

# Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering

Yanlin Feng<sup>♣\*</sup> Xinyue Chen<sup>♣\*</sup> Bill Yuchen Lin<sup>♥</sup> Peifeng Wang<sup>♥</sup> Jun Yan<sup>♥</sup> Xiang Ren<sup>♥</sup>

fengyanlin@pku.edu.cn, kiwisher@sjtu.edu.cn,  
{yuchen.lin, peifengw, yanjun, xiangren}@usc.edu

<sup>♥</sup>University of Southern California

<sup>♣</sup>Peking University

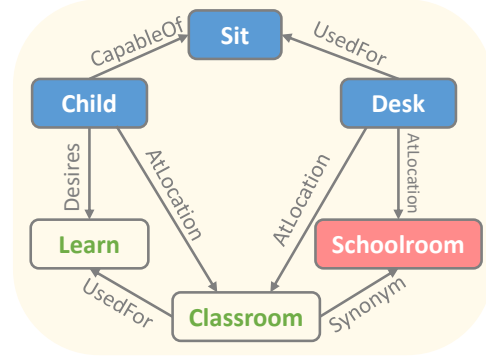
<sup>♣</sup>Shanghai Jiao Tong University

## Abstract

While fine-tuning pre-trained language models (PTLMs) has yielded strong results on a range of question answering (QA) benchmarks, these methods still suffer in cases when external knowledge are needed to infer the right answer. Existing work on augmenting QA models with external knowledge (e.g., knowledge graphs) either struggle to model multi-hop relations efficiently, or lack transparency into the model’s prediction rationale. In this paper, we propose a novel knowledge-aware approach that equips PTLMs with a *multi-hop relational reasoning module*, named multi-hop graph relation networks (MHGRN). It performs multi-hop, multi-relational reasoning over subgraphs extracted from external knowledge graphs. The proposed reasoning module unifies path-based reasoning methods and graph neural networks to achieve better interpretability and scalability. We also empirically show its effectiveness and scalability on CommonsenseQA and OpenbookQA datasets, and interpret its behaviors with case studies. In particular, MHGRN achieves the state-of-the-art performance (76.5% accuracy) on the CommonsenseQA official test set.<sup>1</sup>

## 1 Introduction

Many recently proposed question answering tasks require not only machine comprehension of the question and context, but also *relational reasoning* over entities (concepts) and their relationships based by referencing external knowledge (Talmor et al., 2019; Sap et al., 2019; Clark et al., 2018; Mihaylov et al., 2018). For example, the question in Fig. 1 requires a model to perform relational reasoning over mentioned entities, *i.e.*, to infer latent relations among the concepts: {CHILD, SIT, DESK,



Where does a **child** likely **sit** at a **desk**?

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

Figure 1: **Illustration of knowledge-aware QA.** A sample question from CommonsenseQA can be better answered if a relevant subgraph of ConceptNet is provided as evidence. **Blue nodes** correspond to entities mentioned in the question, **pink nodes** correspond to those in the answer. The other nodes are some associated entities introduced when extracting the subgraph. ★ indicates the correct answer.

SCHOOLROOM}. Background knowledge such as “a child is likely to appear in a schoolroom” may not be readily contained in the questions themselves, but are commonsensical to humans.

Despite the success of large-scale pre-trained language models (PTLMs) (Devlin et al., 2019; Liu et al., 2019b), there is still a large performance gap between fine-tuned models and human performance on datasets that probe relational reasoning. These models also fall short of providing interpretable predictions as the knowledge in their pre-training corpus is not explicitly stated, but rather is implicitly learned. It is thus difficult to recover the knowledge used in the reasoning process.

This has led many works to leverage knowledge graphs to improve machine reasoning ability to infer these latent relations for answering

\*The first two authors contributed equally. The major work was done when both authors interned at USC.

<sup>1</sup><https://github.com/INK-USC/MHGRN>.

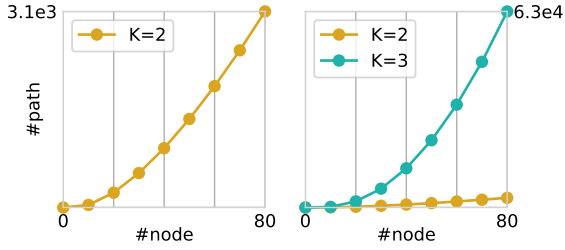


Figure 2: **Number of  $K$ -hop relational paths w.r.t. the node count in extracted graphs on CommonsenseQA.** **Left:** The path count is polynomial w.r.t. the number of nodes. **Right:** The path count is exponential w.r.t. the number of hops.

this kind of questions (Mihaylov and Frank, 2018; Lin et al., 2019; Wang et al., 2019; Yang et al., 2019). Knowledge graphs represent relational knowledge between entities into multi-relational edges, thus making it easier for a model to acquire relational knowledge and improve its reasoning ability. Moreover, incorporating knowledge graphs brings the potential of interpretable and trustworthy predictions, as the knowledge is now explicitly stated. For example, in Fig. 1, the relational path (CHILD  $\rightarrow$  AtLocation  $\rightarrow$  CLASSROOM  $\rightarrow$  Synonym  $\rightarrow$  SCHOOLROOM) naturally provides evidence for the answer SCHOOLROOM.

A straightforward approach to leveraging a knowledge graph is to directly model these relational paths. KagNet (Lin et al., 2019) and MH-PGM (Bauer et al., 2018) extract relational paths from knowledge graph and encode them with sequence models, resulting in multi-hop relations being explicitly modeled. Application of attention mechanisms upon these relational paths can further offer good interpretability. However, these models are hardly scalable because the number of possible paths in a graph is (1) *polynomial* w.r.t. the number of nodes (2) *exponential* w.r.t. the path length (see Fig. 2). Therefore, some (Weissenborn et al., 2017; Mihaylov and Frank, 2018) resort to only using one-hop paths, namely, triples, to balance scalability and reasoning capacities.

Graph neural networks (GNN), in contrast, enjoy better scalability via their message passing formulation, but usually lack transparency. The most commonly used GNN variant, Graph Convolutional Network (GCN) (Kipf and Welling, 2017), performs message passing by aggregating neighborhood information for each node, but ignores the relation types. RGCN (Schlichtkrull et al., 2018) generalizes GCN by performing relation-specific ag-

	GCN	RGCN	KagNet	MHGRN
Multi-Relational Encoding	✗	✓	✓	✓
Interpretable	✗	✗	✓	✓
Scalable w.r.t. #node	✓	✓	✗	✓
Scalable w.r.t. #hop	✓	✓	✗	✓

Table 1: **Properties** of our MHGRN and other representative models for graph encoding.

gregation, making it applicable to encoding multi-relational graphs. However, these models do not distinguish the importance of different neighbors or relation types and thus cannot provide explicit relational paths for model behavior interpretation.

In this paper, we propose a novel graph encoding architecture, *Multi-hop Graph Relation Networks* (MHGRN), which combines the strengths of path-based models and GNNs. Our model inherits scalability from GNNs by preserving the message passing formulation. It also enjoys interpretability of path-based models by incorporating structured relational attention mechanism to model message passing routes. Our key motivation is to perform *multi-hop* message passing within a *single* layer to allow each node to directly attend to its multi-hop neighbours, thus enabling multi-hop relational reasoning with MHGRN. We outline the favorable features of knowledge-aware QA models in Table 1 and compare our MHGRN with representative GNNs and path-based methods.

