文档类别

杭州海康威视数字技术股份有限公司

文档编号

算法调研报告

|  |  |
| --- | --- |
| 编制 |  |
| 审批 |  |

密级级别：[内部公开]

生效时间： 2021年6月30日

保密期：无

杭州海康威视数字技术股份有限公司 版权所有

目录

[1. 绪论 3](#_Toc103882038)

[1.1 编写目的 3](#_Toc103882039)

[1.2 研究背景 3](#_Toc103882040)

[1.3 工作说明 3](#_Toc103882041)

[2. 调研概述 3](#_Toc103882042)

[2.1 算法一 3](#_Toc103882043)

[**2.1.1** **算法原理** 3](#_Toc103882044)

[**2.1.2** **算法流程** 4](#_Toc103882045)

[**2.1.3** **算法实验** 4](#_Toc103882046)

[**2.1.4** **算法总结** 4](#_Toc103882047)

[2.2 算法二 5](#_Toc103882048)

[**2.2.1** **算法原理** 5](#_Toc103882049)

[**2.2.2** **算法流程** 8](#_Toc103882050)

[**2.2.3** **算法实验** 8](#_Toc103882051)

[**2.2.4** **算法总结** 8](#_Toc103882052)

[2.3 参考文献 8](#_Toc103882053)

[3. 调研结论 9](#_Toc103882054)

[3.1 可行技术路线 9](#_Toc103882055)

[4. 修订记录 9](#_Toc103882056)

# 绪论

## 编写目的

本文档旨在调研对于长文本表征的实现方式，为下游的长文本任务提供前提。

## 研究背景

在之前的NLP任务中，bert对于文本的表征有长度512的限制，并且由于bert的位置编码是学习得来的，它不支持对512位以后的位置编码的表达。对于长文本来说，512位以后的文本被忽略了。我们需要对位置编码做出改变使得他们能够支持对长文本的表征。

## 工作说明

本文档预期读者是大数据算法部团队成员。

# 调研概述

本次调研将介绍两种方式来改变位置编码。

## 算法一

层次分解算法

### **算法原理**

具体来说，假设已经训练好的绝对位置编码向量为𝑝\_1,𝑝\_2,…,𝑝\_𝑛，我们希望能在此基础上构造一套新的编码向量𝑞\_1,𝑞\_2,…𝑞\_𝑚,其中𝑚>𝑛。为此我们设

𝑞\_((𝑖−1)∗𝑛+𝑗)=𝛼𝑢\_𝑖+(1−𝛼) 𝑢\_𝑗

其中𝛼∈(0,1)且𝛼≠0.5是一个超参数，𝑢\_1,𝑢\_2,…,𝑢\_𝑛 是该套位置编码的“基底”。这样的表示意义很清晰，就是将位置(𝑖−1)∗𝑛+𝑗层次的表示(𝑖,𝑗)，然后𝑖,𝑗对应的位置编码分别为 𝛼𝑢\_𝑖 和 (1−𝛼) 𝑢\_𝑗 ，而最终(𝑖−1)∗𝑛+𝑗 的编码向量则是两者的叠加。要求 𝛼≠0.5 是为了区分(𝑖,𝑗)和(𝑗,𝑖)两种不同的情况。

我们希望在不超过𝑛时，位置向量保持跟原来的一样，这样就能与已经训练好的模型兼容。换句话说，我们希望𝑞\_1=𝑝\_1,𝑞\_2=𝑝\_2,…,𝑞\_𝑛=𝑝\_𝑛，这样就能反推出各个𝑢\_𝑖了：

𝑢\_𝑖=(𝑝\_𝑖−𝛼𝑝\_1 )/(1−𝛼), 𝑖=1,2,3,…,𝑛

这样一来，我们的参数还是𝑝\_1,𝑝\_2,…,𝑝\_𝑛，但我们能表示出𝑛^2 个位置的编码，并且前𝑛个位置的编码跟原来模型是相容的。

### **算法流程**

初始模型仅为一个基础的roberta中文模型，在build\_transformer\_model()函数中加上hierarchical\_position=True 即可使得模型突破自身maxlen的限制。可以先使用预训练任务对模型进行训练，再进行下游任务的fintune（这里由于显卡限制，原模型的maxlen和下游任务的maxlen皆为1024，所以省去了预训练任务）。训练好的模型仅仅是一个能够生产不止1024位位置编码的模型，针对不同下游任务还需做相应的调整。该模型的input就是一段文本，通过已训练好的模型转换成embedding，然后进行下游任务。

### **算法实验**

本次模拟的下游任务是一个相似度对比任务，来源于CAIL-2019，每个样本有三个文本，分别构成一对正样例和一对负样例，通过一个Triplet Network，对每个样本的正负样例对比相似度，评估结果为accuracy。

