一文读懂BERT

原创 王猛 空天信息 5月7日

2018年前，自然语言处理（Natural Language Processing, NLP）可谓是百花齐放，传统模型在各自擅长领域中发挥着大大小小的作用。自2018年以来，NLP发生里程碑式转折，一种名为BERT的模型在11个NLP经典问题上超越现有模型，尤其在被认为是“NLP皇冠上的明珠”的阅读理解顶级水平测试SQuAD1.1上发挥出超越人类的表现，BERT也当之无愧地获得2019年《NAACL》最佳长论文奖。时至今日，BERT原始论文已被引用近13000次，其各种衍生模型也长期霸占着各大NLP竞赛榜单。

BERT是什么？

BERT全称为Bidirectional Encoder Representation from Transformers[1]，是一种用于语言表征的预训练模型。它强调不再像以往传统单向模型预训练的方式，而是采用新的训练策略，以致能生成深度双向语言表征模型。

图1. BERT也是动画《芝麻街》中的一个人物

为什么选择BERT?

在NLP算法落地过程中遇到最大挑战通常是缺乏足够训练数据。总体而言，获取大量文本并非难事，但帮助算法清楚问题定义需要有足够数量被标记样本。对文本进行人工标注是一件耗时耗力的工作，并且因为每个人对文本内容理解层次不同，产生的标注带有主观倾向也会造成偏差。

为了弥合数据鸿沟，多种仅使用未标记文本语料的语言模型被相继提出，这种模型被称为预训练模型（Pre-trained Models, PTM），BERT就是这种技术开花结的最好的果之一。预训练模型是通过自监督学习从大规模数据中得到、与具体业务无关的模型。如哈工大开源的中文预训练模型BERT-wwm就是在百科、新闻、问答等数据上训练得到，这些数据含有的词汇量达到5.4亿之多。

BERT如何使用？

在NLP中，预训练模型的训练被称为是上游任务；而具体业务，如：情感分析、阅读理解、文本摘要等则被称作是下游任务。通过对预训练模型在相对少量业务数据上进行训练，便可将模型用于不同目的的下游任务，这个过程称为微调（Fine-Tune）。微调可显著提高模型准确性。此外，与重新开始对业务数据进行训练相比，微调仅需要很少样本就可以让模型拥有良好性能。

