关于百度ERNIE及将知识图谱引入Bert

张俊林
你所不知道的事

百度放出了ERNIE，报道内容参考：“中文任务全面超越BERT：百度正式发布NLP预训练模型ERNIE”

从目前报道的内容看，好像百度的ERNIE主要工作是：

1.预训练阶段仍旧采取 字输入，但是Mask对象是单词，如果是单纯的对单词进行Mask，我觉得这改进还好，不过我猜ERNIE很可能还专门挑出一定比例的实体词进行了连续Mask，实体词Mask我觉得是很有意义的，为啥这么感觉等会说。

2.采取了很多知识类的中文语料进行预训练，这个也挺好。

把“知识图谱”加入Bert的模型中，我自己也特别看好这个方向，之前也安排个别同学在尝试这个思路，不过还没啥结果，估计很多同行也正在做。百度的工作可以看做是这个方向的初步探索结果，还仅仅使用了实体概念，没有把实体关系融入进去，后面应该很自然会拓展到“实体关系类”知识的引入。

为啥把“知识图谱”引入Transformer是个好的改进方向呢？我们可以认为Bert的预训练阶段采取的语言模型任务，这算是通用的语言知识，胜在量大，但是因为是自监督的模式，虽然其实里面也大量包含了各种“知识图谱”中的知识，比如“太原是山西的省会”这种句子里的知识应该也能通过语言模型编码到TF参数里。但是毕竟不是专门学习这种知识，所以可能针对这种知识的编码能力不算太强，当然这是纯猜测。

如果我们有量级非常大的"知识图谱"，而明显百度在这个方面是明显有优势的，编码到Bert模型里，估计对于下游的知识类任务或者包含NER相关的任务有

但是我觉得不应该在第一阶段预训练阶段来对"知识图谱"进行编码，第一阶段预训练阶段应该做什么，感觉GPT-2已经说明白了，就是增大数据规模，增加数据多样性，增加数据质量。量大，质好，花样多。只要持续做这三个事情，感觉性能还有提升空间。当然也可以寻找一个非语言模型的任务，换成另外一种任务，但是肯定要是自监督或者无监督的，因为第一阶段的最大优点：量大。这个优势不能放弃。

感觉应该把“知识图谱”的编码放到第二个阶段，比如拿到一个非常大的知识图谱，假设它是三元组表示<实体1，关系R，实体2>，要求Bert去做有监督的训练，比如可以输入<实体1，关系R>要求预测实体2，或者输入<实体1，实体2>，要求Bert预测它们的关系。等等，有不同的方法可以强迫Bert去学习知识图谱里的知识，如果这个图谱量级够大，那么这样的有监督地专门学习阶段学习效率应该是比让它去通过语言模型学习知识更有效。

甚至我觉得应该把Bert改成三阶段的，第一个阶段LM语言模型，追求量大，质好，花样多；第二个阶段，专门对各种知识图谱进行有监督的学习；第三个阶段才是原来的finetuning阶段。如果能够把大量知识编码进去，我相信对于很多下游任务应该会有促进作用。

至于Bert第一阶段应该采用“词输入”还是“字输入”？我个人觉得还是字输入好，我们之前在Bert放出代码，但是还没有放出预训练模型的时候，试着做过单词输入的Bert预训练模型，并和字输入的模型进行过比较，结果是词输入的效果是不如字输入的效果的。当然，预训练数据规模不是特别特别大，随着预训练数据规模的加大，可能两者的差距会减小。使用词输入，相对字输入，我觉得有几个缺点：一个是对分词工具有依赖，尤其是NER、新词等OOV问题，会比较影响模型效果。第二个是在预测的时候，如果是基于字的则预测结果标签集合较小，而如果是基于词的，明显标签空间会大很多，这很可能也会有劣势。而百度仍然采取字输入，但是Mask采取单词的方式，我觉得算是一种折中方案，包括N-gram，也算折中方案，能比较好的平衡两者，其实是挺好的。