最终结果为 final test acc: 0.668620。baseline为0.64。

### **算法总结**

由于此算法仅仅改进了位置编码，所以提升效果并不是特别明显，如果预训练模型能够对专有领域语料库进行pretrain，效果更佳。并且用较新的sbert或者simCSE范式对模型进行fintune，或许能够更好的提升。

## 算法二

Reformer算法

### **算法原理**

简单来说,Roformer是一个绝对位置编码替换成RoPE的WoBert模型。这里说明一下RoPE的原理：

* + - 1. 基本思路

为了达到这个目的, 我们假设通过下述运算来给 添加绝对位置信息:

也就是说, 我们分别为 设计操作 , 使得经过该操作后, 就带有了位置 的绝对位置信息。Attention的核心运算是内积, 所以我们希望的内积的结果带有相对位置信息, 因此假设存在恒等关系：

所以我们要求出该恒等式的一个 (尽可能简单的) 解。求解过程还需要一些初始条件, 显然我们可以合理地设 和

* + - 1. 求解过程

同上一篇思路一样, 我们先考虑二维情形, 然后借助复数来求解。在复数中有 ,  
 代表复数的实部, 所以我们有

简单起见, 我们假设存在复数 , 使, 然后我们用复数的指数形式, 设

那么代入方程后就得到方程组

对于第一个方程, 代入 得到

最后一个等号源于初始条件 和 。所以现在我们可以很简单地设

这里的 是 本身的幅角, 最后一个等号同样源于初始条件。根据上式得到  
, 所以 应该是一个只与 相关、跟 无关的函数, 记为 , 即 。接着代入 , 整理得到

即 是等差数列, 设右端为 , 那么就解得 。

* + - 1. 编码形式

综上, 我们得到二维情况下用复数表示的RoPE:

根据复数乘法的几何意义, 该变换实际上对应着向量的旋转, 所以我们称之为“旋转式位置编码”, 它还可以写成矩阵形式:

由于内积满足线性叠加性, 因此任意偶数维的RoPE, 我们都可以表示为二维情形的拼接, 即  
也就是说, 给位置为 的向量 乘上矩阵 、位置为 的向量 乘上矩阵 , 用变换后的 序列做 Attention, 那么Attention就自动包含相对位置信息了, 因为成立恒等式:

值得指出的是, 是一个正交矩阵, 它不会改变向量的模长, 因此通常来说它不会改变原模型的稳 定性。  
由于 的稀疏性, 所以直接用矩阵乘法来实现会很浪费算力, 推荐通过下述方式来实现RoPE:

其中 是逐位对应相乘, 即Numpy、Tensorflow等计算框架中的 运算。从这个实现也可以看到, Ro PE可以视为是乘性位置编码的变体。

### **算法流程**

初始模型为roformer预训练中文模，可以直接进行下游任务的fintune（这里由于显卡限制，原模型的maxlen和下游任务的maxlen皆为1024）。训练好的模型仅仅是一个能够生产不止1024位位置编码的模型，针对不同下游任务还需做相应的调整。该模型的input就是一段文本，通过已训练好的模型转换成embedding，然后进行下游任务。

### **算法实验**

本次模拟的下游任务是一个相似度对比任务，来源于CAIL-2019，每个样本有三个文本，分别构成一对正样例和一对负样例，通过一个Triplet Network，对每个样本的正负样例对比相似度，评估结果为accuracy。

最终结果为 final test acc: 0.67634。baseline为0.64。

### **算法总结**

由于此算法仅仅改进了位置编码，所以提升效果并不是特别明显，如果预训练模型能够对专有领域语料库进行pretrain，效果更佳。并且用较新的sbert或者simCSE范式对模型进行fintune，或许能够更好的提升。

## 参考文献

[1] 苏剑林. (Mar. 23, 2021). 《Transformer升级之路：2、博采众长的旋转式位置编码 》[Blog post]. Retrieved from https://kexue.fm/archives/8265

[2] 苏剑林. (Dec. 04, 2020). 《层次分解位置编码，让BERT可以处理超长文本 》[Blog post]. Retrieved from https://kexue.fm/archives/7947

# 调研结论

本文档的目的是为bert处理长文本提供一定算法支持支持。

## 可行技术路线

算法第一版本由于显卡的限制，并没有将maxlen设置为更大的数值（如允许，可设置为4096左右）。若设置为更大的maxlen，可以对全文进行embedding（这里下游数据集文本的最大程度为3000左右，1024不能够对全文进行embedding）。在设置为更大的maxlen之后，可以先继续预训练，然后再进行下游任务，效果应会有所提升。

# 修订记录

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **变更时间** | **版本** | **变更人** | **审批人** | **变更说明** |
| 1 | 2022.5.19 | V 0.1.0 | 边圣陶 |  | 新建 |
| 2 |  |  |  |  |  |
| 3 |  |  |  |  |  |
| 4 |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |