Hypergraph Transformer: Weakly-Supervised Multi-hop Reasoning for Knowledge-based Visual Question Answering

Yu-Jung Heo[1,4]**, Eun-Sol Kim**[2]**, Woo Suk Choi**[1]**, and Byoung-Tak Zhang**[1,3]
[1] Seoul National University [2] Department of Computer Science, Hanyang University
[3] AI Institute (AIIS), Seoul National University [4] Surromind
yjheo@bi.snu.ac.kr, eunsolkim@hanyang.ac.kr, {wschoi, btzhang}@bi.snu.ac.kr

Abstract

Knowledge-based visual question answering (QA) aims to answer a question which requires visually-grounded external knowledge beyond image content itself. Answering complex questions that require multi-hop reasoning under weak supervision is considered as a challenging problem since i) no supervision is given to the reasoning process and ii) high-order semantics of multi-hop knowledge facts need to be captured. In this paper, we introduce a concept of hypergraph to encode high-level semantics of a question and a knowledge base, and to learn high-order associations between them. The proposed model, Hypergraph Transformer, constructs a question hypergraph and a query-aware knowledge hypergraph, and infers an answer by encoding inter-associations between two hypergraphs and intra-associations in both hypergraph itself. Extensive experiments on two knowledge-based visual QA and two knowledge-based textual QA demonstrate the effectiveness of our method, especially for multi-hop reasoning problem. Our source code is available at https://github.com/yujungheo/kbvqa-public.

1 Introduction

Visual question answering (VQA) is a semantic reasoning task that aims to answer questions about visual content depicted in images (Antol et al., 2015; Zhu et al., 2016; Hudson and Manning, 2019), and has become one of the most active areas of research with advances in natural language processing and computer vision. Recently, researches for VQA have advanced, from inferring visual properties on entities in a given image, to inferring commonsense or world knowledge about those entities (Wang et al., 2017, 2018; Marino et al., 2019; Shah et al., 2019; Zellers et al., 2019).

In this paper, we focus on the task which is called knowledge-based visual question answering, where a massive number of knowledge facts from a general knowledge base (KB) is given with an image-question pair. To answer the given question as shown in Figure 1, a model should understand the semantics of the given question, link visual entities appearing in the given image to the KB, extract a number of evidences from the KB and predict an answer by aggregating semantics of both the question and the extracted evidences. Following these, there are two fundamental challenges in this task. i) To answer a complex question, multi-hop reasoning over multiple knowledge evidences is necessary. ii) Learning a complex reasoning process is difficult especially in a condition where only QA is provided without extra supervision on how to capture any evidence from the KB and infer based on them. That is, the model should learn which knowledge facts to be attended to and how to combine them to infer the correct answer on its own. Following the previous work (Zhou et al., 2018), we call this setting *under weak supervision*.

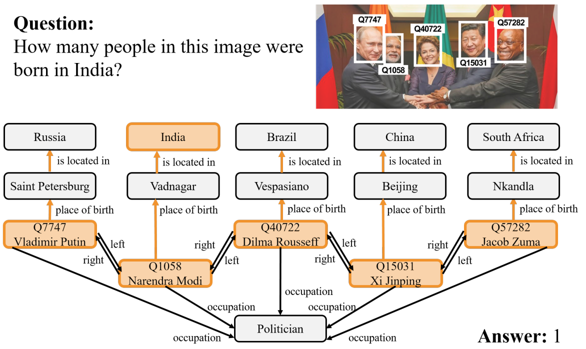


Figure 1: An example of knowledge-based visual question answering. The rectangles and arrows between the rectangles represent the entities and relations from KB. To answer the given question, the multiple reasoning evidences (marked as orange) are required.

Under weak supervision, previous studies proposed memory-based methods (Narasimhan and Schwing, 2018; Shah et al., 2019) and graph-based methods (Narasimhan et al., 2018; Zhu et al., 2020) to learn to selectively focus on necessary pieces of knowledge. The memory-based methods represent knowledge facts in a form of memory and calculate soft attention scores of each memory with respect to a question. Then, it infers an answer by attending to knowledge evidence with high attention scores. On the other hand, to explicitly consider relational structure between knowledge facts, graph-based methods construct a query-aware knowledge graph by retrieving facts from KB and perform graph reasoning for a question. These methods mainly adopt an iterative message passing process to propagate information between adjacent nodes in the graph. However, it is difficult to capture multi-hop relationships containing long-distance nodes from the graph due to the well-known over-smoothing problem, where repetitive message passing process to propagate information across long distance makes features of connected nodes too similar and undiscriminating (Li et al., 2018; Wang et al., 2020).

To address the above limitation, we propose a novel method, Hypergraph Transformer, which exploits hypergraph structure to encode multi-hop relationships and transformer-based attention mechanism to learn to pay attention to important knowledge evidences for a question. We construct a question hypergraph and a knowledge hypergraph to explicitly encode high-order semantics present in the question and each knowledge fact, and capture multi-hop relational knowledge facts effectively. Then, we perform hyperedge matching between the two hypergraphs by leveraging transformer-based attention mechanism. We argue that introducing the concept of hypergraph is powerful for multi-hop reasoning problem in that it can encode high-order semantics without the constraint of length and learn cross-modal high-order associations.

The main contributions of this paper can be summarized as follows. i) We propose Hypergraph Transformer which enhances multi-hop reasoning ability by encoding high-order semantics in the form of a hypergraph and learning inter- and intra-high-order associations in hypergraphs using the attention mechanism. ii) We conduct extensive experiments on two knowledge-based VQA datasets (KVQA and FVQA) and two knowledge-based textual QA datasets (PQ and PQL) and show superior performances on all datasets, especially multi-hop reasoning problem. iii) We qualitatively observe that Hypergraph Transformer performs robust inference by focusing on correct reasoning evidences under weak supervision.

2 Related Work

Knowledge-based visual question answering (Wang et al., 2017, 2018; Shah et al., 2019; Marino et al., 2019; Sampat et al., 2020) proposed benchmark datasets for knowledge-based visual question answering that requires reasoning about an image on the basis of facts from a large-scale knowledge base (KB) such as Freebase (Bollacker et al., 2008) or DBPedia (Auer et al., 2007). To solve the task, two pioneering studies (Wang et al., 2017, 2018) suggested logical parsing-based methods which convert a question to a KB logic query using predefined query templates and execute the generated query on KB for searching an answer. Since then information retrieval-based methods which retrieve knowledge facts associated with a question and conduct semantic matching between the facts and the question are introduced. (Narasimhan and Schwing, 2018; Shah et al., 2019) proposed memory-based methods that represent knowledge facts in the form of memory and calculate soft attention scores of the memory with a question. (Narasimhan et al., 2018; Zhu et al., 2020) represented the retrieved facts as a graph and performed graph reasoning through message passing scheme utilizing graph convolution. However, these methods are complicated to encode inherent high-order semantics and multi-hop relationships present in the knowledge graph. Therefore, we introduce a concept of hypergraph and propose transformer-based attention mechanism over hypergraphs.

Multi-hop knowledge graph reasoning is a process of sequential reasoning based on multiple evidences of a knowledge graph, and has been broadly used in various downstream tasks such as question answering (Lin et al., 2019; Saxena et al., 2020; Han et al., 2020b,a; Yadati et al., 2021), or knowledge-enhanced text generation (Liu et al., 2019; Moon et al., 2019; Ji et al., 2020). Recent researches have introduced the concept of hypergraph for multi-hop graph reasoning (Kim et al., 2020; Han et al., 2020b,a; Yadati et al., 2019, 2021; Sun et al., 2020). These models have a similar motivation to the Hypergraph Transformer proposed in this paper, but core operations are vastly different. These models mainly update node representations in the hypergraph through a message passing process using graph convolution operation. On the

