GreaseLM

HTET ARKAR

Junior, Undergraduate

School of Computer Science and Engineering

Chung-Ang University

16-Jan-25

Graph REASoning Enhanced Language Models for Question Answering (GreaseLM)

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren,
Percy Liang, Christopher D. Manning, Jure Leskovec
Stanford University

ICLR 2022

Contents

- Introduction
- Previous Works
- Proposed Method
- Experiments
- Conclusion

Introduction

❖ Question Answering Task

- A fundamental task
- Demand intricate reasoning and understanding comprehension skills
- To interpret the text and provide appropriate responses to posed question

❖ Multiple Choice QA Systems

- A question q
- A set of answer options A
- Optional context c

Where does a child likely sit at a desk?

A. Schoolroom * B. Furniture store C. Patio
D. Office building E. Library

- Task: selecting the best option to answer the question

Introduction

- **MCQA Task**
  - A challenging task that requires complex reasoning
    - Explicit constraints described in the textual context of the question
    - Unstated, relevant knowledge about the world
- **Complex Reasoning** (Zhang, Xikun, et al.)
  - The number of prepositional phrases in the questions
    - To easier burrow, after prey, in a what
  - The presence of negation terms
    - e.g. no, never
  - The presence of hedging terms
    - Terms indicating uncertainty (e.g. sometimes; maybe)

CommonsenseQA	A weasel has a thin body and short legs to easier burrow after prey in a what? (A) tree (B) mulberry bush (C) chicken coop (D) viking ship **(E) rabbit warren**

MCQA Task

❖ Foundation of most modern QA systems

- Large pre-trained language models fine-tuned on QA datasets
- Learn to implicitly encode broad knowledge about world
  - ➔ Able to leverage when fine-tuned on a domain-specific downstream task

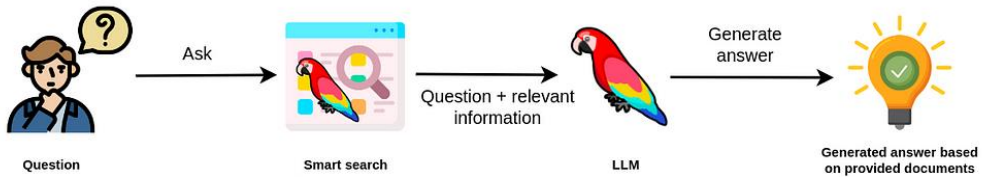


Image credit: neo4j

MCQA Task

DMAIS

❖ Challenges using LMs

- Lack of **Structured Reasoning**
  - LLMs primarily rely on patterns learned from large-scale pretraining
  - Not robustly represent latent relationships between concepts (which is necessary for reasoning)

Where does a child likely sit at a desk?

A. Schoolroom * B. Furniture store C. Patio
D. Office building E. Library

If it is not used for **hair**, a **round brush** is an example of what?

A. hair brush B. bathroom C. **art supplies***
D. shower E. hair salon

QA context

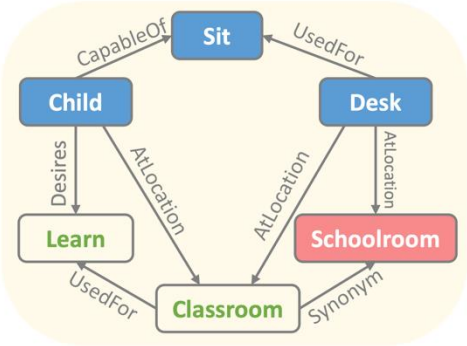
MCQA Task

❖ Massive Knowledge Graphs (KG)

- More suited for structured reasoning
- Enable explainable predictions e.g. by providing reasoning paths (KagNet)
- Capture external knowledge explicitly using triplets

Where does a child likely sit at a desk?

A. Schoolroom * B. Furniture store C. Patio
D. Office building E. Library



Lin, Bill Yuchen, et al. "Kagnet: Knowledge-aware graph networks for commonsense reasoning." *arXiv preprint arXiv:1909.02151* (2019).
Feng, Yanlin, et al. "Scalable multi-hop relational reasoning for knowledge-aware question answering." arXiv preprint arXiv:2005.00646 (2020).

