模型改进思路

日期：2023 年 3 月 18 日

1 T5 模型的理解与生成能力

我认为 T5 模型能够理解给定的文本 (我理解的这里文本包括自然语言问题和数据库 schma 信息)，且生成的 SQL 语句能符合要求，但是仍然有提升空间。首先是在 MIGA 论文中案例分析的部分如图 1 所示[1]，可见 T5-3B 模型在 Text-to-SQL 领域的单轮问题做 zero-shot 甚至二次提问可以有较优的效果，并且在 Spider 数据上可以达到 EM：71.5%，EX：74.4% 的准确率，这已经是超过许多传统 Text-to-SQL 领域模型的效果了。

Case #1 Goal	How many dog pets are raised by female students?
Question #1	Who are the female students?
T5-3B & MIGA	**SELECT** * **FROM** student **WHERE** sex = "f" ✓
Case #2 Goal	What is the average earnings of poker players with height higher than 200?
Question #1	What is the height of each poker player?
T5-3B & MIGA	**SELECT** t2.height **FROM** poker_player **AS** t1 **JOIN** people **AS** t2 **ON** t1.people_id = t2.people_id ✓
Question #2	Who all have heights greater than 200?
T5-3B & MIGA	**SELECT** * **FROM** poker_player **AS** t1 **JOIN** people **AS** t2 **ON** t1.people_id = t2.people_id **WHERE** t2.height > 200 ✓

图 1: T5-3B 模型在 Text-to-SQL 领域表现效果

在面对鲁棒性和泛化性上，T5-3B 模型仍然能达到不错效果。如图 2 所示[2]，T5-3B 模型在 Spider-syn、Spider-realistic 以及 Spider-dk 上表现优于近期 RAT-SQL+Grappa 等模型的效果，即使是参数量较少的 T5-large 也有较好的表现。

MODEL	SYN	DK	REALISTIC
GNN	23.6	26.0	-
IRNet	28.4	33.1	-
RAT-SQL	33.6	35.8	-
RAT-SQL + BERT	48.2	40.9	58.1
RAT-SQL + Grappa	49.1	38.5	59.3
LGESQL + ELECTRA	64.6	48.4	69.2
T5-large	53.6	40.0	58.5
GRAPHIX-T5-large	61.1 (↑ 7.5)	48.6 (↑ 8.6)	67.3 (↑ 8.8)
T5-3B	58.0	46.9	62.0
GRAPHIX-T5-3B	**66.9** (↑ 8.9)	**51.2** (↑ 4.3)	**72.4** (↑ 10.4)

