第3讲 语言模型是如何进化的？

前沿课·吴军讲GPT
今天

进入课程

时长：12分31秒 大小：17.18MB

00:00 / 12:31

宝木 转述

得到用户，你好！

欢迎来到《前沿课·吴军讲GPT》。

上一讲我们说了，ChatGPT的本质是语言模型，就是对人类的语言建立数学模型，也解释了语言模型的工作原理。今天各种语言模型的底层技术，其实出现在50年以前。

不过，不用我说你也知道，像GPT-3和GPT-4这样的语言模型，已经比早期的语言模型强大太多了。如同汽车、移动电话和计算机在不断进步一样，语言模型诞生之后也在持续进步。

这一讲，我们就讲讲语言模型进化的历史，说说自然语言处理的要点。

我知道，大部分同学不会做这方面的工作，但是了解一下自然语言处理的发展趋势和目前方法的局限性，可以帮我们理性看待ChatGPT，消除对于人工智能的迷信。

阶段一

语言模型发展的第一个阶段是上个世纪90年代之前，当时，贾里尼克等人用它解决了语音识别的问题。

在此之前，计算机只能识别百十来个英语单词，识别的错误率还非常高。在采用通信中的算法，并且引入语言模型之后，IBM的语音识别系统可以识别22000个英文单词，而且错误率从过去的30%左右降低到了10%以内。

这意味着什么？当识别的错误率控制在10%以内，也就是每识别十个词出错不到一个词

的时候，就有应用价值了。因为这时候，人可以根据自己的知识纠正那一个错字。

这件事，对当时的一个年轻人触动很大，他就是还在卡内基梅隆大学读博士的李开复。

在征得导师同意后，李开复决定抛开卡内基梅隆大学所走的传统的人工智能道路，利用IBM的方法实现一个实用的语音识别系统。后来他做到了，并且成为当时美国的科技新星，还到CBS电视台向观众展示了他打造的语音识别系统。随后，很多家公司高薪挖他，这让他彻底离开学术界，进入工业界。

但是公平地讲，当时的语音识别系统离实用还差得远，和今天的更是完全无法相比。不过，这件事起码证明了，语言模型确实有用。

上世纪90年代还发生了一件事，就是贾里尼克的一个下属——彼得·布朗——利用语言模型实现了英语和法语之间的自动翻译。

相比语音识别，机器翻译的难点在于要准备两种语言的语言模型，而且最好利用两种语言一一对应的句子来做训练。如果你用狄更斯的英语小说和雨果的法语小说来训练，它们对应不上，效果自然就不会好。由于当时计算机上一一对应的多语种文本只有《圣经》和少量的联合国文件，因此布朗的翻译系统效果并不好。

不久之后，布朗干脆离开学术界，成为著名对冲基金文艺复兴技术公司的技术负责人，他建立数学模型解决问题的本事让基金和他本人都挣了很多钱。

在当时，布朗的成果并没有受到学术界的重视。直到十多年后，德国科学家弗朗茨·奥科在Google用类似的方法比较好地解决了机器翻译问题，布朗最初的那篇论文才成为那个领域被引用数量最多的论文之一。当然，奥科也用了比当时其它研究机构多上万倍、比布朗多上百万倍的数据。不过，这都是后话了。

阶段二

说回语言模型。

进入到90年代后，人们发现，单纯增加训练数据的做法并不是十分有效。贾里尼克自己做了一个统计，发现不论数据如何增加，都会有大量的语言学现象覆盖不了。于是，人们开始考虑用语言中更深的知识和信息建立语言模型。语言模型的发展由此进入到了第

二阶段。

我就是在那个背景下进行语言模型研究的。当时，我们用了语言中的语法和语义信息，可以在不增加数据量的情况下大大提高语言模型的能力。

比如，读音相似的词，在不同的主题中，出现的可能性相差巨大。举个例子，和“事实”（不含音调）同音的词就有四十多个，在过去，到底是哪一个就很不好确定。但是加入语法和语义信息后，我们就知道——“时事”“事实”“失实”这些词会在时政新闻中经常出现，而“失事”“史实”“诗史”等词会在历史题材的文章中经常出现。很多时候，确定了主题，就知道该使用什么词语了。

当语言模型加入了语法和语义信息后，语言模型的训练，以及使用它计算概率时，就不再是简单的统计了。

在此之前的语言模型，稍微有一些概率论和统计学背景，学过一门自然语言处理课程的人基本都能构建。但是此后的语言模型，公式非常复杂，绝大部分人是搞不懂的。

当然，**把更多信息，比如语法信息、语义信息，加入语言模型后，人们面临的更大的挑战是，训练模型的计算量变得特别大。**

关于训练模型的计算量，我们后面还会讲到，这里举一个小例子:

在我之前，IBM有一位科学家，他是世界上第一个成功利用语义信息训练语言模型的人。他主要的工作，不是改进算法或者数学模型，而是写了一个程序，能够把IBM沃森实验室所有空闲的服务器都用上，然后计算了一年，完成了模型的训练。

要知道，沃森实验室是一个有上千名科学家的大型研究机构。这么大的研究机构，这么多的服务器，在空闲时间计算一年，你想想这是多大的计算量。

由于摩尔定律的作用，也就是计算机处理器的性能每18个月翻一番，我们有能力进行越来越复杂的计算，能处理越来越多的数据，也能利用和存储更多的信息。但是，这个计算量的增长是超过计算能力的提升的。

