032 | 文档理解的重要特例：多模文档建模

2017-12-15 洪亮劼

讲述：初明明
时长 07:42　大小 3.53M

本周我们重点分享搜索系统中的一个重要部件，那就是文档理解。周一我们首先分享了文档理解最基本的一个步骤，那就是给文档分类，主要是看不同文档表达什么类别的信息。然后，周三我们聊了聊另外一个重要的文档理解组件，也就是文档聚类的一些基本的概念和技术。今天我就来和你分享一个文档理解的重要特例：**多模文档建模**（Multimodal Modeling）。

多模数据

我们首先来了解一下，到底什么是多模数据。

多模数据，其实就是说数据有多种模式（Modal）的表达途径。而这些多种不同的模式都共同参与描述同一个数据点的不同方面。

比如，有一张照片反映的是美国总统特朗普在华盛顿白宫的致辞。那么照片本身是对这个场景的一个描述，这是一个模式。然后，和照片相应的文字描述，说明这是特朗普在白宫的致辞，又是另外一个模式。这两个模式是相辅相成的，都是对这个场景的描述。很明显，针对这样多种数据模式的建模是多媒体时代、社交媒体时代非常重要的课题。

在文档领域，非常普遍的情况是文字和图片混搭。一般来说，新闻网站一般都有大量的图文信息。而有一些特殊场景，文字和图片则出现很不对称的混合情况。比如，一些社交媒体（例如 Instagram、Pinterest 甚至 Twitter）上很多短文档都仅仅包含图片或者图片和很少的文字。在这些情况中，文字和图片就成了非常重要的互相补充的信息源。

另外，在电子商务网站中，商品的图片正在成为越来越重要的信息途径。用户经常依靠图片来判断是否要购买某个商品。在电子商务网站上已经很难看到只有文字描述的商品信息了。因此，对于文档的搜索来说，对图文信息的理解是一个核心的技术问题。

那么，多模数据的建模难点是什么呢？

不同模式的数据其实是有不同的特征，如何能够有效利用各自的特性来最优地反映到某一个任务中（比如分类或者聚类等），是多模数据建模的难点。

多模数据建模基础

那么，如何对多种模式的数据进行建模呢？

多模数据建模的核心思路就是数据表征（Representation）。我们需要思考的是如何学习到文字的表征，以及图片的表征。然后，又如何把文字和图片的表征能够联系到一起。

一个最直接的思路，应该是文字采用我们熟悉的各种文字特性，然后利用图片相关的特性提取技术来对图片进行表征。得到文字和图片各自的表征之后，直接把两个不同的特征向量（Feature Vector）连接到一起，就得到了一个“**联合表征**”（Joint Representation）。

比如，假设我们学习到了一个 1000 维度的文字特征向量，然后一个 500 维的图片特征向量，那么，联合特征向量就是 1500 维度。

一个相对比较现代的思路是利用两个不同的神经网络分别代表文字和图片。神经网络学习到“**隐含单元**”（Hidden Unit）来表达图片信息以及文字信息之后，我们再把这些“隐含单元”联结起来，组成整个文档的“**联合隐含单元**”。

另外一个思路，那就是并不把多种模式的数据表征合并，而是保持它们的独立。在文字图片这个例子中，那就是保持文字和图片各自的表征或者特征向量，然后通过某种关系来维持这两种表征之间的联系。

有一种假设就是，虽然各种数据模式的表象是不一样的，例如图片和文字的最终呈现不一样，但是内在都是这个核心内容的某种表述。因此，这些数据模式的内在表达很可能是相近的。

这个假设套用到这里，那就是我们假设文字和图片的各自的表征相近，而这个“相近”是依靠某种相似函数来描述，比如这里就经常使用“**余弦相似函数**”（Cosine Similarity）。

有了上述两种思路之后，一种混合的思路就很自然地出现了。混合思路的基本想法是这样的。数据不同的模式肯定是某种内在表征的不同呈现，因此，需要一个统一的内在表征。但是，只采用一种表征来表达不同的数据源，又明显是不够灵活的。所以，在这种混合的思路里，我们依然需要两种不同的特征来表达文字和图片。