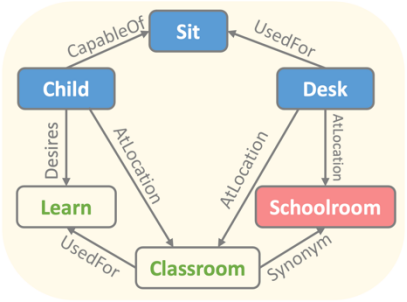
MCQA Task

Reasoning

- Relational reasoning over entities (concepts) and their relationships by referencing external knowledge

Where does a child likely sit at a desk?

A. Schoolroom * B. Furniture store C. Patio
D. Office building E. Library



- Leveraging KGs into QA systems
  - Represent relational knowledge between entities with multi-relational edges for models to acquire
  - Bring the potential of interpretable and trustworthy predictions
  - E.g., relational path:
    - **CHILD** -> AtLocation -> CLASSROOM -> Synonym -> **SCHOOLROOM**

Feng, Yanlin, et al. "Scalable multi-hop relational reasoning for knowledge-aware question answering." arXiv preprint arXiv:2005.00646 (2020).

Previous Works

- **Knowledge-aware QA Framework**

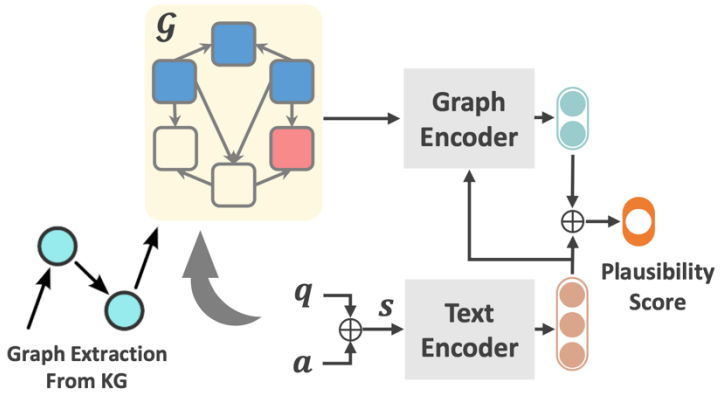
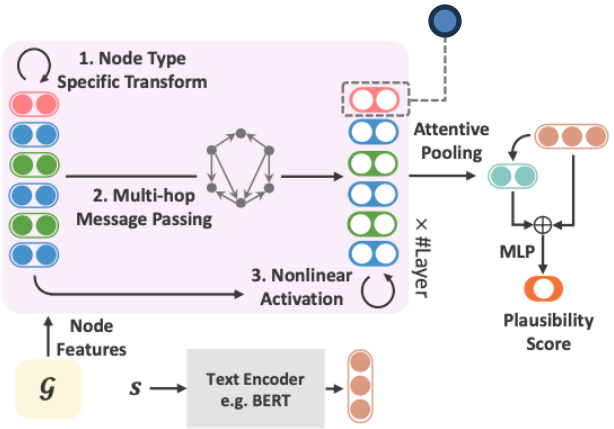


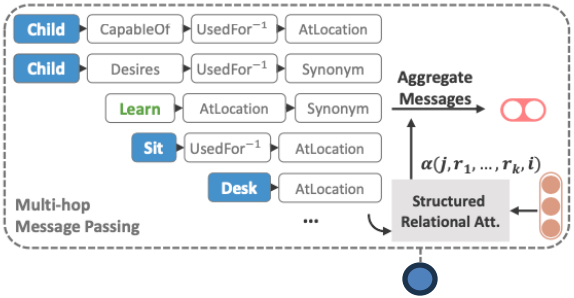
Figure 3: **Overview of the knowledge-aware QA framework.** It integrates the output from graph encoder (for relational reasoning over contextual subgraphs) and text encoder (for textual understanding) to generate the plausibility score for an answer option.

Feng, Yanlin, et al. "Scalable multi-hop relational reasoning for knowledge-aware question answering." arXiv preprint arXiv:2005.00646 (2020).

