，其由 Train,Test 两部分构成，Train 数据集上可以进行验证集划分、建模，Test 数据集上可以进行测试。

4. 解决方案

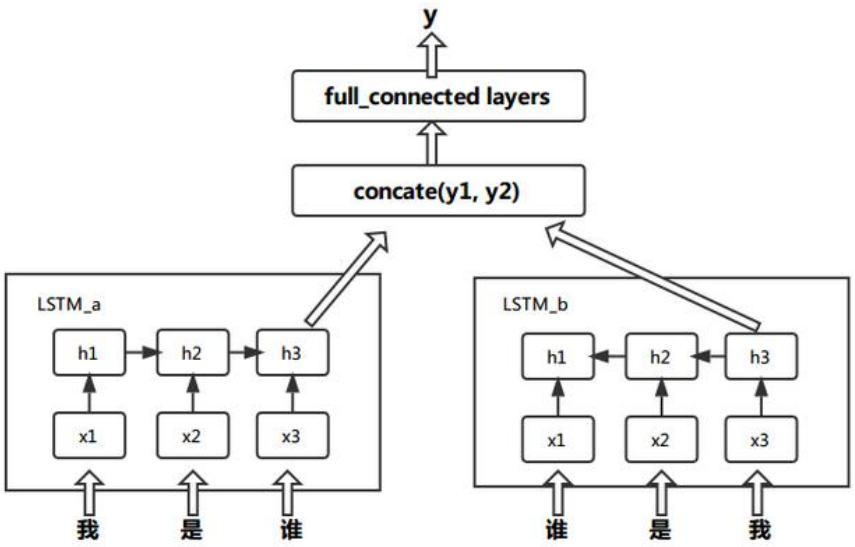
引用

每一个句子是由单词组成的，可以把所有句子中的单词列出来，进行编码，再用单词的编码对句子进行编码。最终得到是一个个的向量，可以用深度学习的方法从这些向量中找出规律，即模型。该模型就可以用来预测句子的相似性。

5. 基准模型

引用 1

本项目中使用的是简单的单层 LSTM+全连接层对数据进行训练。如图：



LSTM 可以用来解决复杂的数据输入，不会导致梯度消失。

6. 评价指标

训练数据集中是一个句子对以及它们是否相似的标签， 相似则值为 1，否则为 0，所以可以有以下两种方式：

第一种，用训练好的模型去预测训练的所有数据，把得到的值和原来的标签作对比，统计出正确统计的标签的占比，占比越高，则效时越好。

第二种，预测所有的测试数据，把预测结果保存到一个 csv 文件中提交到 kaggle，得到损失度，值越低，说明效果越好。

7. 项目设计

引用

7.1：读入数据，数据如下

	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0
4	4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0

7.2：语料编码

把训练数据中的所有文本放到一个列表中，输入分词器 Tokenizer,

texts = []

tokenizer = Tokenizer(num_words=MAX_WORDS, lower=False)

tokenizer.fit_on_texts(texts)

7.3：词语映射

用分词器 tokenizer 把所有单词和它的编码联系起来。、

7.4：搭建一个单层 LSTM+全连接层的网络

7.5：训练网络

model.fit(question1_list, question2_list, ...)

7.6：预测结果

model.predict

先预测训练集上的数据，和标签作对比，直到准确率达到一个比较满意的值（0.8 以上）时，再预测测试集的数据

7.7：保存预测结果

打印训练集上的准确率，把测试集上的预测结果保到一个 csv 文件中，提交到 kaggle 查看损失度，如果结果不满意，重新优化模型，再次训练并预测。