[预训练语言模型专题] BART & MASS 自然语言生成任务上的进步

原创 李大姐 朴素人工智能

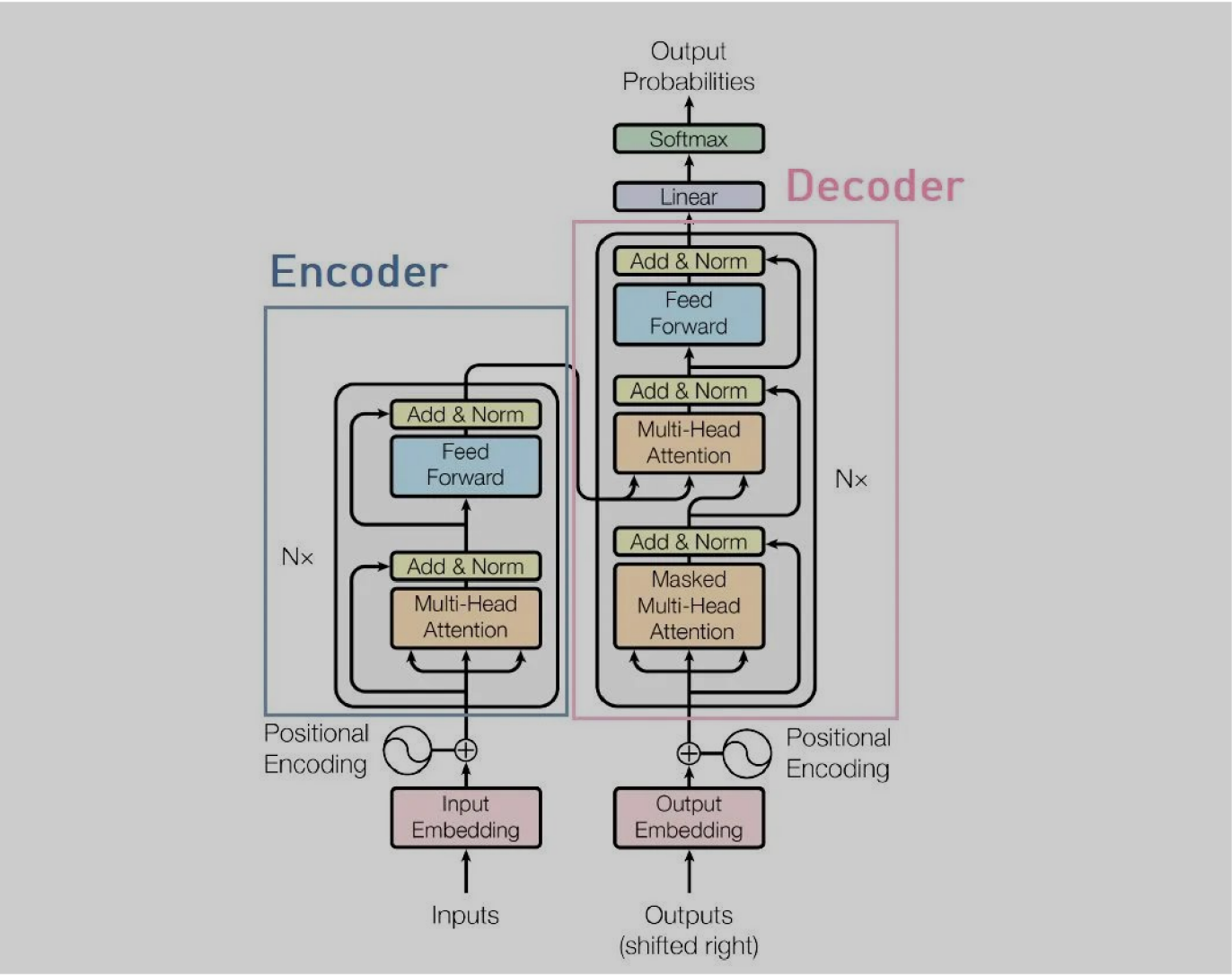
来自专辑

预训练语言模型

点击蓝字关注我们↑↑↑↑

BART和MASS都是2019年发布的，面向生成任务，基于Transformer神经翻译结构的序列到序列模型。分别由Facebook 和微软亚洲研究院提出。他们都对encoder输入的屏蔽(mask)方式进行了改进，并且在生成任务的效果也都比之前有了不少提升。让我们花**10分钟**来一起来看看这两个模型吧。

两个模型都是以Transformer的神经翻译模型作为基础结构，而Transformer的encoder-decoder结构（图 1）的具体讲解可以参考上一篇文章。



