Encoder-only 架构

Encoder-only架构的核心是双向编码模型(Bidirectional Encoder Model)。该模型在处理输入序列时,同时利用从左到右和从右到左的注意力机制,能够全面捕捉每个token的上下文信息,因此也被称为全面注意力机制。这种双向编码的特性使其在自然语言处理任务中表现出色。

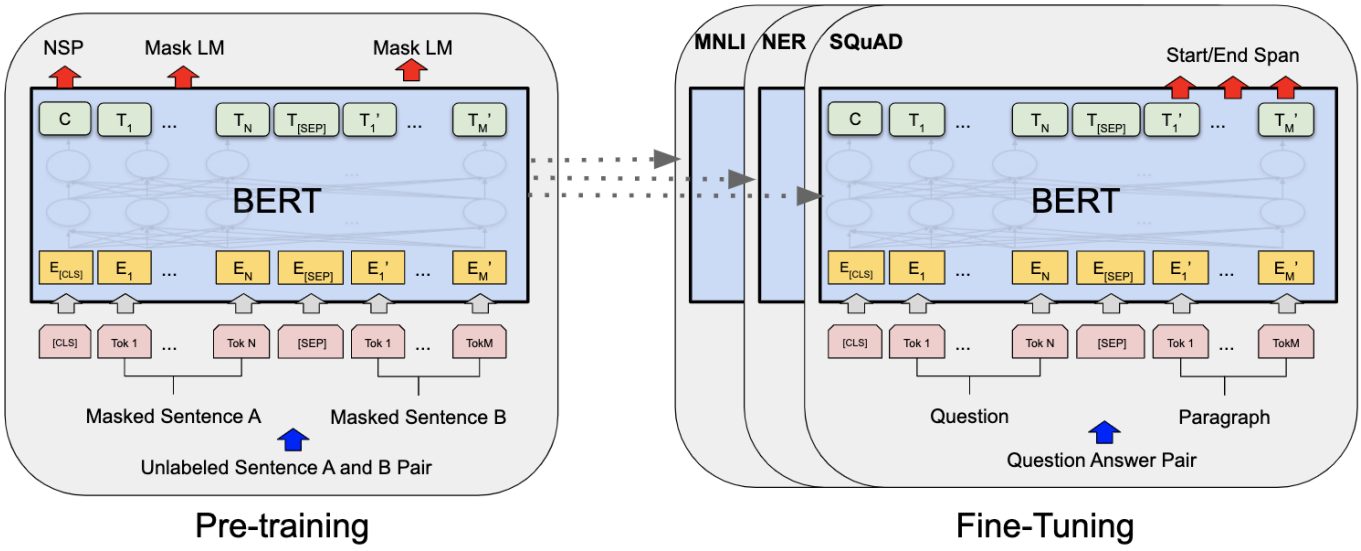
与Word2Vec、GloVe等为每个词提供静态向量表示的传统方法不同,双向编码器为每个词生成动态的上下文嵌入(Contextual Embedding)。这种嵌入会根据具体的上下文动态调整,能够更准确地理解词语间的依赖关系和语义信息,有效解决词语多义性问题。研究表明,这种动态表示方法在句子级(sentence-level)任务上的表现显著优于静态词嵌入方法。

Encoder-only架构基于双向编码模型,采用了Transformer架构中的编码器部分。虽然不直接生成文本,但其产生的上下文嵌入对深入理解输入文本的结构和含义至关重要。这些模型在需要深度理解和复杂推理的NLP任务中展现出卓越能力。目前,BERT及其变体(如RoBERTa、ALBERT等)都是基于Encoder-only架构的主流大语言模型。

BERT语言模型

论文：BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

地址：https://arxiv.org/pdf/1810.04805



BERT由Google于2018年10月提出,是继Word2Vec、ELMo、GPT之后的一个里程碑式模型。它基于Transformer的Encoder结构,通过双向掩码语言模型(Masked LM)和句子级任务(Next Sentence Prediction)进行大规模预训练,在11项NLP任务上创造了当时的最佳成绩,推动了"预训练+微调"范式在NLP领域的广泛应用。

BERT模型结构

BERT的结构与Transformer编码器基本一致,由多个编码模块堆叠而成,每个模块包含多头自注意力层和前馈网络层。BERT有两个版本:

- BERT-Base: 12层编码器,768维隐藏层,12个注意力头,1.1亿参数
- BERT-Large: 24层编码器,1024维隐藏层,16个注意力头,3.4亿参数

BERT预训练方式

BERT使用BookCorpus(8亿token)和英语维基百科(25亿token)数据集进行预训练,总计约33亿token,数据量达15GB。预训练采用两个创新性任务:

掩码语言建模(Masked Language Model,MLM)和前面文章中提到的下一句预测(Next Sentence Prediction,NSP)。具体流程如下:

首先,从原始文本构造样本序列,每个序列包含两个句子。这两个句子有50%概率是连续的,50%概率是随机选取的。然后对序列分词,添加特殊标记`[CLS]`(用于聚合序列信息)和`[SEP]`(用于分隔句子)。

接着进行NSP任务,训练模型判断序列中的两句话是否连续。这帮助模型理解句子间的关系,对问答、推理等任务很有帮助。

最后进行MLM任务,随机遮盖约15%的token(替换为`[MASK]`或随机词),让模型预测被遮盖的原始内容。这类似完形填空,训练模型理解上下文。模型仅对被遮盖的token计算损失并更新参数。

这两个预训练任务的结合使BERT能够同时理解token级别的细节和句子级别的语义,为下游任务提供了坚实的语言理解基础。

BERT 衍生语言模型介绍

BERT的成功催生了一系列衍生模型，它们继承了BERT双向编码的核心特性，并在此基础上进行改进和优化，以提升性能或效率。其中最具代表性的是RoBERTa、ALBERT、SpanBERT、XLNet、ELECTRA和DeBERTa等， 下面将分别介绍这些模型。

RoBERTa

RoBERTa(Robustly Optimized BERT Pretraining Approach)由Facebook AI(现Meta)于2019年7月提出，旨在通过更充分的训练来提升BERT的性能。该模型通过扩大训练数据集、延长训练时间和优化超参数设置来改进预训练过程，从而提高模型在各类NLP任务上的表现。

RoBERTa在结构上与BERT基本一致，同样基于多层堆叠的编码模块。RoBERTa-Base与BERT-Base对标，包含12个编码模块，768维隐藏层，12个注意力头，约1.2亿参数。RoBERTa-Large则与BERT-Large对标，包含24个编码模块，1024维隐藏层，16个注意力头，约3.5亿参数。

在预训练方面，RoBERTa显著扩充了训练语料，除了使用BERT原有的BookCorpus和维基百科数据外，还加入了CC-News、OpenWebText和Stories等数据集，总量达到约160GB。RoBERTa移除了BERT的下文预测任务，并将静态掩码改进为动态掩码机制。通过对训练数据创建多个不同掩码版本，增加了训练的多样性，使模型能够学习更丰富的上下文信息。

> 静态掩码：在BERT中，掩码是静态的，即在预训练过程中，每个token的掩码位置是固定的。
>
> 动态掩码：在RoBERTa中，掩码是动态的，即在预训练过程中，每个token的掩码位置是随机选择的。

ALBERT

论文：ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

地址：https://arxiv.org/pdf/1909.11942

ALBERT(A Lite BERT)由Google Research团队于2019年9月提出，是一个轻量级BERT变体。针对BERT参数量庞大(Base版1.1亿，Large版3.4亿)导致的训练和推理效率问题，ALBERT通过创新的参数共享和嵌入分解技术大幅降低了模型参数量。

ALBERT在保持与BERT相似结构的同时，引入了两项关键优化。

- **参数因子分解**，将词表映射矩阵分解为两个较小的矩阵，显著减少了嵌入层的参数量。例如，通过将嵌入维度设为128，ALBERT的嵌入层参数量仅为BERT的六分之一左右，和LoRA一样，也是通过低秩分解来减少参数。
- **跨层参数共享机制**，让所有编码层共用同一组参数，进一步压缩了模型规模。