比如，2012年，我们在Google做机器问答时，就需要对几百亿个句子做深入的语法分析，提取各种有用的信息，而不是简简单单地进行统计。这个计算量，即便是Google也感到很吃力。

今天，ChatGPT的语言模型利用的语言信息

更加丰富，计算量也更大，即使利用大量的计算能力更强的GPU，计算的成本也很高。

除了挖掘语言中更多的信息，从上世纪90年代起，科学家还提出了自适应的语言模型。**也就是说，一个人使用语言模型的过程中，他个人输入的信息和间接提供的反馈信息，能够被用来改进语言模型。**

比如，我们用语言模型构建一个拼音输入法，使用者在不断输入中文时，他输入的中文文本就被用来做训练数据了。这样一来，这个输入法软件就会越用越聪明。

当然，这种做法不仅需要个人数据，还需要大量的存储空间。因为这时候，不同的人使用的语言模型就已经不同了。

今天的ChatGPT其实就有这个功能，它会根据用户的输入不断迭代优化。你如果用它写几篇文章，然后不断对那些文章进行修改，它就会学到你的行文习惯，然后根据你的行文习惯输出内容。但是，如果人们刻意误导它，输入垃圾信息，它输出的文本的质量也会跟着下降。

刚才说了，挖掘语言中更深层的信息，能有效提高语言模型的能力，但与此同时，也让计算量变得巨大。因此，为了让计算机更有效率地工作，在2000年之后，很多语言模型的研究工作就转到提高算法效率上了。

比如，我在博士工作的后期，花了一年多的时间改进训练方法，将计算量减少了两到三个数量级。也就是说，只需要之前千分之一到百分之一的计算资源，就能做同样的事情。

阶段三

到了2010年前后，Google开发了深度学习的工具Google大脑：**一方面，能够更有效地利用计算资源，这使得语言模型能够越做越大；另一方面，也让模型计算出的概率越来越准确。**语言模型的发展由此进入第三阶段。

在随后的几年里，Google做了一件事，就是把之前的语言模型重新训练一遍。在训练数据没有增加的情况下，模型的准确性提高了很多，相应产品的质量也提高了。

由于数据量的增加和语言模型准确率的提高，到了本世纪的第二个10年，计算机就逐渐做到了从写一句话到写一个完整的段落。

在此之前，无论是语音识别，还是机器翻

译，你都需要给它一个完整的句子，让它生产一个具有同等信息量的句子。比如，你给它一句中文，它产生相应的英文；或者你给它一个句子的语音，它产生相应的文字。输入输出是1:1。

而当语言模型足够强大之后，你给它一点输入，它就可以产生大量输出了。

比如，你问它一个问题，可能只有一句话，它生成的答案可能有一大段文字。你给它几个关键词，让它生成一份简历，它可能会给你产生一页纸的内容。

了解信息论原理的同学肯定会知道，输入的信息量少、输出的信息量多，这中间的差异必须有办法给补回来，否则输出的内容里一定会充满不确定性。

补充的信息从哪里来呢？实际上，都是存在于语言模型中的，需要先把信息输入语言模型。这时的语言模型就是所谓产生式的了。GPT中的G，就代表Generative，产生。

比如，GPT-3的训练数据就包括几十万本图书、几乎全部维基百科的内容。这么多的数据还只是它所用数据的五分之一不到。实际上，单是维基百科的内容，几乎就能回答今天绝大部分人提出的所有有意义的问题了。因此，ChatGPT的表现比很多人好并不奇怪，因为人一辈子学习的内容可能连它的千分之一都没有。

当然，光有信息还不够，为了保证一段文字中各个句子之间的连贯性，还需要语言模型能够计算出几个句子出现在一起的概率。在段落这个层级进行计算，要比在句子这个层级计算，计算量大出成千上万倍，而不是几倍几十倍。

同时，为了让语言模型产生的段落不太发散，通常还要用一些模版对生成的文字进行限制。这也是为什么ChatGPT写汇报和简历能写得很好，写不受限制的文章就显得非常枯燥的原因。

类似的，今天的写作软件通常只能应付有套路的写作要求，比如写财经报道、写大学申请材料等。所谓的“套路”，其实就是计算机使用的模版。今天，美国的作家群体普遍认为，ChatGPT只能按照一定的模版写作。

当然，要让ChatGPT能够回答好问题，或者和人进行对话，还需要让它理解人的问题。这件事相对写作本身要容易一些。

今天，**计算机对自然语言做语法分析，可以**

做得比人还准确。通过句子的语法结构，可以理解常见的实词，也就是名词、动词和形容词的含义。通常来说，句子的含义是这些词的含义的叠加，于是，它就能理解这句话的含义。

当然，在丰富的人类语言中，也经常有例外，就是一句话中所有的词你都懂，但是这句话听不懂。对于这种情况，计算机也没有办法，这时候，它就会胡乱回答，甚至给出完全相反的答案。

下节预告

好，这一讲就到这里。

语言模型的算法其实不仅能处理和自然语言相关的问题，从理论上讲，用其它数据建立类似的模型，就能处理大量的、原本需要利用人类智能才能解决的问题。这些内容，我们下一讲再讲。我是吴军，我们下一讲见。

划重点 添加到笔记

语言模型的发展大致经历了三个阶段：
1、上个世纪90年代之前，当时贾里尼克等人用它解决了语音识别的问题。
2、90年代后，语言模型加入了语法和语义信息，还出现了自适应的语言模型。
3、2010年前后，Google开发了深度学习的工具Google大脑，能够更有效地利用计算资源，也让模型计算出的概率越来越准确。

《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: 今天