图（1） Transformer的encoder-decoder结构

MASS

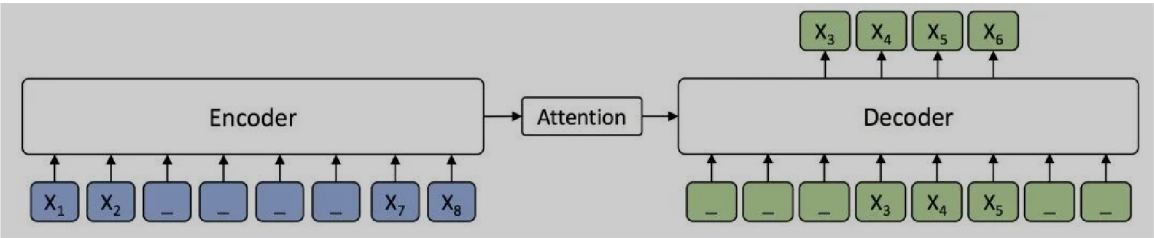
Masked Sequence to Sequence Pre-training for Language Generation

任务类型

面向自然语言生成任务（神经机器翻译、文本摘要和会话响应生成等）的预训练模型。

模型结构

MASS 是一个基于Transformer的序列到序列模型，由attention将encoder和decoder连接在一起。相比基础的Transformer结构，MASS的不同为：***它的encoder的输入是被随机屏蔽了一个长度为k的连续片段的句子（图 2 左侧）。decoder来预测这个被屏蔽的片段，其中decoder的输入会屏蔽在encoder中没有被屏蔽掉的token（图 2 右侧）。***



图（2） MASS的encoder-decoder结构，其中“-”表示被屏蔽掉的token

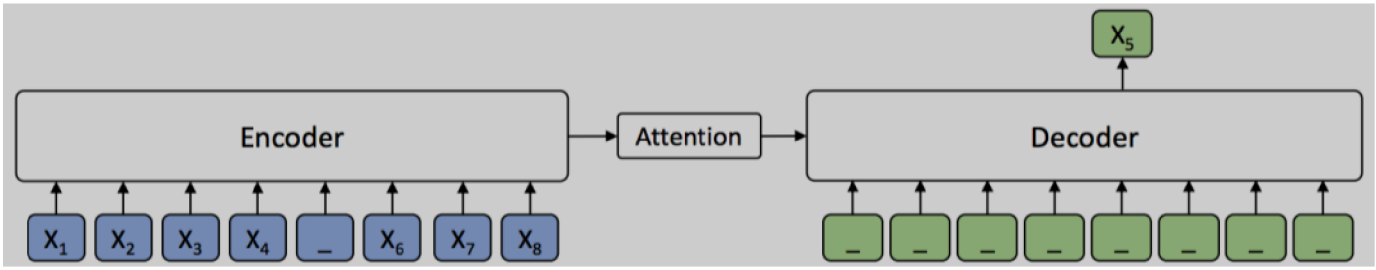
举例说明： 图 2 中，encoder端的输入中，第3-6个token被屏蔽掉，而decoder只预测这3-6这几个连续的token，而屏蔽掉其它token。这里会引入一个超参数 **k**（被屏蔽的连续文段的长度占句子总长度的百分比），通过参数k可以对模型进行灵活的调整。

这种设计的优势有：

1. decoder端的**输入是源输入中被屏蔽的文段**，可以使decoder在预测的时候更加专注利用源输入，而不是目标端要预测的token的前一个token（有点绕口，参考图 2 理解）。
2. **预测encoder端被屏蔽的文段**，迫使encoder要更好的理解encoder输入中没有被屏蔽的文字。
3. 目标端**预测连续文段**，decoder可以建立比预测离散token更好的语言建模能力。
4. **超参数k**，使模型更加的***灵活***。

