|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1. 发明人信息 (如存在外籍发明人需要特别指出，每个发明人的贡献比例将等同于奖励比例) | | | | |
|  | 发明人 | 贡献比例 | 联系方式(手机+邮箱) | 部门 |
| 1 | 魏林林 | 90% | 13281207928+weilinlin004@ke.com |  |
| 2 | 马宝昌 | 10% | 18210063019+mabaochang001@ke.com |  |
| 3 |  |  |  |  |
| 4 |  |  |  |  |
| 5 |  |  |  |  |
| 6 |  |  |  |  |
| 1. **发明完成日期** | | | 具体技术完成时间2021年8月18日 | |
| 1. **发明完成地点** | | | 电子科技大厦 | |
| 1. **第一发明人身份证号码** | | | 371502199404093844 | |
| 1. **技术支持人 姓名、电话及邮箱** | | |  | |
| 1. **方案名称** | | | | |
| **一种用于提升语义匹配任务的基于注意力蒸馏的文本表示模型** | | | | |
| 1. **是否为在研的研发项目，项目名称为？预计上线时间？** | | | | |
| **是；项目名称：SSC智能客服 & 贝问贝答；SSC上线时间：2021年07月15日，贝问贝答上线时间：2021年07月20日** | | | | |
| 1. **之前是否递交过跟该项目相关的技术交底书，或者与该技术交底书相关联的其他交底书（例如某一技术交底书的方案更新版）？技术交底书名称是？** | | | | |

|  |
| --- |
| **否** |
| 1. **背景技术(必填项)** |
| 请用简洁的语言说明该发明所属的技术领域或者直接应用的技术领域；  详细介绍一下与本发明专利申请要求保护的主题相关的现有技术的现状、工作原理、工作方式如何？客观地指出现有技术的缺点和不足，解释一下原因。  **所属的技术领域**：  信息检索，问答系统  **现有技术现状**：  在目前主流的语义匹配方法中，大致分成两个类别。基于表示的和基于交互的语义匹配模型。基于表示的模型主要是利用孪生网络这种双塔结构，如DSSM、SBERT等，也被称为Bi-Encoders。具体做法是将查询语句（query）与候选文本（candidates）单独的映射为两个向量，然后将两个向量拼接通过一个分类层或者直接计算两个向量的相似度。另一种就是基于交互的模型，如主流的BERT，也被称为Cross-Encoders。具体做法是将查询语句与候选文本拼接在一起输入到一个编码器中，在编码的过程中利用注意力机制来挖掘查询语句与候选文本之间细粒度的语义关系。两类框架的具体形式如下图1所示    基于表示的Bi-Encoders框架 基于交互的Cross-Encoders框架  此外还有一些改进模型从数据增强以及延迟交互的角度融合两种框架各自的优势。如Augmented SBERT先利用基于Cross-Encoders框架的模型标注一些训练数据然后送给基于Bi-Encoders框架的模型训练，ColBERT和Poly-Encoders则是在双塔结构的最后一层进行query和candidates的交互。  **现有技术缺点：**   1. 基于表示的Bi-Encoders框架单独的编码query和candidates，没有建模两者之间细粒度的相似性，模型性能不如Cross-Encoders。特别是在数据量不足时，输出的句向量难以充分表达句子的语义。 2. 基于交互的Cross-Encoders框架需要将query与所有的candidates分别构成一对输入到编码器中，导致巨大的计算开销，无法满足工业界需求。 3. 一些改进模型没有充分的建模query与candidates之间的细粒度的语义信息，只是做了浅层的语义信息交互。 |
| 1. **设计要点(必填项)** |
| 针对现有技术的缺点和不足，阐述本发明的发明目的、解决的技术问题是什么？要达到的效果是什么？  为了解决现有问题的不足，请用技术的语言简要列出本发明的设计要点，即：本方案必须具备的技术点，以及每个设计要点所起的作用和达到的效果。  本发明要解决的技术问题&要达到的效果：   1. 解决Bi-Encoders框架中不能建模query和candidates之间细粒度的语义信息。 2. 解决一些改进模型中只是将query和candidates进行浅层的语义交互问题。 3. 通过Cross-Encoders框架中计算得到的交互语义信息增强Bi-Encoders框架输出的句向量表示。   发明设计简要说明   1. 利用知识蒸馏的思想将Cross-Encoders框架中计算的query与candidates之间的注意力矩阵转移到Bi-Encoders框架中。 2. 通过逐层传递的方式实现Bi-Encoders框架中query与candidates之间深层次的语义交互。 |
