BART

介绍

BART是一种采用序列到序列模型构建的降噪自编码器，适用于各种最终任务。它使用基于标准transformer的神经机器翻译架构。该模型具有在受损文本上的双向编码器和从左至右的自回归解码器。BART的预训练包括：

1. 使用噪声函数破坏文本;
2. 学习序列到序列模型以重建原始文本。

架构

- **Encoder**

  Encoder 负责将 source 进行 self-attention 并获得句子中每个词的 representation，最经典的 Encoder 架构就是 BERT，通过 Masked Language Model 来学习词之间的关系。**但是单独 Encoder 结构不适用于生成任务**

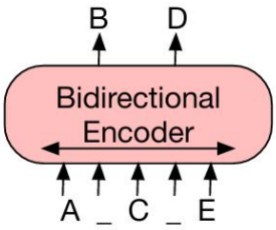

图1 双向编码器:用掩码替换随机Tokens，文档双向编码。丢失的Tokens是独立预测的。

- **Decoder**

  输入与输出之间差一个位置，主要是模拟在 Inference 时，不能让模型看到未来的词，这种方式称为自回归。但是单独 Decoder 结构仅基于左侧上下文预测单词，无法学习双向交互。

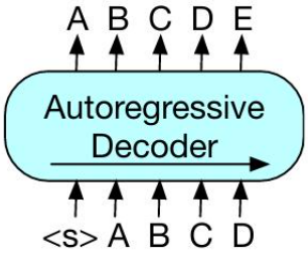

图2 自回归解码器：自回归预测的Tokens，这意味着GPT可以用于生成文本。

BART 使用标准的 Transformer 模型，不过做了一些改变：

1. 同 GPT 一样，将 ReLU 激活函数改为 GeLU，并且参数初始化服从正态分布 N(0,0.02)
2. BART base 模型的 Encoder 和 Decoder 各有 6 层，large 模型增加到了 12 层
3. BART 解码器的各层对编码器最终隐藏层额外执行 cross-attention

Pre-Training BART

BART的训练方法是先破坏文档，然后使用交叉熵损失（解码器的输出与原始文档之间的交叉熵）通过解码器恢复它们。

与其他去噪自编码器(一般需要定制特定的噪声方案)不同的是BART可以使用任何的加噪方式。在极端情况下，源信息可以全部缺失，此时的BART就蜕化成了一个语言模型。

BART 作者尝试了不同的方式来破坏输入：

1.**Token Masking**：像BERT一样，对随机tokens进行采样并替换为[MASK] token。

2.**Token Deletion**：从输入中删除随机tokens。与token屏蔽不同，该模型必须确定哪些位置缺少输入。

3.**Document Rotation**：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)。

4.**Sentence Permutation**：根据句号将文档分为句子。这些句子会随机排列。

5.**Document Rotation**：随机选择一个token，然后对文档进行转换，使其以该token开头。此任务训练模型以识别文档的开始。

Fine-tuning BART

1.序列分类任务：编码器和解码器的输入相同，解码器 token 的最终隐藏状态被输入到多类别线性分类器中。BART 在解码器最后额外添加了一个 token，如下图所示，该 token 位置的输出可以被认为是该句子的 representation

2.Token分类任务：将相同的输入馈送到编码器和解码器，并使用来自解码器的最终隐藏表示对tokens进行分类。

。。。等等一系列对比实验

使用bart模型做一个简单的摘要生成实验

```
tokenizer = BartTokenizer.from_pretrained('bart-large-cnn')
model = BartForConditionalGeneration.from_pretrained('bart-large-cnn')

article_input_ids = tokenizer.batch_encode_plus([LONG_BORING_TENNIS_ARTICLE], return_tensors='pt', max_length=1024)['input_ids'].to(torch_device)
summary_ids = model.generate(article_input_ids,num_beams=4,length_penalty=2.0,max_length=142,min_length=56,no_repeat_ngram_size=3)

summary_txt = tokenizer.decode(summary_ids.squeeze(), skip_special_tokens=True)
display(Markdown('> **Summary: **'+summary_txt))
```

```
Downloading: 100% 899k/899k [00:00<00:00, 2.86MB/s]
Downloading: 100% 456k/456k [09:44<00:00, 781B/s]
Downloading: 100% 1.26k/1.26k [00:27<00:00, 45.8B/s]
Downloading: 100% 1.63G/1.63G [00:27<00:00, 59.9MB/s]
```

> *Summary: *Andy Murray beat Dominic Thiem 3-6 6-4, 6-1 in the Miami Open. The world No 4 is into the semi-finals of the tournament in Florida. Murray was awaiting the winner of Tomas Berdych and Juan Monaco. Thiem lost in the second round of a Challenger event to soon-to-be new Brit Aljaz Bedene.

```
from transformers import GPT2LMHeadModel, GPT2Tokenizer
gpt2_tok = GPT2Tokenizer.from_pretrained('gpt2')
gpt2_model = GPT2LMHeadModel.from_pretrained('gpt2', output_past=True)

enc = gpt2_tok.encode(LONG_BORING_TENNIS_ARTICLE, max_length=1024-155, return_tensors='pt')

source_and_summary_ids = gpt2_model.generate(enc, max_length=1024, do_sample=False)

end_of_source = "An official statement said:"
_, summary_gpt2 = gpt2_tok.decode(source_and_summary_ids[0]).split(end_of_source)
display(Markdown('> **GPT2:** ' + summary_gpt2))
```

```
Downloading: 100% 1.04M/1.04M [00:00<00:00, 2.48MB/s]
Downloading: 100% 456k/456k [00:00<00:00, 3.27MB/s]
Downloading: 100% 224/224 [00:08<00:00, 25.1B/s]
Downloading: 100% 548M/548M [00:08<00:00, 63.1MB/s]
Setting `pad_token_id` to 50256 (first `eos_token_id`) to generate sequence
```

> **GPT2:** 'To have a player like James Ward, Kyle Edmund, Liam Broady and Aljaz Bedene in the top 100 is a huge achievement for the Lawn Tennis Association. The Lawn Tennis Association is committed to the development of the sport and the development of the sport's players. The Lawn Tennis Association is committed to the development of the sport and the development of the sport's players. The Lawn Tennis Association is committed to the development of the sport and the development of the sport's players. The Lawn Tennis Association is committed to the development of the sport and the development of the sport's players. The Lawn Tennis Association is committed to the development of the sport and the development of the sport's players. The Lawn Tennis Association is committed to the development of the sport and the development of the

由此可见，在摘要生成任务上，bart比gpt2的效果更好。

后续打算微调做一些其他下游任务的实验。