BERT 简介及应用

BERT（Bidirectional Encoder Representation from Transformers)是 2018 年底由谷歌发布的语言表征模型，刷新了 11 种各类型 NLP 任务的性能记录，其最大亮点在于模型本身的效果优异以及普适性强。

BERT 模型的特点是利用海量的无标注语料进行自监督学习训练，获得单词的通用特征表示，最终提供给具体 NLP 任务作为词嵌入特征。BERT 模型提供的是一个供其它任务迁移学习的模型，根据任务微调或者固定之后可以作为特征提取器。

NLP 的 4 类主要任务（序列标注，分类，句子关系判断，生成式任务）均可以通过改造 BERT 模型的输入输出部分，使用 BERT 预训练产生的基础模型参数，对效果进行不同程度的提升。

1. BERT 简介

BERT 是通用语言表征模型，通常采用大规模、与特定 NLP 任务无关的文本语料进行训练，其目标是学习语言本身的样子。

原论文的题目是《BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding》，因此很容易得到 BERT 模型的 4 个关键词。

1）Pre-training，预训练。海量不需要经过人工标注的训练语料，经过预训练后得到通用模型，再根据具体应用精加工使之匹配场景。

2）Transformers，引入自注意力机制的模型。不仅可以解决长依赖问题，而且能够提高计算效率，更好地捕获文本的表示。

3）Bidirectional，双向预测。以往只能通过从前向后或者从后向前的进行单向预测，而通过上下文的双向预测拼接，能够更完整的理解整个语句。

4）Deep 深层结构。通过上下文全向预测，启用多个聚焦点，不是两个单向的拼接，而是序列并行处理。

BERT 逐渐调整模型参数，使得模型输出的文本语义表示能够刻画语言的本质，便于后续针对具体 NLP 任务作微调。大致可分为以下两个步骤。

1. 第一步，通过全向预测被遮盖住的词汇，来初步训练 Transformer 模型的参数。具体设置是把每篇文章中 15%的词汇遮盖，让模型根据上下文全向地预测被遮盖的词。例如有 1 万篇文章，每篇文章平均有 100 个词汇，随机遮盖 15%的词汇，则模型的任务是正确地预测这 15 万个被遮盖的词汇。

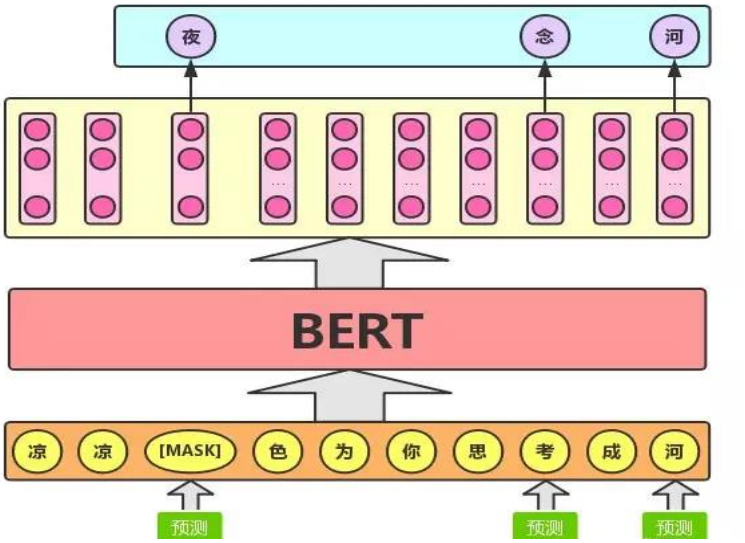

图 1　BERT 训练步骤 1：Masked LM

2. 第二步，通过语句的抽取和重排序，继续训练模型的参数。例如上述的 1 万篇文章中，从中挑选 20 万对语句，总共 40 万条语句。其中 10 万对语句是连续的两条上下文语句，另外 10 万对语句不是连续的语句。然后让 Transformer 模型来识别这 20 万对语句，哪些是连续的，哪些不连续。