| 1. **具体实施方式(必填项)** |
| 结合具体实施例和附图，完整、清楚地写出本发明要实现上述发明目的的具体技术方案，使本领域的普通技术人员能够理解该技术方案，并能按照该技术方案实现本发明的发明目的。本部分非常重要，这是发明专利的核心部分。  **对于一种XXX装置的专利申请来说**，要结合图纸详细说明构成本发明XXX装置都有哪些部分？各部分的连接关系？位置关系？实现发明目的的工作原理。  **对于一种XXX方法的专利申请来说**，要结合反映该方法的程序框图、按照时间顺序详细说明具体的实现步骤，在描述该方法时，要将一切条件、数据、材料、设备、工具和必要的技术规格、型号等进行必要的交代，使同领域的普通技术人员能够按照您的说明，将该发明或者实用新型再现出来，而不能建立在自己已经深刻理解本方案的前提下。 1. 公式/符号说明.    |  |  | | --- | --- | | **公式/符号** | **含义** | | [CLS] | 一个特殊的字符，添加在输入句子的第一个单词之前，编码后的向量表示整个句子的语义。 | | [SEP] | 一个特殊的字符，当输入文本不止一个句子时，添加在句子之间用来分隔各个句子。 | | Q | 代表输入的查询语句query。 | | D | 代表输入的候选集candidates。（candidates既可以是文档，也可以是单个句子）。 | |  | 代表查询语句Q的第个单词。 | |  | 代表查询语句Q在编码器中第行的表示。（同理代表候选文本在编码器中第行的表示。） | |  | 代表Bi-Encoders框架中第层下和计算的第个头的注意力分布矩阵。 | |  | 代表Cross-Encoders框架中第层下计算的第个头的注意力分布矩阵。 | |  | 代表Bi-Encoders框架中第层下和计算的所有注意力头拼接得到的注意力分布矩阵。（同理。） | |  | 代表候选集D的第个单词。 | |  | 模型预测的概率分布与标签之间的交叉熵损失。计算公式： | |  | 模型预测值与标签值之间的误差平方和： |  2. 基于注意力蒸馏的文本表示方法 2.1 相关工作  相关工作包括三部分：注意力计算、预训练模型、知识蒸馏。  注意力机制应用在自然语言处理领域主要是为了计算两个实体之间的相关程度，将相关程度高的实体赋予较高的注意力权重。通常所说的注意力指的是两个不同的句子之间做交互计算，也叫做交互注意力，目的是计算出一个句子内的所有单词和另一个句子的所有单词之间的相关程度。假设第一个句有个单词，第二个句子有个单词，每一个单词用向量表示。记两个句子的向量表示分别为，显然  预训练语言模型（比如BERT）已经在各种NLP任务上达到了显著的效果，是目前最为常用的编码器。以base版本的BERT为例，一共由12层Transformer结构组成，每一层都是由多头自注意力模块和全连接模块组成，如下图2所示。    图2. 一层Transformer结构  从图1中我们可以看出，Bi-Encoders的特点是将query和candidates独立的输入到编码器（如BERT）中进行编码得到各自的向量表示，然后计算两个向量的余弦相似度作为预测的分数。整个编码过程中query和candidates之间是没有任何交互的，这主要是因为不能直接将query和candidates的交互信息作为编码的一部分，否则编码器就不能用来独立的编码一个句子的向量。ColBERT模型的做法是在Bi-Encoders的最后一层取出query和candidates的表示向量计算它们之间的注意力，将注意力分数作为预测值与标签计算交叉熵损失，很显然这只是一种浅层的语义信息交互机制。  知识蒸馏(Knowledge distillation)是指将一个大模型(教师模型)所学习到的知识迁移至小模型(学生模型)，使得学生模型可以学习更多的知识，在参数量小于教师模型的情况下可以逼近教师模型的性能。知识蒸馏的方式可以分为很多种，最初广泛采用的方式是利用教师模型的预测作为一种监督信号让学生模型拟合，由于模型预测习惯称为logits，所以本文将这种蒸馏方式称为基于logits的蒸馏，这种方式有效的原因是教师模型预测分布的熵比较高，所蕴含的信息比one-hot这种标签更加的丰富，基于logits的蒸馏结构图如下图3所示：    图3. 