用最少的数学和术语来解释大模型

蒂莫西·B·李、肖恩·特罗特

来源 https://www.understandingai.org/p/large-language-models-explained-with

ChatGPT 去年秋天推出时，给科技行业和更广阔的世界带来了冲击波。那时，机器学习研究人员已经对大型语言模型（LLM）进行了几年的实验，但公众并没有密切关注，也没有意识到它们已经变得多么强大。

如今，几乎每个人都听说过 LLM，并且有数以千万计的人尝试过。但是，仍然没有多少人了解它们是如何工作的。

如果您对这个主题有所了解，您可能听说过 LLM 接受过 “预测下一个单词” 的训练，并且他们需要大量的文本才能做到这一点。但解释往往就到此为止了。他们如何预测下一个单词的细节通常被视为一个深奥的谜团。

原因之一是这些系统的开发方式不寻常。传统软件是由人类程序员创建的，他们向计算机提供明确的分步指令。相比之下，ChatGPT 建立在使用数万亿普通语言单词进行训练的神经网络之上。

结果，地球上没有人完全理解 LLM 的内部运作。研究人员正在努力获得更好的理解，但这是一个缓慢的过程，需要数年甚至数十年才能完成。

尽管如此，专家们仍然对这些系统的工作原理有很多了解。本文的目标是让广大读者能够了解大量此类知识。我们的目标是解释这些模型的内部工作原理，而不诉诸技术术语或高级数学。

我们将从解释词向量开始，这是语言模型表示和推理语言的令人惊讶的方式。然后我们将深入研究 Transformer，它是 ChatGPT 等系统的基本构建块。最后，我们将解释这些模型是如何训练的，并探讨为什么良好的性能需要如此大量的数据。

艾米莉·本德（Emily Bender）称这些模型为“随机鹦鹉”，这是完全正确的。无论复杂性如何增加，都无法将非理性的、纯粹确定性的数学过程转变为对真理的理性理解。因此，“幻觉”。这些模型就像一面文化镜子：如果当我们凝视它们时，我们看到的看起来很人类，那是因为我们是人类。这绝对不是因为镜子自发地变成了人。

词向量

要了解语言模型的工作原理，您首先需要了解它们如何表示单词。人类用一系列字母来表示英语单词，例如猫的 C-A-T。语言模型使用一长串数字来表示单词，称为词向量。例如，以下是将猫表示为向量的一种方法：

```
[0.0074, 0.0030, -0.0105, 0.0742, 0.0765, -0.0011, 0.0265, 0.0106, 0.0191, 0.0038, -0.0468, -0.0212,
1
0.0091, 0.0030, -0.0563, -0.0 396, -0.0998, -0.0796, …, 0.0002]
```

为什么要使用这种巴洛克符号？这是一个类比。华盛顿特区位于北纬 38.9 度、西经 77 度。我们可以使用向量符号来表示：

华盛顿特区在 [38.9, 77]

纽约在 [40.7, 74]

伦敦在 [51.5, 0.1]

巴黎在 [48.9, -2.4]

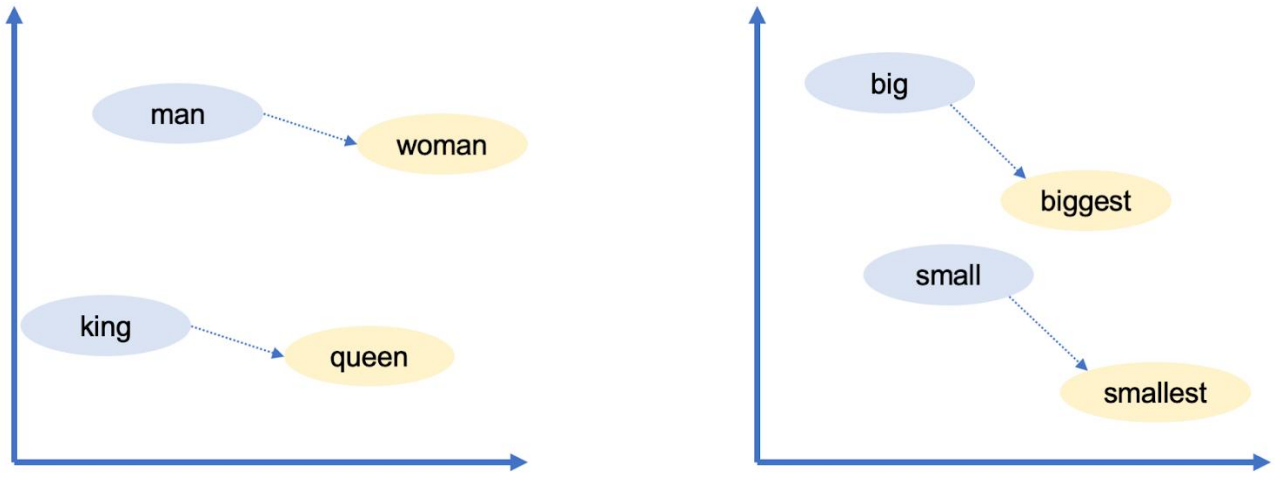
这对于推理空间关系很有用。您可以看出纽约与华盛顿特区接近，因为 38.9 接近 40.7，77 接近 74。同样，巴黎接近伦敦。但巴黎距离华盛顿特区很远。

语言模型采用类似的方法：每个单词向量代表想象的 “单词空间” 中的一个点，并且具有更相似含义的单词被放置得更近。例如，向量空间中与猫最接近的单词包括狗、小猫和宠物。用实数向量（而不是一串字母，如“C-A-T”）表示单词的一个关键优势是数字可以实现字母所不能的操作。

单词太复杂，无法仅用二维表示，因此语言模型使用数百甚至数千维的向量空间。人类思维无法想象具有如此多维度的空间，但计算机完全有能力对它们进行推理并产生有用的结果。

研究人员几十年来一直在尝试词向量，但当谷歌在 2013 年宣布其 word2vec 项目时，这个概念才真正开始流行。谷歌分析了从谷歌新闻中收集的数百万份文档，以找出哪些词往往出现在相似的句子中。随着时间的推移，经过训练来预测哪些单词与哪些其他单词同时出现的神经网络学会了将相似的单词（例如狗和猫）在向量空间中靠近放置。

谷歌的词向量还有另一个有趣的特性：你可以使用向量算术 “推理” 单词。例如，谷歌研究人员将向量取为最大，减去大的，加上小的。最接近结果向量的单词是最小的。



你可以用向量算术来进行类比！在这种情况下，大就是最大，小就是最小。谷歌的词向量捕获了许多其他关系：

- Swiss is to Switzerland as Cambodian is to Cambodia. (nationalities)
- Paris is to France as Berlin is to Germany. (capitals)
- Unethical is to ethical as possibly is to impossibly. (opposites)
- Mouse is to mice as dollar is to dollars. (plurals)
- Man is to woman as king is to queen. (gender roles)

由于这些向量是根据人类使用单词的方式构建的，因此它们最终反映了人类语言中存在的许多偏见。例如，在某些词向量模型中，医生减去男人加女人得到护士。减少这样的偏见是一个活跃的研究领域。

尽管如此，词向量是语言模型的有用构建块，因为它们编码有关词之间关系的微妙但重要的信息。如果一个语言模型了解了关于猫的一些信息（例如：它有时会去看兽医），那么同样的事情也可能适用于小猫或狗。如果模型了解了巴黎和法国之间的关系（例如：他们共享一种语言），那么柏林和德国、罗马和意大利也很可能也是如此。

词义取决于上下文

像这样的简单词向量方案并不能捕捉到有关自然语言的一个重要事实：单词通常具有多种含义。

例如，银行一词可以指金融机构或河边的土地。或者考虑以下句子：

- John picks up a magazine.
- Susan works for a magazine.

这些句子中杂志的含义相关，但略有不同。约翰拿起一本实体杂志，而苏珊在一家出版实体杂志的组织工作。

当一个词有两个不相关的含义时，例如银行，语言学家将它们称为同音异义词。当一个词有两个密切相关的含义时，例如杂志，语言学家将其称为一词多义。

像 ChatGPT 这样的 LLM 能够根据该单词出现的上下文，用不同的向量表示相同的单词。有一个银行（金融机构）向量和一个不同的银行（河流）向量。有一个用于杂志（实体出版物）的矢量，另一个用于杂志（组织）的矢量。正如您所料，LLM 在多义含义上使用比同音含义更多的相似向量。

到目前为止，我们还没有谈到语言模型如何做到这一点——我们很快就会讨论这一点。但我们正在开发这些向量表示，因为它是理解语言模型如何工作的基础。

传统软件旨在对明确的数据进行操作。如果你要求计算机计算“2 + 3”，那么 2、+ 或 3 的含义就不会含糊不清。但自然语言充满了超越同音异义词和一词多义的歧义：

- In “the customer asked the mechanic to fix **his** car” does **his** refer to the customer or the mechanic?
- In “the professor urged the student to do **her** homework” does **her** refer to the professor or the student?
- In “fruit **flies** like a banana” is **flies** a verb (referring to fruit soaring across the sky) or a noun (referring to banana-loving insects)?

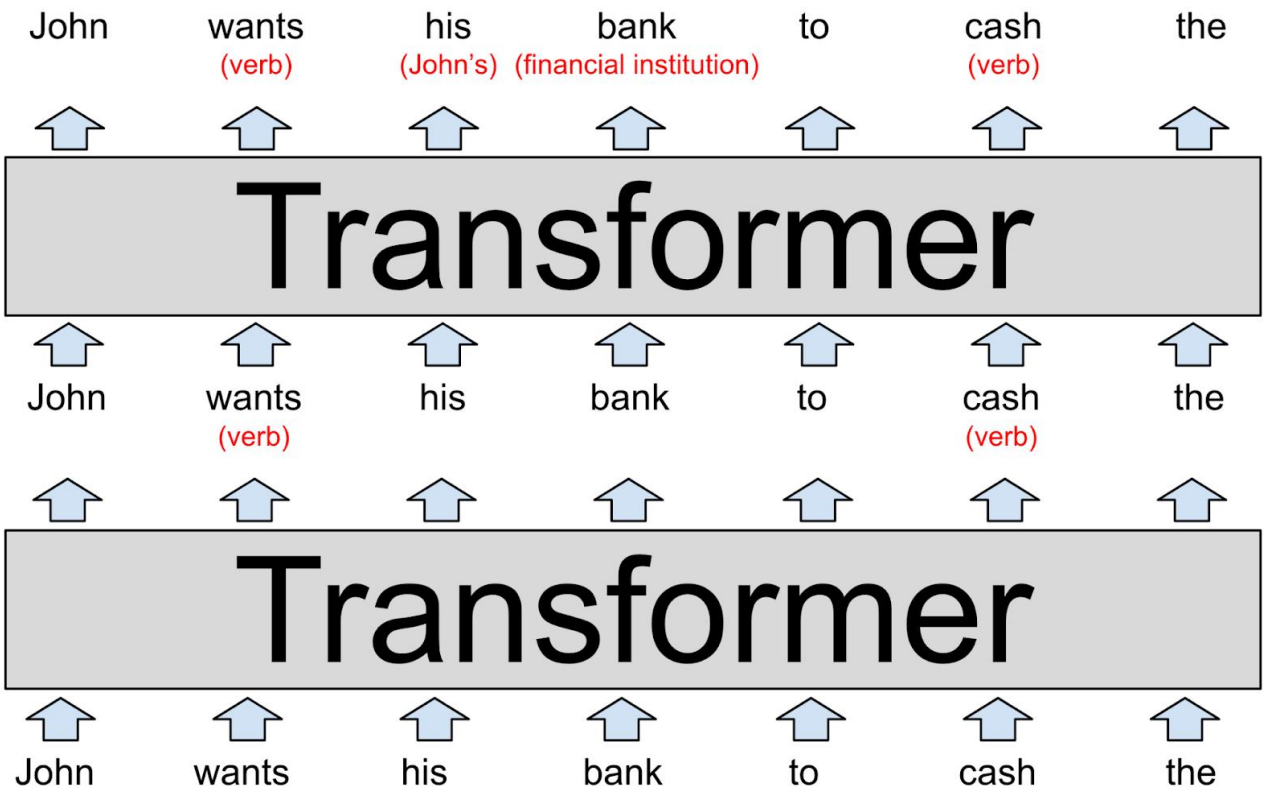
人们根据上下文解决这样的歧义，但是没有简单或确定的规则来做到这一点。相反，它需要了解世界的事实。你需要知道机械师通常会修理客户的汽车，学生通常会做自己的作业，而水果通常不会飞。

词向量为语言模型提供了一种灵活的方式来表示特定段落上下文中每个词的精确含义。现在让我们看看他们是如何做到的。

将词向量转换为词预测

GPT-3 是 ChatGPT2 原始版本背后的模型，分为数十层。每一层都采用一系列向量作为输入（输入文本中的每个单词都有一个向量），并添加信息以帮助阐明该单词的含义并更好地预测下一个可能出现的单词。

让我们首先看一个程式化的例子：



LLM 的每一层都是一个 Transformer，这是 Google 在 2017 年一篇具有里程碑意义的论文中首次引入的神经网络架构。

该模型的输入（如图底部所示）是部分句子“约翰希望他的银行兑现”。这些单词表示为 word2vec 式向量，被输入到第一个 Transformer 中。

Transformer 发现需求和现金都是动词（这两个词也可以是名词）。我们将这个添加的上下文表示为括号中的红色文本，但实际上，模型会通过以人类难以解释的方式修改单词向量来存储它。这些新向量（称为隐藏状态）被传递到堆栈中的下一个转换器。

第二个转换器添加了另外两个上下文：它澄清了银行指的是金融机构而不是河岸，并且 his 是指约翰的代词。第二个 Transformer 产生另一组隐藏状态向量，反映模型到目前为止所学到的一切。

上图描绘了一个纯粹假设的 LLM，所以不要太认真地对待细节。我们很快就会看到对真实语言模型的研究。真正的 LLM 往往有不止两层。例如，最强大的 GPT-3 版本有 96 层。

研究表明，前几层侧重于理解句子的语法并解决歧义，如我们上面所示。后面的层（我们没有展示这些层是为了使图表保持在可管理的大小）致力于对整个段落有一个高层次的理解。

例如，当 LLM “通读”一个短篇小说时，它似乎会跟踪有关故事人物的各种信息：性别和年龄、与其他人物的关系、过去和当前的位置、个性和目标等等..

研究人员并不确切了解 LLM 如何跟踪这些信息，但从逻辑上讲，模型必须通过修改隐藏状态向量来实现这一点，因为它们从一层传递到下一层。在现代 LLM 中，这些向量非常大，这很有帮助。

例如，最强大的 GPT-3 版本使用 12,288 个维度的词向量，即每个词由 12,288 个数字的列表表示。这比 Google 2013 年的 word2vec 方案大 20 倍。您可以将所有这些额外维度视为一种“暂存空间”，GPT-3 可以使用它来为自己写下有关每个单词上下文的注释。后面的层可以读取和修改前面层所做的注释，从而使模型逐渐加深对整个段落的理解。

因此，假设我们更改了上面的图表，以描述解释 1,000 字故事的 96 层语言模型。第 60 层可能包含约翰的向量，并带有附加注释，例如“（主角，男性，嫁给 Cheryl，Donald 的表弟，来自明尼苏达

州，目前在 Boise，试图找到他丢失的钱包）。”同样，所有这些事实（可能还有更多）都会以某种方式编码为与单词 John 相对应的 12,288 个数字的列表。或者，其中一些信息可能被编码在 Cheryl、Donald、Boise、钱包或故事中其他单词的 12,288 维向量中。

网络的第 96 层（即最后一层）的目标是输出最终单词的隐藏状态，其中包括预测下一个单词所需的所有信息。

Transformer 的注意力机制

现在我们来谈谈每个 transformer 内部发生了什么。transformer 有一个两步过程来更新输入通道中每个单词的隐藏状态：

在注意力步骤中，单词“环顾四周”寻找具有相关上下文并相互共享信息的其他单词。

在前馈步骤中，每个单词“思考”在之前的注意力步骤中收集的信息，并尝试预测下一个单词。

当然，执行这些步骤的是网络，而不是单个单词。但我们这样措辞是为了强调转换将单词而不是整个句子或段落视为分析的基本单位。这种方法使 LLM 能够充分利用现代 GPU 芯片的大规模并行处理能力。它还可以帮助 LLM 扩展到数千字的段落。这些都是早期语言模型陷入困境的领域。

你可以把注意力机制想象成单词的匹配服务。每个单词都会生成一个清单（称为查询向量），描述其正在查找的单词的特征。每个单词还创建一个描述其自身特征的清单（称为关键向量）。网络将每个关键向量与每个查询向量进行比较（通过计算点积）以找到最佳匹配的单词。一旦找到匹配项，它就会将信息从生成键向量的单词传输到生成查询向量的单词。

例如，在上一节中，我们展示了一个假设的转换器，它计算出在部分句子“约翰希望他的银行兑现”中，他指的是约翰。这就是幕后的样子。 his 的查询向量可能会有效地表示“我正在寻找：一个描

述男性的名词”。约翰的关键向量可能有效地表示“我是：描述男性的名词”。网络将检测到这两个向量匹配，并将有关约翰的向量的信息移动到他的向量中。

每个注意力层都有几个“注意力头”，这意味着这种信息交换过程在每一层都会发生多次（并行）。每个注意力头专注于不同的任务：

```
1正如我们上面所讨论的，一个注意力头可能会将代词与名词进行匹配。
2另一个注意力头可能致力于解决像银行这样的同音异义词的含义。
3第三个注意力头可能会将两个单词的短语链接在一起，例如“乔·拜登”。
4等等
```

注意力头经常按顺序操作，一层注意力操作的结果成为后续层注意力头的输入。事实上，我们上面列出的每一项任务都可能很容易需要多个注意力头，而不仅仅是一个。

GPT-3 的最大版本有 96 层，每层有 96 个注意力头，因此 GPT-3 每次预测新单词时都会执行 9,216 次注意力操作。

一个真实的例子

在最后两节中，我们展示了注意力头如何工作的程式化版本。现在让我们看一下真实语言模型内部运作的研究。去年，Redwood Research 的科学家研究了 ChatGPT 的前身 GPT-2 如何预测 “When Mary and John went to the store, John gave a drink to. ” 这句话的下一个单词。

