ACENet: Attention Guided Commonsense Reasoning on Hybrid Knowledge Graph

Chuzhan Hao, Minghui Xie, and Peng Zhang*
College of Intelligence and Computing, Tianjin University
{chuzhanhao, minghuixie, pzhang}@tju.edu.cn

Abstract

Augmenting pre-trained language models (PLMs) with knowledge graphs (KGs) has demonstrated superior performance on commonsense reasoning. Given a commonsense based QA context (question and multiple choices), existing approaches usually estimate the plausibility of candidate choices separately based on their respective retrieved KGs, without considering the interference among different choices. In this paper, we propose an Attention guided Commonsense rEasoning Network (ACENet)[1] to endow the neural network with the capability of integrating hybrid knowledge. Specifically, our model applies the multi-layer interaction of answer choices to continually strengthen correct choice information and guide the message passing of GNN. In addition, we also design a mix attention mechanism of nodes and edges to iteratively select supporting evidence on hybrid knowledge graph. Experimental results demonstrate the effectiveness of our proposed model through considerable performance gains across CommonsenseQA and OpenbookQA datasets.

1 Introduction

Commonsense question answering (CSQA) aims to answer questions based on the understanding of context and some background knowledge, which is the critical gap between the human intelligence and machine intelligence (Talmor et al., 2019). This capability of owning prior knowledge and reasoning is a foundation for communication and interaction with the world. Therefore, commonsense reasoning has become an important research task with various datasets and models proposed in this field (Mihaylov et al., 2018; Talmor et al., 2019; Bhagavatula et al., 2020; Feng et al., 2020; Yasunaga et al., 2021; Zhang et al., 2022).

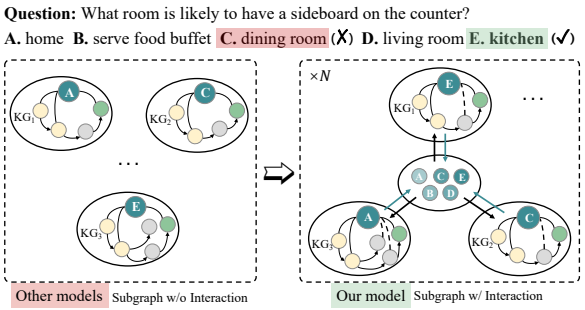

Figure 1: Through the interaction between subgraphs, the correct choice information is continuously reinforced. The subgraph is retrieved from ConceptNet (Speer et al., 2017). The nodes with letter are the q-c pairs and connect to other nodes of their respective subgraphs. Yellow nodes correspond to entities mentioned in the question, green nodes correspond to those in the answer. The other nodes are some associated entities introduced when extracting the subgraph.

Recently, PLMs (Devlin et al., 2019) have made significant progress in many question answering tasks because of its powerful representation capability. Nevertheless, since commonsense knowledge is rarely stated by natural language (Gunning, 2018), this makes it hard for PLMs to learn commonsense knowledge from pre-training corpus. Therefore, many CSQA models augment the PLMs with various external knowledge sources (e.g., structured knowledge ConceptNet (Speer et al., 2017) and unstructured knowledge Wikipedia). Compared with unstructured knowledge, structured knowledge sources have the advantage of being easier to train and recover explicit evidence, which leads many researchers to leverage KGs to reason.

A straightforward approach to leverage a KG is to directly model these relational paths (Santoro et al., 2017; Lin et al., 2019; Feng et al., 2020). Although path-based models have a strong interpretability, they are easily affected by the sparsity and scale of KGs. In addition, graph neural networks (GNNs) have achieved promising perfor-

*Corresponding author.
[1] https://github.com/HAOChuzhan/ACENet.

mance on modeling KGs. Hence, GNNs are widely used to implicitly capture commonsense knowledge from KGs (Feng et al., 2020; Yan et al., 2021; Yasunaga et al., 2021; Zhang et al., 2022).

However, these approaches have two main issues. First, they lack consideration of the interference effects between choices. In common KG-augmented models, the probability scores of candidate choices are calculated based on their respective reasoning subgraphs or paths separately, which is difficult to capture the nuance between the correct choice and distractors in commonsense questions. Second, the retrieved KGs contain a lot of noisy knowledge, which will mislead reasoning. QAGNN (Yasunaga et al., 2021) and JointLK (Sun et al., 2022) usually filter out the noise knowledge based on node features, but ignore the different significance of various edges which contain rich semantics. Wang et al. (2021) also proves the importance of edge features for commonsense reasoning. Therefore, we should capture the important features from many aspects (e.g., node, edge, graph and QA context).