基于logits的蒸馏方式  其中代表softmax的温度，当时便是常见的softmax，随着温度的升高，规范化后的分布越均匀，熵也就越大。将教师模型的预测分布称作软标签(Soft labels)，由于学生模型要拟合教师模型的预测分布，所以规范化时的温度必须与教师模型保持一致，为了防止教师模型的误判引入误差等问题，同时也对one-hot标签计算损失。另一种常见的知识蒸馏方式就是将教师模型中间层的隐藏状态(或者说中间层的表示向量)作为学习信号让学生模型拟合，即将教师模型学习到的表示作为知识迁移给学生模型，如下图4所示：    图4. 基于表示的蒸馏方式  这种蒸馏方式可以采用的KL散度或者均方误差(MSE)来衡量两个向量表示之间的距离，使得学生模型中间层的向量表示尽可能接近教师模型中间层的向量表示，基于注意力蒸馏的方式与基于表示蒸馏的方式类似。  2.2 模型结构  如图所示，Cross-Encoders和Bi-Encoders框架均采用12层Transformer的BERT作为编码器。首先训练好Cross-Encoders框架作为教师模型。在知识蒸馏阶段，将输入句子按照Cross-Encoders框架要求的格式以及Bi-Encoders框架要求的格式分别输入到两个框架中。取出Cross-Encoders中每一层的注意力权重，记为)，其中代表第一层多头注意力中各个头注意力权重的加和平均，作为Cross-Encoders框架第一层的注意力权重，其余层的注意力权重计算方式同理。对于Bi-Encoders框架，分为两个编码器，两个句子独立编码，需要取出编码第一个句子的Encoder中的表示向量和编码第二个句子的Encoder中的表示向量，将和进行交互计算(图5中的mutual attention模块)，从而得到两个句子之间的交互注意力权重。  在训练Bi-Encoders的编码器时，将query和candidates以特定的方式输入到Cross-Encoders和Bi-Encoders中。对于Bi-Encoders，需要有专门计算query和candidates之间注意力的模块，如图3中的Mutual Attention模块，做法按照Transformer模块中的多头缩放点积即可。对于12层的Transformer结构，为了建模深层次的语义交互，取出query的每一层表示与对应层数的candidates的表示之间计算注意力得到（为了实现蒸馏，注意力权重需要进行缩放）。为了能够将Cross-Encoders中已经学习到的丰富的交互语义信息转移到Bi-Encoders中的编码器，我们采用注意力蒸馏的思想，即最小化每一层的和之间的均方误差，如下公式：    图3. 从Cross-Encoders到Bi-Encoders的注意力蒸馏结构  我们将作为最终损失的一部分：  就是模型最终的loss。的计算方式是将query和candidates的向量表示拼接通过一个浅层的分类器后与标签计算得到的交叉熵损失，其作用是学习编码句向量。的作用就是根据Cross-Encoders提供的注意力权重分布学习query和candidates之间的交互语义信息。通过这种机制，基于Bi-Encoders框架的编码器不仅可以单独的编码句向量，而且通过训练过程中学习到的注意力权重分布，可以丰富编码的句向量表示，提升下游任务上的表现。 |
|  |

|  |
| --- |
| 1. **技术效果**   本方案达到的技术效果：   1. 通过逐层交互的方式实现Bi-Encoders中深层次的语义交互，同时将注意力权重只作为损失的一部分，并不参与编码表示，保留了Bi-Encoders能够独立编码的性质。 2. 基于知识蒸馏的思想将Cross-Encoders中学习的交互语义信息转移到Bi-Encoders中，从而丰富Bi-Encoders编码的语义向量表示。 |
| 1. **请将填写好的技术交底书提交至大为专利管理系统，网址**[**http://10.26.24.36/，并通知**](http://10.26.24.36/，并通知)**于立群。** 2. 若发明已被披露或于未来30天内将被披露给销售商、供应商或消费者，请尽快通知IP部加急处理。 3. 未来是否有披露计划？如有，请指明计划披露日期和对象，包括将向其披露该发明的制造商、销售商、承包商、放款人等，无论是否已签订NDA。 4. 如发现近似专利、论文文献请提供。 5. 如有能够帮助相关人员理解本发明的其他技术参考文献请提供。 |