We summarize the main contributions of this work as follows: 1) We propose MHGRN, a novel model architecture tailored to multi-hop relational reasoning. Our model is capable of explicitly modeling multi-hop relational paths at scale. 2) We propose structured relational attention mechanism for efficient and interpretable modeling of multi-hop reasoning paths, along with its training and inference algorithms. 3) We conduct extensive experiments on two question answering datasets and show that our models bring significant improvements compared to knowledge-agnostic pre-trained language models, and outperform other graph encoding methods by a large margin.

## 2 Problem Formulation and Overview

In this paper, we limit the scope to the task of multiple-choice question answering, although it can be easily generalized to other knowledge-guided tasks (e.g., natural language inference). The overall paradigm of knowledge-aware QA is illustrated in Fig. 3. Formally, given an external knowledge graph (KG) as the knowledge source and a

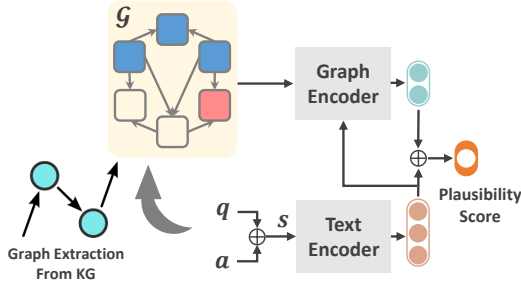


Figure 3: **Overview of the knowledge-aware QA framework.** It integrates the output from graph encoder (for relational reasoning over contextual sub-graphs) and text encoder (for textual understanding) to generate the plausibility score for an answer option.

question  $q$ , our goal is to identify the correct answer from a set  $\mathcal{C}$  of given choices. We turn this problem into measuring the plausibility score between  $q$  and each answer choice  $a \in \mathcal{C}$  then selecting the answer with the highest plausibility score.

We use  $\mathbf{q}$  and  $\mathbf{a}$  to denote the representation vectors of question  $q$  and option  $a$ . To measure the score for  $\mathbf{q}$  and  $\mathbf{a}$ , we first concatenate them to form a statement  $\mathbf{s} = [\mathbf{q}; \mathbf{a}]$ . Then we extract from the external KG a subgraph  $\mathcal{G}$  (i.e., schema graph in KagNet (Lin et al., 2019)), with the guidance of  $\mathbf{s}$  (detailed in §5.1). This contextualized subgraph is defined as a multi-relational graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \phi)$ . Here  $\mathcal{V}$  is a subset of entity nodes in the external KG, containing only entities relevant to  $\mathbf{s}$ .  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$  is the set of edges that connect nodes in  $\mathcal{V}$ , where  $\mathcal{R} = \{1, \dots, m\}$  are ids of all pre-defined relation types. The mapping function  $\phi(i) : \mathcal{V} \rightarrow \mathcal{T} = \{\mathbf{E}_q, \mathbf{E}_a, \mathbf{E}_o\}$  takes node  $i \in \mathcal{V}$  as input and outputs  $\mathbf{E}_q$  if  $i$  is an entity mentioned in  $q$ ,  $\mathbf{E}_a$  if it is mentioned in  $a$ , or  $\mathbf{E}_o$  otherwise. We finally encode the statement to  $\mathbf{s}$ ,  $\mathcal{G}$  to  $\mathbf{g}$ , concatenate  $\mathbf{s}$  and  $\mathbf{g}$ , for calculating the plausibility score.

### 3 Background: Multi-Relational Graph Encoding Methods

We leave encoding of  $\mathbf{s}$  to pre-trained language models which have shown powerful text representation ability, while we focus on the challenge of encoding graph  $\mathcal{G}$  to capture latent relations between entities. Current methods for encoding multi-relational graphs mainly fall into two categories: graph neural networks and path-based models. Graph neural networks encode structured information by passing messages between nodes, directly operating on the graph structure, while path-based methods first decompose the graph into paths and then pool features over all the paths.

**Graph Encoding with GNNs.** For a graph with  $n$  nodes, a graph neural network (GNN) takes a set of node features  $\{h_1, h_2, \dots, h_n\}$  as input, and computes their corresponding node embeddings  $\{h'_1, h'_2, \dots, h'_n\}$  via message passing (Gilmer et al., 2017). A compact graph representation for  $\mathcal{G}$  can thus be obtained by pooling over the node embeddings  $\{h'_i\}$ :

$$\text{GNN}(\mathcal{G}) = \text{Pool}(\{h'_1, h'_2, \dots, h'_n\}). \quad (1)$$

As a notable variant of GNNs, graph convolutional networks (GCNs) (Kipf and Welling, 2017) additionally update node embeddings by aggregating messages from its direct neighbors. RGCNs (Schlichtkrull et al., 2018) extend GCNs to encode multi-relational graphs by defining relation-specific weight matrix  $\mathbf{W}_r$  for each edge type:

$$h'_i = \sigma \left( \left( \sum_{r \in \mathcal{R}} |\mathcal{N}_i^r| \right)^{-1} \sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} \mathbf{W}_r h_j \right), \quad (2)$$

where  $\mathcal{N}_i^r$  denotes neighbors of node  $i$  under relation  $r$ .<sup>2</sup>

While GNNs have proved to have good scalability, their reasoning is done at the node level, thus making them incompatible with modeling path-level reasoning chains—a crucial component for QA tasks that require relational reasoning. This property also hinders the model’s decisions from being interpreted at the path level.

**Graph Encoding with Path-Based Models.** In addition to directly modeling the graph with GNNs, one can also view a graph as a set of relational paths connecting pairs of entities.

Relation Networks (RN) (Santoro et al., 2017) can be adapted to multi-relational graph encoding under QA settings. RNs use MLPs to encode all triples (one-hop paths) in  $\mathcal{G}$  whose head entity is in  $\mathcal{Q} = \{j \mid \phi(j) = \mathbf{E}_q\}$  and tail entity is in  $\mathcal{A} = \{i \mid \phi(i) = \mathbf{E}_a\}$ . It then pools the triple embeddings to generate a vector for  $\mathcal{G}$  as follows.

$$\text{RN}(\mathcal{G}) = \text{Pool} \left( \{ \text{MLP}(h_j \oplus e_r \oplus h_i) \mid j \in \mathcal{Q}, i \in \mathcal{A}, (j, r, i) \in \mathcal{E} \} \right). \quad (3)$$

Here  $h_j$  and  $h_i$  are features for nodes  $j$  and  $i$ ,  $e_r$  is the embedding of relation  $r \in \mathcal{R}$ ,  $\oplus$  denotes vector concatenation.

<sup>2</sup>For simplicity, we assume a single graph convolutional layer. In practice, multiple layers are stacked to enable message passing from multi-hop neighbors.

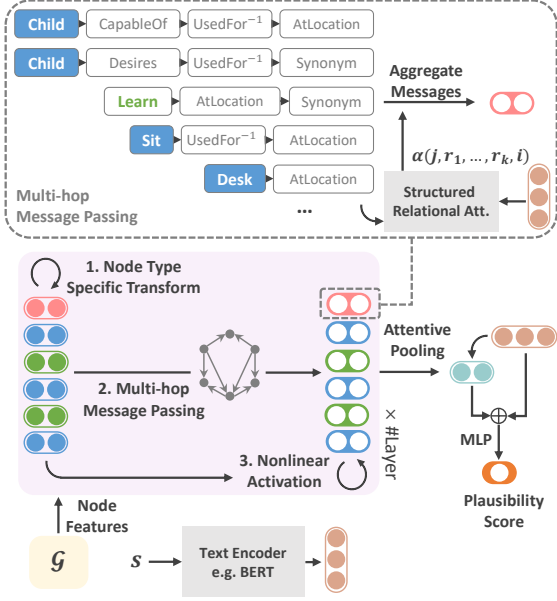


Figure 4: **Our proposed MHGRN architecture for relational reasoning.** MHGRN takes a multi-relational graph  $\mathcal{G}$  and a (question-answer) statement vector  $s$  as input, and outputs a scalar that represent the plausibility score of this statement.

To further equip RN with the ability of modeling nondegenerate paths, KagNet (Lin et al., 2019) adopts LSTMs to encode all paths connecting question entities and answer entities with lengths no more than  $K$ . It then aggregates all path embeddings via attention mechanism:

$$\text{KAGNET}(\mathcal{G}) = \text{Pool}\left(\left\{\text{LSTM}(j, r_1, \dots, r_k, i) \mid (j, r_1, j_1), \dots, (j_{k-1}, r_k, i) \in \mathcal{E}, 1 \leq k \leq K\right\}\right). \quad (4)$$

## 4 Proposed Method: Multi-Hop Graph Relation Networks (MHGRN)

This section presents *Multi-hop Graph Relation Network* (MHGRN), a novel graph neural network architecture that unifies both GNN and RN, for encoding multi-relational graphs to augment text comprehension. MHGRN inherits the capability of path reasoning and interpretability from path-based models, while preserving good scalability of GNN with the message passing formulation.

### 4.1 MHGRN: Model Architecture

Following the introduction to GNN in Sec. 3, we consider encoding a multi-relational graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \phi)$  into a fixed size vector  $g$  conditioned on textual representation  $s$ , by first transforming input

node features  $\{h_1, \dots, h_n\}$  into node embeddings  $\{h'_1, \dots, h'_n\}$ , and then pooling these embeddings. Node features can be initialized with pre-trained weights (details in Appendix A) and we focus on the computation of node embeddings.

**Type-Specific Transformation.** To make our model aware of the node type information  $\phi$ , we first perform node type specific linear transformation on the input node features:

$$x_i = U_{\phi(i)} h_i + b_{\phi(i)}, \quad (5)$$

where the learnable parameters  $U$  and  $b$  are specific to the type of node  $i$ .

**Multi-Hop Message Passing.** As mentioned before, our motivation is to endow GNN with the capability of *directly modeling paths*. To this end, we propose to pass messages directly over all the relational paths of lengths up to  $K$ , where  $K$  is a hyper-parameter. The set of valid  $k$ -hop relational paths is defined as:

$$\Phi_k = \{(j, r_1, \dots, r_k, i) \mid (j, r_1, j_1), \dots, (j_{k-1}, r_k, i) \in \mathcal{E}\} \quad (1 \leq k \leq K). \quad (6)$$

We perform  $k$ -hop ( $1 \leq k \leq K$ ) message passing over these paths, which is a generalization of the single-hop message passing in RGCN (see Eq. 2):

$$z_i^k = \sum_{(j, r_1, \dots, r_k, i) \in \Phi_k} \alpha(j, r_1, \dots, r_k, i) / d_i^k \cdot \mathbf{W}_0^K \cdots \mathbf{W}_0^{k+1} \mathbf{W}_{r_k}^k \cdots \mathbf{W}_{r_1}^1 x_j \quad (1 \leq k \leq K), \quad (7)$$

where the  $\mathbf{W}_t^k$  ( $1 \leq t \leq K, 0 \leq r \leq m$ ) matrices are learnable<sup>3</sup>,  $\alpha(j, r_1, \dots, r_k, i)$  is an attention score elaborated in §4.2 and  $d_i^k = \sum_{(j, \dots, i) \in \Phi_k} \alpha(j, \dots, i)$  is the normalization factor. The  $\{\mathbf{W}_{r_k}^k \cdots \mathbf{W}_{r_1}^1 \mid 1 \leq r_1, \dots, r_k \leq m\}$  matrices can be interpreted as the low rank approximation of a  $\{m \times \dots \times m\}_k \times d \times d$  tensor that assigns a separate transformation for each  $k$ -hop relation, where  $d$  is the dim. of  $x_i$ .

Incoming messages from paths of different lengths are aggregated via attention mechanism (Vaswani et al., 2017):

$$z_i = \sum_{k=1}^K \text{softmax}(\text{bilinear}(s, z_i^k)) \cdot z_i^k. \quad (8)$$

<sup>3</sup>  $\mathbf{W}_0^t$  ( $0 \leq t \leq K$ ) are introduced as padding matrices so that  $K$  transformations are applied regardless of  $k$ , thus ensuring comparable scale of  $z_i^k$  across different  $k$ .



**Non-linear Activation.** Finally, we apply shortcut connection and nonlinear activation to obtain the output node embeddings.

$$\mathbf{h}_i' = \sigma(\mathbf{V}\mathbf{h}_i + \mathbf{V}'\mathbf{z}_i), \quad (9)$$

where  $\mathbf{V}$  and  $\mathbf{V}'$  are learnable model parameters, and  $\sigma$  is a non-linear activation function.

## 4.2 Structured Relational Attention

The remaining problem becomes how to effectively parameterize the attention score  $\alpha(j, r_1, \dots, r_k, i)$  in Eq. 7 for all possible  $k$ -hop paths without introducing  $\mathcal{O}(m^k)$  parameters. We first regard it as the probability of a relation sequence  $(\phi(j), r_1, \dots, r_k, \phi(i))$  conditioned on  $\mathbf{s}$ :

$$\alpha(j, r_1, \dots, r_k, i) = p(\phi(j), r_1, \dots, r_k, \phi(i) \mid \mathbf{s}), \quad (10)$$

which can naturally be modeled by a probabilistic graphical model, such as conditional random field (Lafferty et al., 2001):

$$\begin{aligned} p(\dots \mid \mathbf{s}) &\propto \exp\left(f(\phi(j), \mathbf{s}) + \sum_{t=1}^k \delta(r_t, \mathbf{s}) \right. \\ &\quad \left. + \sum_{t=1}^{k-1} \tau(r_t, r_{t+1}) + g(\phi(i), \mathbf{s})\right) \\ &\triangleq \underbrace{\beta(r_1, \dots, r_k, \mathbf{s})}_{\text{Relation Type Attention}} \cdot \underbrace{\gamma(\phi(j), \phi(i), \mathbf{s})}_{\text{Node Type Attention}}, \quad (11) \end{aligned}$$

where  $f(\cdot)$ ,  $\delta(\cdot)$  and  $g(\cdot)$  are parameterized by two-layer MLPs and  $\tau(\cdot)$  by a transition matrix of shape  $m \times m$ . Intuitively,  $\beta(\cdot)$  models the importance of a  $k$ -hop relation while  $\gamma(\cdot)$  models the importance of messages from node type  $\phi(j)$  to node type  $\phi(i)$  (e.g. the model can learn to pass messages only from question entities to answer entities).

Our model scores a  $k$ -hop relation by decomposing it into both context-aware single-hop relations (modeled by  $\delta$ ) and two-hop relations (modeled by  $\tau$ ). We argue that  $\tau$  is indispensable, without which the model may assign high importance to illogical multi-hop relations (e.g., [AtLocation, CapableOf]) or noisy relations (e.g., [RelatedTo, RelatedTo]).