In response, we propose ACENet to capture the nuance of multiple choices by integrating the QA context and the external commonsense knowledge graphs. Given a QA context and multiple retrieved subgraphs of choices, we encode each q-c pair using PLM. Then the q-c pair is introduced into respective subgraphs as a global node (Ying et al., 2021). Knowledge is transmitted between subgraphs to construct a complete hybrid knowledge graph for reasoning (see § 3.2). First, we apply *knowledge interaction layer* to carry out the information interaction between subgraphs and guide GNN message passing. The layer is stacked to form a hierarchy that enables multi-layer interactions to recursively reinforce the important choice information in message passing (see Figure 1). Additionally, in order to further aggregate key features in the reasoning graph, we design a *mix attention mechanism of nodes and edges* to iteratively select supporting evidence based on the global node. Our model simultaneously leverage the hybrid knowledge of PLM, KGs and different choices to augment the commonsense reasoning ability. In summary, our contributions are as follows:

- We propose a knowledge interaction layer to fuse the knowledge of PLM and different choices. The multi-layer interactions continuously strengthen correct choice information in the hybrid knowledge graph.
- We design a mix attention mechanism of nodes and edges to iteratively select relevant knowledge over multiple layers of GNN. The global information of q-c pair is also introduced to enhance evidence selection.
- Experimental results show that ACENet is superior to current KG-augmented methods. Through multi-layer interactions and multi-head attention guidance over hybrid knowledge graph, ACENet exhibits stronger performance in complex reasoning, such as solving questions with negation or more prepositions.

2 Related Work

Graph Neural Networks (GNNs). GNNs have been widely used to model knowledge graph due to its strong ability to process graph structured data. GNNs often follow a neighborhood aggregation and then message passing scheme (Gilmer et al., 2017). Recently, a lot of works on CSQA use GNN to model external KGs. MHGRN (Feng et al., 2020) transforms single-hop propagation into multi-hop propagation based on RGCN (Schlichtkrull et al., 2018). But it does not take into account the different importance of various nodes. QAGNN (Yasunaga et al., 2021), GreaseLM (Zhang et al., 2022), JointLK (Sun et al., 2022) use Graph Attention Network (GAT) (Velickovic et al., 2018) to represent knowledge graph. GAT is a commonly used variant of GNN, which performs attention-based message passing of node features. According to GSC (Wang et al., 2021), edge features play an essential role for commonsense reasoning. Hence, we design a mix attention mechanism of nodes and edges based on GAT.

Question Answering with LM+KG. Although pre-trained language models have achieved great success in many NLP domains, they do not perform well on reasoning questions yet. Therefore, many works propose LM+KG methods for CSQA, which use knowledge graph as external knowledge source for PLMs. JAKET (Yu et al., 2020) aligns the entities and relations between questions and knowledge graph and fuses the two kind of representations. QAGNN (Yasunaga et al., 2021) introduces a context node as the bridge of PLMs and knowledge graph. The context node is initialized with the encoding of PLM. GreaseLM (Zhang et al., 2022) designs an interactive scheme to bidirectionally transfer the information from both the LM and KG in multiple layers. JointLK (Sun et al.,

2022) calculates the fine-grained attention weight between each question token and each KG node to strengthen the joint reasoning ability. They all focus on enhancing the fusion of two knowledge source, but lack consideration for the interference effects of different choices in QA context.

3 Methodology

The diagram of the proposed ACENet is shown in Figure 2. We assume a setting where each example in our data set contains a question q and a set of answer choices $\{c_1, c_2, ..., c_n\}$. We derive the gold answer from QA context and relevant commonsense knowledge. Therefore, we retrieve a KG $\mathcal{G}$ as the source of commonsense knowledge following prior work (Feng et al., 2020).

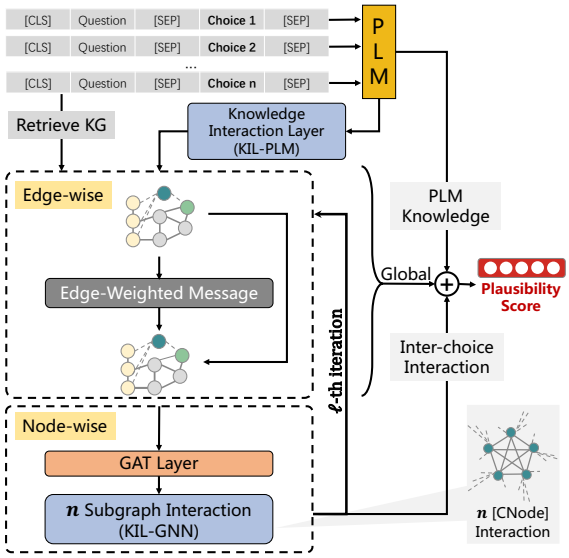

Figure 2: Overall architecture of our proposed ACENet.

3.1 Knowledge Interaction Layer

As shown in Figure 2, given a question and n answer choices, we concatenate them to get n q-c pair $[q; c_i]$ $(i \in [1, n])$ separately. For each q-c pair, they will be as the inputs to feed through PLM. We use the "[CLS]" token output from PLM as a summary vector for each choice.

Although PLMs can learn the general language representation well (Qiu et al., 2020) for each choice, it encodes each q-c pair separately, without considering inter-choice interference effects that are essential for the downstream commonsense question answering task. Our model begins to use the representation of each q-c pair to integrate external commonsense knowledge in respective subgraphs (see Figure 3). How to initialize the summary representation of each choice is crucial in minimizing distracting information being passed to the downstream supporting evidence selection and answer prediction tasks.

