Encoder-Decoder 架构

Encoder-Decoder架构是一种强大的神经网络模型,主要用于处理序列到序列(Sequence to Sequence, Seq2Seq)任务,如机器翻译、文本摘要等。它在Encoder-only架构的基础上引入了Decoder组件,形成了一个完整的编码-解码系统。

架构组成

该架构主要包含两个核心部分:

1.**编码器(Encoder)**

- 由多个编码模块堆叠而成
- 每个编码模块包含:
    - 自注意力模块
    - 全连接前馈模块
- 将输入序列转换为包含丰富语义信息的上下文向量

2.**解码器(Decoder)**

- 由多个解码模块堆叠而成
- 每个解码模块包含:
    - 带掩码的自注意力模块(防止信息泄露)
    - 交叉注意力模块(实现编解码器信息交互)
    - 全连接前馈模块

注意力机制特点

1.**编码器中的自注意力**

- 采用双向注意力机制
- 全面捕捉上下文信息
- 对输入序列进行"通盘考虑"

2.**解码器中的自注意力**

- 采用单向注意力机制
- 仅以上文为条件解码
- 通过掩码操作避免"窥视"未来信息

3.**交叉注意力**

- 连接编码器和解码器
- 将解码器的查询(query)与编码器的键(key)和值(value)结合
- 确保生成过程中能参考输入的全局上下文

T5模型

论文：《Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer》

链接：https://arxiv.org/abs/1910.10683

创新点

T5(Text-to-Text Transfer Transformer)模型由Google Research团队于2019年提出,其主要创新在于:

- 将多种NLP任务统一为文本到文本的转换范式
- 通过输入前缀指示不同任务
- 开创了早期的提示(Prompt)技术应用

模型版本

T5提供了5个不同规模的版本:

版本	编码/解码模块数	隐藏层维度	注意力头数	总参数量
T5-Small	6/6	512	8	6000万
T5-Base	12/12	768	12	2.2亿
T5-Large	24/24	1024	16	7.7亿
T5-3B	24/24	1024	32	28亿
T5-11B	24/24	1024	128	110亿

预训练创新

- 基于C4(Colossal Clean Crawled Corpus)数据集,规模约750GB
- 采用Span Corruption预训练任务:
  - 选择15%的连续Token进行掩码
  - 每次掩码3个连续Token
  - 要求模型预测完整的语义片段

下游任务适配

T5模型支持两种任务适配方式:

1. 零样本(Zero-Shot)学习:
   - 利用Prompt工程技术
   - 无需额外训练数据
   - 适用于一般场景
2. 微调(Fine-Tuning):
   - 需要带标签训练数据
   - 需要更多计算资源

- 适用于高精度要求场景

衍生模型

- mT5: 支持100多种语言的多语言版本
- T0: 增强零样本学习能力的多任务训练版本
- Flan-T5: 通过指令微调提升模型灵活性的改进版本

BART

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

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

BART（Bidirectional and Auto-Regressive Transformers）是由Meta AI研究院于2019年10月提出的Encoder-Decoder架构模型。BART通过设计多样化的预训练任务来同时提升模型在文本生成和理解任务上的表现能力。

模型结构

BART采用标准的Transformer架构,包含编码器和解码器两部分。模型提供两个版本:

- BART-Base: 6层编码器/解码器,768维隐藏层,12个注意力头,1.4亿参数
- BART-Large: 12层编码器/解码器,1024维隐藏层,16个注意力头,4亿参数

预训练数据

BART使用与RoBERTa相同的预训练语料,包括:

- BookCorpus(小说数据集)
- 英语维基百科
- CC-News(新闻数据集)
- OpenWebText(网页数据)
- Stories(故事数据集)

总数据量约160GB。

预训练任务

BART通过五种不同的文本破坏任务来训练模型进行文本重建:

1. **Token遮挡(Token Masking)**

- 随机将部分token替换为[MASK]
- 训练模型恢复被遮挡的内容
- 例如："I love [MASK] and [MASK].", "I [MASK] jogging and [MASK]."

2. **Token删除(Token Deletion)**

- 随机删除部分token
- 训练模型推断缺失位置和内容
- 例如："I love and reading", "I jogging and reading.", "love jogging and reading."

3. **连续文本填空(Text Infilling)**

- 替换多个连续token为单个[MASK]
- span长度服从$\lambda = 3$的泊松分布
- 训练模型恢复完整文本片段
- 例如："I love [MASK] and reading."

