什么是Transformer

12110644 周思呈

1 背景

ChatGPT 3.5发布于2022年11月，发布之后很快就由于其优秀的表现引起了各行各业的AI热潮。通过学习大量现成文本和对话集合，ChatGPT能够像人类那样即时对话，流畅的回答各种问题。无论是英文还是其他语言，从回答历史问题，到写故事，甚至是撰写商业计划书和行业分析，几乎无所不能。ChatGPT名字中的GPT，全称Generative Pretrained Transformer，其中的Transformer是一种基于注意力机制的编码器-解码器结构神经网络模型。本文将对Transformer模型进行一些介绍。

2 循环神经网络RNN

在Transformer被提出之前，自然语言处理常用的方法是循环神经网络。神经网络常常被用于分类和回归等传统机器学习任务，但它通常只能有单个输入。在自然语言处理中，输入的文本是序列化的数据，先后输入之间存在关联性。当我们在理解一句话意思时，孤立的理解这句话的每个词是不够的，我们需要处理这些词连接起来的整个序列。比如，“我吃苹果”这句话是最简单的主谓宾结构的文本，由代词、动词和名词三部分组成。如果模型判断出“我”是主语人称代词，那么“我”后面的“吃”就很有可能是谓语动词，并且“吃”后面的“苹果”很有可能是宾语名词。循环神经网络能够处理这些序列化的信息，因为其隐藏层中的输入不仅来自于前一层，还包括了上一个时刻这个隐藏层自己的输出。比如说，时刻t_1 输入了“我”，在时刻t_2 输入”吃“的时候，隐藏层的输入其实是之前对”我“的分析结果和”吃“这个文本。循环神经网络的输出数量不是固定的，对应于多种自然语言处理的任务。比如，多输入单输出可以应用于文本情感分析，单输入多输出可以应用于图片生成文本。最常用的模型是多输入多输出，又叫Encoder-Decoder模型或者Seq2Seq模型，常用于机器翻译。普通的循环神经网络只能记住前面的输入所产生的影响，所以又出现了双向循环神经网络，能够将后文的信息也作为隐藏层的输入。

但是循环神经网络在训练过程中效率较低。由于每个时刻的隐藏层都需要前一个时刻的信息作为输入，所以这个模型在训练过程中难以并行。为了解决这个问题，Google在2017年提出了基于注意力机制的Transformer模型。它不仅解决了训练时的并行问题，还获得了更完整的上下文信息，因此在许多测试中都取得了非常好的效果，成为大部分自然语言处理模型的一个基本模型。

3 注意力机制

Transformer的基本结构仍然是Seq2Seq，相比于之前的工作，其最重要的改进是引入了注意力机制。在注意力机制中，每个分词都对应了三个向量，分别是query q，key k 和information v。首先拿每个q 对每个k 进行attention操作，每个q^i 和k^j 都得到一个对应的$\alpha_{i,j}$ ，然后用softmax函数将其归一化得到$\hat{\alpha_{i,j}}$ 。接着将$\hat{\alpha_{i,j}}$ 与每一个v 相乘得到b 。这就是输出的序列。我们可以观察到，在产生b^1 的时候，实际上已经用到了输入序列a 中的每一项，也就是说输出序列的每个元素都包含了输入序列的所有信息。并且，由于模型并不依赖于前一时刻的训练结果，它的训练是可以并行的。