Therefore, we propose a knowledge interaction layer (*KIL* shown in Figure 3) to strengthen the correct choice information. First we add a multi-head attention (Vaswani et al., 2017) KIL on top of the "CLS" tokens. This layer is defined as:

$$\boldsymbol{\alpha}_{ij} = \mathrm{MHA}(\boldsymbol{Q}^t, \boldsymbol{K}^t, \boldsymbol{V}^t) \quad (1)$$

$$\boldsymbol{H}^t = \boldsymbol{\eta} \odot \tilde{\boldsymbol{H}}^t + (1 - \boldsymbol{\eta}) \odot (\boldsymbol{\alpha}_{ij} \boldsymbol{V}^t) \quad (2)$$

where Q,K,V are interactive representations of all q-c pairs, which are linear projections from stacked embeddings of q-c pairs. MHA is the multi-head attention mechanism. α_{ij} is the attention weight between choices. $\eta = \sigma(\tilde{H}^t W + b)$, σ denotes the sigmoid activation function, $\odot$ represents the element-wise product, $\tilde{H}^t$ is the choice representations before passing through the t-th KIL layer. Our motivation for adding attention across the q-c pairs generated from different choices is to encourage inter-choice interactions. By allowing choice representations to interact with each other, the model is able to train on a better input signal for message aggregation and passing.

3.2 Hybrid Knowledge Graph

To unify the knowledge of PLM and KGs into the same reasoning space and take advantage of both, we introduce the q-c pair into the extracted subgraphs $\mathcal{G}_i$. Inspired by Gilmer et al. (2017) and Yasunaga et al. (2021), in hybrid knowledge graph, we add the q-c pair as a special node called [CNode] to the $\mathcal{G}_i$, and make connection between [CNode] and each node individually. Each node in the $\mathcal{G}_i$ is divided into four types based on information sources: q-C node, Question entity node, Answer entity node and Retrieved entity node, referred to as $\mathcal{T} = \{\mathcal{C}, \mathcal{Q}, \mathcal{A}, \mathcal{R}\}$.

To further leverage the interference effects of different choices, the [CNode] node replaces various graph pooling functions to represent global information for each subgraph $\mathcal{G}_i$. In the BERT model (Devlin et al., 2019), there is a similar token, i.e., [CLS], which is a special token attached at the beginning of each sequence, to represent the sequence-level feature on downstream tasks. Thus, we use the [CNode] node as a medium of interaction between subgraphs to achieve information transmission between internal choices.

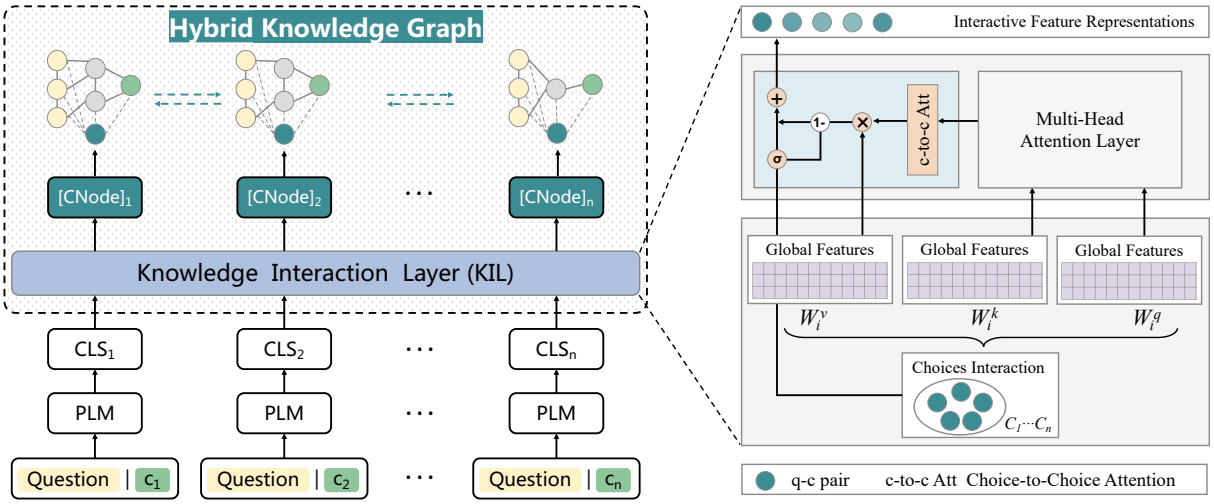

Figure 3: The schematic diagram of Hybrid Knowledge Graph and Knowledge Interaction Layer. The retrieved nodes have been marked in the graph, where the correspondence between knowledge sources and graph nodes has been highlighted in the same color. The grey nodes are some associated entities in subgraph.

We initialize the embedding of [CNode] with the representation of the q-c pair ($\mathcal{C}_i^0 = f_{KIL}(f_{LM}([q; c_i]))$), and other nodes on $\mathcal{G}_i$ by their pre-trained entity embeddings prepared by Feng et al. (2020). In message aggregation and passing stage, the representation of [CNode] is updated as normal nodes in subgraph and the [CNode] aggregates the information from all nodes. Inspired by this, we can realize knowledge interaction between different subgraphs $\mathcal{G}_i$ and define the importance of evidence on $\mathcal{G}_i$ relying on $[\text{CNode}]_i$. Hence, the global node can serve as a hub to help node communications and subgraph interactions, which can make each node more aware of the non-local information. Combining PLM, KGs and inter-choice interaction information, we construct a novel hybrid knowledge graph (see Figure 3).

