0511 汇报

朱睿涵　李宸亦

项目三

2022 年 5 月 11 日

目录

现有问题

目前主流的机器阅读理解模型：抽取式

- 特点：将答案设定为段落中的一个连续片段
- 局限：不自然、不通顺

生成式 vs 抽取式：

- 生成式不再局限于直接从段落中抽取答案
- 参考段落、问题、词表生成答案
- 更完整、更自然、更流畅

现有生成式模型的问题：通常基于整个段落生成答案，**缺乏对答案边界和问题类型信息的理解**

- 本文的解决方案：基于多任务学习的生成式阅读理解框架
- 本文 baseline 模型：UniLMV2 模型

多任务学习

多任务学习机制：可提高模型泛化能力
机制：

- 同时学习多个相关任务，让这些任务同时共享知识
- 利用任务之间的相关性，提升每个任务的泛化性能

一般做法：

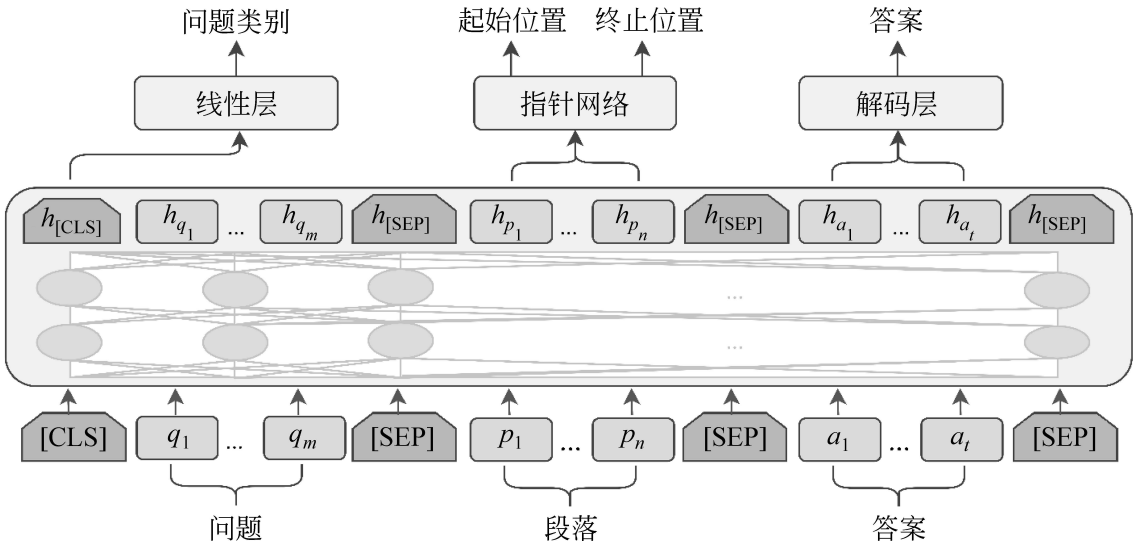
- 在所有任务上共享模型编码层
- 针对特定的任务层有所区别

本文模型的多任务

- 答案生成（主任务）
- 答案抽取（辅助任务）
- 问题分类（辅助任务）

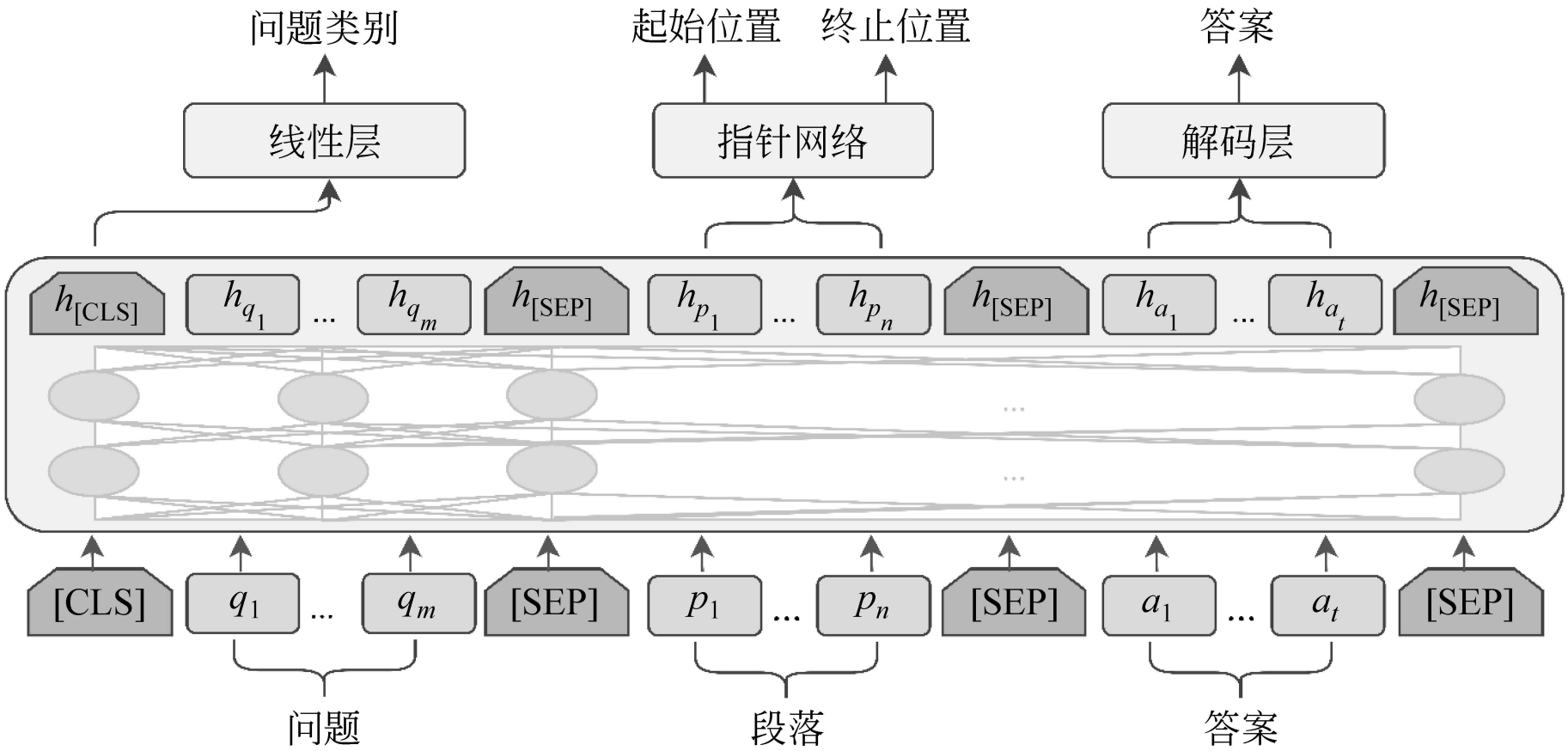
模型概览

本文提出的生成式阅读理解模型由编码层和任务层组成：



- 编码层：
  - 基于深度双向 Transformer 编码器
  - 借鉴 UniLMV2 的自注意力掩码机制控制答案生成过程中的可见信息
- 任务层：
  - 答案生成模型：beam search 解码生成
  - 答案抽取模型：指针网络识别答案位置
  - 问题分类模型：线性层判断问题具体类型

编码层



- 基于预训练模型 UniLMV2 构建编码层
- 采用预训练的 BERT 进行问题和段落的交互
- 改进注意力遮蔽矩阵，采用伪遮蔽语言模型

编码层具体原理和过程

词嵌入：

- 采用 WordPiece 分词工具，将问题、段落和答案分词
- 对答案的部分词项进行一定概率的遮蔽，拼接后作为模型输入词向量 X_i：

$$X_i = WE(w_i) + SE(w_i) + PE(w_i) \tag{1}$$

输入序列表示为：

$$H^0 = [X_1, X_2, ..., X_{|x|}] \tag{2}$$

编码层：UniLMV2 的编码层使用 12 层堆叠的 Transformer 网络

- Transformer 两个子层：

$$LayerNorm(x + SubLayer(x)) \tag{3}$$

  - 多头注意力机制
  - 前向神经网络

编码层具体原理和过程

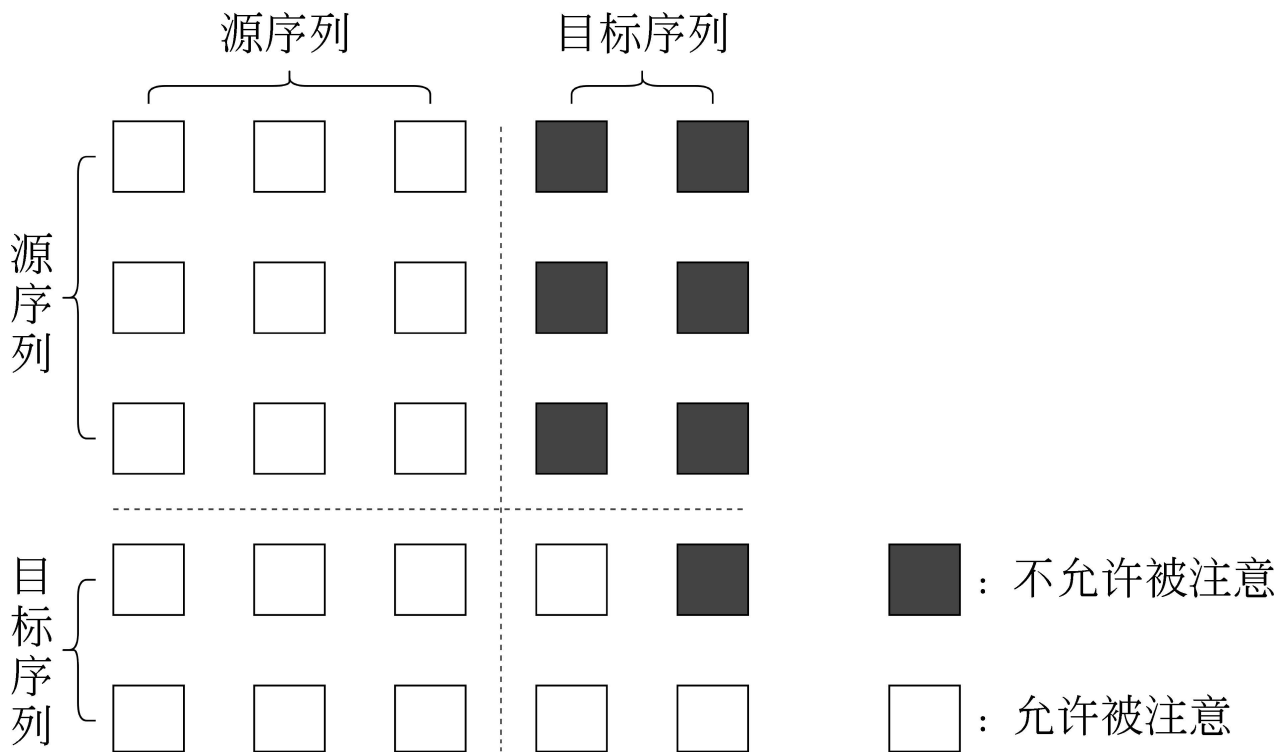
Transformer 的自注意力头 A_l 计算：

$$A_l = softmax(\frac{Q_l K_l^T}{\sqrt{d_k}} + M) V_l \tag{4}$$

$$Q_l = H^{l-1} W_l^Q, K_l = H^{l-1} W_l^K, V_l = H^{l-1} W_l^V \tag{5}$$

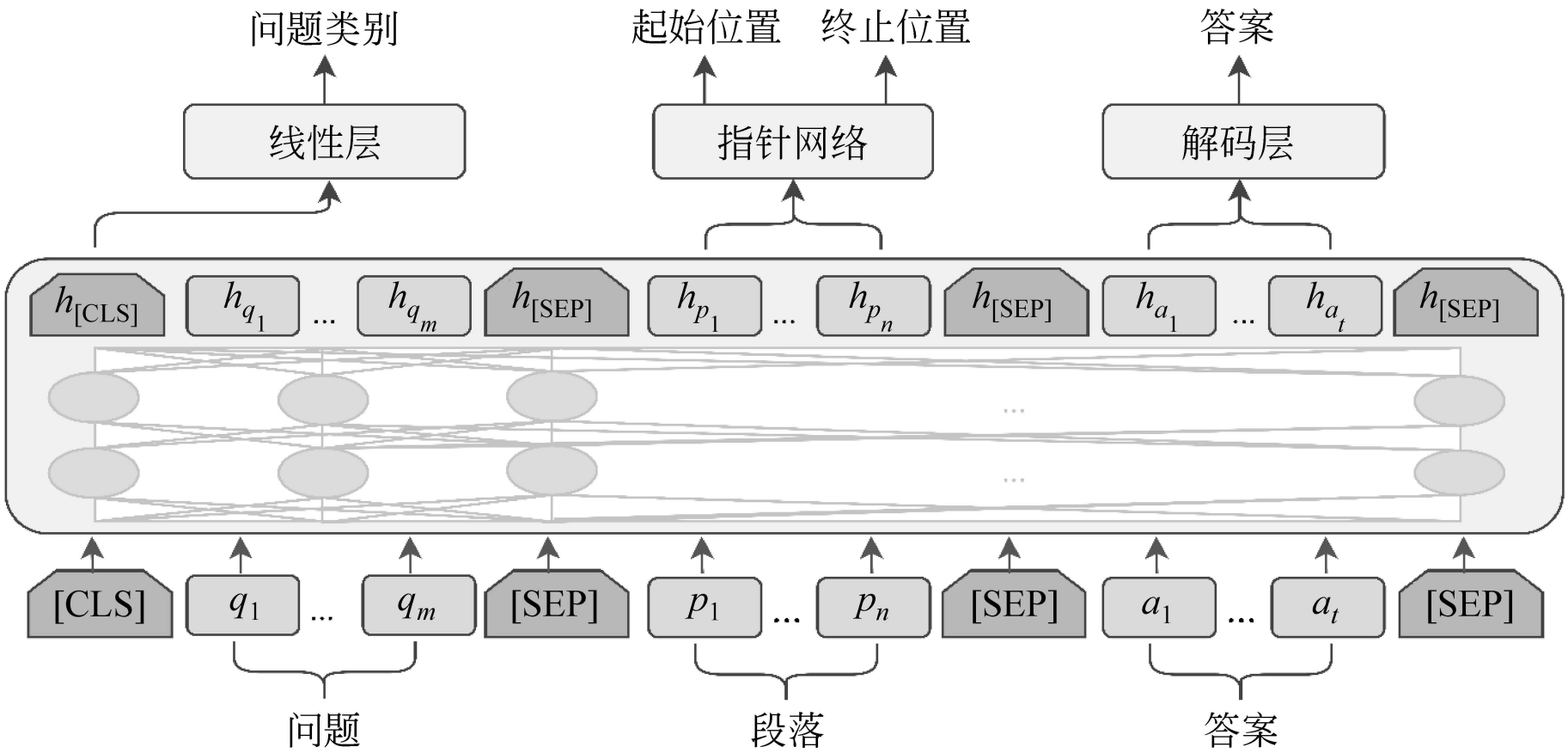
编码层具体原理和过程

M 为注意力遮蔽矩阵：



问题和段落不会和答案进行交互，保证了训练和测试阶段所含信息的一致性

任务层



任务层由答案生成模型、答案抽取模型和问题分类模型三部分组成

任务层

答案生成模型：

- 训练阶段：
  - 真实答案会以一定概率被随机遮蔽
  - 通过解码层对被遮蔽的词项进行预测来生成答案

$$H^a = LayerNorm(Gelu(Linear(H^a))) \tag{6}$$

$$\alpha = Softmax(Linear(H^a)) \tag{7}$$

- 测试阶段：直接采用训练好的解码层和 beam search 对问题和段落进行解码，生成答案

答案抽取模型：通过指针网络对答案的起始和终止位置进行识别

$$s, e = Softmax(Linear(H^p)) \tag{8}$$

问题分类模型：用线性层判断问题的类别

$$c = Softmax(Linear(H^{cls})) \tag{9}$$

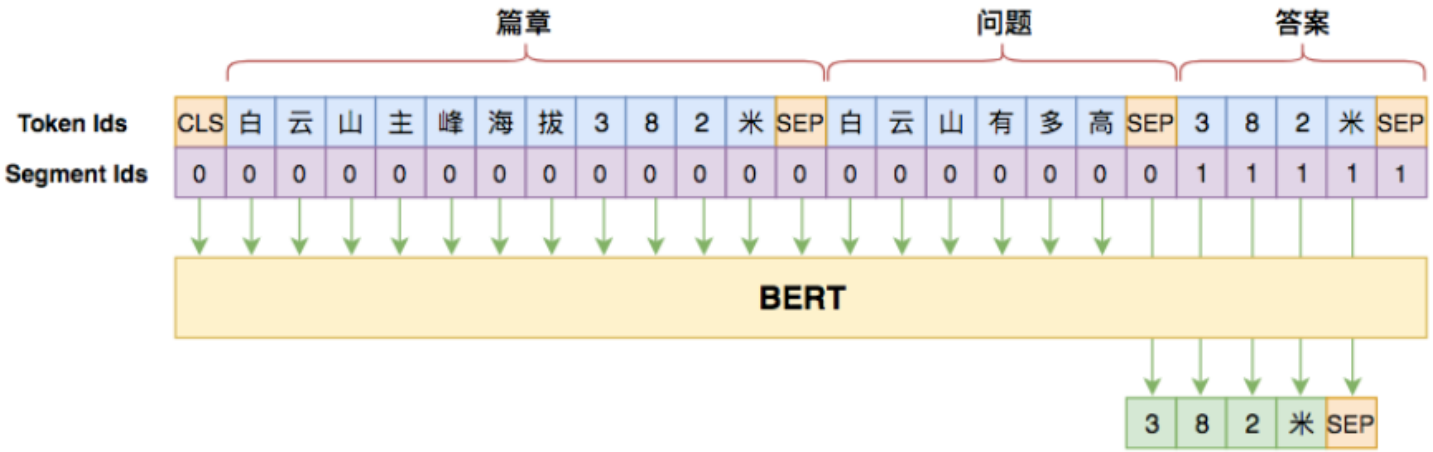
论文总结

- 针对问题：生成式阅读理解模型缺乏答案边界和问题分类信息的理解
- 提出模型：基于多任务学习的生成式阅读理解模型，通过答案抽取模型和问题分类模型优化生成式阅读理解模型
- 实验结果：在 CoQA，MS MARCO(NLG)，NarrativeQA 三个数据集上均取得目前生成式模型的最好性能

目录

Baseline

没有细看，仅初步浏览
模型：UniLM，与上篇论文中的 baseline 相似



输出处理：普通的 beam search 基础上加上按篇章平均
存在问题：模型缺乏答案边界和问题分类信息的理解
解决方法：使用多任务学习进行优化

目录

数据集

仅初步了解

- CoQA：基于多个领域的多轮对话进行构建，保持了人类对话简短的特征，存在大量的指代和省略现象，问题和答案普遍较短
- MS MARCO：多文档问答数据集，其中提供了一个自然语言生成(NLG)的子数据集，答案并非严格匹配文档中的片段
- NarrativeQA：基于书本故事和电影脚本人工编辑构建的生成式阅读数据集

目录

总结

这两天的工作：

- 阅读论文：基于多任务学习的生成式阅读理解
- 初步了解 baseline，还未细看
- 初步了解相关数据集

References