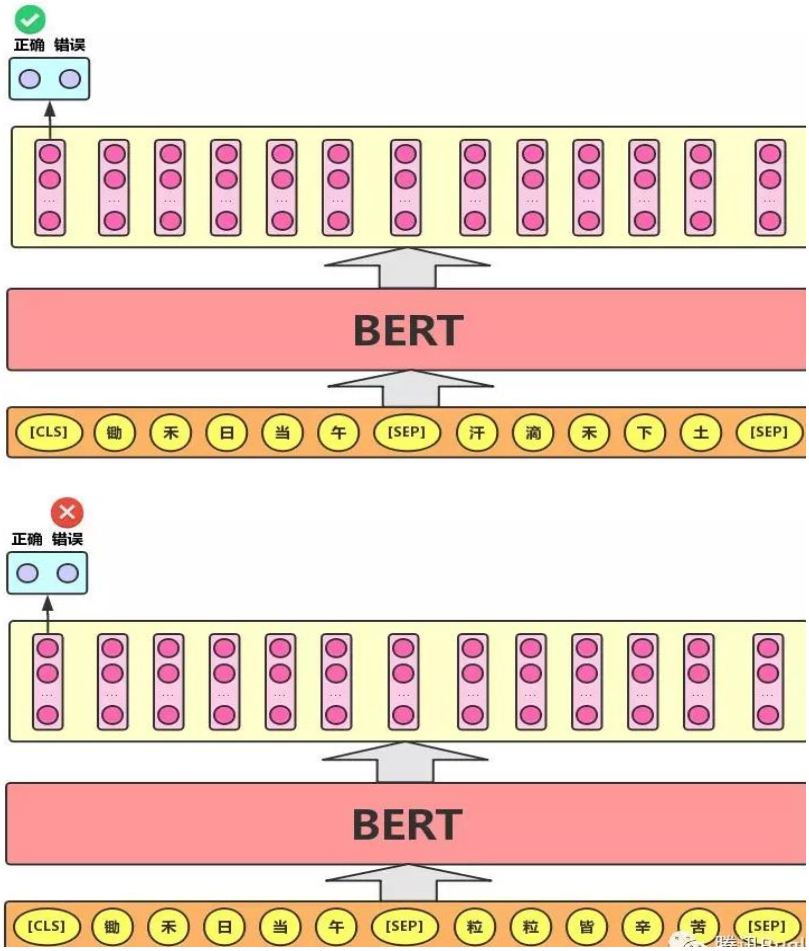

图 2　BERT 训练步骤 2：NextSentence Prediction

这两步类似英文学习中的完形填空和段落重排序。通过这两步训练后，得到的 Transformer 模型及其参数，就是通用的语言表征模型 BERT。

2. BERT 应用

从应用来看，BERT 包括三个部分，一是使用庞大语料训练的字符向量集，二是应用注意力思想的多层特征提取层（或者叫注意力层），三是工程化的模型加载环境和调用接口。

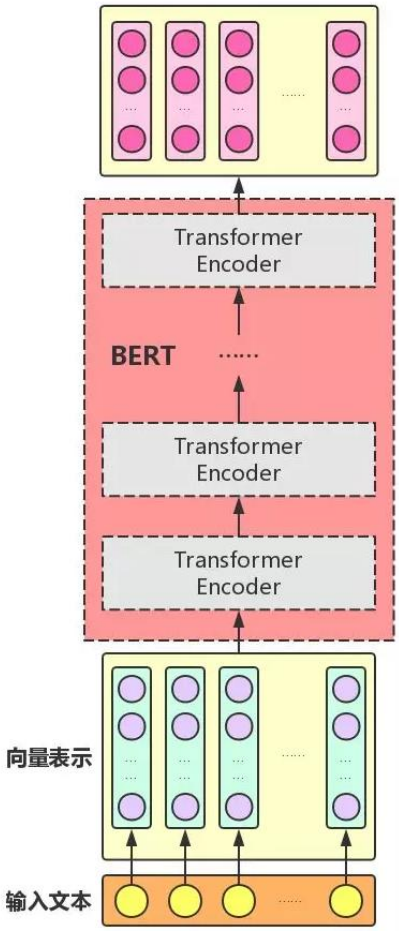

图 3　BERT 基本架构

BERT 应用的优点如下：

1）将语言建模的成本降到最低。谷歌使用亿级语料训练的字符向量，可以作为工作的起点。

2）特征工程的任务量降到最低。使用基础的 BERT 进行分类训练和预测，不做任何特征工程的调试，效果与使用 1500 行代码调试两个月的 TextCNN 分类模型效果一致。

