Week6

RetroReader 概览

RetroReader 是上海交通大学提出的一种回顾式阅读器,基于 Bert 和其他策略的组合阅读理解模型,通过 SQuAD 数据集,判断问题是否可回答和准确回答。它首次提出了显著性检验方法,展示该论文的模型,比 baselines 显著性更好,获取了 SOTA 结果。

RetroReader 模型包括泛读模块(Sketchy Reading)和精读模块(Intensive Reading),泛读模块用于阅读文章和问题,得到初步的判断,精读模块验证可回答性,并给出候选。两个模块的输出汇总在一起,做出最终决定。整体框架如下:

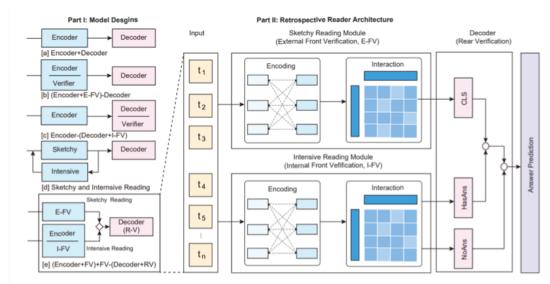


Figure 1: Reader overview. For the left part, models [a-c] summarize the instances in previous work, and model [d] is ours, with the implemented version [e]. In the names of models [a-e], " (\cdot) " represents a module, "+" means the parallel module and "-" is the pipeline. The right part is the detailed architecture of our proposed Retro-Reader.

前人工作:包括直接预测不验证的 Encoder+Decoder 模式,将答案验证放到 Encoder 或 Decoder 阶段的(Encoder+E-FV)-Decoder 模式和 Encoder-(Decoder+I-FV)模式。注意此中 Encoder、Decoder 表示的是 passage 和 question 做的交互,非 transformer 中的 Encoder 和 Decoder:

泛读器: 它是一个基于 Bert 的 External Front Verification, E-FV, 输入的嵌入向量为 Token embedding+Position embedding+Token-type embedding, 通过交互层 Multihead Attention、Add+Norm、FFN 和 Add+Norm 后, 进入 Fine-Tuning 微调阶段,在此阶段构建是否可回答二分类问题,得到输出,使用第一个输出 h0 of [CLS]判断问题是否可回答;

精读器:和泛读器类似,经过 Bert 模型,得到的 H 输出进行 passage 和 question 的拆分,获得 HP 和 HQ。用 HP 和 HQ 进行 cross attention 和 matching attention 计算,得到的结果 H'用来决定 start 和 end indices;

接着汇总泛读器和精读器的结果,综合判断问题是否可回答,再结合精读器的输出,得到有结果和无结果的评分。通过 threshold 来判断最终结果为:有答案或无答案。具体流程图可参照附件一。

RetroReader 是一种新形式的模型,对后续解决问题有很大的启发,但训练较慢。

附件一: 流程图

