BART

全称：Bidirectional and Auto-Regressive Transformers

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

paper：https://arxiv.org/pdf/1910.13461.pdf

预训练模型：

- `bart-large`：基础预训练模型；
- `bart-large-cnn`：基础模型在 CNN/Daily Mail Abstractive Summarization Task微调后的模型；
- `bart-large-mnli`：基础模型在MNLI classification task微调后的模型；

想法：

GPT：使用自回归方式预测 token，这意味着 GPT 可用于生成任务。但是，该模型仅基于左侧上下文预测单词，无法学习双向交互。

BERT：用掩码替换随机 token，双向编码文档。由于缺失 token 被单独预测，因此 BERT 较难用于生成任务。

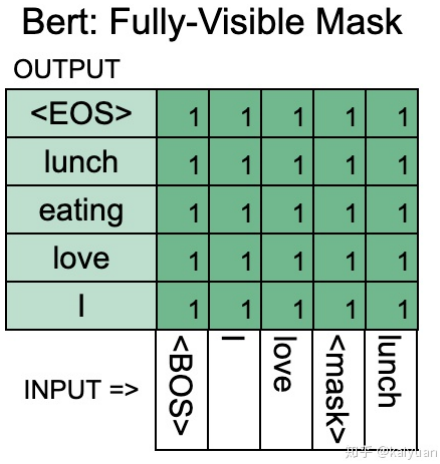
在Encoder端学习BERT的表示方法，在Decoder端学习GPT的生成方法，将两者的优点结合起来。

做法：

- 在BERT的双向编码器架构中添加因果解码器；
- 用更复杂的预训练任务代替BERT的完形填空任务。

Encoder-Decoder

其中Encoder的注意力矩阵是`Fully-visible`的， 和BERT一样，可以获取双向信息。



而Decoder的注意力矩阵是`autoregressive`，

Seq2Seq Decoder

Causal Mask

Output

<EOS>	1	1	1	1	1
lunch	1	1	1	1	0
eating	1	1	1	0	0
love	1	1	0	0	0
I	1	0	0	0	0

编码器和解码器通过`cross attention`连接，**其中每个解码器层都对编码器输出的最终隐藏状态进行attention操作**，这会使得模型生成与原始输入紧密相关的输出。

预训练方式

输入添加噪声后的文档，重建原文档，损失为原文档和输出文档的交叉熵损失

特点： 可以适用于所有的噪声添加方式

添加噪声：

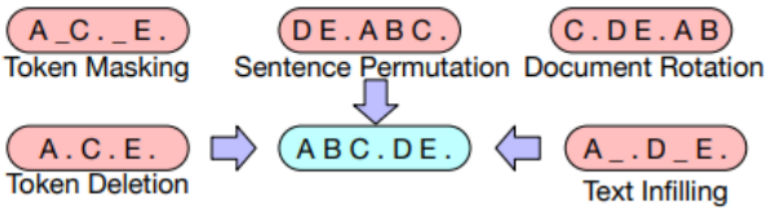


Figure 2: Transformations for noising the input that we experiment with. These transformations can be composed.

- Token Masking
- Token Deletion
- **Token Infilling:**

  对几个文本范围进行采样，并用一个[MASK] token替换(可以是0长度)
- Sentence Premutation

  随机打乱句子顺序
- Document Rotation

  随机选取一个token，将输入移动到以该token 开始

上面这些方式是可以相互组合训练的。

Fine Tuning

序列分类任务

序列分类任务中，编码器和解码器的输入相同，最终解码器 token 的最终隐藏状态被输入到新的多类别线性分类器中。该方法与 BERT 中的 CLS token 类似，不过 BART 在解码器最后额外添加了一个 token，这样该 token 的表征可以处理来自完整输入的解码器状态（见图 3a）。

token 分类任务

对于 token 分类任务，研究人员将完整文档输入到编码器和解码器中，使用解码器最上方的隐藏状态作为每个单词的表征。该表征的用途是分类 token。

序列生成任务

由于 BART 具备自回归解码器，因此它可以针对序列生成任务进行直接微调，如抽象问答和摘要。在这两项任务中，信息复制自输入但是经过了处理，这与去噪预训练目标紧密相关。这里，编码器的输入是输入序列，解码器以自回归的方式生成输出。

机器翻译

研究人员用新的随机初始化编码器替换 BART 的编码器嵌入层。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。

源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。第一步中，研究人员冻结 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。第二步中，研究人员将所有模型参数进行少量迭代训练。