In the following subsections, we will conduct GNN message aggregation and passing over hybrid knowledge graph to score each choice.

3.3 GNN Architecture

Structured data like knowledge graph is much more efficient in representing commonsense compared with unstructured text (Xu et al., 2021). Therefore, we design a mix attention mechanism of nodes and edges to achieve iterative supporting evidence selection based on the reasoning graph $\mathcal{G}_i$. Meanwhile, we also add the KIL between the layers of GNN to enhance global information interaction among choices (see KIL-GNN in Figure 2).

Edge Encoding. To leverage edge information in supporting evdience selection and representation of the whole graph, we should capture the source/target node types and the edge types. Following Yasunaga et al. (2021), we first obtain the type embedding u_t of each node t, as well as the edge embedding r_{st} from node s to node t by

$$\boldsymbol{r}_{st} = f_r(e_{st}, u_s, u_t) \tag{3}$$

where $u_s, u_t \in \mathbb{R}^{\mathcal{T}}$ are one-hot embeddings indicating the node types of s and t, $e_{st} \in \mathbb{R}^{\mathcal{R}}$ is a one-hot embedding indicating the relation type of the edge $s \rightarrow t$. Here we add self-loops for all nodes. $f_r : \mathbb{R}^{|\mathcal{R}|+2|\mathcal{T}|} \rightarrow \mathbb{R}^D$ is a 2-layer MLP. We then compute the importance of each edge depending on [CNode] node in the reasoning process.

Edge-Weighted Message Updating. Wang et al. (2021) points out that edge encoding is of vital importance for commonsense reasoning. To better encode effective edge features into message aggregation, each edge's weight is used to rescale information flow on that edge. Intuitively, an edge's weight signifies the edge's relevance for reasoning about the given task instance. Thus, We also use the global node [CNode] as global context to compute edge attention weights.

Formally, the update rule of edges at layer ℓ is:

$$\boldsymbol{w}_{(i,j)}^{\ell} = f_w^{\ell}([\mathcal{C}^{\ell}, \boldsymbol{r}_{ij}^{\ell}]) \tag{4}$$

$$\boldsymbol{A}_{(i,j)}^{\ell} = \frac{e^{w_{(i,j)}^{\ell}}}{\sum_{(s,t)\in\epsilon} e^{w_{(s,t)}^{\ell}}} \tag{5}$$

$$\tilde{\boldsymbol{r}}_{st}^{\ell} = \sum_{s\in\mathcal{N}_t\cup\{t\}} \boldsymbol{A}_{(s,t)}^{\ell} \boldsymbol{r}_{st}^{\ell} \tag{6}$$

where f_w^ℓ is a 2-layer MLP. $\mathcal{N}_t$ is the set of node t's incoming neighbors. We then compute the complete node message from s to t as

$$\tilde{h}_s^\ell = f_m(h_s^\ell, \tilde{r}_{st}^\ell) \tag{7}$$

where f_m denotes a linear fully connected layer. h_s^0 is the initial embedding for node s.

The embedding of each node s is updated as $\tilde{h}_s^\ell$, which is related to the neighboring edges of node s. For each edge neighbor, edge weight $\boldsymbol{A}_{(i,j)}^\ell$ is used to rescale the edge's influence on message updating of node s. Through this soft pruning method, we integrate the essential edge information into node features. In the following message aggregation and passing, the node features on the hybrid subgraph is strongly contextualized.

Message Aggregation and Passing. For message passing, we use the multi-head attention GAT (Velickovic et al., 2018), which induces node representation through iterative message passing between neighbors on the graph. Specifically, in the ℓ-th layer of ACENet, we update the representation of each node t to:

$$\boldsymbol{h}_t^{\ell+1} = \Big\Vert_{k=1}^{K} f_n\left(\sum_{s\in\mathcal{N}_t\cup\{t\}} \alpha_{st}^k \tilde{\boldsymbol{h}}_s^\ell\right) \tag{8}$$

where $||$ represents concatenation, α_{st}^k are normalized attention coefficients computed by the k-th attention mechanism(α^k), $\mathcal{N}_t$ represents the neighborhood of an arbitrary node t, and f_n is a 2-layer MLP. Note that, in this setting, the final returned output, h_t, will consist of the important edge-wise and node-wise features for each node.

Then, we will use the multi-head attention to compute attention weight α_{st} from node s to node t. The query and key vectors can be obtained by

$$\boldsymbol{q}_s = f_q(\tilde{\boldsymbol{h}}_s^\ell), \boldsymbol{k}_t = f_k(\tilde{\boldsymbol{h}}_t^\ell) \tag{9}$$

where f_q and f_k are linear transformations. Experimental results also show that the degree feature of nodes is also crucial, thus we add the degree feature d_s to the local node attention weight, which is computed as follows:

$$\alpha_{st} = \frac{\exp(\gamma_{st})}{\sum_{t'\in\mathcal{N}_s\cup\{s\}}\exp(\gamma_{st'})}\cdot d_s, \gamma_{st} = \frac{\boldsymbol{q_s}\boldsymbol{k_t}}{\sqrt{D}} \tag{10}$$

Subgraph Information Interaction. In the above process, we execute message aggregation and passing of single layer GAT. [CNode] aggregates the information from other nodes of its subgraph in the message passing process. In order to further strengthen correct choice information and perception of the overall QA context, we add *knowledge interaction layer* between each layer of GAT to fuse the global representation $\mathcal{G}_i$ (shown in Figure 2).