图 2: T5-3B 模型在鲁棒性和泛化性上的效果

[1]图片取自多轮问题领域的 MIGA 论文

[2]图片取自 Graphix-T5 模型对应论文

2 多模态角度的改进思路

2.1 模型改进

此处借鉴模型 ALBEF 在模态融合方面的结构，在 Graphix-T5 模型上改进，模型结构分别如图 3 所示。

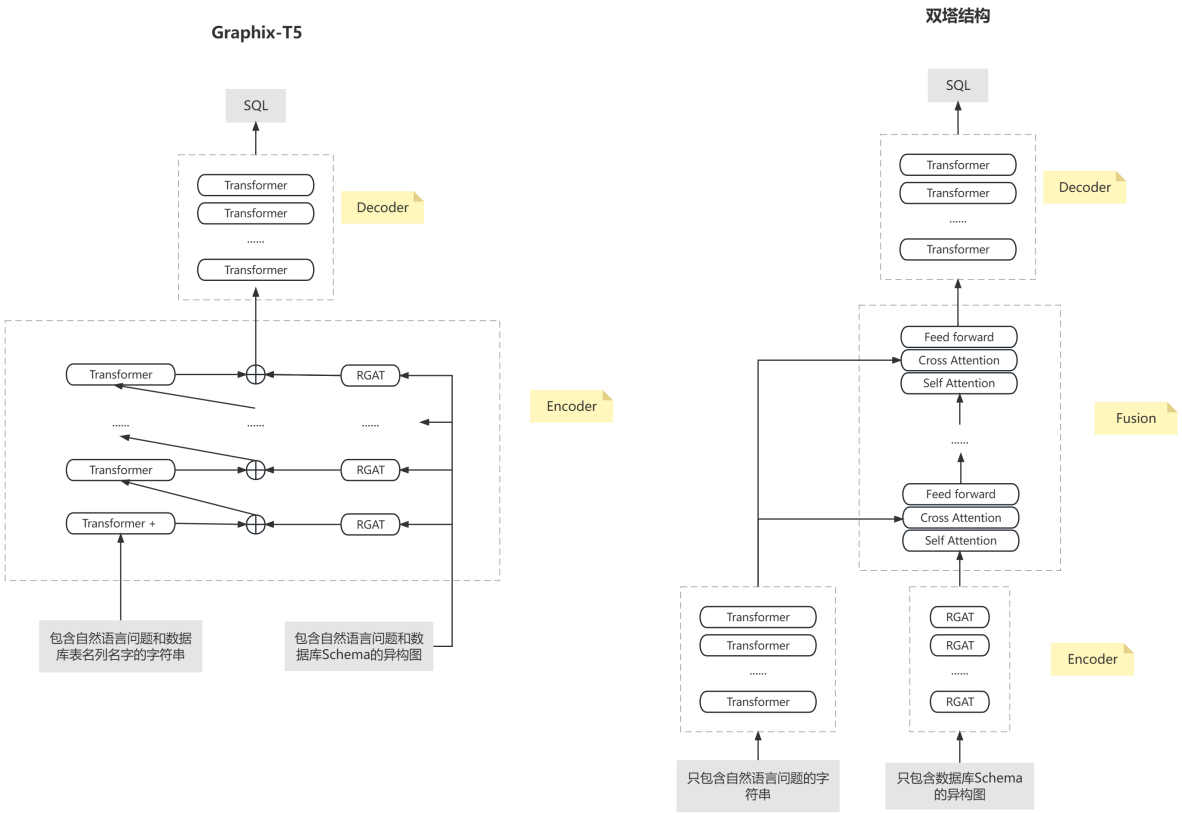

图 3: T5-3B 模型在鲁棒性和泛化性上的效果

Graphix-T5 中，Transformer 部分输入为包含自然语言问题和数据库表名列名的字符串，论文中所给的展示为：$x = \left[q_1, \ldots, q_{|Q|} \,|\mathcal{D}_{\text{name}}|\, t_1 : c_1^{t_1}, \ldots, c_{|C|}^{t_1} | \ldots |t_{|\mathcal{T}|} : c_1^{t_{|\mathcal{T}|}}, \ldots, c_{|C|}^{t_{|\mathcal{T}|}} \,|\, *\right]$，其中 q_i 是自然语言问题中第 i 个 token，t_j 代表数据库 D 的第 j 个 $table$，$c_k^{t_j}$ 代表第 j 个 $table$ 的第 k 个 $column$，$*$ 代表的是数据库中特殊的 $column\ token$。D_{name} 是数据库的名字。RGAT(关系图注意网络) 中输入为包含数据库 Schema 的异构图。图 $G =< V, R >$，其中 $V = Q \cup C \cup T$，Q, C, T 分别表示自然语言问题，数据库的列名以及表名，$R = r_1, ..., r_{|R|}$ 代表的各节点之间的关系。后续每一层加和 transformer 块的结果与 RGAT 的结果，作为下一层 Transformer 的输入，下一层 RGAT 则仍然使用初始异构图输入。Decoding 部分使用 T5 模型原始 Decoder。

在使用双塔结构改进思路中 Encoding 部分区分文本 Encoding 和 Schema Encoding。在输入方面，考虑到 Dr.Spider 以及 Spider syn 中提到的避免模型通过从自然语言问题中寻找与 Schema 匹配的单词而不是理解自然语言问题，在此文本 Encoding 方面仅输入含有自然语言问题的字符串即 $x = [q_1, \ldots, q_{|Q|}]$，Schema 数据 Encoding 方面仅输入包含 Schema 数据的异构图即 $G =< V, R >$，其中 $V = C \cup T$。除此之外，在 Schema 数据 Encoding 部分也使用多个 RGAT 块，多层解析数据库 Schema 信息。后续将两个模态信息使用 cross-attention 层实现模态融合。Decoding 部分与 Graphix-T5 模型一致，均使用 T5 模型原始 Decoder。

