第2讲 ChatGPT的本质是什么？

前沿课·吴军讲GPT
2023年4月26日
进入课程 >

时长：12分37秒　大小：17.31MB

00:00 / 12:37

宝木 转述

得到，你好！

欢迎来到《前沿课·吴军讲GPT》。

很多人喜欢说“某某技术横空出世”这样的话，渲染一种很神秘的氛围。但我要提醒一个事实：新技术往往不神秘。世界上很多爆红的产品和服务，其实使用的技术都是半个世纪之前的。

比如，如果你听过我讲5G的课，会知道5G通信中最关键的编码技术，低密度奇偶检查码（LDPC）是上个世纪60年代提出的，90年代前甚至被认为是无用的，但是后来5G用上了。

ChatGPT所依赖的语言模型技术也是如此，它到今天也有50多年的历史了。换句话说，ChatGPT的底层技术是50年前的。

今天在媒体上，很多没有开发过相应技术的人，甚至很多学了一点技术的研究生都会讲，技术发展了这么多年，不该早就超越了当时的极限了吗？

其实，无论是在工业上，还是在信息产业上，技术确实是可以改变的，但是物理学和信息论的原理和极限是无法改变的。虽然今天人工智能使用的语言模型和50年前在技术上有很大的不同，但是它们的原理是一样的。这一讲，我们就来把ChatGPT的底层技术，也就是语言模型说明白。

什么是语言模型？

什么是语言模型？顾名思义，就是对人类的语言建立数学模型。这里面最重要的关键词

是“数学”。语言模型不是逻辑框架，不是生物学的反馈系统，而是由数学公式构建的模型。

那么解决和语言相关的问题，为什么要用数学模型呢？

因为这是一条捷径。我们想让计算机理解自然语言，让计算机回答问题，或者进行几种语言之间的翻译，第一反应会是，应该让计算机理解人的语言，让它学会语法……但是人们经过很多次尝试，都失败了，或者说至少到今天还没有做到。而换个思路，如果想办法把这些问题变成数学问题，然后通过计算，就能间接解决这些自然语言处理的问题。于是，语言模型的想法就被提出来了。

最初提出语言模型概念的，是我们前面提到的贾里尼克博士。

1972年，他正式离开康奈尔大学，在IBM负责和语言处理有关的各个项目。当时，IBM垄断着全世界的计算机市场，在当时的IT领域可以讲是独孤求败。于是贾里尼克就要挑战一下自己，做一些传统计算机做不到的事情，特别是和人工智能有关的事情。于是贾里尼克为研究部门挑选了语音识别这个课题，也就是让计算机把人说的话识别出来。

为什么是这个课题呢？语音识别从1946年开始就有人研究，这个历史和电子计算机的历史几乎一样长，但研究了二十多年依然不得要领。当时的计算机，只能识别百十来个单词，而且错误率高达30%，这样的技术当然是没有什么使用价值的。

当时就像前面说的，人们主要是从语音学和语言学入手研究这个问题。贾里尼克和之前那些研究人员所不同的是，他不是语音学家，也不是语言学家，而是一个数学基础非常好的通信专家，是当时美国大学信息论课程教科书的作者。因此，贾里尼克的想法不受之前的限制，他以一种独特的视角来看待自然语言处理问题，把它们都看成是一种通信问题。

自然语言和通信问题，这两件事又有什么关联呢？

根据香农确立的现代通信原理，所谓的通信，也被称为信道的编码和解码，无非是信息源先产生一个原始信息，然后在接收方还原一个和原始信息最接近的信息。

比如，你传输一句话，“中国是一个古老的国家”。在传输前要对它进行编码，比如编成010101111000……但是，传输中一定会有噪

音和信号损失，接收方接收到的编码可能是1010111000……这样就翻译不回原来的句子了。

那怎么办呢？我们可以把和接收到的编码相似的句子都列举出来。比如：

国中是一个古老的国家

中国是一个古老的国家

国是一个古老的国家

中国一个古老的国

等等。

然后，通信系统会计算哪一种可能性的概率最大，然后把它选出来。只要传输中的噪音不是太大，而且传输的信息有冗余，我们就都能复原原来的信息。贾里尼克做研究的时候，已经有很多通信的算法来做这件事，我在《数学之美》这本书中已经有所介绍。

从这个角度来看待语音识别也是一样的。当人和人交谈的时候，我说“中国是一个古老的国家”这句话，在空气中或者电话线上传播的是声音的波形，而在接听者那里听到的，其实是带有噪音的声音，他需要接收声音的波形，来还原讲话人说的话。只要噪音不是太大，人是能够做到这件事的。

当然，要让计算机来做这件事，不是让计算机学着人的做法去理解语言，而是最好能够让计算机计算出来哪一种可能的语句概率最大。这种计算自然语言每个句子概率的数学模型，就是语言模型。

比如在刚才讲的例子中：

“国中是一个古老的国家”的概率是0.05，“中国是一个古老的国家”的概率是0.2，其它几个候选句子的概率都是0.01，于是我们就认为，概率最高的“中国是一个古老的国家”，就是讲话人讲的句子。

