043 | 基于深度学习的搜索算法：卷积结构下的隐含语义模型

2018-01-10 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 06:42 大小 3.07M

这个星期，也是我们整个搜索领域分享的最后一周内容，来看一些搜索算法的前沿思考，特别是深度学习对搜索领域的影响。周一我们分享了一篇较早利用深度学习技术来进行搜索建模的论文，论文提出如何使用前馈神经网络来对查询关键字和文档进行信息提取，从而能够学习更有意义的语义信息。

今天我们来看一篇文章《信息检索中结合卷积池化结构的隐含语义模型》（A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval），可以说这是我们周一分享论文的一个后续工作。这篇论文发表在第 23 届世界信息和知识管理大会 CIKM 2014 上。

论文背景介绍

这篇论文的主要目的是探讨深度学习中的卷积神经网络能否应用在搜索中，并取得较好的效果。

下面我们先来了解一下这篇论文作者群的信息。

第一作者 Yelong Shen 是微软研究院的一名资深研究员。

第二作者是何晓冬（Xiaodong He）是微软研究院深度学习组的主任研究员兼经理，发表过一百多篇学术论文，在人工智能领域，特别是近年来在深度学习领域有很突出的贡献。

第三作者高剑峰（Jianfeng Gao）是一名长期在微软研究院工作的研究员和经理。

第四作者邓力（Li Deng）是微软研究院的人工智能学者，曾担任微软的首席人工智能科学家并且领导深度学习中心。2017 年 5 月，邓力离开微软加入 Citadel，美国著名的金融机构，担任首席人工智能官的职位。

最后一位作者格雷古瓦·梅尼尔（Grégoire Mesnil）是来自蒙特利尔大学的一名博士学生。

这篇论文自 2014 年发表后已被引用 180 多次，是探讨深度学习在搜索领域中应用的主要论文之一。

卷积结构下的隐含语义模型详解

我们周一介绍的深度结构化语义模型，其主要思想是希望能够利用前馈神经网络来对查询关键字和文档进行信息提取。这个模型有一个很明显的问题，那就是在第一步对查询关键字或文档进行特征提取时所形成的词向量（Term Vector）是忽略了文字原本的顺序信息的，也就是依然是一个“词袋模型”（Bag of Words）假设，这显然是丢失了很多信息的。

当然，我们今天要分享的卷积结构下的隐含语义模型，也并不是第一个想要解决这个问题的模型。在经典的信息检索领域的研究中，已经有不少这方面的尝试了。那么对于深度学习来说，又有什么优势呢？

近些年来深度学习模型兴起的一个重要动力就是在图像、音频、视频领域的技术突破。而这些突破离不开一个重要的基础模型，**卷积神经网络**的成熟。这个模型对有空间位置结构性的数据，比如图像中每一个像素，有较强的建模能力，成为了探索结构信息建模的一个利器。那么，能不能把在这些领域中已经成熟的经验借鉴到搜索领域呢？

如果把文本的词与词，句子与句子之间的关系看作是一种空间位置关系的话，那么从假设上来看，就很符合卷积神经网络模型的基本设置。接下来，我们就来看看这个模型具体是怎么应用到搜索中的。

首先，模型对查询关键字或者文档的文字进行“**移动窗口**”式（Sliding Window）的扫描。这第一步就和之前的深度结构化语义模型有了本质区别。然后，模型进一步把“移动窗口”下的词转换成为**字母级别的表征向量**（Representation Vector）。这个步骤之后，模型采用了**卷积层**来提取空间位置的特征，也是把数据的维度大幅度降低。卷积层之后就是基本的“**池化层**”（Pooling Layer），这里的模型采用了**最大池化**（Max Pooling），也就是从多个卷积层的结果中，每一个层对应元素中的最大元素。在池化层之后，就是进行一个全部展开的语义层。

更加直白地说，**整个模型就是希望先从原始的文字信息中，利用保留顺序的一个移动窗口提取最基本的特征；然后利用卷积神经网络的标配，卷积层加池化层，来提取空间位置信息；最后利用一个全部的展开层来学习下一步的系数。**卷积层主要抓住的是单词这个级别的特征；而池化层则是希望抓住句子这个层面的语义信息；最后利用句子这个层面的语义信息形成整个文字的内在语义表达。

这个模型是如何被训练出来的呢？事实上，可以说整个模型的训练过程和我们周一分享的深度结构化语义模型的训练过程一模一样。首先，同样是利用用户的点击信息，也就是针对某一个查询关键字，有哪些文档被点击过，作为正例数据，其他文档作为负例数据；然后把整个建模问题看做是一个多类分类问题；这样就可以利用标签信息对整个模型进行学习。

隐含语义模型的实验效果

和深度结构化语义模型一样，隐含语义模型也仅仅使用了查询关键字和文档之间的文字信息，所以也只能和文字型的排序算法进行比较。最终文章在数据集上采用了 Bing 的搜索数据，有 1 万 2 千多的查询关键字以及每个查询关键字所对应的 74 个文档，每个文档又有 4 级的相关标签，用来计算 NDCG 这样的指标。数据虽然和之前一篇不完全一样，但是在数量级上是差不多的。