MHGRN

❖ Multi-Hop Graph Relation Network





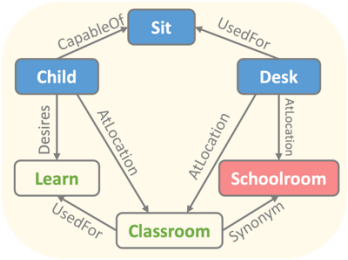


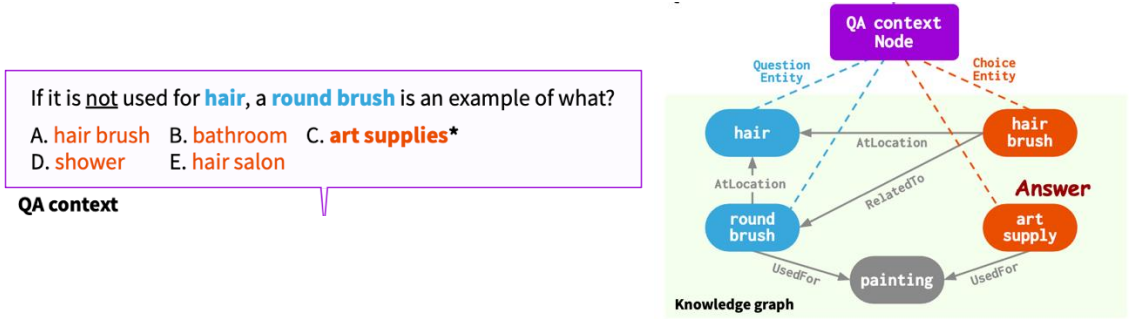
Figure 4: **Our proposed MHGRN architecture for relational reasoning.** MHGRN takes a multi-relational graph $\mathcal{G}$ and a (question-answer) statement vector $\boldsymbol{s}$ as input, and outputs a scalar that represent the plausibility score of this statement.

Feng, Yanlin, et al. "Scalable multi-hop relational reasoning for knowledge-aware question answering." arXiv preprint arXiv:2005.00646 (2020).

QA-GNN

❖ Challenges in QA Sys (LMs + KGs)

- ➢ To identify relevant knowledge from large KGs
- ■ Introducing many entities that are semantically irrelevant to QA context in subgraph
  - □ Retrieving subgraph by taking topic entities and their few-hop neighbors
- ➢ Perform joint reasoning over the QA context and KG



Yasunaga, Michihiro, et al. "QA-GNN: Reasoning with language models and knowledge graphs for question answering." *arXiv preprint arXiv:2104.06378* (2021).

QA-GNN

❖ Solution

- Relevance scoring
  - Using LM to estimate the importance of KG nodes relative to the given QA context
- Joint reasoning
  - Connecting QA context and KG to form a joint graph
  - Mutually updating their representations through GNNs

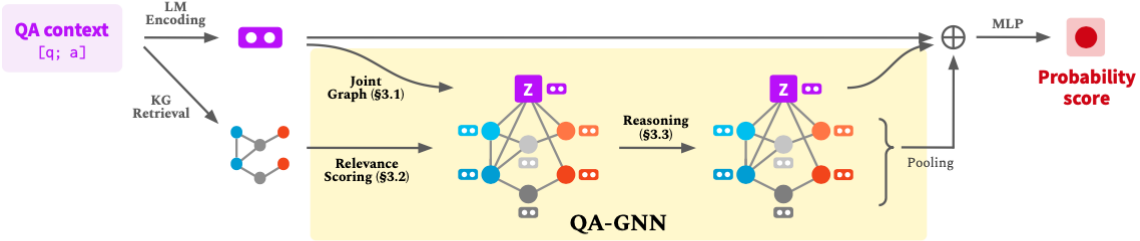


Figure 2: Overview of our approach. Given a QA context (z), we connect it with the retrieved KG to form a joint graph (*working graph*; §3.1), compute the relevance of each KG node conditioned on z (§3.2; node shading indicates the relevance score), and perform reasoning on the working graph (§3.3).