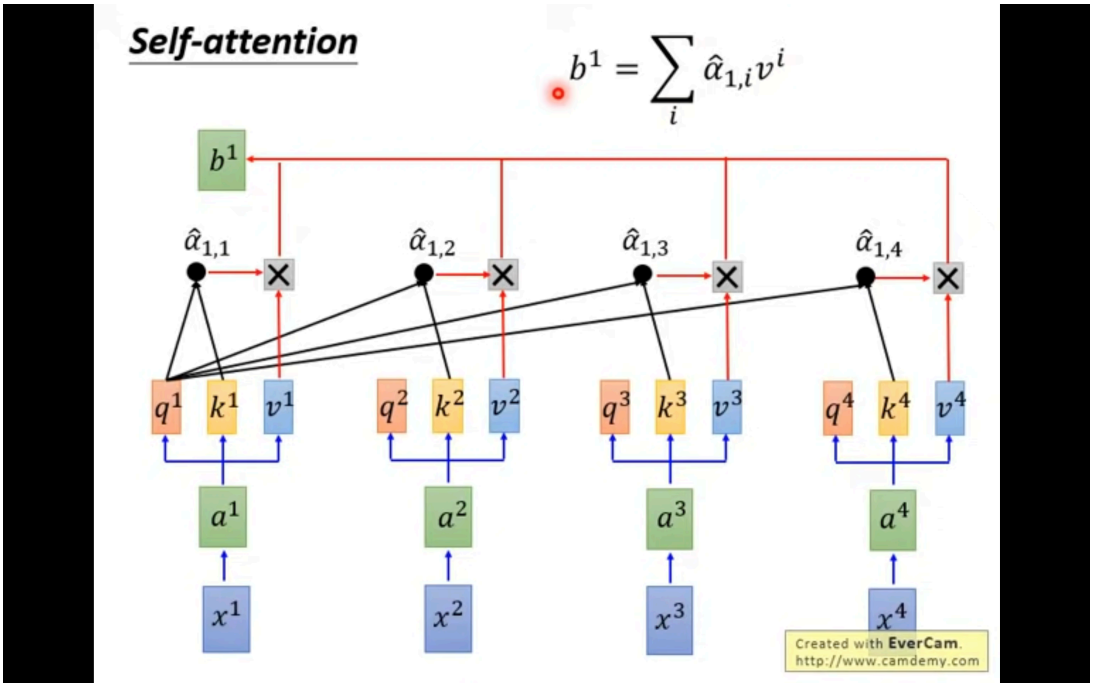

Fig1. 注意力机制

但这样的self-attention层存在一个问题。虽然输入序列的所有信息都被考虑到了，但输入序列的顺序对这个模型来说是不重要的。无论是“我吃苹果”还是“苹果吃我”，在这样的self-attention层中都可以获得一样的输出。所以我们还需要引入位置编码来保留输入序列的顺序信息。这个位置编码信息和x 一起被输入到模型中，可以是加上一个预先设定好的矩阵e ，也可以是在x 后面拼接一个位置向量p 。位置向量p 通过独热码进行编码，只有对应位置的元素是1，其他都是0。这两种方法都可以让模型学习到输入序列的位置信息。

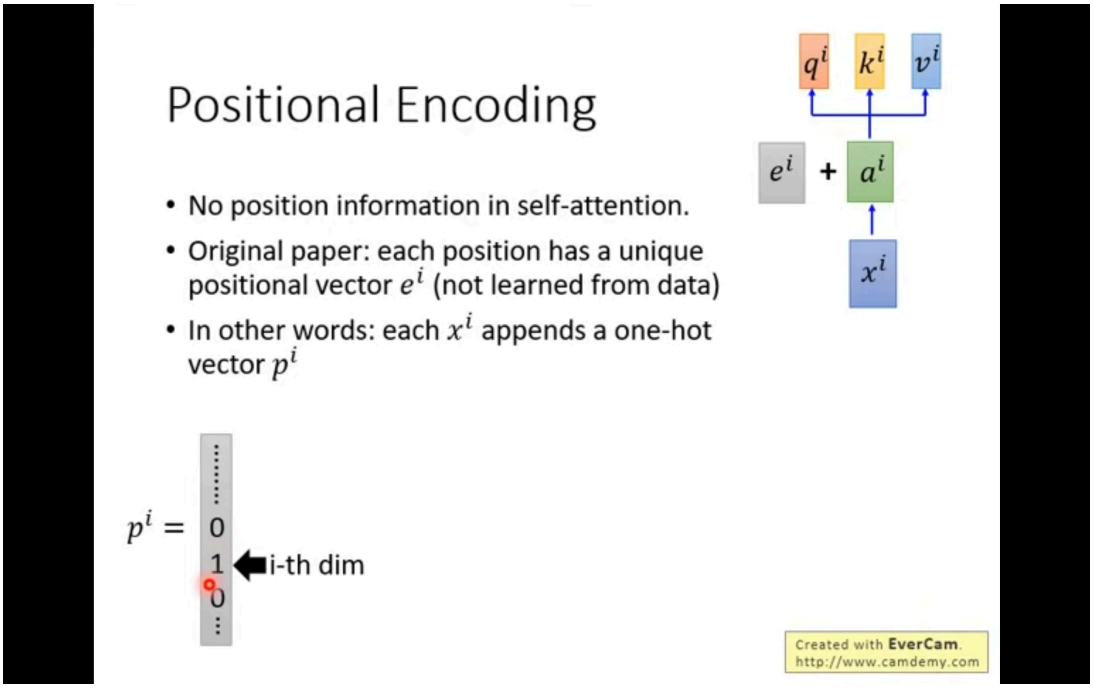

Fig2. 位置编码

Google在提出注意力机制之后进行了一系列实验，下面这张图展示了将模型应用于机器翻译方面之后对于注意力权重的可视化。由于Attention层在训练时每个q 都会对每个k 进行attention操作，所以每两个词之间都会有一个注意力权重。下面这张图中，连线颜色深表示权重较大，颜色浅表示权重较小。左边的输入文本是"The animal didn't cross the street because it was too tired"，我们通常会认为这里的"it"指的是"animal"。可以看到，模型的输出中，"it"对于"animal"的注意力权重较大。而右边的输入文本是"The animal didn't cross the street because it

was too wide"，这个语境中的"it"指的是"street"，对应的模型注意力权重中，"it"对于"street"的注意力权重较大。这说明注意力机制取得了良好的训练效果，且这个模型是具有一定可解释性的。