## 4.3 Computation Complexity Analysis

Although the multi-hop message passing process in Eq. 7 and the structured relational attention module in Eq. 11 handles potentially exponential number of paths, we show that it can be computed in

Model	Time	Space
<i><math>\mathcal{G}</math> is a dense graph</i>		
$K$ -hop KagNet	$\mathcal{O}(m^K n^{K+1} K)$	$\mathcal{O}(m^K n^{K+1} K)$
$K$ -layer RGCN	$\mathcal{O}(mn^2 K)$	$\mathcal{O}(mnK)$
MHGRN	$\mathcal{O}(m^2 n^2 K)$	$\mathcal{O}(mnK)$
<i><math>\mathcal{G}</math> is a sparse graph with maximum node degree <math>\Delta \ll n</math></i>		
$K$ -hop KagNet	$\mathcal{O}(m^K n K \Delta^K)$	$\mathcal{O}(m^K n K \Delta^K)$
$K$ -layer RGCN	$\mathcal{O}(mnK\Delta)$	$\mathcal{O}(mnK)$
MHGRN	$\mathcal{O}(m^2 n K \Delta)$	$\mathcal{O}(mnK)$

Table 2: **Computation complexity** of different  $K$ -hop reasoning models on a dense/sparse multi-relational graph with  $n$  nodes and  $m$  relation types. Despite the quadratic complexity w.r.t.  $m$ , MHGRN’s time cost is similar to RGCN on GPUs with parallelizable matrix multiplications (cf. Fig. 7).

linear time using dynamic programming (see Appendix B). As summarized in Table 2, the time complexity and space complexity of our model on a sparse graph are  $\mathcal{O}(m^2 n K \Delta)$  and  $\mathcal{O}(mnK)$  respectively, both of which are linear with respect to either the path length  $K$  or the number of nodes  $n$ .

## 4.4 Expressive Power of MHGRN

In addition to the efficiency and scalability, we now discuss the modeling capacity of our model. With the message passing formulation and relation-specific transformations, MHGRN is by nature the generalization of RGCN. Furthermore, it is capable of directly modeling paths, making it interpretable as are path-based models like RN and KagNet. To show this, we first generalize RN (Eq. 3) to the multi-hop setting and introduce  $K$ -hop RN (see Appendix C for a formal definition).  $K$ -hop RN models multi-hop relation as the composition of single-hop relations with element-wise multiplication. We show that MHGRN is capable of representing  $K$ -hop RN (see Appendix D for the proof).

## 4.5 Learning, Inference and Path Decoding

We now discuss the learning and inference process of MHGRN instantiated for the task of multiple-choice question answering. Following the problem formulation in Sec. 2, we aim to determine the plausibility of an answer candidate  $a \in \mathcal{C}$  given the question  $q$  with the information from both text  $\mathbf{s}$  and graph  $\mathcal{G}$ . We first obtained the graph representation  $\mathbf{g}$  by performing attentive pooling over the output node embeddings of answer enti-

Methods	BERT-Base		BERT-Large		RoBERTa-Large	
	IHdev-Acc.(%)	IHtest-Acc.(%)	IHdev-Acc.(%)	IHtest-Acc.(%)	IHdev-Acc.(%)	IHtest-Acc.(%)
w/o KG	57.31 ( $\pm 1.07$ )	53.47 ( $\pm 0.87$ )	61.06 ( $\pm 0.85$ )	55.39 ( $\pm 0.40$ )	73.07 ( $\pm 0.45$ )	68.69 ( $\pm 0.56$ )
RGCN (Schlichtkrull et al., 2018)	56.94 ( $\pm 0.38$ )	54.50 ( $\pm 0.56$ )	62.98 ( $\pm 0.82$ )	57.13 ( $\pm 0.36$ )	72.69 ( $\pm 0.19$ )	68.41 ( $\pm 0.66$ )
GconAttn (Wang et al., 2019)	57.29 ( $\pm 0.62$ )	54.41 ( $\pm 0.50$ )	62.98 ( $\pm 0.17$ )	56.94 ( $\pm 0.77$ )	73.38 ( $\pm 0.27$ )	69.88 ( $\pm 0.47$ )
KagNet <sup>†</sup> (Lin et al., 2019)	55.57	56.19	62.35	57.16	-	-
RN (1-hop)	58.27 ( $\pm 0.22$ )	56.20 ( $\pm 0.45$ )	63.04 ( $\pm 0.58$ )	58.46 ( $\pm 0.71$ )	74.57 ( $\pm 0.91$ )	70.08 ( $\pm 0.21$ )
RN (2-hop)	59.81 ( $\pm 0.76$ )	56.61 ( $\pm 0.68$ )	63.36 ( $\pm 0.26$ )	58.92 ( $\pm 0.14$ )	73.65 ( $\pm 3.09$ )	69.59 ( $\pm 3.80$ )
MHGRN	60.36 ( $\pm 0.23$ )	<b>57.23</b> ( $\pm 0.82$ )	63.29 ( $\pm 0.51$ )	<b>60.59</b> ( $\pm 0.58$ )	74.45 ( $\pm 0.10$ )	<b>71.11</b> ( $\pm 0.81$ )

Table 3: **Performance comparison on CommonsenseQA in-house split.** We report in-house Dev (IHdev) and Test (IHtest) accuracy (mean and standard deviation of four runs) using the data split of Lin et al. (2019) on CommonsenseQA. <sup>†</sup> indicates reported results in its paper.

ties  $\{h_i^l \mid i \in \mathcal{A}\}$ . Next we simply concatenate it with the text representation  $s$  and compute the plausibility score by  $\rho(q, a) = \text{MLP}(s \oplus g)$ .

During training, we maximize the plausibility score of the correct answer  $\hat{a}$  by minimizing the *cross-entropy* loss:

$$\mathcal{L} = \mathbb{E}_{q, \hat{a}, \mathcal{C}} \left[ -\log \frac{\exp(\rho(q, \hat{a}))}{\sum_{a \in \mathcal{C}} \exp(\rho(q, a))} \right]. \quad (12)$$

The whole model is trained **end-to-end** jointly with the text encoder (e.g., RoBERTa).

During inference, we predict the most plausible answer by  $\text{argmax}_{a \in \mathcal{C}} \rho(q, a)$ . Additionally, we can decode a reasoning path as evidence for model predictions, endowing our model with the interpretability enjoyed by path-based models. Specifically, we first determine the answer entity  $i^*$  with the highest score in the pooling layer and the path length  $k^*$  with the highest score in Eq. 8. Then the reasoning path is decoded by  $\text{argmax} \alpha(j, r_1, \dots, r_{k^*}, i^*)$ , which can be computed in linear time using dynamic programming.

## 5 Experimental Setup

We introduce how we construct  $\mathcal{G}$  (§5.1), the datasets (§5.2), as well as the baseline methods (§5.3). Appendix A shows more implementation and experimental details for reproducibility.

### 5.1 Extracting $\mathcal{G}$ from External KG

We use *ConceptNet* (Speer et al., 2017), a general-domain knowledge graph as our external KG to test models’ ability to harness structured knowledge source. Following KagNet (Lin et al., 2019), we merge relation types to increase graph density and add reverse relations to construct a multi-relational graph with 34 relation types. To extract an informative contextual graph  $\mathcal{G}$  from the KG, we recognize entity mentions in  $s$  and link them to entities in

Methods	Single	Ensemble
RoBERTa <sup>†</sup>	72.1	72.5
ALBERT <sup>†</sup> (Lan et al., 2019)	-	76.5
XLNet + GR <sup>†</sup> (Lv et al., 2019)	75.3	-
RoBERTa+KEDGN <sup>†</sup>	72.5	74.4
RoBERTa + KE <sup>†</sup>	73.3	-
RoBERTa+HyKAS 2.0 <sup>†</sup> (Ma et al., 2019)	73.2	-
XLNet+DREAM <sup>†</sup>	66.9	73.3
RoBERTa+KEDGN <sup>†</sup>	72.5	74.4
RoBERTa+FreeLB <sup>†</sup> (Zhu et al., 2020)	72.2	73.1
MHGRN (RoBERTa, $K = 2$ )	<b>75.4</b>	<b>76.5</b>

Table 4: **Performance comparison on official test of CommonsenseQA** with leaderboard SoTAs<sup>4</sup> (accuracy in %). <sup>†</sup> indicates reported results on the leaderboard. MHGRN achieves state-of-the-art on both single and ensemble settings with RoBERTa-large encoder.