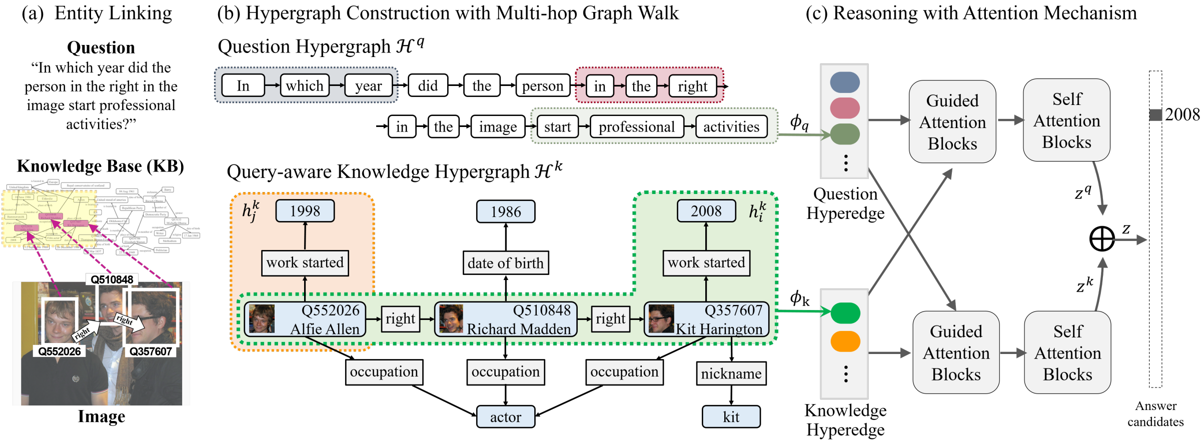


Figure 2: The overview of Hypergraph Transformer. (a) Entity linking module links concepts from query (a given image and a question) to KB. (b) Query-aware knowledge hypergraph $\mathcal{H}^k$ and question hypergraph $\mathcal{H}^q$ are constructed by multi-hop graph walk. (c) Two hyperedge sets are fed into the guided-attention and self-attention blocks to learn inter- and intra-association in them. The joint representation is used to predict an answer.

contrary, our method update node representations via hyperedge matching of hypergraphs instead of message passing scheme. We argue that this update process effectively learns the high-order semantics inherent in each hypergraph and the high-order associations between two hypergraphs.

3 Method

3.1 Notation

To capture high-order semantics inherent in the knowledge sources, we adopt the concept of hypergraph. Formally, directed hypergraph $\mathcal{H} = \{\mathcal{V}, \mathcal{E}\}$ is defined by a set of nodes $\mathcal{V} = \{v_1, ..., v_{|\mathcal{V}|}\}$ and a set of hyperedges $\mathcal{E} = \{h_1, ..., h_{|\mathcal{E}|}\}$. Each node is represented as a w-dimensional embedding vector, i.e., $v_i \in \mathbb{R}^w$. Each hyperedge connects an arbitrary number of nodes and has partial order itself, i.e., $h_i = \{v'_1 \preceq ... \preceq v'_l\}$ where $\mathcal{V}' = \{v'_1, ..., v'_l\}$ is a subset of $\mathcal{V}$ and $\preceq$ is a binary relation which denotes an element (v'_i) precedes the other (v'_j) in the ordering when $v'_i \preceq v'_j$. A hyperedge is flexible to encode different kinds of semantics in the underlying graph without the constraint of length.

3.2 Entity linking

As shown in Figure 2(a), entity linking module first links concepts from query (a given image-question pair) to knowledge base. We detect visual concepts (e.g., objects, attributes, person names) in a given image and named entities in a given question. The semantic labels of visual concepts or named entities are then linked with knowledge entities in the knowledge base using exact keyword matching.

3.3 Hypergraph construction

Query-aware knowledge hypergraph A knowledge base (KB), a vast amount of general knowledge facts, contains not only knowledge facts required to answer a given question but also unnecessary knowledge facts. Thus, we construct a query-aware knowledge hypergraph $\mathcal{H}^k = \{\mathcal{V}^k, \mathcal{E}^k\}$ to extract related information for answering a given question. It consists of a node set $\mathcal{V}^k$ and hyperedge set $\mathcal{E}^k$, which represent a set of entities in knowledge facts and a set of hyperedges, respectively. Each hyperedge connects the subset of vertices $\mathcal{V}'^k \subset \mathcal{V}^k$.

We consider a huge number of knowledge facts in the KB as a huge knowledge graph, and construct a hypergraph by traversing the knowledge graph. Such traversal, called graph walk, starts from the node linked from the previous module (see section 3.2) and considers all entity nodes associated with the start node. We define a triplet as a basic unit of graph walk to preserve high-order semantics inherent in knowledge graph, i.e., every single graph walk contains three nodes *{head, predicate, tail}*, rather than having only one of these three nodes. In addition to the triplet-based graph walks, a multi-hop graph walk is proposed to encode multiple relational facts that are interconnected. Multi-hop graph walk connects multiple facts by setting the arrival node (*tail*) of the preceding walk as the starting (*head*) node of the next walk, thus, n-hop graph walk combines n facts as a hyperedge.

Question hypergraph We transform a question sentence into a question hypergraph $\mathcal{H}^q$ consisting of a node set $\mathcal{V}^q$ and a hyperedge set $\mathcal{E}^q$. We assume that each word unit (a word or named entity) of the question is defined as a node, and has edges to adjacent nodes. For question hypergraph, each word unit is used as a start node of a graph walk. The multi-hop graph walk is conducted in the same manner as the knowledge hypergraph. A n-gram phrase is considered as a hyperedge in the question hypergraph (see Figure 2(b)).

3.4 Reasoning with attention mechanism

To consider high-order associations between knowledge and question, we devise structural semantic matching between the query-aware knowledge hypergraph and the question hypergraph. We introduce an attention mechanism over two hypergraphs based on guided-attention (Tsai et al., 2019) and self-attention (Vaswani et al., 2017). As shown in Figure 2(c), the guided-attention blocks are introduced to learn correlations between knowledge hyperedges and question hyperedges by inter-attention mechanism, and then intra-relationships of in knowledge or question hyperedges are trained with the following self-attention blocks. The details of two modules, guided-attention blocks and self-attention blocks, are described as below. Note that we use Q, K, and V for query, key, value, and q, k as subscripts to represent question and knowledge, respectively.

Guided-attention To learn inter-association between two hypergraphs, we first embed a knowledge hyperedge and a question hyperedge as follows: $e^k = \phi_k \circ f_k(h^k) \in \mathbb{R}^d, e^q = \phi_q \circ f_q(h^q) \in \mathbb{R}^d$ where $h^{[\cdot]}$ is a hyperedge in $\mathcal{E}^{[\cdot]}$. Here, $f_{[\cdot]}$ is a hyperedge embedding function and $\phi_{[\cdot]}$ is a linear projection function. The design and implementation of $f_{[\cdot]}$ are not constrained (e.g., any pooling operation or any learnable neural networks), but we use a simple concatenation operation of node representations in a hyperedge as $f_{[\cdot]}$. The representations of hyperedges in the same hypergraph (e.g., e^k, e^q) are packed together into a matrix E^k and E^q.

We define the knowledge hyperedges E^k and the question hyperedges E^q as a query and key-value pairs, respectively. We set a query $Q_k = E^k W_{Q_k}$, a key $K_q = E^q W_{K_q}$, and a value $V_q = E^q W_{V_q}$, where all projection matrices $W_{[\cdot]} \in \mathbb{R}^{d \times d_v}$ are learnable parameters. Then, scaled dot product attention using the query, key, and value is calculated as $\mathrm{Attention}(Q_k, K_q, V_q) = \mathrm{softmax}(\frac{Q_k K_q^T}{\sqrt{d_v}})V_q$ where d_v is the dimension of the query and the key vector. In addition, the guided-attention which uses the question hyperedges as query and the knowledge hyperedges as key-value pairs is performed in a similar manner: $\mathrm{Attention}(Q_q, K_k, V_k)$.

Self-attention The only difference between guided-attention and self-attention is that the same input is used for both query and key-value within self-attention. For example, we set query, key, and value based on the knowledge hyperedges E_k, and the self-attention for knowledge hyperedges is conducted by $\mathrm{Attention}(Q_k, K_k, V_k)$. For question hyperedges E_q, self-attention is performed in a similar manner: $\mathrm{Attention}(Q_q, K_q, V_q)$.

Following the standard structure of the transformer, we build up guided-attention block and self-attention block where each block consists of each attention operation with layer normalization, residual connection, and a single feed-forward layer. By passing the guided-attention blocks and self-attention blocks sequentially, representations of knowledge hyperedges and question hyperedges are updated and finally aggregated to single vector representation as $z_k \in \mathbb{R}^{d_v}$ and $z_q \in \mathbb{R}^{d_v}$, respectively.

3.5 Answer predictor

To predict an answer, we first concatenate the representation z_k and z_q obtained from the attention blocks and feed into a single feed-forward layer (i.e., $\mathbb{R}^{2d_v} \mapsto \mathbb{R}^w$) to make a joint representation z. We then consider two types of answer predictor: multi-layer perceptron and similarity-based answer predictor. Multi-layer perceptron as an answer classifier $p = \psi(z)$ is a prevalent for visual question answering problems. For similarity-based answer, we calculate a dot product similarity $p = zC^T$ between z and answer candidate set $C \in \mathbb{R}^{|\mathcal{A}| \times w}$ where $|\mathcal{A}|$ is a number of candidate answers and w is a dimension of representation for each answer. The most similar answer to the joint representation is selected as an answer among the answer candidates. For training, we use only supervision from QA pairs without annotations for ground-truth reasoning paths. To this end, cross-entropy between prediction p and ground-truth t is utilized as a loss function.

