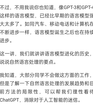
第3讲 语言模型是如何进化的？


前沿课 黄天培NLP

已上新 ｜ 已学习



倍速应用更广泛



定义 转述

得到用户，你好！

前课来到《前沿课 黄天培NLP》，

上一讲我们讲了，ChatGPT的主流底座是语言模型，就是对人类的语言建立数学模型，也就是用一个数学函数来表示语言。

我们知道，语言模型可以简单分为统计语言模型和神经网络语言模型。

不过，不管是统计语言模型，还是GPT-3和GPT-4这样的大语言模型，它们的核心任务都是一样的，就是预测一个词出现在某句话中的概率。

不同点在于，统计语言模型是通过计算词频来估计概率，而神经网络语言模型则是通过训练神经网络来学习概率。

这一讲，我们来聊聊语言模型的进化史，说说自然语言处理是如何一步步发展的。

阶段一

语言模型发展的第一个阶段，是统计语言模型。

它的思想很简单，就是假设一个词出现的概率，只和它前面的 n 个词有关。

这个 n 通常不会太大，一般是 2 到 5。

比如 n 取 2，我们就说一个词出现的概率，只和它前面一个词有关。

这种假设叫做马尔可夫假设，也叫 n 元语法模型。

n 元语法模型的优点是简单，计算起来很快。

只要统计一下词频，就能估算出概率。

但缺点也很明显，就是没有记忆。

它只看前面 n 个词，超过这个范围的词，就完全不管了。

这就导致它很难理解长句子的含义。

比如，“我昨天在公园遇到了一只可爱的小狗，它正在追逐一只蝴蝶”，这句话里，“它”指的是“小狗”，而不是“公园”或“蝴蝶”。

统计语言模型就很难处理这种指代关系。

为了克服这个缺陷，研究者们提出了神经网络语言模型。

它的核心思想是，用一个神经网络来学习词语之间的关系，从而预测下一个词出现的概率。

这样，模型就有了记忆能力，可以记住更长的上下文信息。

阶段二

神经语言模型，这个概念最早是在 2003 年，由蒙特利尔大学的 Yoshua Bengio 等人提出的。

他们使用了一个简单的前馈神经网络，来预测给定前面几个词之后，下一个词出现的概率。

这个模型虽然结构简单，但已经展现出了神经网络在语言建模方面的巨大潜力。

随后，Bengio 团队又在 2006 年提出了一个更深层的神经网络模型，进一步提升了模型的性能。

这两篇论文，为后续的研究奠定了基础，也开启了神经语言模型的时代。

2013 年，Google 团队发表了一篇论文，介绍了一种名为 Word2Vec 的词向量模型。

这个模型，可以将每个词语，映射到一个低维的向量空间中，使得语义相近的词语，在向量空间中的距离也更近。

比如，“国王”和“女王”的向量，在空间中的距离，就比“国王”和“苹果”的向量，要近得多。

Word2Vec 的出现，极大地推动了自然语言处理领域的发展，也为后来的预训练模型，奠定了重要的基础。

阶段三

2017 年，Google 团队发表了一篇论文，介绍了一种名为 Transformer 的神经网络架构。

这个架构，彻底改变了自然语言处理领域的格局，也为后来的大语言模型，比如 GPT 系列，奠定了基础。

Transformer 架构的核心，是自注意力机制。

它可以让模型在处理每个词时，都关注到句子中所有其他词的信息，从而更好地理解上下文关系。

这种机制，使得模型能够捕捉到长距离的依赖关系，从而更准确地理解句子的含义。

2018 年，OpenAI 团队发布了 GPT-1，这是第一个基于 Transformer 架构的预训练语言模型。

GPT-1 的出现，标志着预训练语言模型时代的开始。

随后，OpenAI 团队又陆续发布了 GPT-2、GPT-3 和 GPT-4，不断扩大模型的规模和训练数据，模型的性能也得到了显著提升。

下节预告

好，这一讲就到这里。

划重点 AI 助你划重点

语言模型的发展大致经历了三个阶段：

1. 统计语言模型：基于词频，简单但记忆力差。
2. 神经网络语言模型：引入神经网络，提升了模型的表达能力。
3. Transformer 架构：引入自注意力机制，能够捕捉长距离依赖关系，成为当前主流。