3.4 Answer and Explain

We now discuss the learning and interactive process of ACENet instantiated for Commonsense QA tasks. By integrating the knowledge of PLM, the retrieved KGs and the interaction information of choices, we compute the probability of c_i being the correct answer as:

$$p(c_i|q,c) \propto exp(MLP(\mathcal{C}^{LM}, \mathcal{G}^{KIL}, \mathcal{G})) \tag{11}$$

where c^{LM} is the initial embedding of the q-c pair through PLM, $\mathcal{G}^{KIL}$ is the knowledge interaction representation of q-c pair over different subgraphs, and $\mathcal{G}$ denotes attention-based pooling for last layer of GNN representation.

The whole model is trained end-to-end jointly with the PLM (e.g., RoBERTa (Liu et al., 2019)) using the cross entropy loss. Finally, we choose the choice with the highest probability score as our answer choice.

4 Experiments

In this section, we conducted experiments over two commonsense QA benchmarks by answering the following research questions.

- **RQ1**: Does ACENet outperform state-of-the-art baselines?
- **RQ2**: How do each model module and training data affect ACENet?
- **RQ3**: What is the performance of ACENet on different types of complex questions?
- **RQ4**: What is the intuitive performance of ACENet in the process of reasoning?

4.1 Experimental Settings

4.1.1 Datasets

We conduct experiments to evaluate our approach on two commonsense QA benchmarks: *CommonsenseQA* and *OpenBookQA*.

CommonsenseQA (Talmor et al., 2019) is a 5-way multiple-choice question answering dataset

of 12,102 questions that require background commonsense knowledge beyond surface language understanding. The test set of *CommonsenseQA* is not publicly available, and model predictions can only be evaluated every two weeks via the official leaderboard. We perform our experiments using the in-house (IH) data split of Lin et al. (2019) to compare to baseline methods.

OpenBookQA (Mihaylov et al., 2018) is a 4-way multiple-choice question answering dataset that tests elementary scientific knowledge. It contains 5,957 questions along with an open book of scientific facts. We use the official data split. Additionally, *OpenBookQA* also provides a collection of background facts in a textual form. We use the correspondence between these facts and their questions, prepared by Clark et al. (2020), as an additional input to the context module.

4.1.2 Implementation Details

Following previous work (Yasunaga et al., 2021), we use *ConceptNet* (Speer et al., 2017), a general-domain knowledge graph, as our structured knowledge source. Node embeddings are initialized using the entity embeddings prepared by Feng et al. (2020), which applies pre-trained LMs to all triples in ConceptNet and then obtains a pooled representation for each entity. Given each q-c pair (question and answer choice), we retrieve the top 200 nodes and adjacent edge according the node relevance score following Yasunaga et al. (2021). We set the dimension (D=200) and number of our GNN layers (L=5), with dropout rate 0.2 applied to each layer (Srivastava et al., 2014). The batch size on CommonsenseQA and OpenBookQA is set from {64, 128}. We train the model with the RAdam optimizer (Liu et al., 2020) using two GPUs (Tesla V100), which takes about 20 hours on average. We use separate learning rates for the LM module and the GNN module, which are set from {1e-5, 2e-5,3e-5} and {5e-4, 1e-3, 2e-3}. The above hyperparameters are tuned on the development set.

4.1.3 Compared Methods

Although text corpus can provide complementary knowledge except for knowledge graphs, our model focuses on exploiting the ability of KG and the joint reasoning among different choices, LM and KG, so we choose LM+KG as the comparison methods.

To further investigate the enhancement effects of KGs on CSQA tasks, we compare with a vanilla fine-tuned LM, which does not use the KG. We use RoBERTa-large for *CommonsenseQA*, and RoBERTa-large and AristoRoBERTa for *OpenBookQA*. In addition, the LM+KG methods share a similar high-level framework with our methods. They usually use LM as a text encoder, GNN or RN as the tool of KG message aggregation and passing. But the specific used knowledge and the joint reasoning methods are different: (1) RN (Santoro et al., 2017), (2) RGCN (Schlichtkrull et al., 2018), (3) GconAttn (Wang et al., 2019), (4) KagNet (Lin et al., 2019), (5) MHGRN (Feng et al., 2020), (6) HGN (Yan et al., 2021),(7) JointLK (Sun et al., 2022), (8) QAGNN (Yasunaga et al., 2021), (9) GREASELM (Zhang et al., 2022). (1), (2), (3) are relation-aware GNNs for KGs, and (4), (5) further model paths in KGs. (6) generates the missing edge of subgraphs for reasoning. (7), (8), (9) construct a joint reasoning graph, which can enhance the interaction of multi-modal knowledge. To be fair, we use the same LM for all comparison methods and our model. The key difference between ACENet and these are that they do not simultaneously consider the interference effects among choices or the importance of different edge and node features.