Yasunaga, Michihiro, et al. "QA-GNN: Reasoning with language models and knowledge graphs for question answering." *arXiv preprint arXiv:2104.06378* (2021).

QA-GNN

❖ KG node relevance scoring

- Irrelevant nodes may result in overfitting or introduced unnecessary difficulty in reasoning, especially sub-graph is large

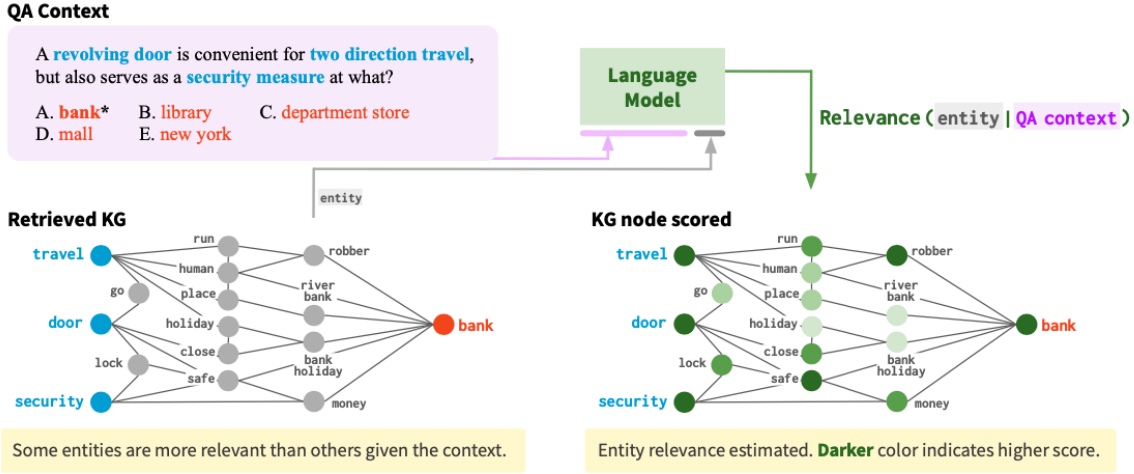


Figure 3: Relevance scoring of the retrieved KG: we use a pre-trained LM to calculate the relevance of each KG entity node conditioned on the QA context (§3.2).

QA-GNN

❖ Behavior for Structured Reasoning

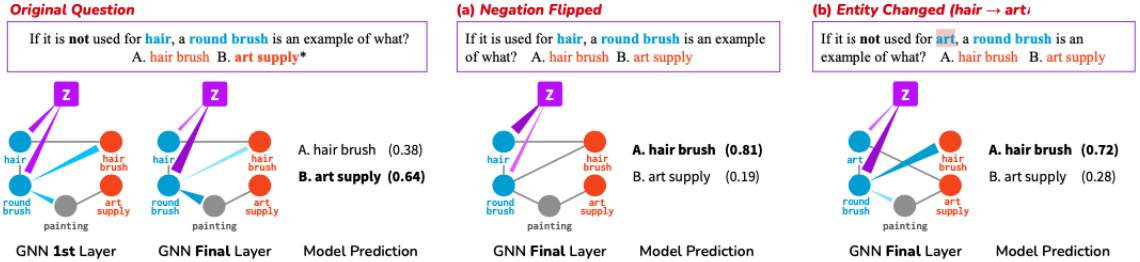


Figure 5: **Analysis of QA-GNN's behavior for structured reasoning**. Given an original question (left), we modify its negation (middle) or topic entity (right): we find that QA-GNN adapts attention weights and final predictions accordingly, suggesting its capability to handle structured reasoning.

Example (Original taken from *CommonsenseQA* Dev)	**RoBERTa Prediction**	**Our Prediction**
[Original] If it is **not** used for hair, a round brush is an example of what? A. hair brush B. art supply	A. hair brush (✗)	B. art supply (✓)
[Negation flip] If it is used for hair, a round brush is an example of what?	A. hair brush (✓ just no change?)	A. hair brush (✓)
[Entity change] If it is not used for art a round brush is an example of what?	A. hair brush (✓ just no change?)	A. hair brush (✓)

Yasunaga, Michihiro, et al. "QA-GNN: Reasoning with language models and knowledge graphs for question answering." *arXiv preprint arXiv:2104.06378* (2021).