在这篇文章里，作者们也比较了一系列的方法，比如 TF-IDF、BM25，以及传统的 PLSA 和 LDA。简单来说，隐含语义模型在最后的比较中取得了不错的结果，NDCG 在第 10 位的表现是接近 0.45，而之前提出的深度结构化语义模型达到了差不多 0.44。虽然利用卷积的效果要好一些，但是差距并不大。在这个数据集上，传统方法要差很多，比如 BM25 的

表现仅有 0.38 左右，而传统的 PLSA 和 LDA 也只有 0.40 左右的表现。应该说在这篇文章中展示出来的效果还是有比较大的差距的。

小结

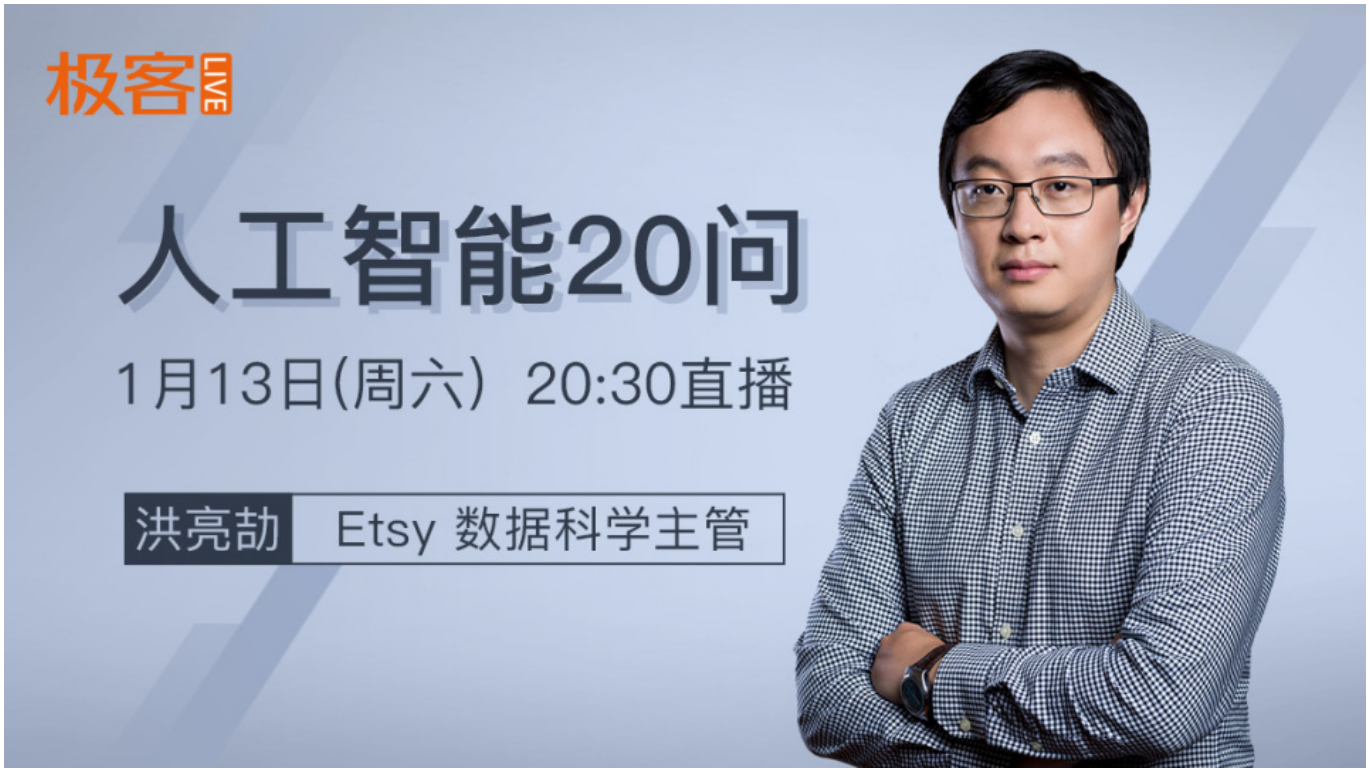
今天我为你讲了卷积结构下的隐含语义模型的一些基本原理，这个模型是利用深度学习技术对搜索算法进行改进的另一个很有价值的尝试，揭开了用深度学习模型，特别是用在图像处理中非常成功的卷积神经网络技术来表征查询关键字和文档会达到的效果。

一起来回顾下要点：第一，我们简要介绍了隐含语义模型提出的历史。第二，我们详细介绍了隐含语义模型的核心思路以及实验结果。

给你留一个思考题，为什么顺序信息并没有像我们想象中的那样，给文档搜索提升带来很大的效果呢？有没有什么解释？

欢迎你给我留言，和我一起讨论。

最后，预告一个小活动，本周六（1 月 13 日）晚 8:30 我会在极客时间做一场直播，欢迎你参加。主题是“人工智能 20 问”，如果你有想交流的问题，欢迎给我留言，我们周六直播见！



上一篇　042 | 基于深度学习的搜索算法：深度结构化语义模型

下一篇　044 | 基于深度学习的搜索算法：局部和分布表征下的搜索模型

精选留言

写留言

> 由作者筛选后的优质留言将会公开显示，欢迎踊跃留言。