EEE-QA: Exploring Effective and Efficient Question-Answer Representations

Zhanghao HU*

Yijun Yang*

Junjie Xu*

Yifu Qiu

Pinzhen(Patrick) Chen

THE UNIVERSITY of EDINBURGH informatics

*These authors contribute equally

Motivation&Contribution

THE UNIVERSITY of EDINBURGH
informatics

Knowledge base question answering:

1. ***Inference efficiency***: Scoring Q and all A in one go instead of scoring each Q-A pair

2. ***Semantic modelling:*** Using (max) pooling instead of CLS for Q-A matching

Methodology

THE UNIVERSITY of EDINBURGH informatics

Previous

Encoder() X 5 times

KG-subgraph

<s> If I am hungry, where should I go? </s> School</S>	<CLS1>	<school>	score = 0.04
<s> If I am hungry, where should I go? </s> Restaurant</S>	<CLS2>	<restaurant>	Score = 0.85
…	…	…	…
…	…	…	…
<s> If I am hungry, where should I go? </s> Library</S>	<CLS5>	<library>	Score = 0.02

Methodology

Now(ours)

<s> If I am hungry, where should I go? </s> School</S>Restaurant</S>..... Library</S>

1. Single Pass Inference with Gate
2. Pooling

Encoder() X 1 times

<Pooling1>
<Pooling2>
...
...
<Pooling5>

+

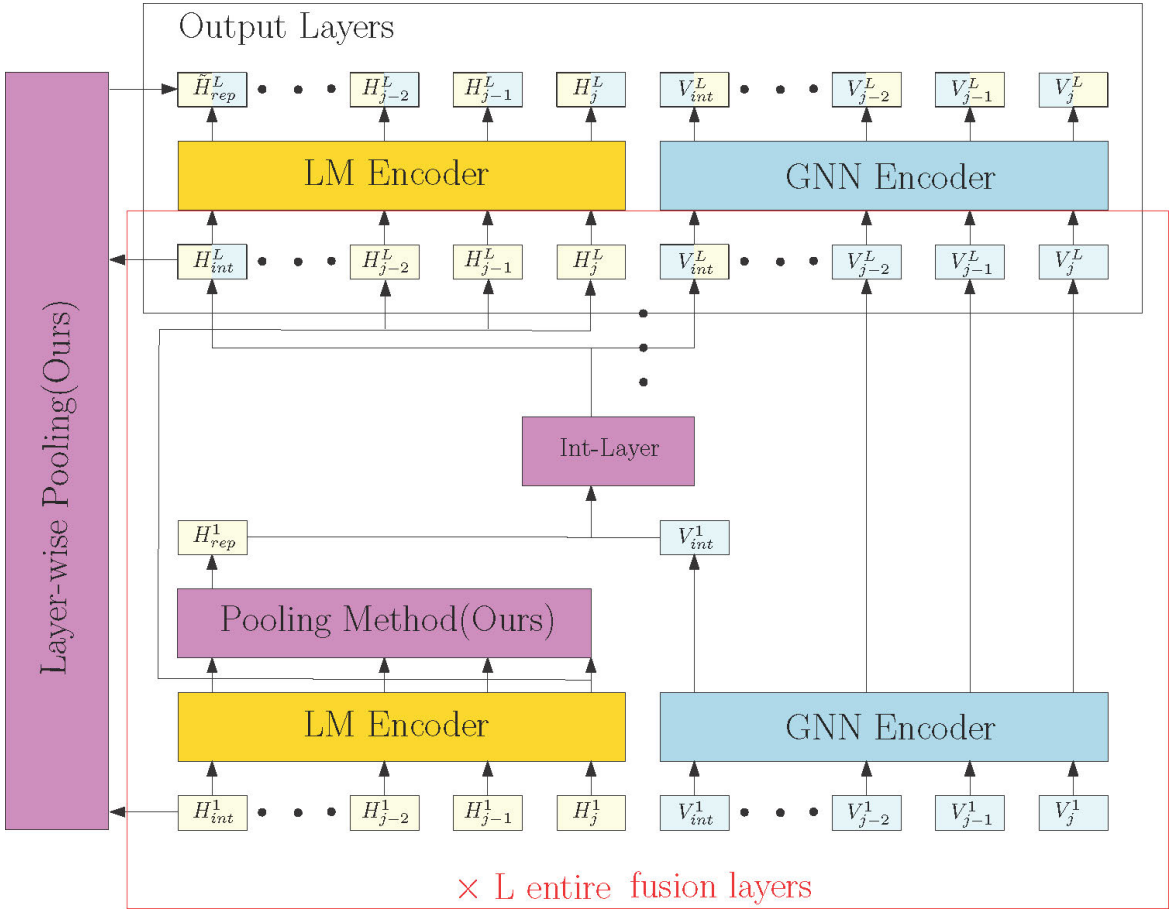
<school>
<restaurant>
...
...
<library>

KG-subgraph

=

score = 0.04
Score = 0.85
...
...
Score = 0.02

Methodology



Experiment

THE UNIVERSITY of EDINBURGH informatics

Pool Performance

System	Accuracy (std.)
RoBERTa-large (Liu et al. 2019b)	68.69 (±0.56)
QA-GNN (Yasunaga et al. 2021)	73.41 (±0.92)
JointLK (Sun et al. 2022)	74.43 (±0.83)
ACENet (Hao, Xie, and Zhang 2022)	74.72 (±0.70)
GreaseLM (Zhang et al. 2022)	74.20 (±0.40)
GreaseLM (Ye et al. (2023)'s re-run)	73.60 (unknown)
GreaseLM (our re-run)	73.57 (±0.08)
+ mean pooling	73.73 (±0.29)
+ max pooling	**75.42 (±0.52)**
+ attentive pooling	73.97 (±0.51)