GPT-2 预测下一个词是 Mary。研究人员发现，三种类型的注意力头促成了这一预测：

> 他们称为 Name Mover Heads 的三个头将信息从 Mary 向量复制到最终输入向量（单词 to）。GPT-2 使用最右边向量中的信息来预测下一个单词。

网络如何决定 “Mary” 是复制的正确词？通过 GPT-2 的计算过程逆向工作，科学家们发现了一组由四个注意力头组成的组，他们将其称为主题抑制头，它们以阻止名称移动头复制 John 名字的方式标记第二个 John 向量。

对象抑制头怎么知道 John 不应该被复制？进一步向后追溯，团队发现了两个注意力头，他们称之为“重复令牌头”。他们将第二个 John 向量标记为第一个 John 向量的副本，这帮助对象抑制头决定不应复制 John 。

简而言之，这九个注意力头使 GPT-2 能够发现“John 给 John 喝了一杯”没有意义，并选择“John 给 Mary 喝了一杯”。

我们喜欢这个例子，因为它说明了完全理解 LLM 是多么困难。Redwood 团队的五名成员发表了一篇 25 页的论文，解释了他们如何识别和验证这些注意力头。然而，即使他们完成了所有这些工作，我们仍然远远无法全面解释为什么 GPT-2 决定预测 Mary 作为下一个词。

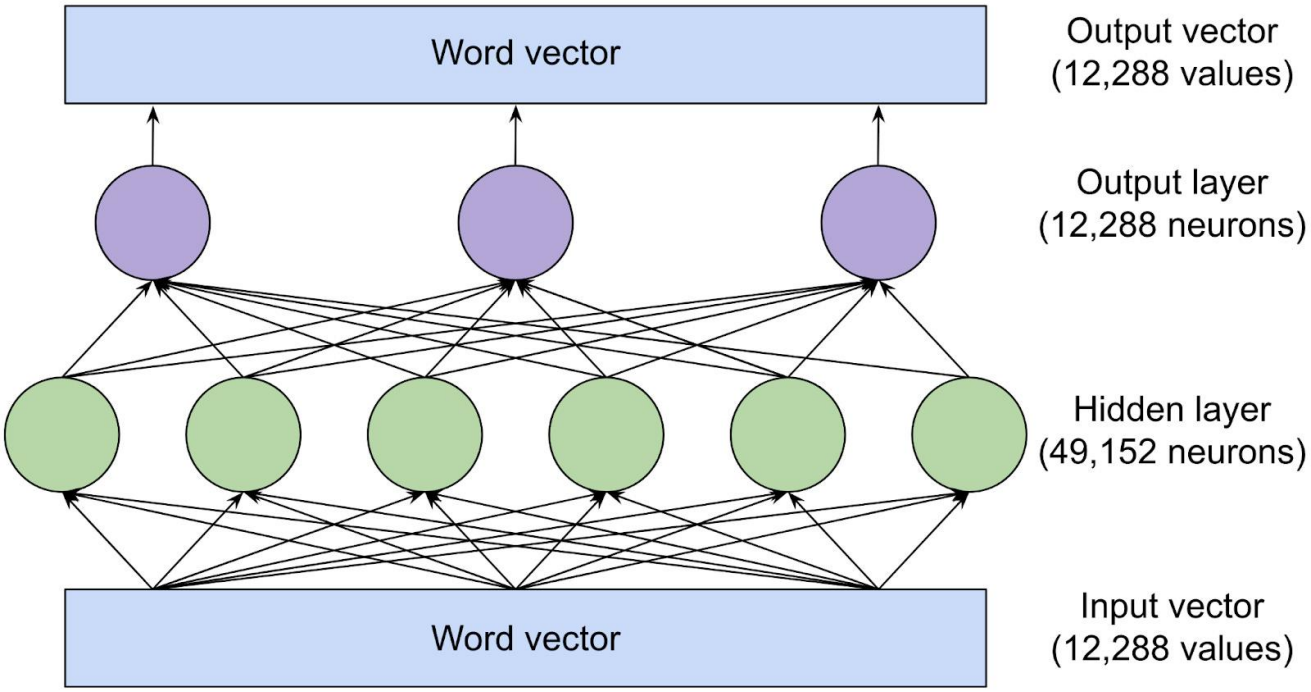
例如，模型如何知道下一个单词应该是某人的名字而不是其他类型的单词？人们很容易想到类似的句子，但 Mary 不能很好地预测下一个单词。例如，在“when Mary and John went to the restaurant, John gave his keys to, ”这句话中，逻辑上的下一个单词将是 “the valet.”。

据推测，通过足够的研究，计算机科学家可以发现并解释 GPT-2 推理过程中的其他步骤。最终，他们可能能够全面理解 GPT-2 如何确定 Mary 是这句话中最有可能的下一个单词。但仅仅理解单个单词的预测就可能需要数月甚至数年的额外努力。

ChatGPT 底层的语言模型（GPT-3.5 和 GPT-4）比 GPT-2 更大、更复杂。它们能够进行比红木团队研究的简单句子完成任务更复杂的推理。因此，充分解释这些系统如何工作将是一个人类不太可能很快完成的巨大项目。

前馈步骤

在注意力头在词向量之间传输信息后，有一个前馈网络会“思考”每个词向量并尝试预测下一个词。在此阶段，单词之间不交换任何信息：前馈层单独分析每个单词。然而，前馈层确实可以访问先前由注意力头复制的任何信息。以下是最大版本的 GPT-3 中前馈层的结构：



绿色和紫色圆圈是神经元：计算其输入的加权和的数学函数。

前馈层的强大之处在于其大量的连接。我们绘制的网络在输出层有 3 个神经元，在隐藏层有 6 个神经元，但 GPT-3 的前馈层要大得多：输出层有 12,288 个神经元（对应于模型的 12,288 维单词）矢量）和隐藏层中的 49,152 个神经元。

