091 | Word2Vec算法有哪些应用?

2018-05-02 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 05:23 大小 2.47M

周一，我们分享了三个比较有代表意义的 Word2Vec 的扩展模型，主要有两种思路，从词的上下文入手重新定义上下文，或者对完全不同的离散数据进行建模。

今天，我们来看一看**Word2Vec 在自然语言处理领域的应用**。如果我们已经通过 SG 模型、CBOW 模型或者其他的算法获得了词向量，接下来我们可以把这些词向量用于什么样的任务中呢?

Word2Vec 的简单应用

最直接的也是最常见的 Word2Vec 应用就是去计算词与词之间的相似度。当我们的数据还是原始的“词包”（Bag of Word），这时候是没法计算词与词之间的相似度的，因为每个词都被表示为某个元素为 1 其余元素都为 0 的离散向量。按照定义，两个离散向量之间

的相似度都是 0。因此，从词包出发，我们无法直接计算词与词之间的相似度，这是从定义上就被限制了的。

Word2Vec 就是为了跨越这个障碍而被发明的，这一点我们在前面就已经提到过了。所以，当我们可以用 Word2Vec 的词向量来表示每一个单词的时候，我们就可以用“**余弦相关度**”（Cosine Similarity）来对两个词向量进行计算。**余弦相关度其实就是计算两个向量的点积，然后再归一化。**如果针对已经归一化了的向量，我们就可以直接采用**点积**来表达两个向量的相关度。不管是余弦相关度还是点积，我们都假设计算结果的值越大，两个词越相关，反之则不相关。

既然我们可以计算两个词的相关度，那么很多依赖相关度的任务就都能够轻松完成。比如，我们希望把词进行聚类，也就是说把相关的词都聚合在一起。通常的聚类算法都可以直接使用，比如我们熟悉的“K 均值”算法。这些算法的核心是计算两个数据点的距离，就可以利用我们刚刚讲的余弦相关度来实现。

我们在谈 Word2Vec 扩展模型的时候，曾经提到了一些扩展模型，可以用于表达比词这个单位更大的文本单元，比如段落和文档向量的获取。其实，当时我们就提到了一种可以得到这些单元向量的简单方法，那就是**直接利用 Word2Vec 来进行加权平均。**在获得了词向量之后，我们就可以用一个文档里所有词的加权平均，甚至是简单的叠加来达到表达文档的目的。这个时候，我们也就可以利用诸如余弦相关度来计算文档之间的相关度了。

另外一个随着 Word2Vec 的推出而大放异彩的应用则是“**词语的类比**”。Word2Vec 的原作者们用类比来表达，这种词向量能够完成一些与众不同的任务。词向量本质上就是一个连续空间的向量，因此从数学上来说，这种向量其实可以进行任何“合规”的运算，比如加、减、乘、除。于是，作者们就利用向量的加减关系，来看能否得到有意义的结果，而得到的结果令人吃惊。

比如，如果我们把“国王”（King）这个单词的向量减去“男人”（Man）这个向量然后加上“女人”（Woman）这个向量，得到的结果，竟然和“王后”（Queen）这个向量非常相近。类似的结果还有“法国”（France）减去“巴黎”（Paris）加上“伦敦”（London）等于“英格兰”（England）等。对于传统的方法来说，这样的行为是无法实现的。因此，Word2Vec 的流行也让这种词语的类比工作变得更加普遍起来。

Word2Vec 的其他使用

在自然语言处理中有一系列的任务，之前都是依靠着“词包”这个输入来执行的。当我们有了 Word2Vec 之后，这些任务都可以相对比较容易地用“词向量”来替代。

一类任务就是利用词来进行某种分类任务。比如，我们希望知道某些文档是属于什么类别，或者某些文档是不是有什么感情色彩，也就是通常所说的“基于监督学习的任务”。**词向量会成为很多文本监督学习任务的重要特性。**在诸多的实验结果中，得到的效果要么好于单独使用词包，要么在和词包一起使用的情况下要好于只使用词包。

在进行监督学习的任务中，词向量的重要性还体现于其**对深度学习架构的支持。**众所周知，即便是最简单的深度学习架构，比如多层感知器，都需要输入的数据是连续的。如果我们直接面对离散的文本数据，常常需要把这些离散的文本数据学习成为连续的向量，其实就是在学习 Word2Vec。经过了这一层的转换之后，我们往往再利用多层的神经网络结果对这些信号进行处理。

在很多实践中人们发现，与其利用不同的任务来学习相应的词向量，还不如直接拿在别的地方学好的词向量当做输入，省却学习词向量这一个步骤，而结果其实往往会有效果上的提升。这种使用词向量的方法叫作“**提前训练**”（Pre-trained）的词向量。其实，不仅仅是在简单的多层感知器中，甚至是在 RNN 等更加复杂的深度架构中，也都更加频繁地利用提前训练的词向量。

总结

今天我为你介绍了 Word2Vec 模型在各种实际任务中的应用。

一起来回顾下要点：第一，我们聊了 Word2Vec 这个模型的一些简单应用，比如如何计算词与词之间的相关度，以及如何进行词的类比计算；第二，我们讨论了如何利用词向量进行更加复杂的自然语言任务的处理。

最后，给你留一个思考题，Word2Vec 和主题模型提供的向量，是互补的还是可以相互替换的呢？

欢迎你给我留言，和我一起讨论。

上一篇　090 | Word2Vec算法有哪些扩展模型?

下一篇　092 | 序列建模的深度学习利器：RNN基础架构

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。