*ConceptNet*, with which we initialize our node set  $\mathcal{V}$ . We then add to  $\mathcal{V}$  all the entities that appear in any two-hop paths between pairs of mentioned entities. Unlike KagNet, we do not perform any pruning but instead reserve all the edges between nodes in  $\mathcal{V}$ , forming our  $\mathcal{G}$ .

### 5.2 Datasets

We evaluate models on two multiple-choice question answering datasets, CommonsenseQA and OpenbookQA. Both require world knowledge beyond textual understanding to perform well.

**CommonsenseQA** (Talmor et al., 2019) necessitates various commonsense reasoning skills. The questions are created with entities from *ConceptNet* and they are designed to probe latent compositional relations between entities in *ConceptNet*.

**OpenBookQA** (Mihaylov et al., 2018) provides elementary science questions together with an open

<sup>4</sup>Models based on ConceptNet are no longer shown on the leaderboard, and thus we got ours’ results directly from the organizers.

book of science facts. This dataset also probes general commonsense knowledge beyond the provided facts. As our model is orthogonal to text-form knowledge retrieval, we do not utilize the provided open book and instead use *ConceptNet*. Consequently, we do not compare our methods with those using the open book.

### 5.3 Compared Methods

We implement both knowledge-agnostic fine-tuning of pre-trained LMs and models that incorporate KG as external sources as our baselines. Additionally, we directly compare our model with the results from corresponding leaderboard. These methods typically leverage textual knowledge or extra training data, as opposed to external KG. In all our *implemented* models, we use pre-trained LMs as text encoders for s for fair comparison. We stick to our focus of encoding *structured* KG and therefore do not compare our models with those (Pan et al., 2019; Zhang et al., 2018; Sun et al., 2019; Banerjee et al., 2019) augmented by other text-form external knowledge (e.g., Wikipedia).

Specifically, we fine-tune BERT-BASE, BERT-LARGE (Devlin et al., 2019), and ROBERTA (Liu et al., 2019b) for multiple-choice questions. We take RGCN (Eq. 2 in Sec. 3), RN<sup>5</sup> (Eq. 3 in Sec. 3), KagNet (Eq. 4 in Sec. 3) and GconAttn (Wang et al., 2019) as baselines. GconAttn generalizes match-LSTM (Wang and Jiang, 2016) from sequence modeling to entity set modeling and achieves success in language inference tasks.

## 6 Results and Discussions

In this section, we present the results of our models in comparison with baselines as well as methods on the leaderboards for both CommonsenseQA and OpenbookQA. We also provide analysis of models' components and characteristics.

### 6.1 Main Results

For CommonsenseQA (Table 3), we first use the in-house data split of Lin et al. (2019) to compare our models with our implemented baselines. In this setting, we take 1,241 examples from official training examples as our in-house test examples and regard the remaining 8,500 ones as our in-house training examples. Almost all KG-augmented models achieve performance gain over vanilla pre-trained

<sup>5</sup>We use mean pooling for 1-hop RN and attentive pooling for 2-hop RN (detailed in Appendix C).

Methods	Dev (%)	Test (%)
BERT (w/o KG) <sup>†</sup>	60.4	60.4
BERT Multi-Task (w/o KG) <sup>†</sup>	-	63.8
GapQA <sup>†</sup> (Khot et al., 2019)	-	59.4 (±1.30)
RoBERTa-Large (w/o KG)	66.76 (±1.14)	64.80 (±2.37)
+ RGCN	64.65 (±1.96)	62.45 (±1.57)
+ GconAttn	66.85 (±1.82)	64.75 (±1.48)
+ RN (1-hop)	64.85 (±1.11)	63.65 (±2.31)
+ RN (2-hop)	67.00 (±0.71)	65.20 (±1.18)
+ MHGRN ( $K = 3$ )	<b>68.10</b> (±1.02)	<b>66.85</b> (±1.19)

Table 5: **Dev and Test accuracy on OpenbookQA.** <sup>†</sup> indicates reported results on the leaderboard.

Methods	IHdev-Acc. (%)
MHGRN ( $K = 3$ )	<b>74.45</b> (±0.10)
- Type-specific transformation (§4.1)	73.16 (±0.28)
- Structured relational attention (§4.2)	73.26 (±0.31)
- Relation type attention (§4.2)	73.55 (±0.68)
- Node type attention (§4.2)	73.92 (±0.65)

Table 6: **Ablation study on model components** (removing one component each time) using ROBERTA-LARGE as the text encoder. We report the IHdev accuracy on CommonsenseQA.

LMs, demonstrating the value of external knowledge on this dataset. Additionally, we evaluate our MHGRN (with the text encoder being ROBERTA-LARGE) on official split (Table 4) for fair comparison with other methods on leaderboard, in both single-model setting and ensemble-model setting. In both cases, we achieve state-of-the-art performances across all existing models.

For OpenbookQA (Table 5), we use official split and build models with ROBERTA-LARGE as text encoder. Overall, external knowledge still brings benefit to this task. Our model surpasses all baselines, with an absolute increase of ~2% on Test. As we seek to evaluate the performance of models in terms of encoding external knowledge graphs, we did not compare with other models based on document retrieval and reading comprehension. Note that other submissions on OBQA leaderboard usually were fine-tuned on external question answering datasets or using large corpus via information retrieval. As we seek to systematically examine knowledge-aware reasoning methods by reasoning over interpretable structures, we do not include comparisons with models that are beyond the scope. Other fine-tuned PTLMs can be simply incorporated as well whenever they are available.

## 6.2 Performance Analysis

**Ablation Study on Model Components.** We assess the impact of our models’ components, shown in Table 6. Disabling type-specific transformation results in  $\sim 1.3\%$  drop in performance, demonstrating the need for distinguishing node type for QA tasks. Our structured relational attention mechanism is also critical, with its two sub-components contributing almost equally.

**Impact of the Amount of Training Data.** We use different fractions of training data of CommonsenseQA and report results of fine-tuning text encoders alone and jointly training text encoder and graph encoder in Fig. 5. Regardless of training data fraction, our model shows consistently more performance improvement over knowledge-agnostic fine-tuning compared with the other graph encoding methods, indicating MHGRN’s complementary strengths to text encoders.

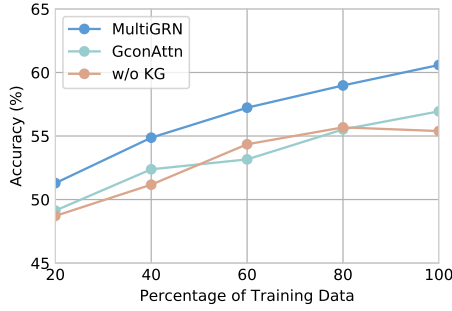


Figure 5: Performance change (accuracy in %) w.r.t. the amounts of training data on CommonsenseQA IHT-test set (same as Lin et al. (2019)).