因此，在 GPT-3 的最大版本中，隐藏层中有 49,152 个神经元，每个神经元有 12,288 个输入（因此有 12,288 个权重参数）。每个神经元有 12,288 个输出神经元，每个神经元有 49,152 个输入值（因此有 49,152 个权重参数）。这意味着每个前馈层有 49,152 * 12,288 + 12,288 * 49,152 = 12 亿

个权重参数。并且有 96 个前馈层，总共 12 亿 * 96 = 1160 亿个参数！这几乎占 GPT-3 1750 亿个参数总数的三分之二。

在 2020 年的一篇论文中，特拉维夫大学的研究人员发现前馈层通过模式匹配发挥作用：隐藏层中的每个神经元都与输入文本中的特定模式匹配。以下是 16 层版本的 GPT-2 中神经元匹配的一些模式：

第一层的神经元匹配以“替代品”结尾的单词序列。

第 6 层中的神经元匹配与军事相关的序列，并以“base”或“bases”结尾。

第 13 层中的神经元匹配以时间范围结尾的序列，例如“下午 3 点到 7 点之间”或“从周五晚上 7:00 到”。

第 16 层的神经元匹配与电视节目相关的序列，例如“原始 NBC 日间版本，已存档”或“时移观看增加了剧集的 57%”。

正如您所看到的，模式在后面的层中变得更加抽象。早期的层倾向于匹配特定的单词，而后面的层则匹配属于更广泛语义类别的句子，例如电视节目或时间间隔。

之所以有趣，是因为如前所述，前馈层一次仅检查一个单词。因此，当它将序列“原始 NBC 日间版本，存档”分类为与电视相关时，它只能访问存档向量，而不能访问 NBC 或日间等单词。据推测，前馈层可以知道存档是电视相关序列的一部分，因为注意力头先前将上下文信息移动到存档向量中。

当神经元匹配这些模式之一时，它就会向词向量添加信息。虽然这些信息并不总是容易解释，但在许多情况下，您可以将其视为对下一个单词的尝试预测。

前馈网络用向量数学进行推理

布朗大学最近的研究揭示了一个很好的例子，说明前馈层如何帮助预测下一个单词。之前我们讨论了 Google 的 word2vec 研究，该研究表明可以使用向量算术进行类比推理。例如，柏林 − 德国 + 法国 = 巴黎。

布朗大学的研究人员发现，前馈层有时会使用这种精确的方法来预测下一个单词。例如，他们研究了 GPT-2 如何响应以下提示：“问：法国的首都是哪里？ A：巴黎 问：波兰的首都是哪？有：”

该团队研究了 24 层的 GPT-2 版本。在每一层之后，布朗大学的科学家都会对模型进行探索，以观察其对下一个标记的最佳猜测。对于前 15 层，最上面的猜测是一个看似随机的单词。在第 16 层和第 19 层之间，模型开始预测下一个词将是“波兰”——不正确，但正在变暖。然后在第 20 层，最上面的猜测变成了华沙——正确答案——并在最后四层保持这种状态。

布朗大学的研究人员发现，第 20 前馈层通过添加一个将国家向量映射到相应首都的向量，将波兰转换为华沙。将相同的向量添加到中国就产生了北京。

同一模型中的前馈层使用向量算术将小写单词转换为大写单词，将现在时单词转换为过去时态单词。

注意力层和前馈层有不同的工作

到目前为止,我们已经研究了 GPT-2 单词预测的两个现实示例:注意力头帮助预测 John 给 Mary 喝了一杯酒，前馈层帮助预测华沙是波兰的首都。

在第一种情况下，Mary 来自用户提供的提示。但在第二种情况下，华沙并不在提示中。相反，GPT-2 必须“记住”华沙是波兰首都的事实——它从训练数据中学到的信息。

当布朗大学的研究人员禁用将波兰转换为华沙的前馈层时，该模型不再预测华沙为下一个词。但有趣的是，如果他们随后在提示的开头添加“波兰的首都是华沙”这句话，那么 GPT-2 就可以再次回答这个问题。这可能是因为 GPT-2 使用注意力头复制了提示中前面的名称 Warsaw。

这种分工更普遍：注意力头从提示中较早的单词中检索信息，而前馈层使语言模型能够“记住”提示中没有的信息。

事实上，考虑前馈层的一种方法是将其视为模型从训练数据中学到的信息数据库。早期的前馈层更有可能编码与特定单词相关的简单事实，例如“特朗普经常出现在唐纳德之后”。后面的层编码更复杂的关系，例如“添加此向量以将一个国家转换为其首都”。

语言模型是如何训练的

许多早期的机器学习算法需要人类手动标记训练示例。例如，训练数据可能是狗或猫的照片，每张照片都带有人类提供的标签（“狗”或“猫”）。人类标记数据的需要使得创建足够大的数据集来训练强大的模型变得困难且昂贵。

LLM 的一个关键创新是他们不需要明确标记的数据。相反，他们通过尝试预测普通文本段落中的下一个单词来学习。几乎所有书面材料（从维基百科页面到新闻文章再到计算机代码）都适合训练这些模型。

例如，LLM 可能会收到输入“我喜欢加奶油的咖啡和”，并应该预测“糖”作为下一个单词。新初始化的语言模型在这方面会表现得非常糟糕，因为它的每个权重参数（在最强大的 GPT-3 版本中有 1750 亿个权重参数）一开始都是一个本质上随机的数字。

但随着模型看到更多的例子（数千亿个单词），这些权重会逐渐调整，以做出越来越好的预测。

这是一个类比来说明它是如何工作的。假设您要洗澡，并且希望温度恰到好处：不太热，也不太冷。您以前从未使用过这个水龙头，因此您将旋钮指向任意方向并感受水温。如果太热，就向一侧转动；如果太热，就向一侧转动；如果太热，就向一侧转动。如果太冷，你就把它转向另一个方向。越接近正确的温度，所做的调整就越小。

现在让我们对这个类比做一些改变。首先，假设有 50,257 个水龙头，而不是只有一个。每个水龙头对应一个不同的单词，例如“the”、“cat”或“bank”。您的目标是让水仅从与序列中的下一个单词相对应的水龙头中流出。