为了有更直观的理解，图2展示了机器学习中一个经典监督学习问题——垃圾邮件分类的微调过程。

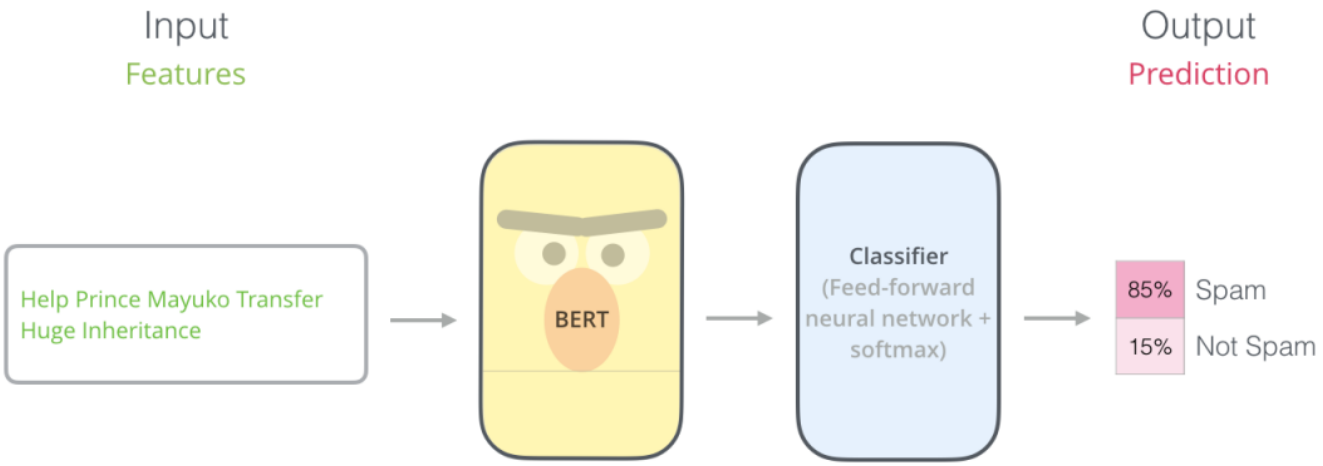

图2. BERT垃圾邮件分类流程

BERT的原理是什么？

上游进行语言模型的预训练，下游微调并应用到具体业务中，这种模式被称为迁移学习（Transfer Learning）。在架构方面，BERT使用大量迁移模型Transformer中的编码器，并对输入文本进行位置编码，结合BERT独特的训练策略来得到预训练模型。

01 迁移学习

Transformer是谷歌于2017年年底在论文《Attention is all you need》[2]中提出的一种序列到序列（Seq2seq）模型，它主要由编码器（Encoder）和解码器（Decoder）两部分组成，详见图3。

介绍Transformer模型最好例子就是它在机器翻译中的应用。在英文翻译成中文过程中，编码器负责阅读与学习输入的英文文本，通过捕捉载体所包含信息学会一定的语言概念，这被称为是上下文（Context）。上下文信息本质是语义在向量空间的一种映射，即语义的数学化表达，一个著名例子就是：国王-男性+女性=女王。之后，编码器将所学内容以隐藏层（Hidden Layer）形式传递给解码器，解码器再利用这些知识进行文本翻译工作。具体来说，在生成中文文本过程中，解码器会对当前中文单词根据上下文信息来预测下一个中文单词，之后再根据下一个词预测下一个词的下一个词，循环往复，直至生成完整句子，这种做法也体现了序列模型的特性。