**Impact of Number of Hops ( $K$ ).** We investigate the impact of hyperparameter  $K$  for MHGRN by its performance on CommonsenseQA (shown in Fig. 6). The increase of  $K$  continues to bring benefit until  $K = 4$ . However, performance begins to drop when  $K > 3$ . This might be attributed to exponential noise in longer relational paths in knowledge graph.

## 6.3 Model Scalability

Fig. 7 presents the computation cost of MultiRGN and RGCN (measured by training time). The computation cost of both models grow linearly w.r.t. to  $K$ . Although the theoretical complexity of MultiRGN is  $m$  times that of RGCN, the ratio of their empirical cost only approaches 2, demonstrating that our model can be parallelized.

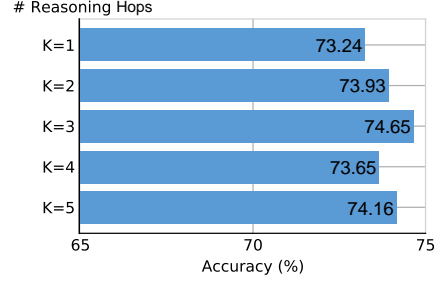


Figure 6: **Effect of  $K$  in MHGRN.** We show IHDev accuracy of MHGRN on CommonsenseQA using the data split of Lin et al. (2019) with different # hops.

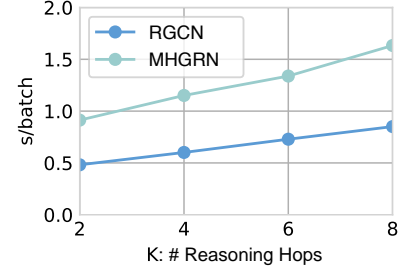


Figure 7: **Analysis of model scalability.** Comparison of per-batch training efficiency w.r.t. # hops  $K$ .

## 6.4 Model Interpretability

We can analyze our model’s reasoning process by decoding the reasoning path using the method described in §4.5. Fig. 8 shows two examples from CommonsenseQA, where our model correctly answers the questions and provides reasonable path evidences. In the example on the left, the model links question entities and answer entity in a chain to support reasoning, while the example on the right provides a case where our model leverage unmentioned entities to bridge the reasoning gap between question entity and answer entities, in a way that is coherent with the implied relation between CHAPEL and the desired answer in the question.

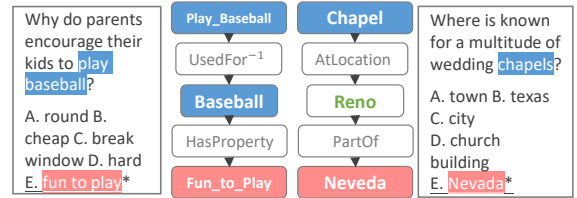


Figure 8: **Case study on model interpretability.** We present two sampled questions from CommonsenseQA with the reasoning paths output by MHGRN.

## 6.5 Potential Compatibility with Other Methods

In theory, our approach is naturally compatible with the methods that utilize textual knowledge



or extra data (such as leaderboard methods in Table 4), because in our paradigm the encoding of textual statement and graph are structurally-decoupled (Fig. 3). We can take, for example, the fine-tuned RoBERTa+KE<sup>6</sup> system as our text encoder and leave the rest of our model architecture unchanged.

## 7 Related Work

**Knowledge-Aware Methods for NLP** Various work have investigated the potential to empower NLP models with external knowledge. Many attempt to extract *structured* knowledge, either in the form of nodes (Yang and Mitchell, 2017; Wang et al., 2019), triples (Weissenborn et al., 2017; Mihaylov and Frank, 2018), paths (Bauer et al., 2018; Kundu et al., 2019; Lin et al., 2019), or sub-graphs (Li and Clark, 2015), and encode them to augment textual understanding.

Recent success of pre-trained LMs motivates many (Pan et al., 2019; Ye et al., 2019; Zhang et al., 2018; Li et al., 2019; Banerjee et al., 2019) to probe LMs’ potential as latent knowledge bases. This line of work turn to *textual* knowledge (e.g. Wikipedia) to directly impart knowledge to pre-trained LMs. They generally fall into two paradigms: 1) Fine-tuning LMs on large-scale general-domain datasets (e.g. RACE (Lai et al., 2017)) or on knowledge-rich text. 2) Providing LMs with evidence via information retrieval techniques. However, these models cannot provide explicit reasoning and evidence, thus hardly trustworthy. They are also subject to the availability of in-domain datasets and maximum input token of pre-trained LMs.

**Neural Graph Encoding** Graph Attention Networks (GAT) (Velickovic et al., 2018) incorporates attention mechanism in feature aggregation, RGCN (Schlichtkrull et al., 2018) proposes relational message passing which makes it applicable to multi-relational graphs. However they only perform single-hop message passing and cannot be interpreted at path level. Other work (Abu-El-Haija et al., 2019; Nikolentzos et al., 2019) aggregate for a node its  $K$ -hop neighbors based on node-wise distances, but they are designed for non-relational graphs. MHGRN addresses these issues by reasoning on multi-relational graphs and being interpretable via maintaining paths as reasoning chains.

## 8 Conclusion

We present a principled, scalable method, MHGRN, that can leverage general knowledge by multi-hop reasoning over interpretable structures (e.g. ConceptNet). The proposed MHGRN generalizes and combines the advantages of GNNs and path-based reasoning models. It explicitly performs multi-hop relational reasoning and is empirically shown to outperform existing methods with superior scalability and interpretability. Our extensive experiments systematically compare MHGRN and other existing methods on knowledge-aware methods. Particularly, we achieve the state-of-the-art performance on the CommonsenseQA dataset.

## References

- Sami Abu-El-Haija, Bryan Perozzi, Amol Kapoor, Nazanin Alipourfard, Kristina Lerman, Hrayr Harutyunyan, Greg Ver Steeg, and Aram Galstyan. 2019. [Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing](#). In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 21–29. PMLR.
- Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2019. [Careful selection of knowledge to solve open book question answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6120–6129, Florence, Italy. Association for Computational Linguistics.
- Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. [Commonsense for generative multi-hop question answering tasks](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4220–4230. Association for Computational Linguistics.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the AI2 reasoning challenge](#). *CoRR*, abs/1803.05457.
- 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.