对于中文任务来说，我觉得对于很多任务来说，分词是不必要的，随着Transformer的能力越来越强大，绝大多数任务应该以字作为输入，而连续的几个字是否应该是个单词，理论上应该让Transformer当做内部特征去学习，所以感觉中文分词是不必要存在的。当然，这个纯属个人猜想，目前无证据。

发布于 2019-03-17

「真诚赞赏，手留余香」

赞赏

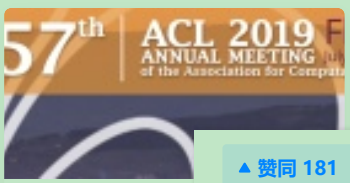
还没有人赞赏，快来当第一个赞赏的人吧!

深度学习（Deep Learning） 人工智能 自然语言处理

文章被以下专栏收录

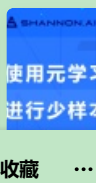
深度学习前沿笔记 关注专栏

推荐阅读



RoBERTa中文预训练模型，你离中文任务的「SOTA」只差...

有了中文文本和实现模型后，我们



▲ 赞同 181 ▼ 18 条评论 分享 收藏 ···

ACL2019参会分享（一）Dialogue篇
武博文

的预训练语言模型，我们最常用的就是 BERT 了，这并不是说它的效
机器之心 发表于机器之心

香侬慧语

18 条评论 ⇌ 切换为时间排序

写下你的评论...

小黄 8 个月前
不知道老师怎么看待现在学界对于训练自然语言到代码的映射的方式呢?
赞

张俊林 (作者) 回复 小黄 8 个月前
挺好的应用方向
赞

逍垚 8 个月前
第一阶段、第二阶段这个想法是bert+tranE的思路么
1

张俊林 (作者) 回复 逍垚 8 个月前
Bert本身就是两个阶段的
赞

ZRX 回复 逍垚 5 个月前
我觉得ERNIE就是follow bert+transE思路，bert和transE都脱胎于w2v，只不过bert使用transformer比w2v的BOW能更好地提取文本特征，transE也是<S,P,O>知识三元组训练集利用w2v BOW词频共现机理进行建模。ERNIE是两者集成，针对masked S|P|O in context进行bert语义表示学习embedding
赞

henryWang 8 个月前
老师是认为第二阶段专门bert学knowledge represent嘛～～其实个人有点觉得ERNIE的knowledge有点数据量硬怼的感觉，paper没发出来不好说，希望baidu有更详细的模型本身的对比证明这种mask
赞

张俊林 (作者) 回复 henryWang 8 个月前
感觉百度目前做法可能还是比较简单直接的思路
赞

杨浩 8 个月前
concat（word embedding, knowledge embedding），一方面理解语言模型，一方面理解知识；然后mask一部分word,或者mask一部分knowledge,使模型更robust,哈，想象空间真大
2

张俊林 (作者) 回复 杨浩 8 个月前
嗯 这个方向感觉有潜力
赞

张翔 回复 杨浩 7 个月前
有意思啊，这个方向有对应工作出来吗
赞

展开其他 2 条回复

瑞教授 8 个月前

字嵌入比词嵌入除了您讲的几点，可以理解成词时信息量损失。让模型内部去计算成词的概率远比预处理的信息损失更小。这一点我在做词性标注、分词系统的时候已经试验过了。

3

张俊林 (作者) 回复 瑞教授 8 个月前

是吗，挺好

赞

kaijien 8 个月前

如果如老师所说分成三个阶段，那么第一个阶段的目标并不存在知识，那么第一阶段输出的特征又怎么保证不丢失各种知识的信息呢？如果第一阶段丢失，又如何进行下一阶段？。。。如果两个阶段联合在一起训练，那么和ernie这种方法感觉就一样了哇。。。？

赞

张俊林 (作者) 回复 kaijien 8 个月前

如果第二阶段做得好的话，第一阶段的信息未必会丢，这个说不准

赞

summer 4 个月前

Bert和知识图谱一起怎么做消歧？

赞

Light 2 个月前

bert again