如何消除不确定性？

当然，爱动脑筋的人会马上指出，如果我想讲的就是那些小概率的事情呢？你用最大的概率来预测不是就错了吗？

这个观察很敏锐，这确实是语言模型的问题。而解决这个问题的办法就是利用更多的上下文信息，消除所有的不确定性。

比如，第一代语言模型用的上下文信息就很少，但是到了GPT-3，就用到了前后2000个

词的信息，包括标点符号等，都算成是词。

由于自然语言中有信息冗余，在这么多上下文里，几乎就不存在不确定性了。这也是为什么今天ChatGPT产生的语句，已经很像人说的话了。但**从本质上讲，它的原理依然是在很多种候选中，选择一个概率或者是可能性最大的句子**。这一点是没有改变的。

模型的概率怎么计算？

接下来的问题是，这个概率该怎么计算，或者说好坏该如何评估呢？

早期的语言模型只是看上下文。还拿“中国是一个古老的国家”举例子，在这句话里，“中国”这个词就比“国中”放到开头要来的通顺。什么叫通顺？就是大家都这么说，用的时候多，并不一定要符合语法。事实上，在生活中很多常见的说法其实都是病句，但是语言模型不考虑这种情况，它只是认为人们说得多的就是好的句子。

那怎么准确计算这个概率呢？就要做一些统计了，统计一下在相同的上下文的条件下，每个词具体出现了多少次。因此，语言模型也被称为是统计语言模型，因为它模型的参数，也就是用来计算各种概率的参数，都是靠统计得出的。

讲到统计之前，我们先要做一个说明，就是**今天对于语言模型参数的统计并不是简单的数数，而是要用很复杂的机器学习方法反复计算**。我们后面会讲语言模型的三个发展阶段，就是根据如何得到模型参数来区分的。但是为了简单起见，我们可以先把它理解为数数。

为了统计出语言模型的参数，就需要事先准备好大量的文本供统计使用。

比如有两个句子，“天为什么是蓝色的”和“天为什么是绿色的”，哪一个概率更大？

我们很容易想到，“蓝色”的概率更大，因为我们经常会在文本中看到这句话，比如在1亿篇文章和书籍的章节中看到了100次，它的概率就是百万分之一。而后一句话没有看到，概率就认为是零。

但是，如果某句话，比如“天为什么是黄色的”在统计的文本中出现了一次，是否我们可以认为它比那些没有出现过的句子概率大呢？这就不好说了，因为这些小概率事件，出现不出现，都有很大的随意性。出现两次的随机事件，也未必能说明它比出现一次的

随机事件发生的概率更大。

为了避免这种所谓小概率事件所带来的噪音，我们能做的就是增加数据量。

2000年前后，我在训练语音识别所使用的语言模型时，只用到了几千万个英语的句子。但是到了2012年，我负责开发计算机问答时，训练的数据就扩大到当时互联网上能找到的全部、上百亿个句子，也就是说，十多年增加了上千倍。

今天，ChatGPT的语言模型所用的训练数据量也是很大的，第一个版本使用的GPT-3用了大约5000亿个词，换算成句子大约是500亿个。GPT-4因为模型规模增加了将近一个数量级，训练数据可能增幅更大了。

当然，提高语言模型的准确性，光增加数据量不够，还需要保证数据的质量，这一点我们后面会讲到。

如何利用语言模型写唐诗？

好，语言模型介绍得差不多了。现在，我们用上刚刚讲的知识，解释一下如何利用语言模型写唐诗。

唐诗大多属于韵律诗，它不仅最后一个字押韵，而且每一句诗都是按照一组两个字或者三个字的单元构成的。

比如王之涣的《登鹳雀楼》：“白日依山尽，黄河入海流，欲穷千里目，更上一层楼。”每一句都可以拆成“2-3”组合。杜甫的《登高》：“无边落木萧萧下，不尽长江滚滚来。”每一句都可以拆成“2-2-3”组合。

于是，我们就把所有的绝句和律诗拆成“2-3”或者“2-2-3”组合，然后以两个字或者三个字作为基本的统计单元进行统计，得到它们在上下文中的概率，然后就可以写出一句句概率比较大的诗句了。

当然，我们还要再根据每一个词的语义，把唐诗中出现的词归类，让一首诗中的每一句符合同一个主题。这里面还有一些具体的细节，我就省略了。总之，只要搞懂了语言模型的道理，有现成的诗供学习，让计算机写诗并不是一件很难的事情。

类似地，让计算机写每周汇报，方法也大致相似，因为这类文章几乎都遵循固定的模版。美国很多作家在测试了ChatGPT后，对它进行了逆向工程，认为它是按照五段论写作的，也就是包括开头和结论，以及中间的三个要点，当然这个要点可以增加或者减

少，要点之间再做一些承接和转折。五段论是美国初中教学生们写作的基本方法，美国大部分篇幅不长的公文，比如电子邮件，都是这么写的。

下节预告

你可能会好奇：一开始，语言模型只能完成语音识别、手写体输入的纠错、拼写纠错这些任务。而现在，ChatGPT已经可以回答我们提出的问题，写出很像样的文章了。两相比较，真是强太多了。这当中发生了什么呢？

下一讲，我们就来讲讲语言模型的发展过程，以及它对人工智能带来的影响。我是吴军，我们下一讲见。

划重点 添加到笔记

1.语言模型不是逻辑框架，不是生物学的反馈系统，而是由数学公式构建的模型。
2.今天对于语言模型参数的统计并不是简单的数数，而是要用很复杂的机器学习方法反复计算。

《前沿课·吴军讲GPT》
成为搞懂GPT的少数人
版权归得到App所有，未经许可不得转载
作家，计算机科学家
吴军

首次发布: 2023年4月26日