<sup>6</sup><https://github.com/jose77/csqa/>

- 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.
- Tushar Khot, Ashish Sabharwal, and Peter Clark. 2019. [What’s missing: A knowledge gap guided approach for multi-hop question answering](#). 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 2814–2828, Hong Kong, China. Association for Computational Linguistics.
- Thomas N. Kipf and Max Welling. 2017. [Semi-supervised classification with graph convolutional networks](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Souvik Kundu, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2019. [Exploiting explicit paths for multi-hop reading comprehension](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2737–2747, Florence, Italy. Association for Computational Linguistics.
- John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. [Conditional random fields: Probabilistic models for segmenting and labeling sequence data](#). In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, Williams College, Williamstown, MA, USA, June 28 - July 1, 2001, pages 282–289. Morgan Kaufmann.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. [RACE: Large-scale ReAding comprehension dataset from examinations](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.
- Zhen-Zhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. [Albert: A lite bert for self-supervised learning of language representations](#). *ArXiv*, abs/1909.11942.
- Shiyang Li, Jianshu Chen, and Dian Yu. 2019. [Teaching pretrained models with commonsense reasoning: A preliminary kb-based approach](#). *CoRR*, abs/1909.09743.
- Yang Li and Peter Clark. 2015. [Answering elementary science questions by constructing coherent scenes using background knowledge](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2007–2012, Lisbon, Portugal. Association for Computational Linguistics.
- 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. 2019a. [On the variance of the adaptive learning rate and beyond](#). *CoRR*, abs/1908.03265.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2019. [Graph-based reasoning over heterogeneous external knowledge for commonsense question answering](#). *CoRR*, abs/1909.05311.
- Kaixin Ma, Jonathan Francis, Quanyang Lu, Eric Nyberg, and Alessandro Oltramari. 2019. [Towards generalizable neuro-symbolic systems for commonsense question answering](#). In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 22–32, Hong Kong, China. Association for Computational Linguistics.
- 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.
- Todor Mihaylov and Anette Frank. 2018. [Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832, Melbourne, Australia. Association for Computational Linguistics.
- Giannis Nikolentzos, George Dasoulas, and Michalis Vazirgiannis. 2019. [k-hop graph neural networks](#). *CoRR*, abs/1907.06051.
- Xiaoman Pan, Kai Sun, Dian Yu, Heng Ji, and Dong Yu. 2019. [Improving question answering with external knowledge](#). *CoRR*, abs/1902.00993.

- 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, 4-9 December 2017, Long Beach, CA, USA*, pages 4967–4976.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. [Social IQa: Commonsense reasoning about social interactions](#). 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.
- Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. [Modeling relational data with graph convolutional networks](#). In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, 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.
- Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019. [Improving machine reading comprehension with general reading strategies](#). 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 2633–2643, Minneapolis, Minnesota. 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, 4-9 December 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.
- Shuohang Wang and Jing Jiang. 2016. [Learning natural language inference with LSTM](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1442–1451, San Diego, California. Association for Computational Linguistics.
- Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael Witbrock. 2019. [Improving natural language inference using external knowledge in the science questions domain](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7208–7215. AAAI Press.
- Dirk Weissenborn, Tomáš Kočiský, and Chris Dyer. 2017. [Dynamic integration of background knowledge in neural nlu systems](#). *arXiv preprint arXiv:1706.02596*.
- An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019. [Enhancing pre-trained language representations with rich knowledge for machine reading comprehension](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2346–2357, Florence, Italy. Association for Computational Linguistics.
- Bishan Yang and Tom Mitchell. 2017. [Leveraging knowledge bases in LSTMs for improving machine reading](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1436–1446, Vancouver, Canada. Association for Computational Linguistics.
- Zhi-Xiu Ye, Qian Chen, Wen Wang, and Zhen-Hua Ling. 2019. [Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models](#). *CoRR*, abs/1908.06725.
- Yuyu Zhang, Hanjun Dai, Kamil Toraman, and Le Song. 2018. [Kg<sup>2</sup>: Learning to reason science exam questions with contextual knowledge graph embeddings](#). *CoRR*, abs/1805.12393.
- Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. [Freelb: Enhanced adversarial training for natural language understanding](#). In *International Conference on Learning Representations*.

## A Implementation Details

	CommonsenseQA	OpenbookQA
BERT-BASE	$3 \times 10^{-5}$	-
BERT-LARGE	$2 \times 10^{-5}$	-
ROBERTA-LARGE	$1 \times 10^{-5}$	$1 \times 10^{-5}$

Table 7: Learning rate for text encoders on different datasets.

	CommonsenseQA	OpenbookQA
RN	$3 \times 10^{-4}$	$3 \times 10^{-4}$
RGCN	$1 \times 10^{-3}$	$1 \times 10^{-3}$
GconAttn	$3 \times 10^{-4}$	$1 \times 10^{-4}$
KV-Memory	$1 \times 10^{-3}$	$1 \times 10^{-3}$
MultiGRN	$1 \times 10^{-3}$	$1 \times 10^{-3}$

Table 8: Learning rate for graph encoders on different datasets.

Our models are implemented in *PyTorch*. We use *cross-entropy* loss and *RAdam* (Liu et al., 2019a) optimizer. We find it beneficial to use separate learning rates for the text encoder and the graph encoder. We tune learning rates for text encoders and graph encoders on two datasets. We first fine-tune ROBERTA-LARGE, BERT-LARGE, BERT-BASE on *CommonsenseQA* and ROBERTA-LARGE on *OpenbookQA* respectively, and choose a dataset-specific learning rate from  $\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 6 \times 10^{-5}, 1 \times 10^{-4}\}$  for each text encoder, based on the best performance on development set, as listed in Table 7. We report the performance of these fine-tuned text encoders and also adopt their dataset-specific optimal learning rates in joint training with graph encoders. For models that involve KG, the learning rate of their graph encoders<sup>7</sup> are chosen from  $\{1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}, 3 \times 10^{-3}\}$ , based on their best development set performance with ROBERTA-LARGE as the text encoder. We report the optimal learning rates for graph encoders in Table 8. In training, we set the maximum input sequence length to text encoders to 64, batch size to 32, and perform early stopping.

For the input node features, we first use templates to turn knowledge triples in *CommonsenseQA* into sentences and feed them into pre-trained BERT-LARGE, obtaining a sequence of

<sup>7</sup>KV-memory actually encode triples to enhance token representations of  $s$  instead of encoding graph alone. But for simplicity, we also refer to it as a graph encoder.

token embeddings from the last layer of BERT-LARGE for each triple. For each entity in *CommonsenseQA*, we perform mean pooling over the tokens of the entity’s occurrences across all the sentences to form a 1024d vector as its corresponding node feature. We use this set of features for all our implemented models.

We use 2-layer RGCN and single-layer MultiGRN across our experiments.

## B Dynamic Programming Algorithm for Eq. 7

To show that multi-hop message passing can be computed in linear time, we observe that Eq. 7 can be re-written in matrix form:

$$\begin{aligned} Z^k &= (D^k)^{-1} \sum_{(r_1, \dots, r_k) \in \mathcal{R}^k} \beta(r_1, \dots, r_k, s) \\ &\quad \cdot G A_{r_k} \cdots A_{r_1} F X W_{r_1}^{1\top} \cdots W_{r_k}^{k\top} \\ &\quad \cdot W_0^{k+1\top} \cdots W_0^{K\top} \quad (1 \leq k \leq K), \end{aligned} \quad (13)$$

where  $G = \text{diag}(\exp([g(\phi(v_1), s), \dots, g(\phi(v_n), s)]))$  ( $F$  is similarly defined),  $A_r$  is the adjacency matrix for relation  $r$  and  $D^k$  is defined as follows:

$$\begin{aligned} D^k &= \text{diag} \left( \sum_{(r_1, \dots, r_k) \in \mathcal{R}^k} \beta(r_1, \dots, r_k, s) \right. \\ &\quad \left. \cdot G A_{r_k} \cdots A_{r_1} F X \right) \quad (1 \leq k \leq K) \end{aligned} \quad (14)$$

Using this matrix formulation, we can compute Eq. 7 using dynamic programming:

## C Formal Definition of $K$ -hop RN

**Definition 1** (*K-hop Relation Network*) A multi-hop relation network is a function that maps a multi-relational graph to a fixed size vector:

$$\begin{aligned} KHopRN(\mathcal{G}; \tilde{W}, \tilde{E}, \tilde{H}) &= \sum_{k=1}^K \sum_{\substack{(j, r_1, \dots, r_k, i) \in \Phi_k \\ j \in \mathcal{Q} \quad i \in \mathcal{A}}} \tilde{\beta}(j, r_1, \dots, r_k, i) \cdot \tilde{W}(\tilde{h}_j \oplus (\tilde{e}_{r_1} \circ \cdots \circ \tilde{e}_{r_k}) \oplus \tilde{h}_i), \\ &\quad (15) \end{aligned}$$

where  $\circ$  denotes element-wise product and  $\tilde{\beta}(\dots) = 1/(K|A| \cdot |\{(j', \dots, i) \in \mathcal{G} \mid j' \in \mathcal{Q}\}|)$  defines the pooling weights.