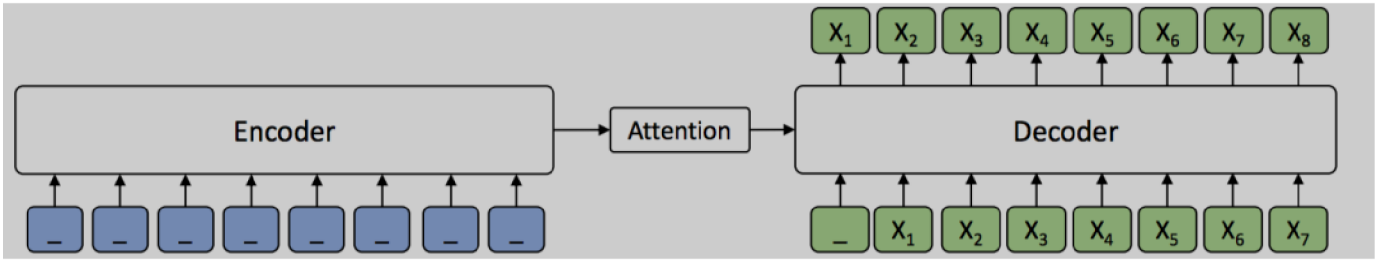
超参数 k

这里单独把超**参数k**拎出来讲，是因为参数k的设计使得MASS模型可以被看作一个统一的预训练框架，BART（k =1）和GPT（k=m）都可以被包含在这种框架里面。这里 **k 是指被屏蔽的连续文段的长度占句子总长度的百分比，除此之外，再加上 k = 1 和 k = m (m 为句子长度)**。



图（3） k=1 --> BERT

这时decoder输入为空，可视为一个非线性分类器，类似于BERT中使用的softmax矩阵，MASS类似于由Transformer encoder 块累加起来的BERT。



图（4） k=m --> GPT

这时encoder输入为空，没有信息输入到encoder中，MASS类似于由Transformer decoder 块累加起来的GPT。

在翻译任务，摘要任务以及对话相应生成任务上的实验结果表明，**k = 50%**时，模型效果最好。因为此时encoder和decoder之间达到一个很好的**平衡**，如果encoder或者decoder端的输入token太少，会造成模型偏向某一边，不利于encoder-decoder框架提取encoder中的句子表示以及在decoder中建模和生成句子的语言生成任务。

BART

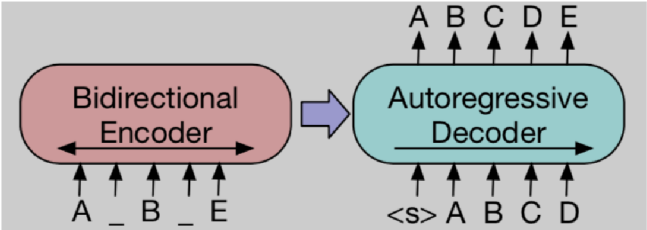
Denoising Sequence-to-Sequence Pre-training for
Natural Language Generation, Translation, and Comprehension

任务类型

BART是一个面向自然语言**生成、翻译**和**理解**任务的 序列到序列 预训练 降噪自编码器（降噪自编码器的原理在公众号介绍UniLM 2.0有具体讲解）。它的训练分为（1）**用任意的噪声函数（选择被屏蔽token的方法）来破坏输入文本。**（2）**训练模型重现未被破坏的文本。**

模型结构

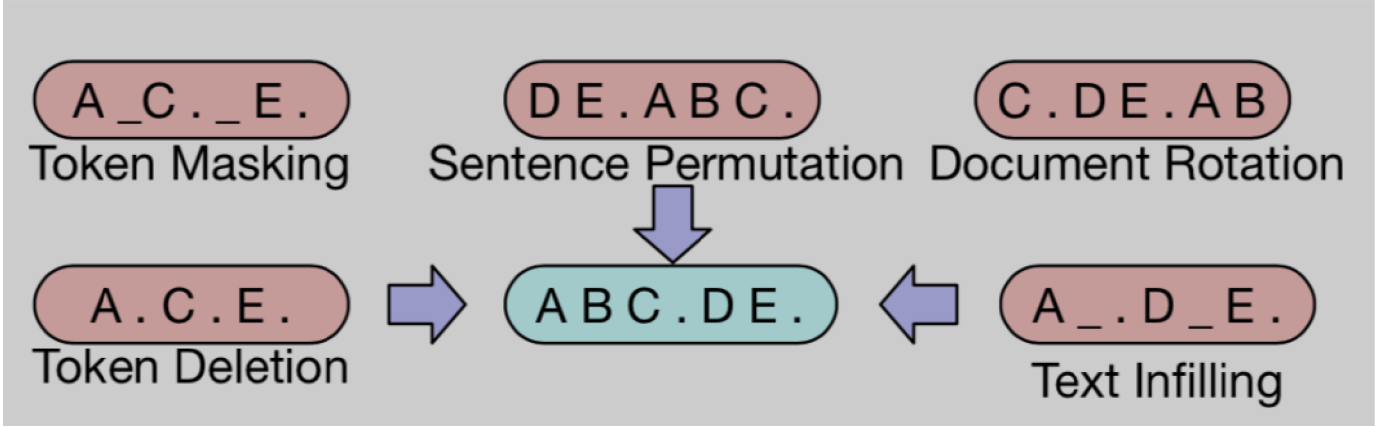
和MASS一样，BART也是基于标准Transformer神经翻译模型的网络结构（做了一点改动，参考GPT，将激活函数ReLU换成 GeLUs），同样也是在选择被屏蔽token的方法（噪声函数）上进行了改变（图 5）。不同于MASS的是，BART对decoder没有进行改变。