4.2 Main Results (RQ1)

The results on CommonsenseQA in-house split dataset are shown in Table 1. The results on OpenBookQA test dataset are shown in Table 2. We repeat each experiment 4 times and report the mean and standard deviation of accuracy.

Methods	IHdev-Acc. (%)	IHtest-Acc. (%)
RoBERTa-large (w/o KG)	73.07 (±0.45)	68.69 (±0.56)
+RGCN	72.69 (±0.19)	68.41 (±0.66)
+GconAttn	72.61 (±0.39)	68.59 (±0.96)
+RN	74.57 (±0.91)	69.08 (±0.21)
+KagNet	73.47 (±0.22)	69.01 (±0.76)
+MHGRN	74.45 (±0.10)	71.11 (±0.81)
+HGN	-	73.64 (±0.30)
+QA-GNN	76.54 (±0.21)	73.41 (±0.92)
+JointLK	77.88 (±0.25)	74.43 (±0.83)
+GREASELM	78.50 (±0.50)	74.20 (±0.40)
+ACENet (**Ours**)	**78.54** (±0.45)	**74.72** (±0.70)

Table 1: Performance comparison on CommonsenseQA in-house split. We follow the data division method of Lin et al. (2019) and report the in-house Dev (IHdev) and Test (IHtest) accuracy.

As show in both datasets, our proposed model ACENet outperforms previous methods. We observe consistent improvements over fine-tuned LMs and existing LM+KG models. The boost over QA-GNN suggests that ACENet makes a better use

of inter-choice interaction information than existing LM+KG methods.

Methods	RoBERTa-Large	AristoRoBERTa
Fine-tuned LMs (w/o KG)	64.80 (±2.37)	78.40 (±1.64)
+RGCN	62.45 (±1.57)	74.60 (±2.53)
+GconAttn	64.75 (±1.48)	71.80 (±1.21)
+RN	65.20 (±1.18)	75.35 (±1.39)
+MHGRN	66.85 (±1.19)	80.60
+JointLK	70.34 (±0.75)	**84.92** (±1.07)
+QA-GNN	67.80 (±2.75)	82.77 (±1.56)
+GREASELM	-	84.80
+ACENet (**Ours**)	**70.47** (±0.12)	83.40 (±0.14)

Table 2: Test accuracy comparison on OpenBookQA. Methods with AristoRoBERTa use the textual evidence by Clark et al. (2020) as an additional input to the QA context.

4.3 Ablation Studies (RQ2)

We further conduct specific experiments to investigate the effectiveness of different components in our model.

Impact of Model Components. We add each model component individually and report the accuracy on the CommonsenseQA IHdev set in Table 3. Adding the edge&node attention mechanism leads to 0.79% improvement in performance which shows that some nodes and edges are not conductive to reasoning. Additionally, when we add the KIL (GNN) module, the results have a significant improvement: 76.33% → 77.56% (+1.23%), suggesting that the interaction of different choices is essential in the process of message passing. Meanwhile, our KIL (PLM) provides a better initial representation for the q-c pairs, which is also critical.

Model	Dev Acc.
None	76.33
(a) w/ KIL(PLM)	76.67
(b) w/ KIL(GNN)	77.56
(c) w/ Edge&Node Attention	77.12
(d) w/all (**final**)	**78.54**

Table 3: Ablation study of our model components (adding one component each time), using the CommonsenseQA IHdev set.

Impact of Less Labeled Training Data. Table 4 shows the results of our model and baselines when trained with less training data on CommonsenseQA. Even in the case of less training data, our model still achieves the best test accuracy, which suggests that incorporating the knowledge of external KGs and multiple choices are helpful for commonsense reasoning under the low-resource setting.

Methods	**RoBERTa-Large**	
	60%Train	100%Train
LM Finetuning	65.56 (±0.76)	68.69 (±0.56)
RN	66.16 (±0.28)	70.08 (±0.21)
MHGRN	68.84 (±1.06)	71.11 (±0.81)
HGN	71.10 (±0.11)	73.64 (±0.30)
QA-GNN	70.27 (±0.35)	73.41 (±0.92)
GREASELM	71.08 (±0.52)	74.20 (±0.40)
ACENet (**Ours**)	**71.31** (±0.42)	**74.72** (±0.70)

Table 4: Performance change (accuracy in the amounts of training data on CommonsenseQA IHtest set (same as Lin et al. (2019)).

Impact of Number of Layers (L) and Heads (H). To give further insight into the factors for the capacity of our models, we investigate the impact of the number of layers and heads in the reasoning process. The Figure 4 shows the performance of our model with different numbers of layers and heads. We can observe that increasing the number of layers and heads in a certain range improves the performance of our model. The intuitive explanation is that multiple heads help the model to focus multiple knowledge rules and at the same time multiple layers help the model to recursively select the relevant knowledge rules (Paul and Frank, 2020).