## D Expressing K-hop RN with MultiGRN

**Theorem 1** Given any  $\tilde{W}, \tilde{E}, \tilde{H}$ , there exists a parameter setting such that the output of the model becomes  $K\text{HopRN}(\mathcal{G}; \tilde{W}, \tilde{E}, \tilde{H})$  for arbitrary  $\mathcal{G}$ .

*Proof.* Suppose  $\tilde{W} = [\tilde{W}_1, \tilde{W}_2, \tilde{W}_3]$ , where  $\tilde{W}_1, \tilde{W}_3 \in \mathbb{R}^{d_3 \times d_1}, \tilde{W}_2 \in \mathbb{R}^{d_3 \times d_2}$ . For *MultiGRN*, we set the parameters as follows:  $\mathbf{H} = \tilde{H}, \mathbf{U}_* = [I; \mathbf{0}] \in \mathbb{R}^{(d_1+d_2) \times d_1}, \mathbf{b}_* = [\mathbf{0}, \mathbf{1}]^\top \in \mathbb{R}^{d_1+d_2}, \mathbf{W}_r^t = \text{diag}(\mathbf{1} \oplus \tilde{e}_r) \in \mathbb{R}^{(d_1+d_2) \times (d_1+d_2)} (r \in \mathcal{R}, 1 \leq t \leq K), \mathbf{V} = \tilde{W}_3 \in \mathbb{R}^{d_3 \times d_1}, \mathbf{V}' = [\tilde{W}_1, \tilde{W}_2] \in \mathbb{R}^{d_3 \times (d_1+d_2)}$ . We disable the relation type attention module and enable message passing only from  $\mathcal{Q}$  to  $\mathcal{A}$ . By further choosing  $\sigma$  as the identity function and performing pooling over  $\mathcal{A}$ , we observe that the output of *MultiGRN* becomes:

$$\begin{aligned}
& \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \mathbf{h}_i' \\
&= \frac{1}{|\mathcal{A}|} (\mathbf{V} \mathbf{h}_i + \mathbf{V}' \mathbf{z}_i) \\
&= \frac{1}{K|\mathcal{A}|} \sum_{k=1}^K (\mathbf{V} \mathbf{h}_i + \mathbf{V}' \mathbf{z}_i^k) \\
&= \sum_{k=1}^K \sum_{\substack{(j, r_1, \dots, r_k, i) \in \Phi_k \\ j \in \mathcal{Q}, i \in \mathcal{A}}} \tilde{\beta}(j, r_1, \dots, r_k, i) (\mathbf{V} \mathbf{h}_i + \\
&\quad \mathbf{V}' \mathbf{W}_{r_k}^k \dots \mathbf{W}_{r_1}^1 \mathbf{x}_j) \\
&= \sum_{k=1}^K \sum_{\substack{(j, r_1, \dots, r_k, i) \in \Phi_k \\ j \in \mathcal{Q}}} \tilde{\beta}(\dots) (\mathbf{V} \mathbf{h}_i + \mathbf{V}' \mathbf{W}_{r_k}^k \\
&\quad \dots \mathbf{W}_{r_1}^1 \mathbf{U}_{\phi(j)} \mathbf{h}_j + \mathbf{V}' \mathbf{W}_{r_k}^k \dots \mathbf{W}_{r_1}^1 \mathbf{b}_{\phi(j)}) \\
&= \sum_{k=1}^K \sum_{\substack{(j, r_1, \dots, r_k, i) \in \Phi_k \\ j \in \mathcal{Q}, i \in \mathcal{A}}} \tilde{\beta}(\dots) (\tilde{W}_3 \mathbf{h}_i + \tilde{W}_1 \mathbf{h}_j \\
&\quad + \tilde{W}_2 (\tilde{e}_{r_1} \circ \dots \circ \tilde{e}_{r_k})) \\
&= \sum_{k=1}^K \sum_{\substack{(j, r_1, \dots, r_k, i) \in \Phi_k \\ j \in \mathcal{Q}, i \in \mathcal{A}}} \tilde{\beta}(\dots) \tilde{W} (\tilde{h}_j \oplus \\
&\quad (\tilde{e}_{r_1} \circ \dots \circ \tilde{e}_{r_k}) \oplus \tilde{h}_i) \\
&= \text{RN}(\mathcal{G}; \tilde{W}, \tilde{E}, \tilde{H})
\end{aligned}$$

(16)

---

**Algorithm 1** Dynamic programming algorithm for multi-hop message passing.

---

**Input:**  $\mathbf{s}, \mathbf{X}, \mathbf{A}_r (1 \leq r \leq m), \mathbf{W}_r^t (r \in \mathcal{R}, 1 \leq t \leq k), \mathbf{F}, \mathbf{G}, \delta, \tau$   
**Output:**  $\mathbf{Z}$

---

```

1:  $\hat{\mathbf{W}}^K \leftarrow \mathbf{I}$ 
2: for  $k \leftarrow K-1$  to 1 do
3:    $\hat{\mathbf{W}}^k \leftarrow \mathbf{W}_0^{k+1} \hat{\mathbf{W}}^{k+1}$ 
4: end for
5: for  $r \in \mathcal{R}$  do
6:    $\mathbf{M}_r \leftarrow \mathbf{F} \mathbf{X}$ 
7: end for
8: for  $k \leftarrow 1$  to  $K$  do
9:   if  $k > 1$  then
10:    for  $r \in \mathcal{R}$  do
11:       $\mathbf{M}_r' \leftarrow e^{\delta(r, \mathbf{s})} \mathbf{A}_r \sum_{r' \in \mathcal{R}} e^{\tau(r', r, \mathbf{s})} \mathbf{M}_{r'} \mathbf{W}_r^{k \top}$ 
12:    end for
13:    for  $r \in \mathcal{R}$  do
14:       $\mathbf{M}_r \leftarrow \mathbf{M}_r'$ 
15:    end for
16:  else
17:    for  $r \in \mathcal{R}$  do
18:       $\mathbf{M}_r \leftarrow e^{\delta(r, \mathbf{s})} \cdot \mathbf{A}_r \mathbf{M}_r \mathbf{W}_r^{k \top}$ 
19:    end for
20:  end if
21:   $\mathbf{Z}^k \leftarrow \mathbf{G} \sum_{r \in \mathcal{R}} \mathbf{M}_r \hat{\mathbf{W}}^k$ 
22: end for
23: Replace  $\mathbf{W}_r^t (0 \leq r \leq m, 1 \leq t \leq k)$  with identity matrices and  $\mathbf{X}$  with  $\mathbf{1}$  and re-run line 1 - line 19 to compute  $\mathbf{d}^1, \dots, \mathbf{d}^K$ 
24: for  $k \leftarrow 1$  to  $K$  do
25:    $\mathbf{Z}^k \leftarrow (\text{diag}(\mathbf{d}^k))^{-1} \mathbf{Z}^k$ 
26: end for
27: return  $\mathbf{Z}^1, \mathbf{Z}^2, \dots, \mathbf{Z}^k$ 

```

---