图（5）BART模型结构，“-”表示被屏蔽的token

噪声函数：

1. **Token Masking** 和BERT一样，随机选择**token**用[MASK] 代替。
2. **Token Deletion** 随机删除**token**，模型必须确定哪些**位置**缺少输入。
3. **Text Filling** 屏蔽一个**文段**，文段长度服从泊松分布（λ=3）。每个文段被**一个[MASK]**标记替换。如果文段长度为0，意味插入一个[MASK]标记（灵感来自Span-BERT）。
4. **Sentence Permutation** 以句号作为分割符，将一篇文章分成多个**句子**，并随机打乱。
5. **Document Rotation** 随机均匀地选择一个**token**，以这个token为中心，旋转文档，选中的这个token作为新的开头，此任务训练模型以识别文档的开头。



图（6）添加噪声的方法，这些方法可以组合

特点：

不同于一些只能针对特定的噪声的降噪自编码器，BART可以使用**任意**的方法去破坏文档，最极端的情况下，源文本信息全部丧失，BART这时就相当于一个语言模型。

微调

BART生成的表达可以用通过多种方式，用于下游应用。

1. **序列分类任务**
   encoder和decoder的输入输出相同，decoder最后token的最终隐藏状态输入到新的多类线性分类器中。这种方法与BERT中的CLS token相似。但是，BART是将额外token添加到末尾，这个token的表达包含了输入decoder 全部状态信息。
2. **token分类任务**
   对于token分类任务，例如判断是否是SQuAD的答案的终点，将完整的文档输入到encoder和decoder中，并使用decoder的顶部隐藏状态作为每个单词的表示。用这个表示对token进行分类。
3. **序列生成任务**
   由于BART是一个自回归解码器，它直接微调，就可以适应序列生成任务，如抽象问题回答和摘要。在这两种任务中，信息来自于输入，这与去噪预训练目标密切相关。这里，输入送入encoder，decoder以自回归方式生成输出。
4. **机器翻译**
   我们将整个BART（包括编码器和解码器）作为一个单独的解码器，增加一个新的encoder块（双向学习）。更准确的说，我们用随机初始化编码器替换 BART的编码器embedding 层。
   该模型以端到端的方式接受训练，即训练一个新的编码器将外来词映射到输入（BART 可将其去噪为英文）。新的编码器可以使用不同于原始 BART 模型的词汇。新编码器的训练分两步，均需要将来自 BART 模型输出的交叉熵损失进行反向传播。（1）冻结 BART 的大部分参数，仅更新随机初始化的源编码器、BART 位置嵌入和 BART 编码器第一层的自注意力输入投影矩阵。（2）所有模型参数进行少量迭代训练。

结果

在 **SQuAD**（抽取式问答的任务）**MNLI**（推理任务）**ELI5**（抽象问题回答生成任务）**XSum**（摘要生成任务）**ConvAI2**（对话反应生成任务）**CNN/DM**（摘要生成任务）等数据集上进行测试， 不同的噪声函数结果差距比较大，总结可以得到以下的结论：

1. 预训练中**Token masking**是非常重要的，没有token mask 的document rotation 和sentence shuffling 这两种方法的结果比较差。

2. 从左到右的预训练有助于提高生成模型的效果。
3. 对于SQuAD这种抽取式问答的，双向encoder 要更加重要。
4. 预训练模型的性能在不同任务中有显著的差异。

体会

BART和MASS对生成任务的效果都有提升。MASS专注于生成任务，BART在保证理解任务性能的前提下，生成任务的结果也得到提升。总体来看，这两个模型有相似也有不同，MASK的方式对预训练模型的结果影响很大。

论文代码

MASS文章：https://arxiv.org/pdf/1905.02450.pdf
MASS代码：https://github.com/microsoft/MASS
BART文章：https://arxiv.org/pdf/1910.13461.pdf
BART代码：huggingface的transformer库最近更新了BART模型。

参考资料

[1] BART \ Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension(2019)

[2] MASS \ Masked Sequence to Sequence Pre-training for Language Generation(2019)

[3] Transformer \ Attention Is All You Need (2017)

[4] Span-BERT \ Improving Pre-training by Representing and Predicting Spans()

文章已于修改