其次，水龙头后面有一个相互连接的管道迷宫，这些管道上也有一堆阀门。因此，如果水从错误的水龙头流出，您不能只是调节水龙头上的旋钮。你派遣一群聪明的松鼠向后追踪每根管道，并调整他们沿途找到的每个阀门。

这变得很复杂，因为同一根管道经常连接到多个水龙头。因此，需要仔细考虑以确定哪些阀门需要拧紧，哪些阀门需要松开，以及松开多少。

显然，如果你太从字面上理解，这个例子很快就会变得愚蠢。建立一个拥有 1750 亿个阀门的管道网络既不现实，也不有用。但由于摩尔定律，计算机能够而且确实以这种规模运行。

到目前为止，我们在本文中讨论的 LLM 的所有部分（前馈层中的神经元和在单词之间移动上下文信息的注意力头）都被实现为一系列简单的数学函数（主要是矩阵乘法），其行为由可调整的权重参数确定。正如我故事中的松鼠通过松开和收紧阀门来控制水流一样，训练算法也会增加或减少语言模型的权重参数来控制信息如何流经神经网络。

训练过程分两步进行。首先是“向前通过”，打开水，检查水是否从正确的水龙头流出。然后水被关闭，有一个“向后通道”，松鼠沿着每根管道赛跑，拧紧和松开阀门。在数字神经网络中，松鼠的角

色是由一种称为反向传播的算法扮演的，该算法在网络中“向后行走”，使用微积分来估计每个权重参数需要改变多少。

完成这一过程（对一个示例进行前向传递，然后进行反向传递以提高网络在该示例上的性能）需要数千亿次数学运算。训练像 GPT-3 这样大的模型需要重复该过程数十亿次 – 每个训练数据字一次。OpenAI 估计训练 GPT-3 需要超过 3000 亿万亿次浮点计算 – 这是数月的工作适用于数十种高端计算机芯片。

GPT-3 令人惊讶的性能

您可能会惊讶地发现训练的效果如此之好。 ChatGPT 可以执行各种复杂的任务——撰写论文、进行类比，甚至编写计算机代码。那么如此简单的学习机制是如何产生如此强大的模型的呢？

原因之一是规模。像 GPT-3 这样的模型所看到的例子数量之多都不为过。 GPT-3 在大约 5000 亿个单词的语料库上进行训练。相比之下，一个典型的人类儿童在 10 岁时会遇到大约 1 亿个单词。

在过去五年中，OpenAI 不断扩大其语言模型的规模。 OpenAI 在一篇被广泛阅读的 2020 年论文中报告称，其语言模型的准确性“与模型大小、数据集大小和用于训练的计算量呈幂律关系，其中一些趋势跨越了七个数量级以上”。

他们的模型越大，他们在涉及语言的任务上就越好。但只有当他们将训练数据量增加类似倍数时，这才是正确的。要使用更多数据训练更大的模型，您需要更多的计算能力。

OpenAI 的第一个 LLM GPT-1 于 2018 年发布。它使用 768 维的词向量，有 12 层，总共 1.17 亿个参数。几个月后，OpenAI 发布了 GPT-2。它最大的版本有 1600 维词向量，48 层，总共 15 亿个参数。

2020 年，OpenAI 发布了 GPT-3，其特征为 12288 维词向量和 96 层，总共 1750 亿个参数。

最后，今年 OpenAI 发布了 GPT-4。该公司尚未公布任何架构细节，但人们普遍认为 GPT-4 比 GPT-3 大得多。

每个模型不仅比其较小的前辈学到了更多的事实，而且在需要某种形式的抽象推理的任务上也表现得更好：

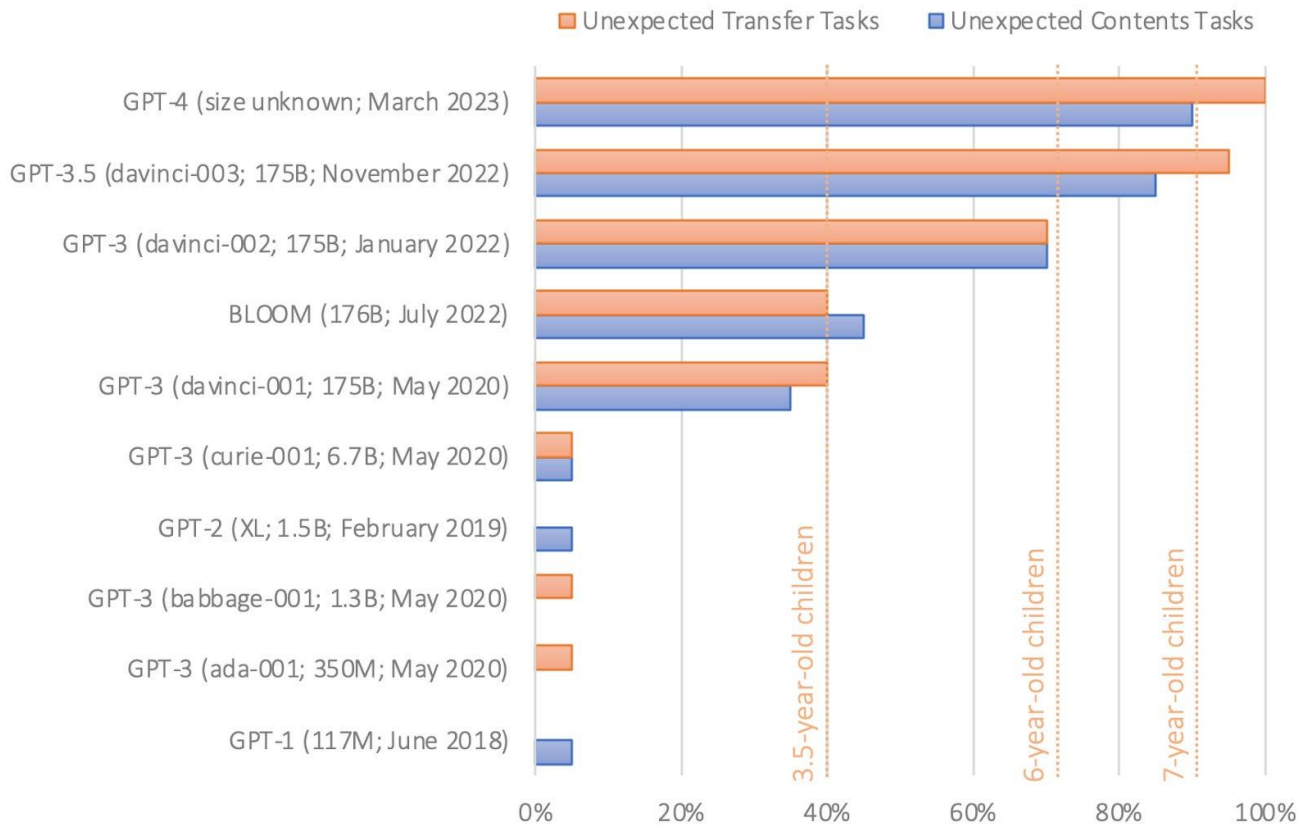
例如，考虑以下故事：

```
这是一个装满爆米花的袋子。袋子里没有巧克力。然而，袋子上的标签上写着“巧克力”，而不是
“爆米花”。山姆找到了袋子。她以前从未见过这个包。她看不到袋子里装的是什么。她读了标签。
```