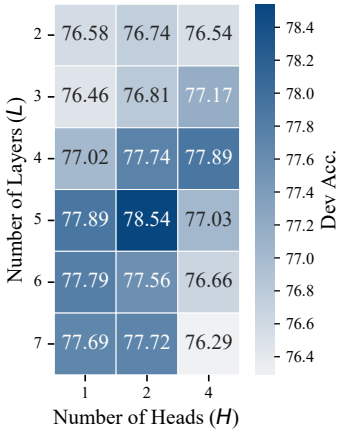

Figure 4: Performance of ACENet model with different numbers of Heads and numbers of GNN Layers on CommonsenseQA IHdev set.

However, performance begins to drop gradually when H=1, 2 and L>5 or H=4 and L>4. A widely accepted explanation for the performance degradation with increasing the layers of GNN is the over-smoothing effect (Chien et al., 2020). Therefore, we set L=5, H=2 to optimally balance their utility. Compared with the baselines, our model achieves better results at different number of layers

Model	Negation Term		Number of Question Prepositions			Number of Question Entities	
	w/o negation	w/ negation	0	1	≥2	≤10 entities	>10 entities
Number	1107	114	551	464	206	1012	209
QA-GNN	77.78	71.93	77.86	76.51	77.18	76.98	78.47
GREASELM	79.31	74.56	79.31	76.94	80.58	77.57	**83.73**
ACENet (**Ours**)	**79.49**	**75.44**	**79.49**	**77.59**	**81.56**	**78.66**	81.34

Table 5: Performance on different types of complex questions. The questions are retrieved from the CommonsenseQA IHdev set.

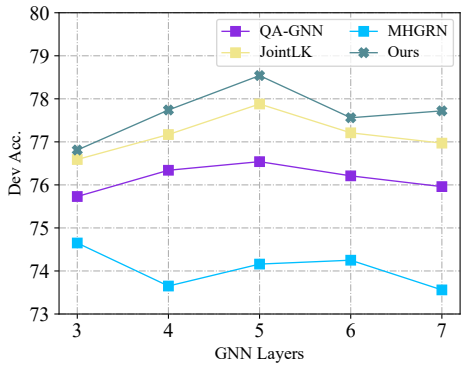

Figure 5: Ablation study on stacked of GNN layers.

(shown in Figure 5).

4.4 Quantitative Analysis (RQ3)

Given these overall performance improvements, we further analyze whether performance improvements were reflected in questions that required more complex reasoning. We define the reasoning complexity of different questions, such as questions with negation and complex questions with more prepositions and entities. We compare our model with the prior better baselines in Table 5.

First, our model exhibits a big boost (+3.51%, +0.88%) over QA-GNN and GREASELM for the questions with negation term (e.g., no, not, never, etc.), suggesting its strength in negative reasoning. Alternatively, the number of prepositions (e.g., in, on, of, with, etc.) in a question usually represents the number of explicit reasoning constraints. Our results in Table 5 demonstrate that our model generally outperforms the baselines for all questions with different number of prepositions. Additionally, the number of the question entities approximately indicates the scale of the retrieved reasoning graph. Our model achieves better results (+1.68%, +1.09%) over QA-GNN and GREASELM for most of the questions (≤10 entities). At the same time, our model and the prior best model, GREASELM perform comparably when aiming at larger scale retrieved graphs.

4.5 Qualitative Analysis (RQ4)

Figure 6 shows the choice-to-choice attention weights induced by the KIL layers of our model in different stages. Our model can strengthen the correct choice information in multi-layer interactions using external KGs to get the right answer, while QA-GNN and GREASELM make the incorrect predictions. We analyze whether different heads focus on multiple knowledge rules. In Figure 6, we observe that two heads focus the different choice-related knowledge in the message aggregation and passing process. First, the attention of two heads represent the key reasoning information in the first several KILs, but gradually averages out by the final layer. The head_1 primarily focuses on "pay bills" in the different KILs, which provides strong evidence of reasoning for the correct answer. In addition, the attention weights on "buy food" and "get things" become higher in head_2. It also helps our model to select the relevant knowledge. As a whole, our model integrates the different knowledge rules mined by each head to realize the correct prediction.

4.6 Analysis of Experimental Results

To explain why ACENet outperforms other baselines, our hypothesis is because of the receptive field of the subgraph nodes expanded with the interaction of multi-layer Knowledge Interaction Layers. And through the aggregation and propagation of multi-layer graph neural networkeach node can more aware of the non-local information. However, the work to explain the result of neural networks requires strenuous efforts. We can think differently and extend this method into more general settings in other tasks (e.g., document modeling, reading comprehension, information extraction, etc.)