4. **句子打乱(Sentence Permutation)**

- 随机打乱句子顺序
- 训练模型理解句间逻辑关系
- 例如："jogging and reading. I love"

5. **文档旋转(Document Rotation)**

- 随机选取token作为新起点重排文本
- 训练模型识别合理的文本开头
- 例如："and reading. I love jogging "

这些任务通过破坏和重建文本的方式,增强了模型对文本结构和语义的理解能力,提高了模型在处理不完整或受损文本时的鲁棒性。

MASS

论文：《MASS: Masked Sequence to Sequence Pre-training for Language Generation》

链接：https://arxiv.org/abs/1905.02450

MASS（Masked Sequence to Sequence）是由微软亚洲研究院于2019年提出的预训练语言模型。它通过序列到序列框架中的部分掩码（mask）策略，巧妙地融合了自编码与自回归的优点，以在机器翻译、文本摘要等生成任务上获得更强表现。

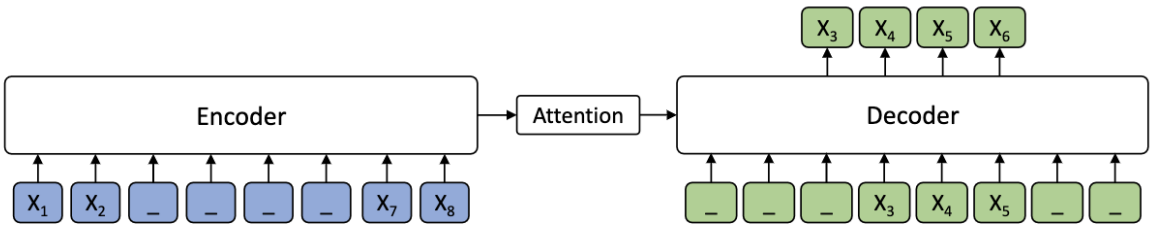

Figure 1. The encoder-decoder framework for our proposed MASS. The token "_" represents the mask symbol [$\mathbb{M}$].

模型架构

MASS采用标准的Transformer encoder-decoder架构:

- Encoder: 12层Transformer编码器
- Decoder: 12层Transformer解码器
- 隐藏层维度: 1024
- 注意力头数: 16
- 总参数量: 约3.2亿

预训练任务设计

MASS提出了一种创新的预训练目标：将输入序列的某个连续片段（span）进行掩码（mask），而仅让Decoder重建被掩码部分，非掩码部分则由Encoder提供给Decoder。这样一来：

- Encoder能理解上下文全局信息
- Decoder只需学会根据部分可见文本（Encoder输出）来生成掩盖的token序列
- 兼具自编码和自回归的优势

给定输入句子 $\mathbf{x} = (x_1, x_2, \ldots, x_n)$，MASS的训练目标为：

$$\mathcal{L}_{\text{MASS}} = \mathbb{E}_{(\mathbf{x}\sim\mathcal{D})} \left[-\log P_\theta \left(\mathbf{x}_{a:b} \mid \tilde{\mathbf{x}} \right) \right] \tag{1}$$

其中：

- $\tilde{\mathbf{x}}$ 在 $\mathbf{x}_{a:b}$ 位置用［MASK］替代，其他位置保持原token
- Decoder自回归地生成 $\mathbf{x}_{a:b}$ 的token序列

统一的预训练框架

MASS通过调整掩码长度k，可以统一多种预训练方法：

- 当`k=1`时：等价于BERT的掩码语言模型
- 当`k=m`时（`m`为序列长度）：等价于GPT的标准语言模型
- 当`k`在`(1,m)`之间：介于两者之间的新方法

Length	Probability	Model
$k=1$	$P(x^u \mid x^{\setminus u};\theta)$	masked LM in BERT
$k=m$	$P(x^{1:m} \mid x^{\setminus 1:m};\theta)$	standard LM in GPT
$k \in (1,m)$	$P(x^{u:v} \mid x^{\setminus u:v};\theta)$	methods in between

预训练数据

MASS使用大规模单语语料进行预训练：

- WMT News Crawl数据
- Wikipedia数据
- 总数据量约100GB

应用场景

MASS在以下任务上表现优异：

- 机器翻译
- 文本摘要
- 对话生成
- 问答系统

通过在目标任务数据上进行微调，MASS可以快速适应各类生成任务。

总结

Encoder-Decoder架构的大语言模型在生成任务上展现出优异性能。这些模型凭借其双编码器-解码器结构和庞大的参数规模,在机器翻译、文本摘要、问答等任务上都取得了优于Encoder-only架构的表现。