+ layerwise CLS pooling	73.97 (±0.16)

Table 2: Performance (accuracy, %) of our pooling investigation compared with previous works on CommonsenseQA.

Experiment

THE UNIVERSITY of EDINBURGH informatics

Efficient inference

System	Pooling	CommonsenseQA			OpenBookQA		
		PLM	+ KG	+ KG + Int	PLM	+ KG	+ KG + Int
1AnP	CLS pool	70.02	72.82	73.97	80.20	81.80	81.60
	Max pool	70.51	73.41	75.42	82.40	82.40	82.60
nAnP	CLS pool	67.12	69.62	68.82	79.40	78.40	81.80
	Max pool	67.12	68.90	69.14	79.40	82.60	82.20
nA1P	CLS pool	67.77	69.30	69.38	78.80	79.40	80.20
	Max pool	68.25	68.65	70.91	79.00	80.60	80.40
	Max pool + Gate	69.62	71.88	70.91	79.60	80.60	81.40

Table 1: Performance (accuracy, %) of our systems on CommonsenseQA and OpenBookQA.

Experiment

THE UNIVERSITY of EDINBURGH informatics

Throughput

GPU Model	Mem. (GB)	Batch size (↑)			Infer. time (↓)		
		1AnP	nA1P	(Δ%)	1AnP	nA1P	(Δ%)
RTX A5000	24	100	160	(+60%)	4.6s	3.3s	(-28%)
RTX 3090	24	100	160	(+60%)	2.45s	1.82s	(-26%)
RTX 2080 Ti	11	30	45	(+50%)	1.13s	0.76s	(-33%)
GTX 1080 Ti	11	30	45	(+50%)	2.64s	1.27s	(-52%)
Titan X Pascal	12	40	55	(+38%)	3.56s	1.55s	(-56%)
GTX 1080	8	10	20	(+100%)	2.21s	0.77s	(-65%)

Table 3: An efficiency comparison between 1AnP and our nA1P: usable batch size and total inference time when solving 1000 QA on a single Nvidia GPU.

Conclusion

1. ***Pooling*** has better representations than [CLS].

2. We propose a ***gated single-pass inference*** approach to encourage ***answer inter-actions*** and ***enhance efficiency***.

3. Experiments demonstrate (1) ***substantial gains with max pooling***, surpassing state-of-the-art KBQA models. (2) We ***maintains a similar performance to the baseline while incurring less computation by 28-65%***.