3）加速实现从想法到产品的工程化。面向任务的修改，集中在用户输入层，代码工作量变得较少。

目前出现了大量使用 BERT 来在 NLP 各个领域进行直接应用的工作，方法都很简单直接，效果总体而言比较好。当然也需要分具体的领域，不同领域受益于 BERT 的程度不太相同。

应用领域 1：问答系统与信息检索

BERT 应用在问答系统领域，目前最成熟也最成功，阅读理解的准确率通常有 30%以上提升。

首先，将长文档切割成段落或者句子片段，然后利用搜索里的倒排索引建立快速查询机制；接着，将候选片段和用户问句作为 BERT 的输入，查询答案会存在候选段落或者句子；最终，BERT 通过分类指出当前片段是否包括问句的正确答案，或者输出答案的起始终止位置。

应用领域 2：对话系统 / 聊天机器人

BERT 应用在聊天机器人领域，潜力比较大。无论是单轮会话还是多轮会话，BERT 已经能做到对效果的显著提升。

聊天机器人主要面临以下两种技术挑战：1）对于单轮对话来说，主要挑战是用户意图分类；2）对于多轮对话来说，主要挑战是历史信息的融合和正确使用。

在多轮对话中，BERT 相对基准方法已经有 10%以上提升。针对复杂的应用场景，未来挖掘潜力巨大。

应用领域 3：文本摘要

BERT 应用在文本摘要领域，目前的提升有限。

文本摘要分为生成式摘要和抽取式摘要，区别在于输出的内容是否局限于原文的句子。生成式任务受 BERT 结构所限，因此表现并不够好；抽取式摘要本质是文本分类任务，BERT 能较好完成。

应用领域 4：文本分类

BERT 应用在文本分类领域，应该说效果能够达到以及超过之前的各种方法，但提升幅度不算太大。

文本分类是把长句子或文档分到一个类别里，特征偏语言的浅层，而且指示性的单词也比较多。任务难度偏简单，BERT 的潜力感觉不太容易发挥出来，基本提升幅度在 3%到 6%之间。

应用领域 5：序列标注

BERT 应用在文本分类领域，能够把任务效果都能做到当前最好。

序列标注包括分词，词性标注，语义角色标注等等，并非具体应用领域，而是 NLP 中一种问题解决模式。其特点是对于句子中任意一个单词，都会有一个对应的分类输出结果。BERT 能提升效果，在具体的应用领域中提升程度不同。

3. 总结

尽管有不同的 NLP 任务，其利用 BERT 的过程是基本一样的。核心过程都是用 Transformer 作为特征抽取器，用 BERT 预训练模型初始化 Transformer 的参数，然后再用当前任务 Fine-tuning，就可以得到不错的结果。

BERT 能够引入绝大多数领域，但对于不同应用效果促进作用是不同的，其擅长的场景如下：

1. BERT 适合解决仅仅依赖语言本身答案的任务。例如问答系统和阅读理解，包含的特征仅仅是语言，因此理解能力越强，解决得就越好。相反，例如搜索和推荐中的用户行为和内容质量等任务，这类任务在文本外的判断因素也很关键，BERT 就无法体现语言理解的优势。

2. BERT 适合解决句子或者段落的匹配类任务。BERT 在预训练阶段学习了句间关系的知识，并且自注意力机制包含句子之间细粒度的匹配，其本质决定 BERT 在解决句子关系判断任务时表现较好。

3. BERT 适合解决需要深层语义特征的任务。Transformer 层深比较深，因此可以逐层捕获不同层级不同深度的特征。相反，分词、词性标注和文本分类等任务，只需较短的上下文，以及浅层的非语义的特征，就可以较好地解决问题，BERT 能够发挥作用的余地不太大。

4. BERT 适合解决输入长度不太长的 NLP 任务。Transformer 的自注意力机制需要对任意两个单词做计算，因此约束 BERT 的输入长度不能太长，否则训练和推理速度会比较久。BERT 更适合解决句子级别或者段落级别的 NLP 任务。