Proposed Method

GreaseLM

DMAIS

❖ LMs' Challenges in QA Systems

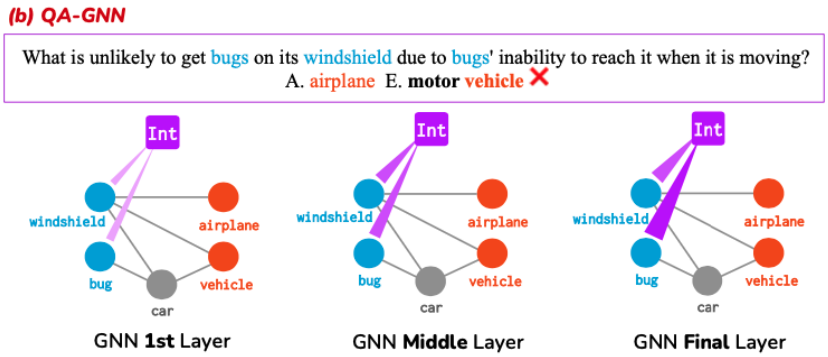
- Struggles with Distributional Shifts
  - LLMs often underperform on examples that are **distributionally different** from those seen during training or fine-tuning
  - Especially on unfamiliar domains, OR questions requiring complex, multi-hop reasoning or nuanced context interpretation
  - Learned behavior relies on simple pattern to offer shortcuts to an answer

What is unlikely to get bugs on its windshield due to bugs' inability to reach it when it is moving?
A. airplane E. **motor vehicle** ✗

GreaseLM

❖ KG for language modeling

- Require finding the right integration of knowledge from KG with the **information and constraints** provided by the QA example
- Leveraging both modalities: Expressive LLMs and structured KGs
- **Fusing in a shallow and non-interactive manner**
  - Encoding both separately and fusing them at the output for a prediction
  - Using one to augment the input of the other



Lin, Bill Yuchen, et al. "Kagnet: Knowledge-aware graph networks for commonsense reasoning." *arXiv preprint arXiv:1909.02151* (2019).
Feng, Yanlin, et al. "Scalable multi-hop relational reasoning for knowledge-aware question answering." arXiv preprint arXiv:2005.00646 (2020).

GreaseLM

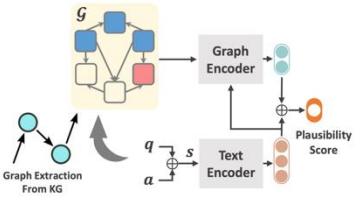
A new model that enables fusion and exchange of information from both the LM and KG in multiple layers of its architecture