Model	Original (ORG)			Paraphrased (PRP)			Mean
	1-hop	2-hop	3-hop	1-hop	2-hop	3-hop	
BLSTM	-	-	-	-	-	-	51.0
MemNN (Sukhbaatar et al., 2015)	-	-	-	-	-	-	59.2
GCN (Kipf and Welling, 2017)	65.7	67.4	66.9	65.8	67.5	67.0	66.7
GGNN (Li et al., 2016)	72.9	74.5	74.0	72.9	74.6	74.1	73.8
MemNN† (Sukhbaatar et al., 2015)	78.1	77.8	76.1	78.0	78.1	76.0	77.3
HAN (Kim et al., 2020)	77.5	77.5	77.2	77.1	77.4	76.9	77.3
BAN (Kim et al., 2018)	83.5	84.0	83.7	83.7	84.3	83.8	83.8
Ours	**88.1**	**90.2**	**91.0**	**87.8**	**90.5**	**90.7**	**89.7**

Table 1: QA accuracy on oracle setting in KVQA under weak supervision. ORG and PRP are a type of question and 1-hop, 2-hop, and 3-hop are the number of graph walks to construct a knowledge hypergraph. The performance of BLSTM and MemNN is reported in (Shah et al., 2019) and we re-implemented MemNN† for a fair comparison.

4 Experimental Settings

4.1 Datasets

In this paper, we evaluate our model across various benchmark datasets: Knowledge-aware VQA (KVQA) (Shah et al., 2019), Fact-based VQA (FVQA) (Wang et al., 2018), PathQuestion (PQ) and PathQuestion-Large (PQL) (Zhou et al., 2018). KVQA, a large-scale benchmark dataset for complex VQA, contains 183,007 pairs for 24,602 images from Wikipedia and corresponding captions, and provides 174,006 knowledge facts for 39,414 unique named entities based on Wikidata (Vrandečić and Krötzsch, 2014) since it requires world knowledge beyond visual content. KVQA consists of two types of questions: original (ORG) and paraphrased (PRP) question generated from the original question via the online paraphrasing tool. FVQA, a representative dataset for commonsense-enabled VQA, considers external knowledge about common nouns depicted in a given image, and contains 5,826 QA pairs for 2,190 images and 4,216 unique knowledge facts from DBPedia (Auer et al., 2007), ConceptNet (Liu and Singh, 2004), and WebChild (Tandon et al., 2014). The last two datasets, PQ and PQL, focus on evaluating multi-hop reasoning ability in the knowledge-based textual QA task. PQ and PQL contain 7,106 and 2,625 QA pairs on 4,050 and 9,844 knowledge facts from the subset of Freebase (Bollacker et al., 2008), respectively. The detailed statistics of the datasets are shown in Appendix A.

4.2 Implementation details

Each node in the knowledge hypergraph and the question hypergraph is represented as a 300-dimensional vector (i.e., $w = 300$) initialized using GloVe (Pennington et al., 2014). Random initialization is applied when a word for a node does not exist in the vocabulary of GloVe. Mean pooling is applied when a node consists of multiple words. For entity linking for KVQA, we apply the well-known pre-trained models for face identification: RetinaFace (Deng et al., 2020) for face detection and ArcFace (Deng et al., 2019) for face feature extraction. For all datasets, we follow the experimental settings as in previous works. We use the similarity-based answer predictor for KVQA, and MLP for the others. We adopt Adam (Kingma and Ba, 2015) to optimize all learnable parameters in the model. We describe details of the experimental settings and the tuned hyperparameters for each dataset in Appendix D.

5 Quantitative Results

5.1 Knowledge-aware visual question answering

We compare the proposed model, Hypergraph Transformer, with other comparative state-of-the-art methods. We report performances on original (ORG) and paraphrased (PRP) questions according to the number of graph walk. For comparative models, three kinds of methods are considered, which are graph-based, memory-based and attention-based networks. The detailed description about the comparative models is described in Appendix E. To evaluate a pure reasoning ability of the models regardless of the performance of entity linking, we first conduct experiments in the oracle setting which ground-truth named entities in an image are given.

As shown in Table 1, our model outperforms comparative models with a large margin across

	PathQuestion			PathQuestion-Large		
	PQ-2H	PQ-3H	PQ-M	PQL-2H	PQL-3H	PQL-M
Seq2Seq (Sutskever et al., 2014)	89.9	77.0	-	71.9	64.7	-
MemNN (Sukhbaatar et al., 2015)	89.5	79.2	86.8	61.2	53.6	55.8
KV-MemNN (Miller et al., 2016)	91.5	79.4	85.2	70.5	63.4	68.6
IRN (Zhou et al., 2018)	96.0	87.7	-	72.5	71.0	-
Embed (Bordes et al., 2014b)	78.7	48.3	-	42.5	22.5	-
Subgraph (Bordes et al., 2014a)	74.4	50.6	-	50.0	21.3	-
MINERVA (Das et al., 2018)	75.9	71.2	73.1	71.8	65.7	66.9
IRN-weak (Zhou et al., 2018)	91.9	83.3	85.8	63.0	61.8	62.4
SRN (Qiu et al., 2020)	96.3	89.2	89.3	78.6	77.5	78.3
Ours	**96.4**	**90.3**	**89.5**	**90.5**	**77.9**(*)	**94.5**

(*) For PQL-3H-More data (2x QA pairs on the same KB as PQL-3H), our model shows 95.4% accuracy.

Table 2: Accuracy on PathQuestion (PQ) and PathQuestion-Large (PQL). 2H and 3H represent the number of multi-hops in ground-truth reasoning paths to answer given questions, and M represents the mixture of 2H and 3H. The models in the first block employ a ground-truth reasoning path as extra supervision (i.e., fully-supervised), and the models in the second block including our model are under weak supervision.

all settings. From the results, we find that the attention mechanism between question and knowledge is crucial for complex QA. Since GCN (Kipf and Welling, 2017) and GGNN (Li et al., 2016) encode question and knowledge graph separately, they do not learn interactions between question and knowledge. Thus, GCN and GGNN show quite low performance under 74% mean accuracy. On the other hand, MemNN† (Weston et al., 2015), HAN (Kim et al., 2020), and BAN (Kim et al., 2018) achieve comparatively high performance because MemNN† adopts question-guided soft attention over knowledge memories. HAN and BAN utilize multi-head co-attention between question and knowledge.

Entity linking setting We also present the experimental results on the entity linking setting where the named entities are not provided as the oracle setting, but detected by the module as described in Section 3.2. As shown in Table 7 of Appendix E, our model shows the best performances for both original and paraphrased questions. For all comparative models, we use the same knowledge hypergraph extracted by the 3-hop graph walk. In entity linking setting, the constructed knowledge hypergraph can be incomplete and quite noisy due to the undetected entities or misclassified entity labels. However, Hypergraph Transformer shows robust reasoning capacity over the noisy inputs. Here, we remark that the upper bound of QA performance is 72.8% due to the error rate of entity linking module. We expect that the performance will be improved when the entity linking module is enhanced.

5.2 Fact-based visual question answering

We conduct experiments on Fact-based Visual Question Answering (FVQA) as an additional benchmark dataset for knowledge-based VQA. Different from KVQA focusing on world knowledge for named entities, FVQA considers commonsense knowledge about common nouns in a given image. Here, we assume that the performance of entity linking is perfect, and evaluate the pure reasoning ability of our model. As shown in Table 8 of Appendix D, Hypergraph Transformer shows comparable performance in both top-1 and top-3 accuracy in comparison with the state-of-the-art methods. We confirm that our model works effectively as a general reasoning framework without considering characteristics of different knowledge sources (i.e., Wikidata for KVQA, DBpedia, ConceptNet, WebChild for FVQA).