2.2 类似领域

在验证该思路是否会有效时，我调研到图片文本多模态领域的 VQA 问题 (Visual Question Answering) 与此处的 Text-to-SQL 领域较为相似。VQA 是指视觉问答，即给定一张图片和一个与该图片相关的自然语言问题，计算机能产生一个正确的回答。这是一个典型的多模态问题，融合了计算机视觉与自然语言处理的技术，计算机需要同时学会理解图像和文字。将此处的图片替换为数据库 Schema 就是当前的单轮 Text-to-SQL 问题。在huggingface上有如图 4 所示的样例。样例中针对问题"What's the animal doing?"，模型从图中获取信息并做出准确回答。思路中借鉴的 ALBEF 模型在 VQA 这一下游任务中表现良好，论文提出时达到 2021 年的 sota 效果，有理由认为借鉴多模态领域的模型结构能提升 Text-to-SQL 领域模型的效果。

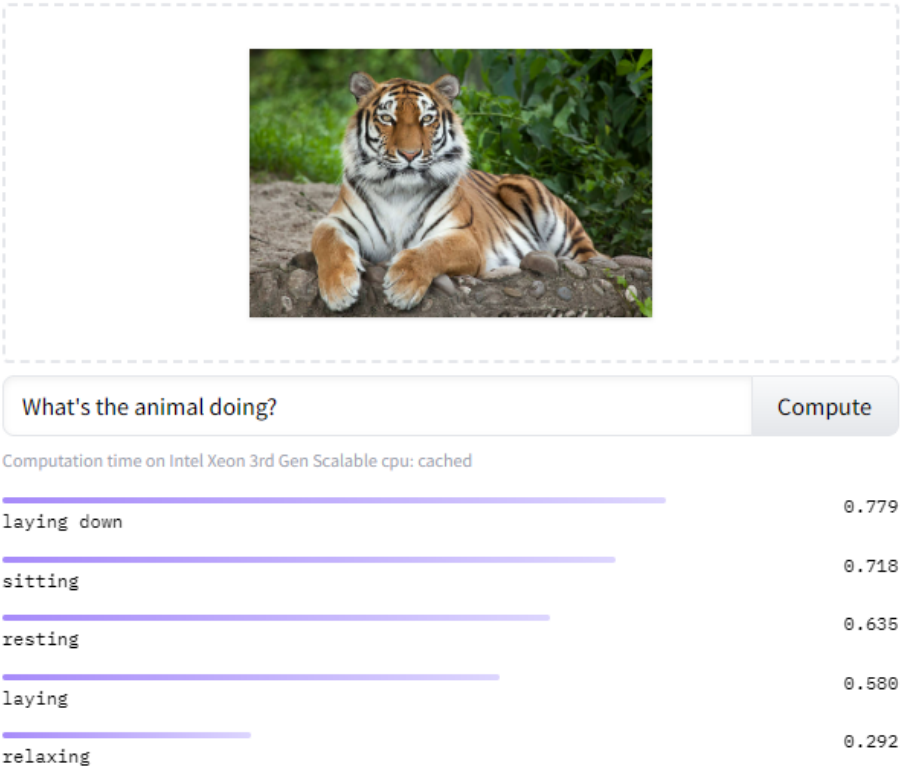

图 4: VQA 问题样例

2.3 思路补充

对于图 3 中提出的模型结构还存在可能的提升空间，一是 Encoder 中的输入，此处我将文本和数据库 Schema 两种模态完全区分，效果可能不如在文本 Encoding 中加上数据库表名列名信息，或者在 Schema 的 Encoding 中额外加上自然语言问题；二是 Schema Encoding 中或许更换其他模型如 TaBERT、或者只用一层 RGAT 等会有更好效果；三是除了模态融合 (fusion) 外，额外加上 Encoding 后的模态对齐 (align) 或是采用多模态领域最近技术等都可能对模型效果有进一步的提升。这些则需要后续实验对比才能得到答案。