GreaseLM

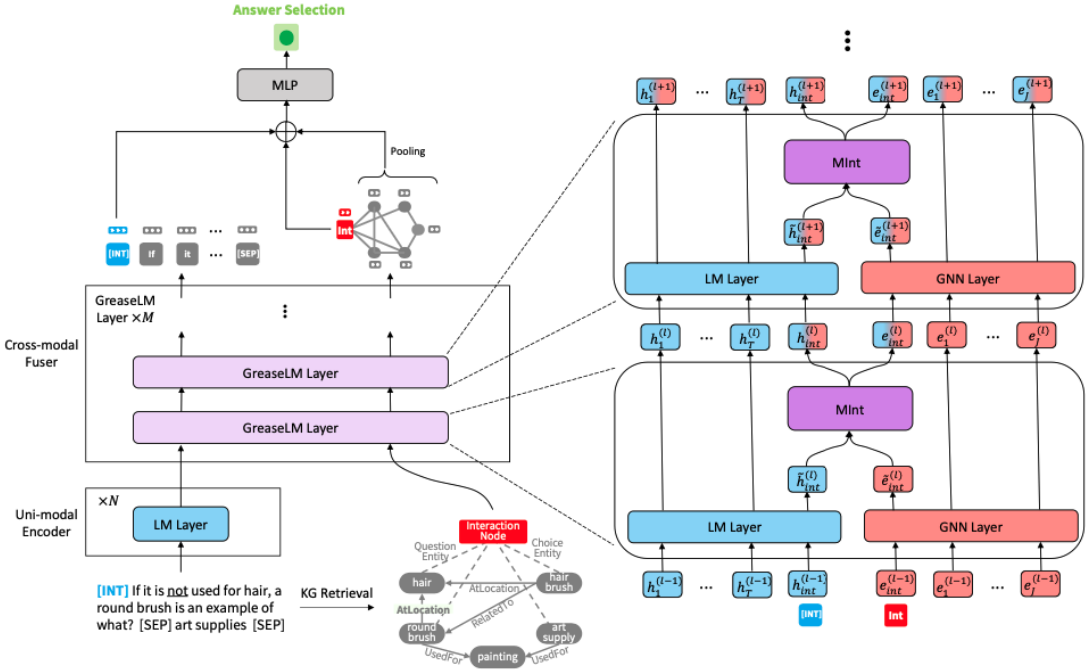
- **Given a QA example (c, q, A) and KG**
- **Input Representation**
  - QA context : [c;q;a] and tokenize the combined sequence
  - KG Retrieval
    - A set of **matched entities** via entity liking to match the tokens in QA context to the entities in KG (using ConceptNet API)
    - Add any **bridge entities** that are in a 2-hop path between any paired of linked entities to get a set of retrieved entities
    - Prune the set of retrieved nodes using **relevance score**
    - Retrieve all the **edges** that connect any two nodes in V_{sub}
    - Form the retrieved **subgraph** G_{sub}

Proposed Method

GreaseLM



❖ Architecture



GreaseLM

❖ Language Pre-encoding

- Using a single layer

$$\{\boldsymbol{h}_{int}^{(\ell)}, \boldsymbol{h}_1^{(\ell)}, \ldots, \boldsymbol{h}_T^{(\ell)}\} = \text{LM-Layer}(\{\boldsymbol{h}_{int}^{(\ell-1)}, \boldsymbol{h}_1^{(\ell-1)}, \ldots, \boldsymbol{h}_T^{(\ell-1)}\}) \quad \text{for } \ell = 1, \ldots, N$$

❖ GreaseLM layer

- Language Representation

$$\{\tilde{\boldsymbol{h}}_{int}^{(N+\ell)}, \tilde{\boldsymbol{h}}_1^{(N+\ell)}, \ldots, \tilde{\boldsymbol{h}}_T^{(N+\ell)}\} = \text{LM-Layer}(\{\boldsymbol{h}_{int}^{(N+\ell-1)}, \boldsymbol{h}_1^{(N+\ell-1)}, \ldots, \boldsymbol{h}_T^{(N+\ell-1)}\}) \quad \text{for } \ell = 1, \ldots, M$$

GreaseLM

❖ GreaseLM layer

- Graph Representation
  - e^0_{int} : initialized randomly
  - Fed into the layer to perform a round of information propagation between nodes in the graph

$$\{\tilde{\boldsymbol{e}}^{(\ell)}_{int}, \tilde{\boldsymbol{e}}^{(\ell)}_1, \ldots, \tilde{\boldsymbol{e}}^{(\ell)}_J\} = \text{GNN}(\{\boldsymbol{e}^{(\ell-1)}_{int}, \boldsymbol{e}^{(\ell-1)}_1, \ldots, \boldsymbol{e}^{(\ell-1)}_J\})$$
$$\text{for } \ell = 1, \ldots, M$$

$$\tilde{\boldsymbol{e}}^{(\ell)}_j = f_n\left(\sum_{e_s \in \mathcal{N}_{e_j} \cup \{e_j\}} \alpha_{sj} \boldsymbol{m}_{sj}\right) + \boldsymbol{e}^{(\ell-1)}_j$$