5.3 PathQuestion and PathQuestion-Large

To verify multi-hop reasoning ability of our model, we conduct experiments on PathQuestion (PQ) and PathQuestion-Large (PQL). PQ and PQL datasets have annotations of a ground-truth reasoning path to answer a given question. Specifically, {PQ, PQL}-{2H, 3H} denotes a split of PQ and PQL with respect to the number of hops in ground-truth reasoning paths (i.e., 2-hop or 3-hop). {PQ, PQL}-M is a mixture of the 2-hop and 3-hop questions in both dataset, and used to evaluate the more general scenario where the number of reasoning path

Model	Inputs		Original (ORG)			Paraphrased (PRP)			Mean
	Knowledge	Question	1-hop	2-hop	3-hop	1-hop	2-hop	3-hop	
(a) SA	Word	Word	79.4	79.6	77.6	77.1	77.7	77.7	78.2
(b) SA+GA	Word	Word	80.9	82.3	81.5	80.7	82.2	81.8	81.6
(c) SA+GA	Word	Hyperedge	82.1	84.2	82.8	81.1	83.5	82.3	82.7
(d) SA+GA	Hyperedge	Word	87.0	89.9	88.9	87.3	89.7	89.2	88.7
(e) SA+GA **(Ours)**	Hyperedge	Hyperedge	**88.1**	**90.2**	**91.0**	**87.8**	**90.5**	**90.7**	**89.7**
(f) **Ours**-SA	Hyperedge	Hyperedge	85.2	88.8	88.3	85.0	88.3	88.4	87.1
(g) **Ours**-GA	Hyperedge	Hyperedge	82.6	83.6	85.0	82.7	83.6	84.9	83.7

Table 3: (a-e) Validation for the effectiveness of using hypergraph. Here, we compare the results with respect to the different types of the input format (i.e., Single Word or Hyperedge) used to represent knowledge and question which are fed into the attention mechanism. (e-g) Ablation study for attention blocks of Hypergraph Transformer. GA and SA are abbreviations of guided-attention and self-attention, respectively.

required to answer a given question is unknown.

The experimental results on diverse split of PQ and PQL datasets are provided in Table 2. The first section in the table includes fully-supervised models which require a ground-truth path annotation as an additional supervision. The second section contains weakly-supervised models learning to infer the multi-hop reasoning paths without the ground-truth path annotation. Hypergraph Transformer is involved in the weakly-supervised models because it only exploits an answer as a supervision.

Our model shows comparable performances on PQ-{2H, 3H, M} to the state-of-the-art weakly-supervised model, SRN. Especially, Hypergraph Transformer shows significant performance improvement (78.6% → 90.5% for PQL-2H, 78.3% → 94.5% for PQL-M) on PQL. We highlight that PQL is more challenging dataset than PQ in that PQL not only covers more knowledge facts but also has fewer QA instances. We observe that the accuracy on PQL-3H is relatively lower than the other splits. This is due to the insufficient number of training QA pairs in PQL-3H. When we use PQL-3H-More which has twice more QA pairs (1031 → 2062) on the same knowledge base as PQL-3H, our model achieves 95.4% accuracy.

6 Validation for Hypergraph Transformer

We verify the effectiveness of each module in Hypergraph Transformer. To analyze the performances of the variants in our model, we use KVQA which is a representative and large-scale dataset for knowledge-based VQA. Here, we mainly focus on two aspects: i) effect of hypergraph and ii) effect of attention mechanism. To evaluate a pure reasoning ability of the models, we conduct experiments in the oracle setting.

6.1 Effect of hypergraph

To analyze the effectiveness of hypergraph-based input representation, we conduct comparative experiments on the different types of input formats for Transformer architecture. Here, we consider the two types of input format, which are single-word-unit and hyperedge-based representations. Compared to hyperedge-based inputs considering multiple relational facts as a input token, single-word-unit takes every entity and relation tokens as separate input tokens. We note that using single-word-unit-based input format for both knowledge and question is the standard settings for the Transformer network and using hyperedge-based input format for both is the proposed model, Hypergraph Transformer. We set the Transformer (SA+GA) as a backbone model, and present the results in Table 3(b-e). When hypergraph-based representations are used for both knowledge and question, the results show the best performance across all settings over question types (ORG and PRP) and a number of graph walk (1-hop, 2-hop, and 3-hop). As shown in Table 3, the mean accuracy of QA achieves 89.7% when both are encoded using hyperedges, while using single-word-unit-based representation causes performance to drop to 81.6%. Especially, when we convert the one of both hyperedge-level representation to single-word-unit-based representation, the mean accuracy of QA is 82.7% and 88.7%, respectively. These results validate that it is meaningful to consider not only knowledge but also question as hypergraphs.

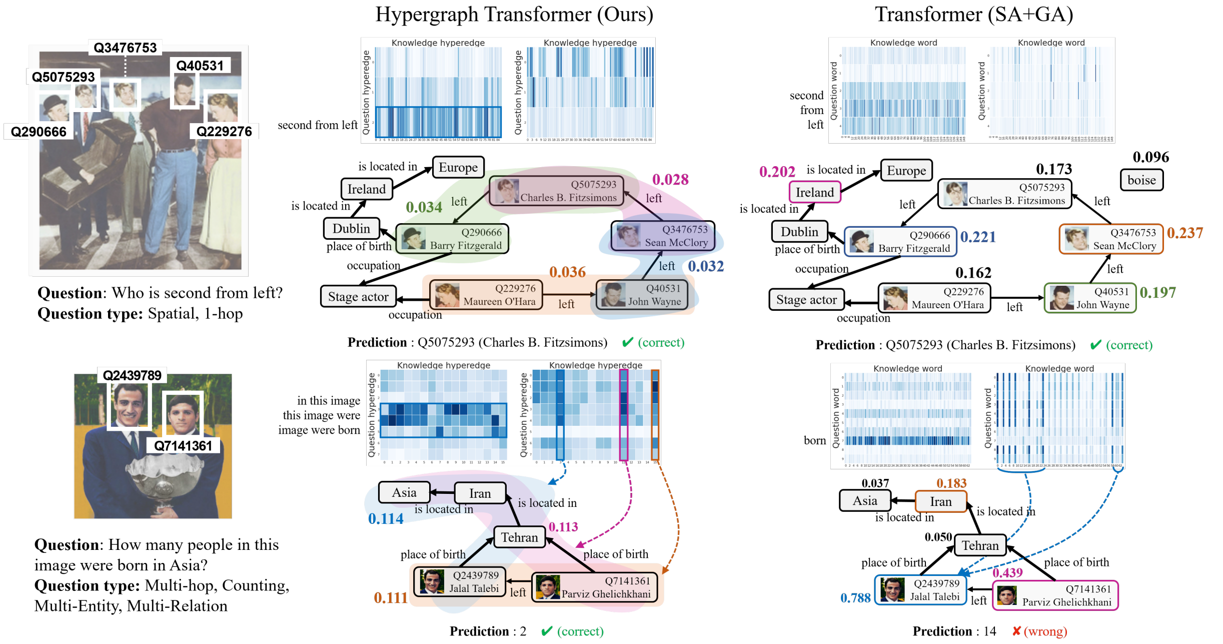


Figure 3: Qualitative analysis on effectiveness of using hypergraph as input format to Transformer architecture. Here, we visualize attention maps for Hypergraph Transformer and the Transformer (SA+GA). All attention scores are averaged over multi-heads and multi-layers. Each x and y axis represent indices of question and knowledge hyperedges in Hypergraph Transformer, and indices of question and knowledge word in Transformer (SA+GA). In the attention maps, the dark colors represent high values. The hyperedges with high attention scores are visualized.

Effect of multi-hop graph walk We compare the performances with different number of graph walks used to construct a knowledge hypergraph (i.e., 1-hop, 2-hop, and 3-hop). All models except ours show slightly lower performance on the 3-hop graph than on the 2-hop graph. We observe that the number of extracted knowledge facts increases when the number of graph walk increases, and unnecessary facts for answering a given question are usually included. Nonetheless, our model shows robust reasoning performance when a large and noisy knowledge facts are given.

6.2 Effect of attention mechanism

To investigate the impacts of each attention block (i.e., GA and SA), ablation studies are shown in Table 3(e-g). The scores across all settings drop when GA or SA is removed. Particularly, the mean accuracy of QA is decreased by 6.0% (89.7% → 83.7%), 2.6% (89.7% → 87.1%) for cutting out the GA and the SA block, respectively. Based on the two experiments, we identify that not only the guided-attention which captures inter-relationships between question and knowledge but also the self-attention which learns intra-relationship in them are crucial to the complex QA. To sum up, Hypergraph Transformer takes graph-level inputs, i.e., hyperedge, and conducts semantic matching between hyperedges by the attention mechanism. Due to the two characteristics, the model shows better reasoning performance focusing on the evidences necessary for reasoning under weak supervision.