由于BERT目标是生成用于语言表示的预训练模型，因此只需要编码器部分即可。所学到上下文信息在BERT中以768维度向量存储。

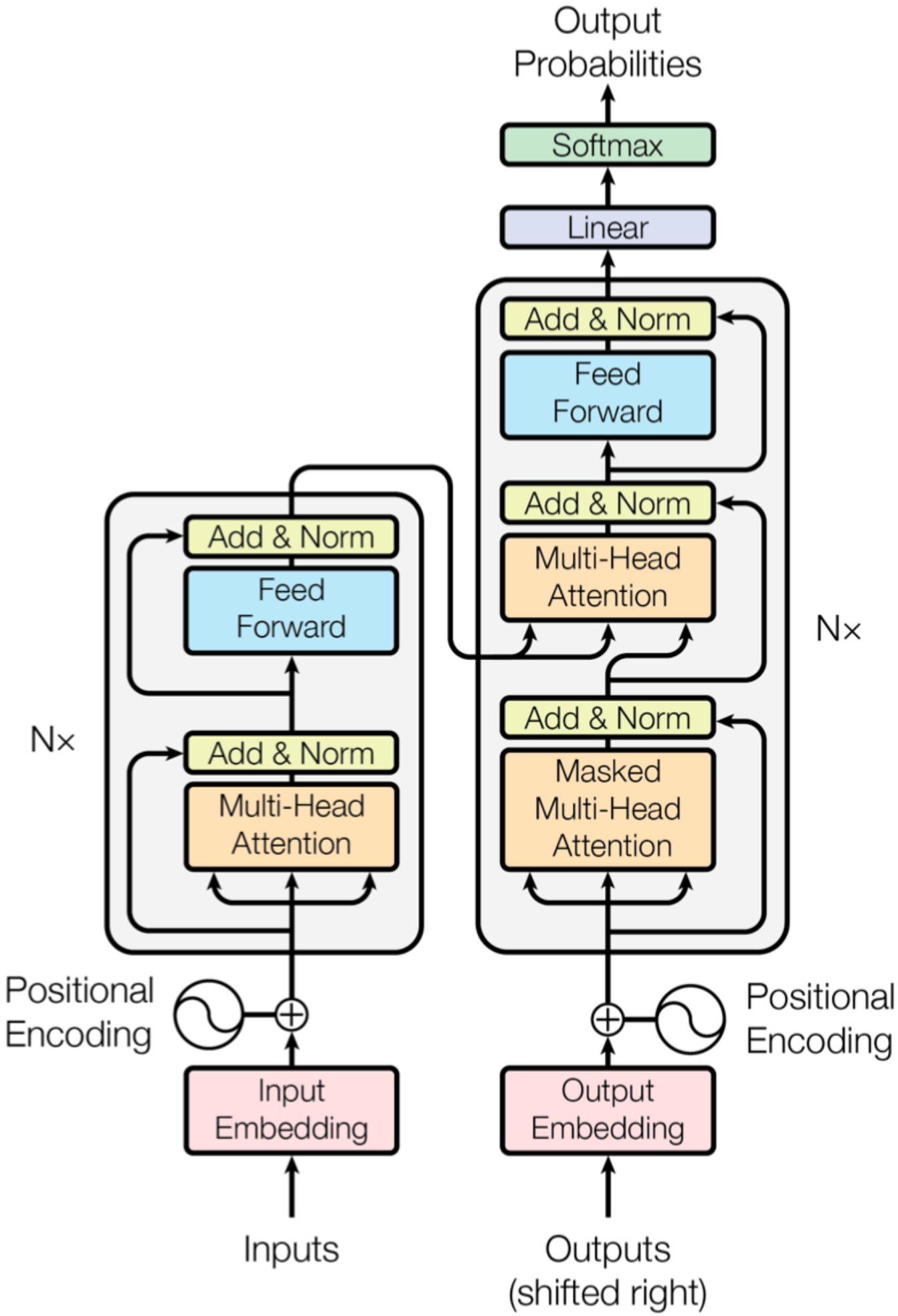

图3. Transformer模型结构

02 位置嵌入

在序列模型中，词在句中位置对语料学习尤为重要。在训练文本时，长短时记忆网络（Long Short Term Memory Network, LSTM）通过一定的顺序读取文本：从左向右或是从右向左，以此来得到词在文中的位置信息，这样的模型称为单向模型。

与单向模型不同的是，在词向量（Word Embedding）进入编码器之前，Transformer模型使用位置编码（Positional Encoding）来提供语序信息。这种设计通过为词向量加入独一无二的纹理信息来表征词在句中的位置，纹理信息则通过sin函数与cos函数的线性变换（公式1-2）生成。因此，编码器可以一次读取整个文本序列，这样的语境化特性使Transformer可以基于词的所有周围环境来学习上下文，并且可以接收更庞大的数据量。

$$PE_{(pos, 2i)}= sin(pos/10000^{2i/d_{model}})\tag{1}$$
$$PE_{(pos, 2i+1)}= cos(pos/10000^{2i/d_{model}})\tag{2}$$

03 训练策略

BERT在模型的训练过程中，会同时结合以下两种策略：

– 遮蔽预测（Masked LM）：BERT会随机遮蔽掉句中部分词，然后通过未遮蔽掉的词提供上下文来预测这些被遮蔽的词是什么。这种训练模式在以往单向模型中很难实现，这意味着与单向模型相比，BERT对上下文有着更深刻的感知。

– 下一句预测（Next Sentence Prediction）：为了让模型能够预测两个给定句子在顺序上是否有逻辑关系，在BERT的训练过程中，模型接收成对的句子作为输入，并预测第二句是否是第一句的后续。通过这样的训练，模型不仅能学习句内信息，还能清楚地捕捉到句间逻辑。这种独特的学习模式也使其在问答系统、阅读理解等问题上有出色的发挥。

毫无疑问，BERT在使用机器学习进行自然语言处理方面取得巨大技术突破。谷歌团队也在GitHub上开源了BERT源代码与预训练模型，涵盖103种语言，可以轻松地使用开源预训练模型进行下游任务训练也让它有更广泛的应用。本文试图描述BERT主要思想同时又不想夹杂过多复杂难懂的概念，若想深入了解BERT，建议您阅读引用中的文章。

引用

[1] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30: 5998-6008.

以上内容由苏州研究院王猛提供。