$$\boldsymbol{r}_{sj} = f_r(\tilde{\boldsymbol{r}}_{sj}, \boldsymbol{u}_s, \boldsymbol{u}_j)$$
$$\boldsymbol{m}_{sj} = f_m(\boldsymbol{e}^{(\ell-1)}_s, \boldsymbol{u}_s, \boldsymbol{r}_{sj})$$
$$\boldsymbol{q}_s = f_q(\boldsymbol{e}^{(\ell-1)}_s, \boldsymbol{u}_s)$$
$$\boldsymbol{k}_j = f_k(\boldsymbol{e}^{(\ell-1)}_j, \boldsymbol{u}_j, \boldsymbol{r}_{sj})$$

$$\gamma_{sj} = \frac{\boldsymbol{q}_s^\top \boldsymbol{k}_j}{\sqrt{D}}$$

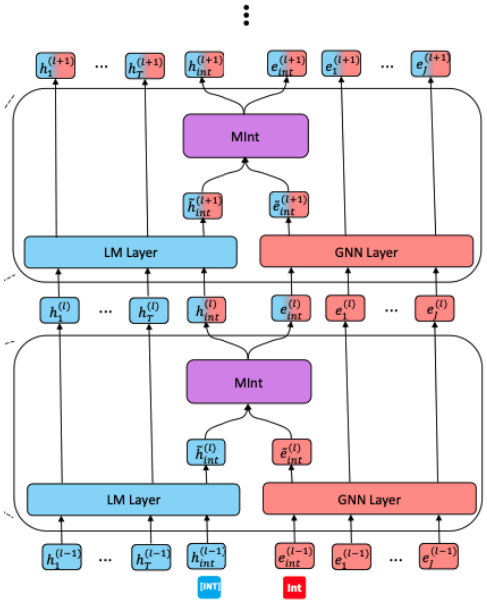
$$\alpha_{sj} = \frac{\exp(\gamma_{sj})}{\sum_{e_s \in \mathcal{N}_{e_j} \cup \{e_j\}} \exp(\gamma_{sj})}$$

GreaseLM

❖ Modality Interaction

- Mixing operation **MInt**

$$[\boldsymbol{h}_{int}^{(\ell)}; \boldsymbol{e}_{int}^{(\ell)}] = \text{MInt}([\tilde{\boldsymbol{h}}_{int}^{(\ell)}; \tilde{\boldsymbol{e}}_{int}^{(\ell)}]),$$



❖ Learning & Inference

- Compute the probability of $\boldsymbol{a}$ being the correct answer

$$p(a \mid q, c) \propto \exp(\text{MLP}(\boldsymbol{h}_{int}^{(N+M)}, \boldsymbol{e}_{int}^{(M)}, \boldsymbol{g}))$$

- Inference time : $\arg\max_{a \in \mathcal{A}} p(a \mid q, c).$

Proposed Method

GreaseLM

Table 2: **Performance comparison on *CommonsenseQA* in-house split** (controlled experiments). As the official test is hidden, here we report the in-house Dev (IHdev) and Test (IHtest) accuracy, following the data split of Lin et al. (2019). Experiments are controlled using same seed LM.

Methods	IHdev-Acc. (%)	IHtest-Acc. (%)
RoBERTa-Large (w/o KG)	73.1 (±0.5)	68.7 (±0.6)
RGCN (Schlichtkrull et al., 2018)	72.7 (±0.2)	68.4 (±0.7)
GconAttn (Wang et al., 2019)	72.6 (±0.4)	68.6 (±1.0)
KagNet (Lin et al., 2019)	73.5 (±0.2)	69.0 (±0.8)
RN (Santoro et al., 2017)	74.6 (±0.9)	69.1 (±0.2)
MHGRN (Feng et al., 2020)	74.5 (±0.1)	71.1 (±0.8)
QA-GNN (Yasunaga et al., 2021)	76.5 (±0.2)	73.4 (±0.9)
GREASELM (**Ours**)	**78.5** (±0.5)	**74.2** (±0.4)

Table 3: **Test Accuracy comparison on *OpenBookQA***. Experiments are controlled using the same seed LM for all LM+KG methods.