7 Qualitative Analysis

Figure 3 provides the qualitative analysis on effectiveness of using a hypergraph as an input format to Transformer architecture. We present the attention map from the guided-attention block, and visualize top-k attended knowledge facts or entities with the attention scores. In the first example, both model, Hypergraph Transformer and Transformer (SA+GA), infer the correct answer, *Q5075293*. Our model responds by focusing on $\{second \preceq from \preceq left\}$ phrase of the question and four facts having a *left* relation among 86 knowledge hyperedges. In comparison, Transformer (SA+GA) strongly attends to the knowledge entities which appear repetitive in the knowledge facts. Especially, the model attends to *Q3476753*, *Q290666* and *Ireland* with the high attention score 0.237, 0.221, and 0.202. In the second example, our model attends to the correct knowledge hyperedges considering the multi-hop facts about *place of birth* of the people shown in the given image, and infers the correct answer. On the other hand, Transformer (SA+GA) strongly attends to the knowledge entity of person (*Q2439789*) presented in the image with undesired attention score 0.788. The second and third attended knowledge entities are the other person (*Q7141361*) and *Iran*. Transformer (SA+GA) fails to focus on the multi-hop facts required to answer the given question and predicts the answer with the wrong number at the end.

8 Discussion and Conclusion

In this paper, we proposed Hypergraph Transformer for multi-hop reasoning over knowledge graph under weak supervision. Hypergraph Transformer adopts hypergraph-based representation to encode high-order semantics of knowledge and questions and considers associations between a knowledge hypergraph and a question hypergraph. Here, each node representation in the hypergraphs is updated by inter- and intra-attention mechanisms in two hypergraphs, rather than by iterative message passing scheme. Thus, Hypergraph Transformer can mitigate the well-known over-smoothing problem in the previous graph-based methods exploiting the message passing scheme. Extensive experiments on various datasets, KVQA, FVQA, PQ, and PQL validated that Hypergraph Transformer conducts accurate inference by focusing on knowledge evidences necessary for question from a large knowledge graph. Although not covered in this paper, an interesting future work is to construct heterogeneous knowledge graph that includes more diverse knowledge sources (e.g. documents on web).

Acknowledgements

We would like to thank Woo Young Kang, Kyoung-Woon On, Seonil Son, Gi-Cheon Kang, Christina Baek, Junseok Park, Min Whoo Lee, Hwiyeol Jo and Sang-Woo Lee for their helpful comments and discussion. This work was partly supported by the IITP (2015-0-00310-SW.StarLab/20%, 2017-0-01772-VTT/20%, 2019-0-01371-BabyMind/10%, 2021-0-02068-AIHub/10%, 2021-0-01343-GSAI/10%, 2020-0-01373/10%) grants, the KIAT (P0006720-ILIAS/10%) grant funded by the Korean government, and the Hanyang University (HY-202100000003160/10%).

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Antoine Bordes, Sumit Chopra, and Jason Weston. 2014a. Question answering with subgraph embeddings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 615–620, Doha, Qatar. Association for Computational Linguistics.

Antoine Bordes, Jason Weston, and Nicolas Usunier. 2014b. Open question answering with weakly supervised embedding models. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 165–180. Springer.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. 2018. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5202–5211. IEEE.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*

2019, Long Beach, CA, USA, June 16-20, 2019, pages 4690–4699. Computer Vision Foundation / IEEE.

Jiale Han, Bo Cheng, and Xu Wang. 2020a. Open domain question answering based on text enhanced knowledge graph with hyperedge infusion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1475–1481, Online. Association for Computational Linguistics.

Jiale Han, Bo Cheng, and Xu Wang. 2020b. Two-phase hypergraph based reasoning with dynamic relations for multi-hop KBQA. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3615–3621. ijcai.org.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 725–736, Online. Association for Computational Linguistics.

Eun-Sol Kim, Woo-Young Kang, Kyoung-Woon On, Yu-Jung Heo, and Byoung-Tak Zhang. 2020. Hypergraph attention networks for multimodal learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 14569–14578. IEEE.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1571–1581.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3538–3545. AAAI Press.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. 2016. Gated graph sequence neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

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.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge aware conversation generation with explainable reasoning over augmented graphs. 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 1782–1792, Hong Kong, China. Association for Computational Linguistics.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas. Association for Computational Linguistics.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander G. Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2659–2670.

Medhini Narasimhan and Alexander G Schwing. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *Proceedings of the European conference on computer vision (ECCV)*, pages 451–468.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Yunqi Qiu, Yuanzhuo Wang, Xiaolong Jin, and Kun Zhang. 2020. Stepwise reasoning for multi-relation question answering over knowledge graph with weak supervision. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 474–482. ACM.

Shailaja Keyur Sampat, Yezhou Yang, and Chitta Baral. 2020. Visuo-linguistic question answering (VLQA) challenge. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4606–4616, Online. Association for Computational Linguistics.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4498–4507, Online. Association for Computational Linguistics.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: knowledge-aware visual question answering. 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 8876–8884. AAAI Press.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2440–2448.

Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. 2020. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 222–229.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Niket Tandon, Gerard de Melo, Fabian M. Suchanek, and Gerhard Weikum. 2014. Webchild: harvesting and organizing commonsense knowledge from the web. In *Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY, USA, February 24-28, 2014*, pages 523–532. ACM.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. 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.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2018. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2017. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1290–1296. ijcai.org.

Xu Wang, Shuai Zhao, Jiale Han, Bo Cheng, Hao Yang, Jianchang Ao, and Zhenzi Li. 2020. Modelling long-distance node relations for KBQA with global dynamic graph. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2572–2582, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha P. Talukdar. 2019. Hypergcn: A new method for training graph convolutional networks on hypergraphs. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1509–1520.

Naganand Yadati, Dayanidhi R S, Vaishnavi S, Indira K M, and Srinidhi G. 2021. Knowledge base question answering through recursive hypergraphs. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 448–454, Online. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Mantong Zhou, Minlie Huang, and Xiaoyan Zhu. 2018. An interpretable reasoning network for multi-relation question answering. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2010–2022, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4995–5004. IEEE Computer Society.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 1097–1103. ijcai.org.

Appendix. This supplementary material provides additional information not described in the main text due to the page limit. The contents of this appendix are as follows: In Section A, we show the detailed statistics for the diverse splits of four benchmark datasets, i.e., KVQA, FVQA, PQ and PQL. In Section B and C, we present the additional quantitative and qualitative analyses on KVQA and PQ datasets, respectively. In Section D, we describe the experimental details for each dataset. In Section E, we depict the implementation details of comparative models for KVQA.

A Data Statistics

The diverse split statistics for four benchmark datasets, KVQA (Shah et al., 2019), FVQA (Wang et al., 2018), PQ and PQL (Zhou et al., 2018), are shown in Table 4. Here, we highlight four aspects as follows: 1) KVQA dataset covers the large number of entities (at least 5 times more) and knowledge facts (at least 17 times more) than FVQA, PQ and PQL. 2) PQ and PQL datasets have annotations of a ground-truth reasoning path to answer a given question. 2H and 3H denote the number of hops (i.e., 2-hop and 3-hop) in ground-truth reasoning paths. Also, M denotes a mixture of the 2H and 3H questions. 3) PQL covers more knowledge facts including a large number of entities and relations than PQ, but has fewer QA pairs. 4) PQL-3H has a quite limited number of QA pairs (1,031). PQL-3H-More has twice more QA pairs (2,062) with the same number of entities, relations, knowledge facts and answers as PQL-3H.

B Additional Analysis on KVQA

Here, we analyze more in-depth on KVQA dataset concerning i) categories of question, and ii) types of answer selector. All models are under the same setting of ORG+3-hop reported in Table 1.

B.1 Analysis on question categories

We analyze QA performances over different question categories in Table 5. Hypergraph Transformer achieves the best accuracy in all categories except Multi-hop (slightly low at second-best). Our model shows notable strengths especially on complex problems such as Comparison, Multi-entity or Subtraction. To draw inferences for these question categories, the model needs to attend to multiple knowledge facts related to a given question, and conducts multi-hop reasoning based on the facts. Also, our model shows significant improvement in spatial question compared to other models. Whereas spatial question is quite simple, it is required to understand a correct spatial relationship between multiple entities in a given image. Examples of QA on diverse question categories are depicted in Figure 4. Answers, inferred by five comparative models and the proposed model, are presented with corresponding image and question. The qualitative results indicate that our model draws reasonable inferences across diverse question categories.