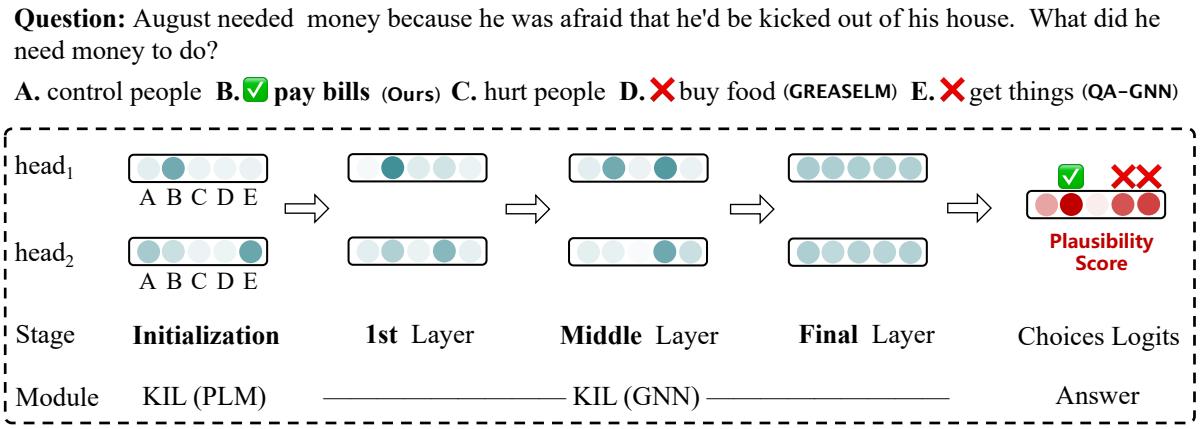

Figure 6: Qualitative analysis of ACENet's inter-choice attention weight changes across multiple knowledge interaction layers in different heads.

5 Conclusions

In this paper, we propose a *multi-head attention knowledge interaction layer* to enhance correct choice information and capture nuances in different choices. Meanwhile, the *mix attention mechanism of nodes and edges* is introduced into message passing to iteratively select relevant knowledge in *hybrid knowledge graph*. Experimental results on CommonsenseQA and OpenBookQA demonstrate the superiority of ACENet over other LM+KG methods and the strong performance in handling complex questions. In future work, we plan to further investigate augmenting effects of knowledge graph for reasoning, and integrate neural and symbolic reasoning system to achieve dual system cognitive intelligence.

Limitations

Although our model achieves competitive performance in commonsense question answering tasks, there are some methods and limitations that can be improved. The limitations of our study are summarized as follows:

1) GNNs incorporates implicit external knowledge in the process of message aggregation and passing. Therefore, existing KG-augmented methods are usually not interpretable enough.

2) The optimal number of GNN layers in our model depends on experimental results. However, the scale of the knowledge graphs is often uncertain in real application scenarios. We can not guarantee that the specific number of GNN layers will achieve the appropriate performance. How to design the depth-adaptive GNNs for a balance between efficiency and effectiveness is a key challenge.

3) At present, our model of using the interaction between choices to strengthen correct choice information is only suitable for question answering tasks with the limited scope.

Ethics Statement

This paper proposes a general approach to fuse QA context, language models and external knowledge graphs for commonsense reasoning. We work within the purview of acceptable privacy practices and strictly follow the data usage policy. In all the experiments, we use public datasets and consist of their intended use. We have also described our experimental settings in detail which ensure the reproducibility of our method. We neither introduce any social/ethical bias to the model nor amplify any bias in the data, so we do not foresee any direct social consequences or ethical issues.

Acknowledgments

This work is supported in part by Natural Science Foundation of China (grant No.62276188 and No.61876129), the Beijing Academy of Artificial Intelligence(BAAI), TJU-Wenge joint laboratory funding, and MindSpore [2].

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Han-

[2] https://www.mindspore.cn/

nah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. 2020. Adaptive universal generalized pagerank graph neural network. *ArXiv preprint*, abs/2006.07988.

Peter Clark, Oren Etzioni, Tushar Khot, Daniel Khashabi, Bhavana Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, et al. 2020. From 'f'to 'a'on the ny regents science exams: An overview of the aristo project. *AI Magazine*, 41(4):39–53.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, Online. Association for Computational Linguistics.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR.

David Gunning. 2018. Machine common sense concept paper. *ArXiv preprint*, abs/1810.07528.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the variance of the adaptive learning rate and beyond. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Debjit Paul and Anette Frank. 2020. Social commonsense reasoning with multi-head knowledge attention. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2969–2980, Online. Association for Computational Linguistics.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Tim Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4967–4976.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Yueqing Sun, Qi Shi, Le Qi, and Yu Zhang. 2022. JointLK: Joint reasoning with language models and knowledge graphs for commonsense question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5049–5060, Seattle, United States. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Kuan Wang, Yuyu Zhang, Diyi Yang, Le Song, and Tao Qin. 2021. Gnn is a counter? revisiting gnn for question answering. *ArXiv preprint*, abs/2110.03192.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, et al. 2019. Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7208–7215.

Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. 2021. Fusing context into knowledge graph for commonsense question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1201–1207, Online. Association for Computational Linguistics.

Jun Yan, Mrigank Raman, Aaron Chan, Tianyu Zhang, Ryan Rossi, Handong Zhao, Sungchul Kim, Nedim Lipka, and Xiang Ren. 2021. Learning contextualized knowledge structures for commonsense reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4038–4051, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems*, 34.

Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. 2020. Jaket: Joint pre-training of knowledge graph and language understanding. *ArXiv preprint*, abs/2010.00796.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models for question answering. *ArXiv preprint*, abs/2201.08860.