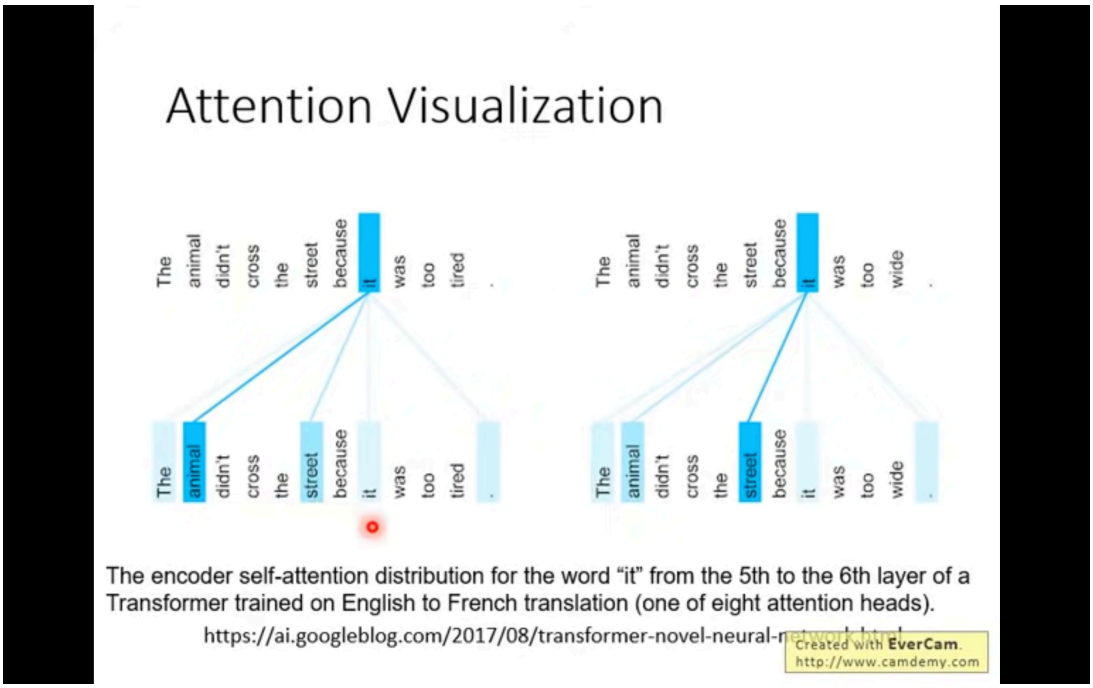

Fig3. Attention可视化

4 整体架构

了解了注意力机制之后，再介绍一下Transformer的整体架构。总的来说，它由多个Encoder和多个Decoder堆叠组成，分别对应下图中的左侧和右侧。Encoder中，文本序列经过位置编码之后输入注意力模型，然后对训练结果的每一层按照正态分布的方式进行归一化处理。这个Encoder模块会重复N次，输出的结果作为右侧模型输入的一部分传给Decoder的注意力模型。Decoder的另外一个输入是序列化数据的上一个输出，同样经过了位置信息的编码，并且经过了Mask处理，只保留当前已经产生出的序列信息。这个模块同样会重复N次。最后，Decoder的输出进行归一化处理，通过softmax函数输出最后的概率。

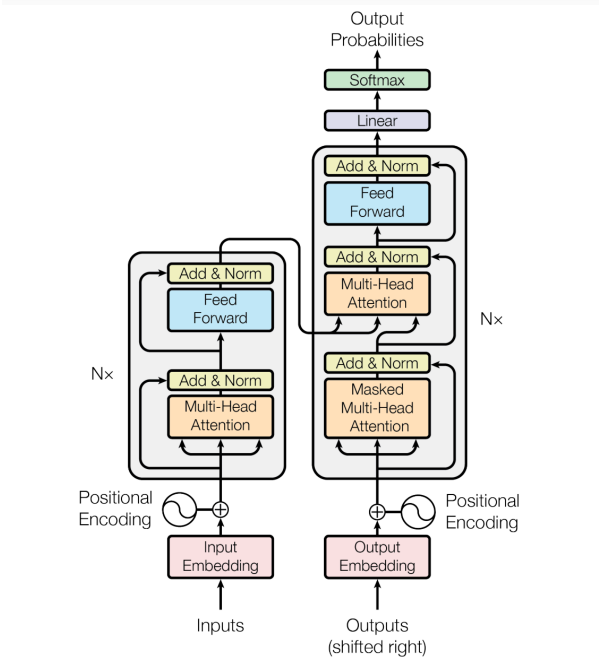

Fig4. Transformer整体架构

参考资料：

[1] Attention Is All You Need

[2] 李宏毅，Transformer

[3] 陈巍：ChatGPT发展历程、原理、技术架构详解和产业未来

[4] Transformer模型详解