B.2 Effect of similarity-based answer selector

To validate the impact of similarity-based answer selector, we replace the similarity-based answer selector (SIM) with a multi-layer perceptron (MLP). We first note that KVQA dataset includes a large number of unique answers (19,360), and contains a lot of zero-shot and few-shot answers in test phase. As shown in Table 6, the MLP fails to infer zero-shot answers which are not appeared in the training phase at all. Besides, the performance difference between SIM and MLP in one-shot answer (appeared in the only one time in training phase) is more than 18%. The MLP uses 17% more parameters than SIM because KVQA has a large number of answer candidates (19,360). When the number of candidate answers increases, the MLP needs more parameters, but SIM does not. To sum up, the similarity-based answer selector (SIM) contributes to infer few-shot and zero-shot answers in parameter-efficient manner.

C Qualitative Analysis on PathQuestion

Figure 5 shows the qualitative analysis of Hypergraph Transformer and Transformer (SA+GA) on PathQuestion. In Figure 5(a), Hypergraph Transformer attends to the second question hyperedge $\{the \preceq ethnicity \preceq of\}$ and the fourth knowledge hyperedge $\{Alice\ Betty\ Stern \preceq children \preceq Otto\ Frank \preceq ethnicity \preceq Germans\}$ to reason based on the multi-hop evidence about ethnicity. On the other hand, Transformer (SA+GA) focuses on the third question word *ethnicity* correctly, but attends to *Otto Frank*, *Jew*, *Male* with the high attention score 0.461, 0.242, and 0.204, not the exact knowledge entity, *Germans*. In Figure 5(b), both model, Hypergraph Transformer and Transformer (SA+GA), fail to infer the correct answer. The predicted answer of Hypergraph Transformer

	KVQA	FVQA	PQ-2H	PQ-3H	PQ-M	PQL-2H	PQL-3H	PQL-M
# Entities	39,414	3,391	1,057	1,837	2,257	5,035	6,506	6,506
# Relations	18	13	14	14	14	364	412	412
# Knowledge facts	174,006	4,216	1,211	2,839	4,050	4,247	5,597	9,844
# Words	63,164	6,663	1,180	1,929	2,407	5,505	7,001	7,034
# QA pairs	183,007	5,826	1,908	5,198	7,106	1,594	1,031	2,625
# Answers	19,360	500	305	1,009	1,107	380	292	438

(*) PQL-3H-More has twice more QA pairs (2,062) with the same number of entities, relations, knowledge facts and answers as PQL-3H.

Table 4: Statistics of four benchmark datasets: Knowledge-aware Visual Question Answering (KVQA), Fact-based Visual Question Answering (FVQA), PathQuestion (PQ) and PathQuestion-Large (PQL).

	Bool	Comp.	Multi	Multi	Multi	1-hop	1-hop	Spatial	Subtract.
			entity	hop	relation		subtract		
MemNN	75.1	50.5	43.5	53.2	45.2	61.0	-	48.1	40.5
GCN	86.8	87.7	87.7	96.7	77.7	61.4	53.7	29.4	37.7
GGNN	86.6	88.8	88.6	95.1	90.0	70.4	55.2	32.6	26.1
HAN	98.1	93.8	93.6	98.2	92.8	73.5	51.5	29.6	29.0
BAN	98.5	94.8	94.5	**99.3**	98.6	81.2	56.7	39.1	39.2
Ours	**99.1**	**96.9**	**96.8**	99.2	**99.3**	**89.9**	**73.3**	**90.1**	**42.4**

Table 5: Analysis of QA accuracy over different question categories of original (ORG) questions in oracle setting. All models use 3-hop graph reported in Table 1. Comp. and Subtract. are abbreviations of Comparison and Subtraction. The best performance of each question type is highlighted in bold.

is wrong even though it attends correctly to the first knowledge hyperedge $\{$*Wallace Reid* $\preceq$ *spouse* $\preceq$ *Dorothy Davenport* $\preceq$ *parents* $\preceq$ *Harry Davenport* $\preceq$ *cause of death* $\preceq$ *Myocardial Infarction*$\}$. However, Transformer (SA+GA) attends to only the second and seventh word (*Dorothy Davenport*) and the fourth and ninth word (*Harry Davenport*) in knowledge with high attention score, not the answer entity, *Myocardial Infarction*. We consider that the reason why Hypergraph Transformer failed to infer the correct answer despite focusing on the exact knowledge fact is that the correct answer word (*Myocardial Infarction*) appears rarely in QA pairs.

D Experimental details

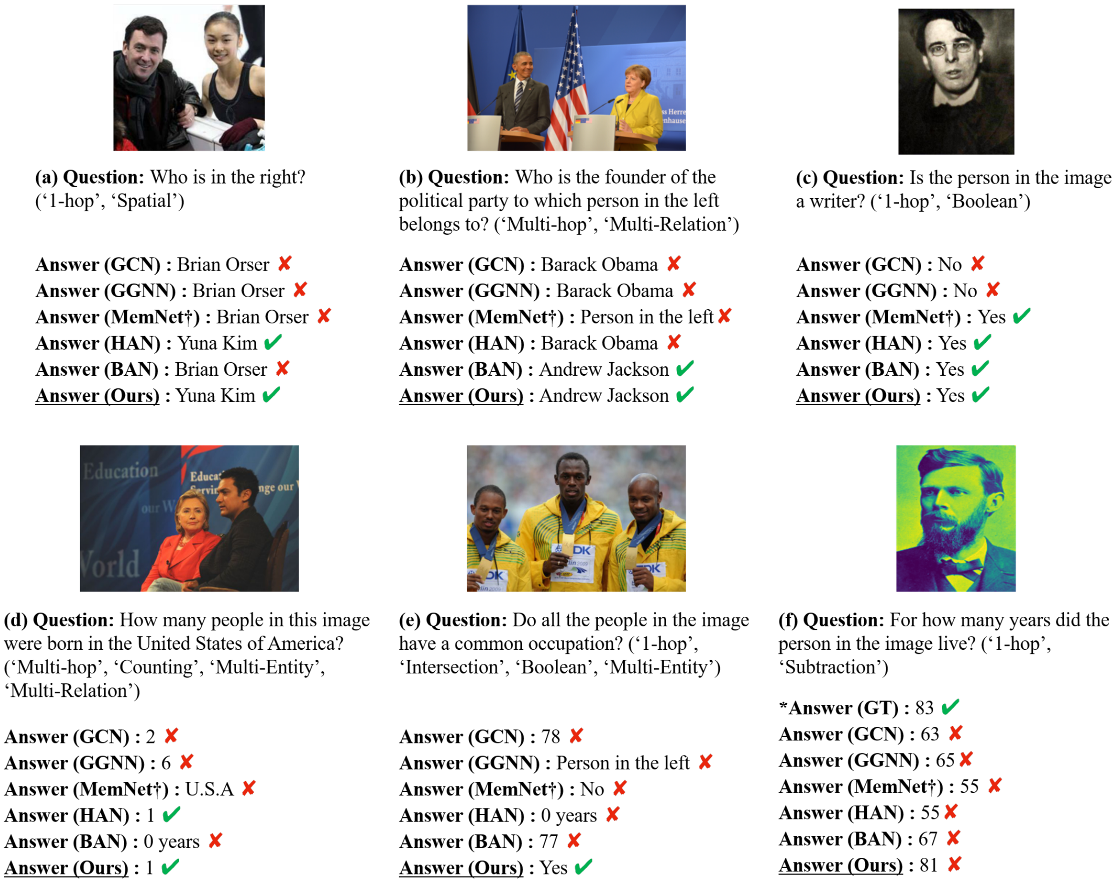
D.1 Knowledge-aware VQA

We follow the experimental settings suggested in (Shah et al., 2019). For entity linking, we apply well-known pre-trained models for face identification: RetinaFace (Deng et al., 2020) for face detection and ArcFace (Deng et al., 2019) for face feature extraction. We first assign a name of the detected faces with the label of the closest distance compared to all of the face embeddings of 18,880 named entities. In addition, we refine a list of detected named entities by matching the associated image caption (i.e., Wikipedia caption). By doing so, we obtain the result of entity linking with top-1 precision 65.0% and top-1 recall 72.8%. QA performances in the entity linking setting on KVQA are shown in Table 7. Here, we note that BLSTM and MemNN of the first section in the table are based on the different entity linking modules with top-1 precision 81.1% and top-1 recall 82.2%[1]. It is more accurate than ours around 9.4% in the recall metric.

D.2 Fact-based VQA

We follow the experimental settings suggested in (Wang et al., 2018). Following the paper, the dataset provides five splits of train and test data. We report the average accuracy of five repeated runs on different data split: 76.55 as top-1 accuracy (average of 76.93, 75.92, 76.24, 76.16, and 77.50) and 82.20 as top-3 accuracy (average of 82.90, 81.45, 81.70, 81.74 and 83.20). The experimental results are shown in Table 8.