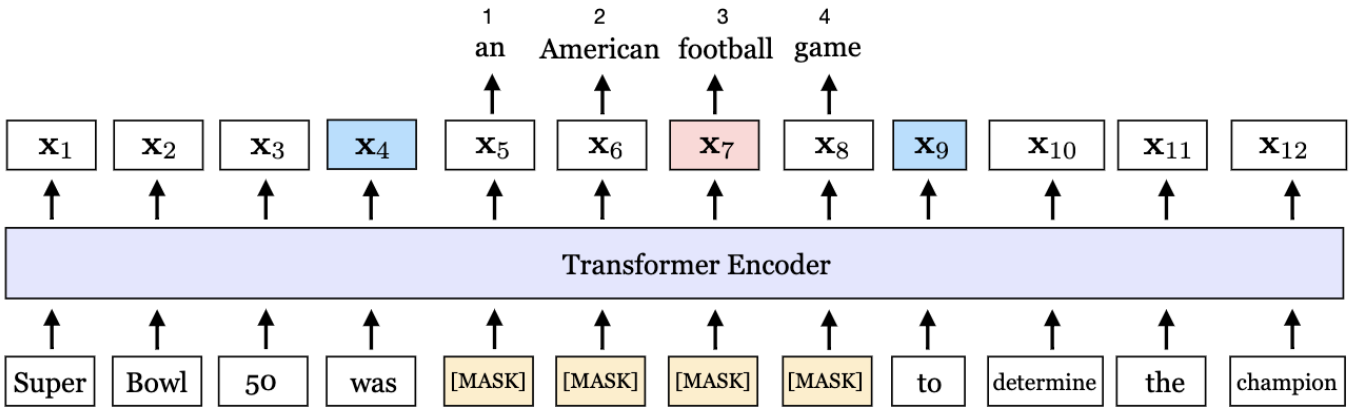
通过这些创新设计，ALBERT在保持较好性能的同时显著提升了效率，特别适合资源受限环境下的部署应用。其参数共享和压缩技术为大规模语言模型的轻量化提供了重要思路。

SpanBERT

论文：SpanBERT: Improving Pre-training by Representing and Predicting Spans
地址：https://arxiv.org/pdf/1907.10529

SpanBERT由Google于2020年10月提出，是一个专门针对跨度(span)级别文本理解的BERT变体。该模型通过创新的预训练目标和掩码策略，显著提升了在抽取式问答、指代消解等需要理解连续文本片段的任务上的表现。

$$\begin{aligned}\mathcal{L}(\text{football}) &= \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football}) \\ &= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)\end{aligned}$$



SpanBERT主要包含三个关键改进:

1 Span级别掩码(Span Masking)

相比BERT的单token掩码，SpanBERT采用了更有挑战性的连续跨度掩码:

- 首先随机采样一个span长度n
- 然后在文本中选择起始位置，对连续n个token进行掩码
- 这种策略迫使模型必须理解更长的上下文才能准确预测被掩码内容
- 更符合实际应用中对连续文本片段的理解需求

2 边界目标函数(Span Boundary Objective, SBO)

在MLM基础上引入了新的预训练目标SBO:

$$\mathrm{SBO}\left(p_i\right)=f\left(x_{s-1}, x_{e+1}, p_i\right) \tag{1}$$

其中:

- x_{s-1}和x_{e+1}分别是被掩码span的前后边界token表示
- p_i是span内第i个位置的相对位置编码
- 通过边界信息预测span内部token，增强了模型对连续文本片段的建模能力

3. 移除NSP任务

与RoBERTa等模型一样，SpanBERT也去掉了原BERT中的下一句预测(NSP)任务。实验表明这种简化不仅没有损害性能，反而提升了模型效果。

通过以上优化，SpanBERT在需要深入理解文本span的任务上取得了显著进展，如:

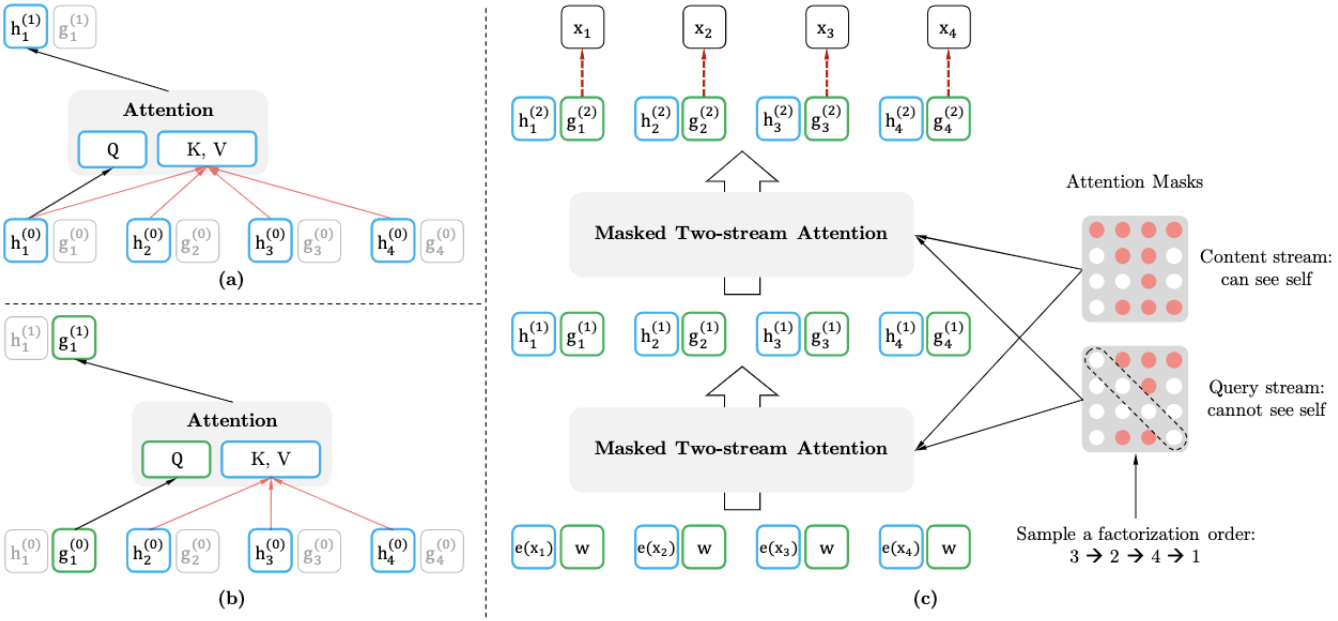
- 抽取式问答: 准确定位答案片段
- 指代消解: 识别指代关系
- 实体抽取: 提取连续的命名实体

XLNet

论文：XLNet: Generalized Autoregressive Pretraining for Language Understanding

地址：https://arxiv.org/pdf/1906.08237

XLNet是由CMU和Google Brain联合提出的一个创新性预训练语言模型，它巧妙地结合了自回归语言模型(AR LM)和自编码语言模型(AE LM)的优势，提出了全新的排列语言建模(Permutation Language Modeling, PLM)预训练目标。



模型设计动机源于对现有方法的深入思考:

- 自回归语言模型(如GPT)：采用从左到右的预测方式，虽然天然适合生成任务，但只能获取单向上下文信息
- 自编码语言模型(如BERT)：通过掩码预测获取双向信息，但[MASK]标记导致预训练和推理阶段的不一致性

核心创新：排列语言建模(PLM)

对于长度为T的输入序列$[x_1, x_2, \ldots, x_T]$，XLNet的做法是：

1. 训练时随机生成序列的一个排列顺序π
2. 按照π的顺序进行自回归预测
3. 在注意力机制中，位置$\pi(j)$只能访问$\pi(k)$中$k < j$的token
4. 通过多次随机排列，使每个token都有机会处于不同位置，从而获得完整的上下文信息

为进一步提升模型性能，XLNet还引入了两个重要优化：

1 部分预测(Partial Prediction)
- 只对序列后部分token进行预测
- 避免因前期上下文信息不足影响模型训练的稳定性

2 整合Transformer-XL架构
- 引入相对位置编码和片段循环机制(Segment Recurrence)
- 显著增强模型处理长序列文本的能力

通过这些创新设计，XLNet在多个NLP任务上取得了超越BERT的性能。其保持自回归特性的同时又能获取双向上下文的设计，使其在理解任务和生成任务上都表现出色。

ELECTRA

论文：ELECTRA: Pre-training with Masked Token Prediction and Sentence Reordering

地址：https://arxiv.org/pdf/2003.10555

ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately)是由Google Brain和斯坦福大学研究人员于2020年3月提出的BERT变体，旨在解决大规模预训练语言模型的效率和可扩展性问题。该模型通过生成器-判别器架构，能更高效地利用预训练数据，提升下游任务表现。