具体来说，混合思路是这样的。首先，我们从文字和图片的原始数据中学习到一个统一的联合表征。然后，我们认为文字和图片各自的表征都是从这个联合表征“发展”或者是“产生”的。很明显，在这样的架构中，我们必须要同时学习联合表征以及两个模式的、产生于联合表征的、单独的各自表征。

值得注意的是，不管是从原始数据到联合表征，还是从联合表征到各自表征，这些步骤都可以是简单的模型，不过通常是多层的神经网络模型。

值得一提的是，在需要多种不同的表征，不管是联合表征还是各自表征的情况中，文字和图片的原始输入甚至是最开始的表征，不一定非要“端到端”（End-to-End）地从目前的数据中学习到。实际上，利用提前从其他数据集中训练好的文字嵌入向量表达来作为文字的输入，是一个非常流行也是非常高效的做法。

有了数据表征之后，很自然地就是利用这些学习到的表征来进行下一步的任务。我们这里就拿文档分类为例。有了联合表征之后，下一步就是利用这个新的表征当做整个文档的特征，学习分类器来进行分类任务。而对于独立的数据表征来说，通常的方法是针对各自表征分别学习一个分类器。这样，我们就有了两个独立的分类器，一个用于文字信息，一个用于图片信息。

有了这两个分类器之后，我们再学习第三个分类器，根据前面两个分类器的分类结果，也就是说这个时候分类结果已经成为了新的特征，来进行第三个分类器的分类。很明显，这个过程需要训练多个不同的分类器，为整个流程增加了不少复杂度。

其他多模数据建模应用

除了我刚才所说的表征的学习以及如何构建分类器以外，多模数据还有一些其他的富有挑战性的任务。

在有文字和图片的情况下，我们经常还需要在这两种模式之间进行转换，或者叫做“翻译”。比如，在已知图片的情况下，如何能够产生一段准确的文字来描述这个图片；或者是在已经有文字的情况下，如何找到甚至产生一张准确的图片。当然，这样的“翻译”并不仅仅局限于文字图片之间，在其他的数据模式中，例如文字和语音之间、语音和图像之间等等，也是普遍存在的。

在这种“翻译”的基础上，更进一步的则是把文字和图片等信息“对接”（Align）起来。比如，针对一组图片，我们能够根据图片的变化产生图片的描述信息。

还有一种应用叫做“**可视化问答**”（Visual Question & Answering），是指利用图片和文字一起回答问题。很显然，要想能够回答好问题，我们需要同时对图片和文字信息进行建模。

不管是“翻译”还是“可视化问答”这些任务，都是近些年来大量利用深度学习所带来的**序列模型**（Sequential Modeling），特别是类似于 RNN 或者 LSTM 等模型的领域。

小结

今天我为你讲了文档理解中的多模数据建模问题。你可以看到这是一个非常火热的领域，如何理解多媒体数据是现代数据处理的一个重要问题 。

一起来回顾下要点：第一，简要介绍了什么是多模数据。第二，详细介绍了多模数据建模的一些基本思路，包括如何获取文档的表征、什么是联合表征和什么是独立表征。然后，我们还讲了如何构建不同的分类器。第三，简要地提及了其他的多模数据建模任务以及这些任务所依靠的基本的深度学习趋势。

最后，给你留一个思考题，多模建模带来了丰富的特性，由这些丰富特性所训练的分类器，就一定能比单一数据源所训练得到的分类器表现得更好吗？

欢迎你给我留言，和我一起讨论。

上一篇　031 | 文档理解的关键步骤：文档聚类

下一篇　033 | 大型搜索框架宏观视角：发展、特点及趋势

精选留言 (1)

写留言

黄德平
2018-12-13

个人觉得不一定，具体要看数据量的大小。数据量少时，使用多模态数据增加了特征的维度，训练很容易过拟合，对于预测没有好处。