[1] The code for the entity linking module has not been released publicly. As such, we implement the module based on the open-source: https://github.com/deepinsight/insightface. We use the pre-trained model named retinaface-mnet025-v2 and LResNet100E-IR,ArcFace@ms1m-refine-v2.



(a) Question: Who is in the right? (‘1-hop’, ‘Spatial’)

Answer (GCN) : Brian Orser ✗
Answer (GGNN) : Brian Orser ✗
Answer (MemNet†) : Brian Orser ✗
Answer (HAN) : Yuna Kim ✓
Answer (BAN) : Brian Orser ✗
Answer (Ours) : Yuna Kim ✓

(b) Question: Who is the founder of the political party to which person in the left belongs to? (‘Multi-hop’, ‘Multi-Relation’)

Answer (GCN) : Barack Obama ✗
Answer (GGNN) : Barack Obama ✗
Answer (MemNet†) : Person in the left ✗
Answer (HAN) : Barack Obama ✗
Answer (BAN) : Andrew Jackson ✓
Answer (Ours) : Andrew Jackson ✓

(c) Question: Is the person in the image a writer? (‘1-hop’, ‘Boolean’)

Answer (GCN) : No ✗
Answer (GGNN) : No ✗
Answer (MemNet†) : Yes ✓
Answer (HAN) : Yes ✓
Answer (BAN) : Yes ✓
Answer (Ours) : Yes ✓

(d) Question: How many people in this image were born in the United States of America? (‘Multi-hop’, ‘Counting’, ‘Multi-Entity’, ‘Multi-Relation’)

Answer (GCN) : 2 ✗
Answer (GGNN) : 6 ✗
Answer (MemNet†) : U.S.A ✗
Answer (HAN) : 1 ✓
Answer (BAN) : 0 years ✗
Answer (Ours) : 1 ✓

(e) Question: Do all the people in the image have a common occupation? (‘1-hop’, ‘Intersection’, ‘Boolean’, ‘Multi-Entity’)

Answer (GCN) : 78 ✗
Answer (GGNN) : Person in the left ✗
Answer (MemNet†) : No ✗
Answer (HAN) : 0 years ✗
Answer (BAN) : 77 ✗
Answer (Ours) : Yes ✓

(f) Question: For how many years did the person in the image live? (‘1-hop’, ‘Subtraction’)

***Answer (GT) :** 83 ✓
Answer (GCN) : 63 ✗
Answer (GGNN) : 65 ✗
Answer (MemNet†) : 55 ✗
Answer (HAN) : 55 ✗
Answer (BAN) : 67 ✗
Answer (Ours) : 81 ✗

Figure 4: Qualitative results on KVQA dataset. GCN, GGNN, MemNN†, HAN, BAN and our model infer answers to a question about a given image. Green and red marks indicate correct and incorrect answers, respectively.

	Original (ORG)				Paraphrased (PRP)			
	Zero-shot	One-shot	Multi-shot	ALL	Zero-shot	One-shot	Multi-shot	ALL
MLP	0.0	78.3	87.2	76.0	0.0	76.9	86.8	75.6
SIM	**93.9**	**96.7**	**90.1**	**91.0**	**92.4**	**96.3**	**89.9**	**90.7**

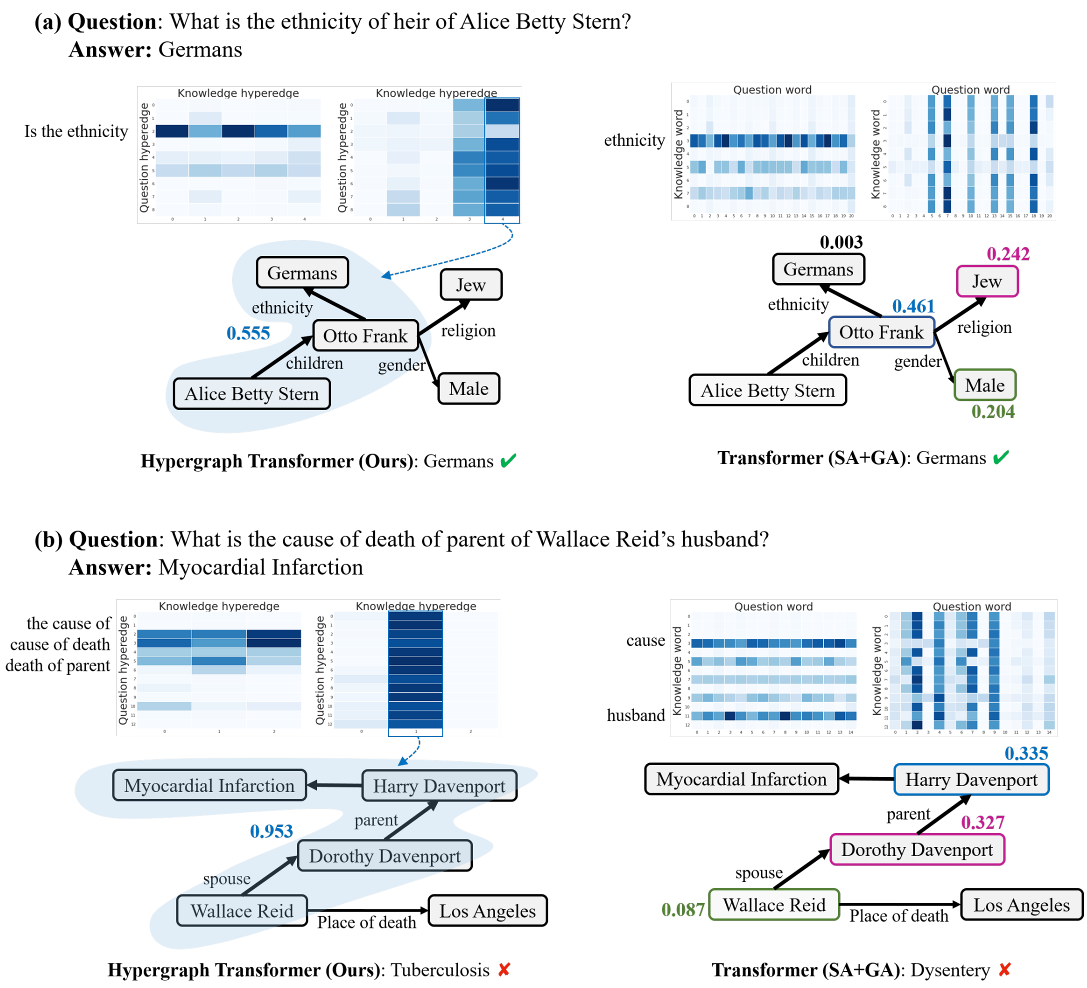
Table 6: Analysis for answer selector with the frequency of answers in the test split. SIM and MLP represent similarity-based answer selector and multi-layer perceptron.

Model	ORG	PRP	Mean
BLSTM	48.0	27.2	37.6
MemNN	50.2	34.2	42.2
GCN	48.9	48.2	48.5
GGNN	50.9	50.9	50.9
MemNN†	54.0	53.9	54.0
HAN	53.4	53.3	53.3
BAN	59.6	60.0	59.8
Transformer (SA)	57.5	58.9	58.3
Transformer (SA+GA)	60.4	59.8	60.1
Ours	**62.0**	**62.8**	**62.4**

Table 7: QA accuracy on entity linking setting in KVQA. The performances of BLSTM and MemNN are reported in (Shah et al., 2019).

	Accuracy	
	@1	@3
Human	77.99	-
LSTM-Q+I (Pre-VQA)	24.98	30.30
Hie-Q+I (Pre-VQA)	43.14	59.44
FVQA-Top3-QQmaping	56.91	64.65
STTF-Q+VConcept	62.20	75.60
RC (pre-SQuAD)	62.94	70.08
Out of the Box	69.35	80.25
Mucko	73.06	**85.94**
Ours	**76.55**	82.20

Table 8: Accuracy on Fact-based Visual Question Answering (FVQA). Top-1 and top-3 accuracy are used as evaluation metrics.



(a) Hypergraph Transformer (Ours): Germans ✓ Transformer (SA+GA): Germans ✓

(b) Hypergraph Transformer (Ours): Tuberculosis ✗ Transformer (SA+GA): Dysentery ✗

Figure 5: Qualitative analysis on effectiveness of using hypergraph as input format to Transformer architecture. Here, we visualize attention maps ($\mathrm{Attention}(Q_k, K_q, V_q)$ and $\mathrm{Attention}(Q_q, K_k, V_k)$) for Hypergraph Transformer and the Transformer (SA+GA). All attention scores are averaged over multi-heads and multi-layers. Each x and y axis represent indices of question and knowledge hyperedges in Hypergraph Transformer, and indices of question and knowledge word in Transformer (SA+GA). In the attention maps, the dark colors represent high values. We also visualize the top-3 attended knowledge hyperedges in Hypergraph Transformer, and top-3 attended knowledge fact in Transformer (SA+GA) with the attention score.