在预训练方面，ELECTRA结合了生成对抗网络(GAN)的思想，采用生成器-判别器结构。其中生成器负责掩码预测，类似BERT模型，将掩码后的文本恢复原状；判别器则执行替换词检测(RTD)任务，检测生成器输出中每个token是否为原文内容。ELECTRA通过扩充数据集，包括ClueWeb、CommonCrawl和Gigaword等，将训练数据量扩充至330亿个token，帮助模型学习更广泛的语言表示。

与BERT相比，ELECTRA的一个重要改进是判别器会对所有token进行真伪判断，而不是仅限于15%的掩码token。这使得模型能够更全面地学习文本的上下文表示。不同于RoBERTa和ALBERT主要在模型结构和预训练数据规模上优化，ELECTRA通过引入生成器-判别器架构和替换语言模型任务，显著提升了训练效率和效果。

DeBERTa

论文：DeBERTa: Decoding-enhanced BERT with Disentangled Attention

地址：https://arxiv.org/pdf/2006.03654

Algorithm 1 Disentangled Attention

DeBERTa(Decoding-enhanced BERT with Disentangled Attention)是一个基于BERT的改进模型,主要通过解耦注意力机制和增强掩码解码器来提升性能。其核心创新点包括:

解耦注意力机制

传统Transformer中的自注意力机制将词语内容和位置信息混合处理,可能导致信息干扰。DeBERTa通过将内容和位置信息解耦,分别计算:

- 内容注意力: 基于词语语义信息计算注意力分数
- 位置注意力: 基于相对位置信息计算注意力分数

最终的注意力分数为:
$Attention(Q^C, K^C, V^C) + Attention(Q^P, K^P, V^P)$

这种解耦设计使模型能更好地捕获语义和位置依赖关系。

增强掩码解码器

DeBERTa对BERT的MLM任务进行了多项改进:

- 采用全词掩码策略,掩盖完整词组而非单个token
- 增加解码器层数,提升恢复被掩码内容的能力
- 优化损失函数设计,提高预测准确率

其他创新点

- 使用相对位置编码替代绝对位置编码
- 通过参数共享提升计算效率
- 采用句子顺序预测(SOP)替代NSP任务

DeBERTa模型结构

DeBERTa基于Transformer编码器架构,主要包含:

1 输入层

- 词嵌入:编码语义信息
- 位置嵌入:编码位置信息
- 分段嵌入:区分不同句子

2 解耦注意力层

- 分别处理内容和位置信息
- 融合两类注意力分数

3 前馈网络层

- 标准Transformer前馈结构

4 输出层

- 根据下游任务配置不同的输出头

DeBERTa预训练方法

DeBERTa采用两个预训练任务:

- **增强版MLM**
- 结合全词掩码和深层解码器
- 提升对上下文的理解能力
- **句子顺序预测(SOP)**
- 判断句子顺序是否正确
- 学习句间逻辑关系

总损失函数:
$\mathcal{L} = \mathcal{L}_{MLM} + \mathcal{L}_{SOP}$

DeBERTa模型变体

1 基础版本

- Base: 适用于一般任务
- Large: 适用于复杂任务

- XLarge: 适用于大规模场景

2.DeBERTaV2

- 更深的解码器结构
- 优化的训练策略
- 改进的位置编码
- 在多个NLP任务上取得SOTA性能

通过这些创新设计,DeBERTa显著提升了BERT的性能,特别是在问答和阅读理解等需要深度语言理解的任务上表现优异。

总结

本章介绍了几种主要的Encoder-only模型,下表总结了它们的主要特点:

模型	主要创新	预训练任务	优势
BERT	双向编码器架构	MLM + NSP	通用语言理解能力强
RoBERTa	优化训练策略	动态掩码MLM	更稳定的训练过程
ALBERT	参数共享机制	MLM + SOP	更小的模型体积
SpanBERT	Span掩码和预测	Span MLM + SOP	更好的实体和关系抽取
XLNet	排列语言建模	PLM	更好的长文本建模
ELECTRA	生成器-判别器架构	RTD	更高效的预训练
DeBERTa	解耦注意力机制	增强MLM + SOP	更好的语言理解能力

这些模型都基于Transformer编码器架构,通过不同的创新设计和训练方法,在自然语言处理任务上取得了显著进展。未来的研究方向包括进一步提升模型效率、探索新的预训练任务以及针对特定领域的优化等。