您可能会猜到，山姆相信袋子里装有巧克力，并且会惊讶地发现里面有爆米花。心理学家将这种推理他人心理状态的能力称为“心理理论”。大多数人从小学起就具备这种能力。专家们对于非人类动物（如黑猩猩）是否具有心理理论存在分歧，但普遍认为心理理论对人类社会认知很重要。

今年早些时候，斯坦福大学心理学家米哈尔·科辛斯基（Michal Kosinski）发表了一项研究，探讨了 LLM 解决心理理论任务的能力。他给了各种语言模型段落，就像我们上面引用的那样，然后要求他们完成一个句子，比如“她相信袋子里装满了”。正确的答案是“巧克力”，但不复杂的语言模型可能会说“爆米花”或其他东西。

GPT-1 和 GPT-2 未通过此测试。但 2020 年发布的第一个版本的 GPT-3 的正确率几乎达到了 40%——Kosinski 的表现水平相当于一个三岁孩子的水平。去年 11 月发布的最新版本的 GPT-3 将这一问题提高到了 90% 左右，与 7 岁儿童的水平相当。 GPT-4 正确回答了大约 95% 的心理理论问题。



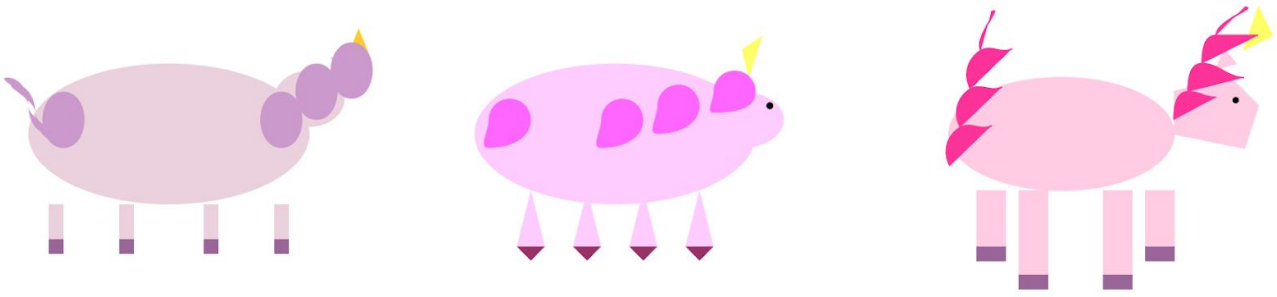
“鉴于既没有迹象表明类 ToM 的能力是被故意设计到这些模型中的，也没有研究表明科学家知道如何实现这一点，类 ToM 的能力很可能自发地、自主地出现，作为模型不断增强的语言能力的副产品，”科辛斯基写道。

值得注意的是，研究人员并不都同意这些结果表明了心理理论的证据：例如，错误信念任务的微小变化导致 GPT-3 的表现更差； GPT-3 在测量心理理论的其他任务中表现出更多的可变性能。正如我们中的一个人（肖恩）所写的，成功的表现可能归因于任务中的混杂——一种“聪明的汉斯”效应，只出现在语言模型中，而不是马身上。

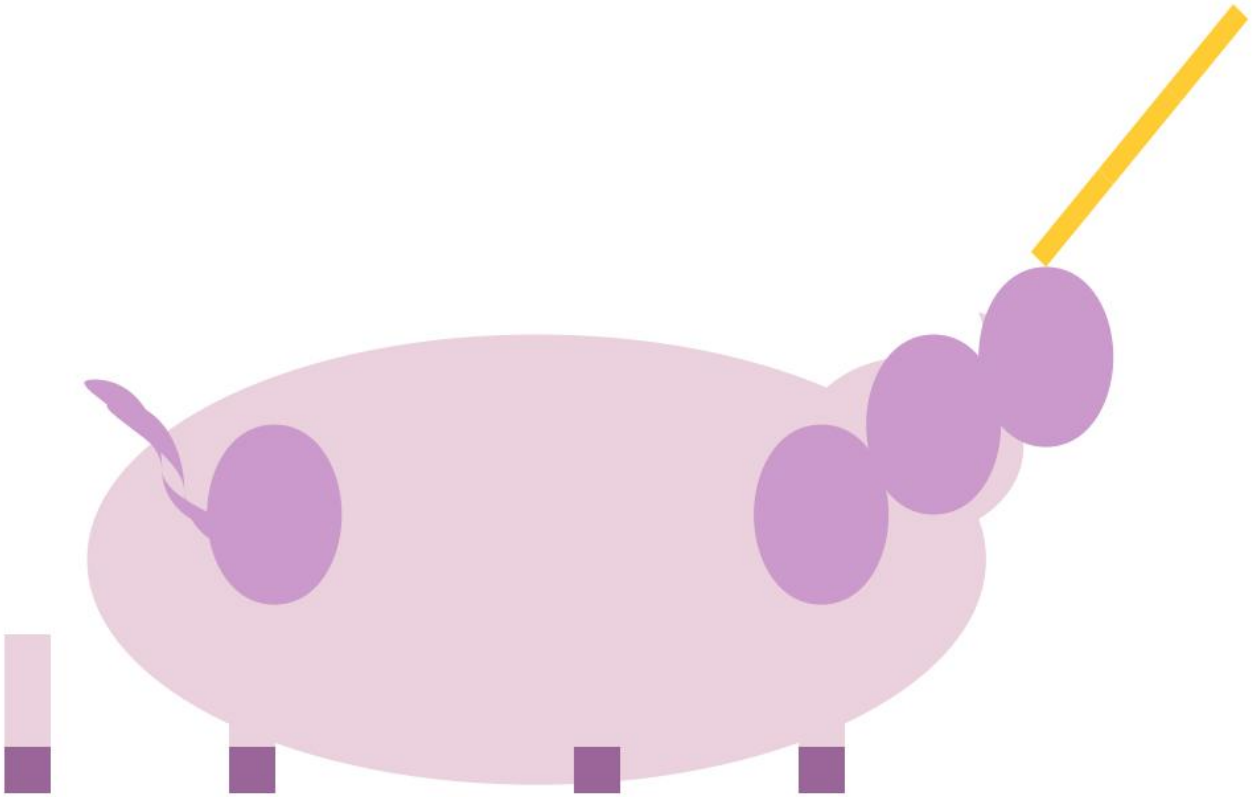
尽管如此，GPT-3 在旨在测量心理理论的多项任务上的接近人类的表现在几年前还是不可想象的，并且与较大的模型通常更适合需要高级推理的任务的想法是一致的。