D.3 PathQuestion and PathQuestion-Large

We follow the same experimental settings suggested in (Zhou et al., 2018). Following the paper, we split the dataset into train, validation, and test sets with a proportion of 8:1:1, and report the average accuracy of five repeated runs on different data split.

E Implementation Details of Comparative Models for KVQA

For comparative models for KVQA, three kinds of methods are considered, which are graph-based, memory-based and attention-based networks.

Graph-based networks. Graph convolutional networks (GCN) (Kipf and Welling, 2017) and gated graph neural networks (GGNN) (Li et al., 2016) are representative models of graph-based neural networks. Both learn node representations of a knowledge and question graph (not a hypergraph), propagating information between neighborhoods. After propagation, node representations in a graph are aggregated to encode a graph-level representation. Joint representation is obtained based on the two graph representations.

Memory-based networks. Memory network (MemNN) (Weston et al., 2015) is a de facto baseline for fact-based question answering. Each fact is embedded into a memory slot, and soft attention is calculated between memory slots and a given question. Joint representation is obtained based on the attention.

Attention-based networks. Bilinear attention networks (BAN) (Kim et al., 2018) and hypergraph attention networks (HAN) (Kim et al., 2020) consider interactions between knowledge and question based on co-attention mechanism. BAN calculates soft attention scores between knowledge entities and question words. Meanwhile, HAN employs stochastic graph walk in a knowledge and question graph to encode high-order semantics (e.g., knowledge facts and question phrases), and considers attention scores between knowledge facts and question phrases. Joint representation is obtained based on the attention as well. The more implementation details of the above comparative models is described as follows.

E.1 Graph convolutional networks

The knowledge and question graph are encoded separately by two graph convolutional networks (GCN) (Kipf and Welling, 2017). Each GCN model consists of two propagation layers and a sum pooling layer across the nodes in the graph. The operation of the propagation layer is as follows: $f(H^{(l)}, A) = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)})$ where $\hat{A} = A + I$, A is an adjacency matrix of the graph, I is an identity matrix, D is a degree matrix of A, $W^{(l)}$ is the model parameters of l-th layer, and $H^{(l)}$ is the representations of the graph in the l-th layer. Here, $H^{(0)}$ is the word embeddings of each entity in the knowledge and question graph. After propagation and aggregation phase, the knowledge and question graph representations are obtained. Then, the two graph representations are concatenated and fed into a single layer feed-forward layer to get joint representation.

E.2 Gated graph neural networks

As the same as graph convolutional networks, the knowledge and question graph are encoded separately by two gated graph neural networks (GGNN). Each GGNN model consists of three gated recurrent propagation layers and a graph-level aggregator. Motivated by Gated Recurrent Units (Cho et al., 2014), GGNN adopts a update gate and a reset gate to renew each node's hidden state. The detailed equation of gated recurrent propagation is as follows: $\mathbf{h}_v^{(1)} = [\mathbf{x}_v^T, \mathbf{0}]^T$ where $\mathbf{x}_v$ is the v-th word embedding of each entity in the knowledge and question graph, $\mathbf{a}_v^{(t)} = A_{v:}^T [\mathbf{h}_1^{(t-1)^T} \cdots \mathbf{h}_{|\mathcal{V}|}^{(t-1)^T}]^T + \mathbf{b}$ where the matrix A determines how nodes in the graph communicate each other and $\mathbf{b}$ is a bias vector. Then, the update gate and reset gate are computed as follows: $\mathbf{z}_v^t = \sigma(W^z\mathbf{a}_v^{(t)} + U^z\mathbf{h}_v^{(t-1)})$, $\mathbf{r}_v^t = \sigma(W^r\mathbf{a}_v^{(t)} + U^r\mathbf{h}_v^{(t-1)})$ where σ is a logistic sigmoid function, and $W^{[\cdot]}$ and $U^{[\cdot]}$ are learnable parameters. Finally, the hidden states of nodes in the given graph are updates as $\mathbf{h}_v^{(t)} = (1 - \mathbf{z}_v^t) \odot \mathbf{h}_v^{(t-1)} + \mathbf{z}_v^t \odot \tilde{\mathbf{h}}_v^{(t)}$ where $\tilde{\mathbf{h}}_v^{(t)} = \tanh(W^h\mathbf{a}_v^{(t)} + U^h(\mathbf{r}_v^t \odot \mathbf{h}_v^{(t-1)})$. After the propagation phase, the nodes in the graph are aggregated to a graph-level representation as $\mathbf{h}_{\mathcal{G}} = \tanh(\sum_{v\in\mathcal{V}} \sigma(i(\mathbf{h}_v^{(T)}, \mathbf{x}_v)) \odot \tanh(j(\mathbf{h}_v^{(T)}, \mathbf{x}_v))$ where i and j are a single layer feed-forward layer, respectively. Then, the two aggregated graph representations are concatenated and fed into another single layer feed-forward layer to get joint representation of question and knowledge graph.

E.3 Memory networks

We reproduce end-to-end memory networks (Sukhbaatar et al., 2015) proposed as a baseline model in (Shah et al., 2019). First, we use Bag-of-words (BoW) representation for knowledge facts and a question. The soft attention over the knowledge facts and the given question is computed as follows: $p_{ij} = \mathrm{softmax}(q_{i-1}^T m_{ij})$ where m is the embeddings of knowledge facts, i is a number of layer and j is an index of knowledge facts. The output representation of i-th layer is $O_i = \sum_j p_{ij}o_{ij}$ where o is the another embeddings of knowledge facts different from m. The updated question representation is $q_{k+1} = O_{k+1} + q_k$, and based on the output representation and question representation, answer is predicted as follows: $\hat{a} = \mathrm{softmax}(f(O_K + q_{K-1}))$ where f is a single layer feed-forward layer. Here, we set up the model as three layers with adjacent and layer-wise weight tying.

E.4 Bilinear attention networks

Bilinear attention networks exploit a multi-head co-attention mechanism between knowledge and question. BAN calculates soft attention scores between knowledge entities and question words as follows: $\mathcal{A} = \mathrm{softmax}(W^h \circ (M^q W^q)(M^k W^k)^\top)$ where M^q, M^k are a row-wise concatenated question words and knowledge entities, $W^{[\cdot]}$ is learnable matrices, and $\circ$ is element-wise multiplication. Based on the attention map $\mathcal{A}$, the joint feature is obtained as follows: $z_i = (M^q W^q)_i{}^\top \mathcal{A}(M^k W^k)_i$ where the subscript i denotes the i-th index of column vectors in each matrix. For multi-head attention, the attended outputs with different heads are concatenated and fed into a single layer feed-forward layer to make a final representation. Here, we use four attention heads as multi-head.

E.5 Hypergraph attention networks

The model architecture and detailed operation of hypergraph attention networks are similar to that of BAN. The difference between BAN and HAN is the abstraction level of the input. For HAN, the hyperedges sampled by stochastic graph walk are fed into the co-attention mechanism. What HAN and our model have in common is introducing a hypergraph to consider high-order relationships in question graph and knowledge graph. Both models share the similar motivation, but the core operations are quite different. Especially, HAN employs stochastic graph walk to construct question and knowledge hypergraph. Due to the randomness of the stochasticity, misinformed or incomplete hyperedges can be extracted.

E.6 Transformer Variants

The model architectures of Transformer (SA) and Transformer (SA+GA) presented in this paper are the same as Hypergraph Transformer. The only difference is the abstraction level of input. The Transformer (SA) and Transformer (SA+GA) take single-word-unit as input tokens, and Hypergraph Transformer takes hyperedges as input tokens. Following (Vaswani et al., 2017; Tsai et al., 2019), we apply positional embeddings to the input sequence of both models. We stack two guided-attention blocks and three self-attention blocks, respectively. Each attention block has multi-head attention with four attention heads followed by layer normalization, residual connections and a single multi-layer perceptron. We set the dropout applied on the token embedding weights, query and key-value embedding weights, attention weights and residual connections from 0.05 to 0.2. We minimize negative log-likelihood using Adam optimizer (Kingma and Ba, 2015) with an initial learning rate from $1e-4$ to $1e-5$ with batch size from 128 to 256. All transformer variant models described in this paper have the same fixed-number of sequence length as follows: 300 for 1-hop, 1,000 for 2-hop and 1,800 for 3-hop graphs.