Model	Acc.
AristoRoBERTa (no KG)	78.4
+ RGCN	74.6
+ GconAttn	71.8
+ RN	75.4
+ MHGRN	80.6
+ QA-GNN	82.8
GREASELM (**Ours**)	**84.8**

Table 4: **Test accuracy comparison to public *OpenBookQA* model implementations**. *UnifiedQA (11B params) and T5 (3B) are 30x and 8x larger than our model.

Model	Acc.	# Params
ALBERT (Lan et al., 2020) + KB	81.0	~235M
HGN (Yan et al., 2020)	81.4	≥355M
AMR-SG (Xu et al., 2021)	81.6	~361M
ALBERT + KPG (Wang et al., 2020)	81.8	≥235M
QA-GNN (Yasunaga et al., 2021)	82.8	~360M
T5* (Raffel et al., 2020)	83.2	**~3B**
T5 + KB (Pirtoaca)	85.4	**≥11B**
UnifiedQA* (Khashabi et al., 2020)	**87.2**	**~11B**
GREASELM (**Ours**)	84.8	~359M

GreaseLM

❖ Quantitative Analysis

- GREASELM performs better than the baselines across all questions with prepositional phrases
- Perform comparably on questions with no prepositional phrases (QA-GNN)
- Increasing complexity of questions requires deeper cross-modal fusion between language and knowledge representations

Table 5: Performance of GREASELM on the *CommonsenseQA* IH-dev set on complex questions with semantic nuance such as prepositional phrases, negation terms, and hedge terms.

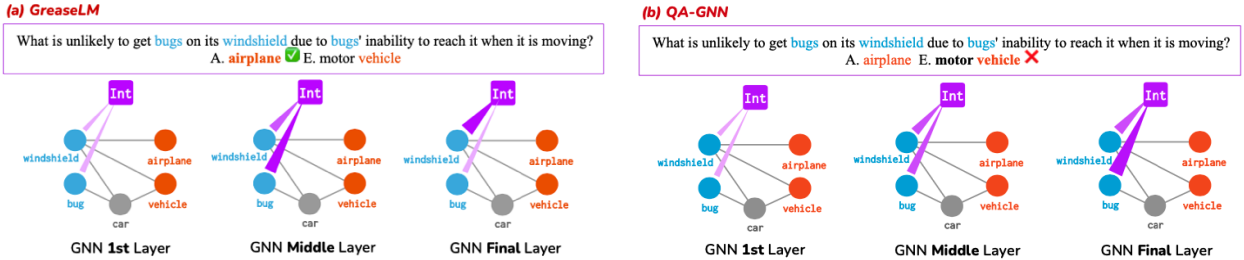
Model	# Prepositional Phrases					Negation Term	Hedge Term
	0	1	2	3	4		
n	210	429	316	171	59	83	167
RoBERTa-Large	66.7	72.3	76.3	74.3	69.5	63.8	70.7
QA-GNN	**76.7**	76.2	79.1	74.9	81.4	66.2	76.0
GREASELM (**Ours**)	75.7	**79.3**	**80.4**	**77.2**	**84.7**	**69.9**	**78.4**

GreaseLM

❖ Qualitative Analysis

- Examine node-to-node attention weights induced by GNN layers
- Analyze whether reflecting more expressive reasoning

➢ As “bug” is mentioned multiple times in the context, it may be well-represented in QA-GNN’s context node initialization, which is never reformulated by language representations, unlike in GREASELM



Future Works

❖ Ethics Statement

- GreaseLM: a method to fuse the (LM+KG) representations for effective reasoning about textual reasoning
  - ➢ Reflect many of the same biases and toxic behaviors exhibited by LMs and KGs that are used to initialize it
  - ➢ e.g. biases about race, gender, and other demographic attributes

- KG: encode stereotypes rather than completely clean commonsense knowledge
  - ➢ Unethical relationships in its relationships in its knowledge resource to arrive at conclusions

Conclusion

- **GreaseLM:** a new model that enables interactive fusion through joint information exchange between knowledge from language models and knowledge graphs
- Experimental results demonstrate superior performance compared to prior KG+LM and LM-only baselines
- Improved capability modeling questions exhibiting textual nuances, such as negation and hedging
- Ethical Concerns

Thank You!

HTET ARKAR
hak3601@cau.ac.kr