这只是语言模型似乎自发发展高级推理能力的众多例子之一。今年 4 月，微软的研究人员发表了一篇论文，认为 GPT-4 显示了通用人工智能的早期诱人迹象——以复杂的、类似人类的方式思考的能力。

例如，一位研究人员要求 GPT-4 使用一种名为 TiKZ 的晦涩图形编程语言绘制独角兽。 GPT-4 回复了几行代码,然后研究人员将其输入到 TiKZ 软件中。生成的图像很粗糙,但它们清楚地表明 GPT-4 对独角兽的样子有一定的了解。



研究人员认为 GPT-4 可能以某种方式记住了从训练数据中绘制独角兽的代码，因此他们给了它一个后续挑战：他们修改了独角兽代码以移除角并移动一些其他身体部位。然后他们要求 GPT-4 重新打开喇叭。 GPT-4 的回应是将喇叭放在正确的位置：



即使作者测试的版本的训练数据完全基于文本，GPT-4 也能够做到这一点。也就是说，它的训练集中没有图像。但 GPT-4 显然在接受大量书面文本训练后学会了推理独角兽身体的形状。

目前，我们对 LLM 如何实现这样的壮举还没有任何真正的了解。有些人认为，这样的例子表明模型开始真正理解训练集中单词的含义。其他人坚持认为，语言模型是“随机鹦鹉”，它们只是重复日益复杂的单词序列，而没有真正理解它们。

这场辩论指出了一种可能无法解决的深刻的哲学张力。尽管如此，我们认为关注 GPT-3 等模型的实证表现很重要。如果语言模型能够始终如一地获得特定类型问题的正确答案，并且研究人员确信他们已经控制了混杂因素（例如，确保语言模型在训练期间不会暴露于这些问题），那么无论它是否以与人类完全相同的方式理解语言，这都是一个有趣且重要的结果。

使用下一个标记预测进行训练如此有效的另一个可能原因是语言本身是可预测的。语言中的规律性常常（尽管并非总是）与物理世界中的规律性相关。因此，当语言模型学习单词之间的关系时，它通常也在隐式地学习世界上的关系。

此外，预测可能是生物智能和人工智能的基础。在安迪·克拉克等哲学家看来，人脑可以被认为是一台“预测机器”，其主要工作是对我们的环境进行预测，然后用于成功地驾驭该环境。直观地说，做出好的预测得益于良好的表示——使用准确的地图比使用不准确的地图更有可能成功导航。世界很大而且很复杂，做出预测可以帮助生物体有效地定位和适应这种复杂性。

传统上，构建语言模型的一个主要挑战是找出表示不同单词的最有用的方式，特别是因为许多单词的含义在很大程度上取决于上下文。下一个单词预测方法使研究人员能够通过将其转化为实证问题来回避这个棘手的理论难题。事实证明，如果我们提供足够的数据和计算能力，语言模型最终只需弄清楚如何最好地预测下一个单词，就能学到很多关于人类语言如何工作的知识。缺点是我们最终得到的系统的内部运作方式我们并不完全了解。

Tim Lee 于 2017 年至 2021 年在 Ars 工作。他最近推出了一份新的时事通讯《Understanding AI》。它探讨了人工智能的工作原理以及它如何改变我们的世界。您可以在这里订阅他的时事通讯。

Sean Trott 是加州大学圣地亚哥分校的助理教授，主要研究人类语言理解和大型语言模型。他在时事通讯《反事实》中讨论了这些主题以及其他主题。

1. 从技术上讲，LLM 对称为标记的单词片段进行操作，但我们将忽略此实现细节，以使文章的长度保持在可管理的范围内。

2. 从技术上讲，ChatGPT 的原始版本基于 GPT-3.5，它是 GPT-3 的后续版本，经历了名为“人类反馈强化学习”(RLHF) 的过程。 OpenAI 尚未发布该模型的所有架构细节，因此在本文中我们将重点关注 OpenAI 详细描述的最新版本 GPT-3。
3. 前馈网络也称为多层感知器。自 20 世纪 60 年代以来，计算机科学家一直在尝试这种类型的神经网络。
4. 从技术上讲，神经元计算其输入的加权和后，会将结果传递给激活函数。我们将忽略这个实现细节，但如果您想了解神经元如何工作的完整解释，您可以阅读 Tim 的 2018 年解释器。
5. 如果您想了解有关反向传播的更多信息，请查看 Tim 的 2018 年关于神经网络如何工作的解释。
6. 在实践中，为了计算效率，训练通常是分批进行的。因此，软件可能会在进行向后传递之前对 32,000 个令牌进行前向传递。