HyperMR: Hyperbolic Hypergraph Multi-hop Reasoning for Knowledge-based Visual Question Answering

Bin Wang[1], Fuyong Xu[1], Zhenfang Zhu[2*], Peiyu Liu[1*]
[1]School of Information Science and Engineering, Shandong Normal University
[2]School of Information Science and Electrical Engineering, Shandong Jiaotong University
wang_bean_068@163.com, fyxu0908@outlook.com
zhuzf_sdjtu@126.com, liupy@sdnu.edu.cn

Abstract

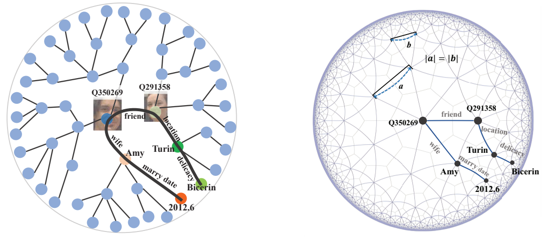
Knowledge-based Visual Question Answering (KBVQA) is a challenging task, which aims to answer an image related question based on external knowledge. Most of the works describe the semantic distance using the actual Euclidean distance between two nodes, which leads to distortion in modeling knowledge graphs with hierarchical and scale-free structure in KBVQA, and limits the multi-hop reasoning capability of the model. In contrast, the hyperbolic space shows exciting prospects for low-distortion embedding of graphs with hierarchical and free-scale structure. In addition, we map the different stages of reasoning into multiple adjustable hyperbolic spaces, achieving low-distortion, fine-grained reasoning. Extensive experiments on the KVQA, PQ and PQL datasets demonstrate the effectiveness of HyperMR for strong-hierarchy knowledge graphs. The code is publicly available at https://github.com/WANGBEAN068/HyperMR/.

Keywords: KBVQA, Multi-hop Reasoning, Hyperbolic Space

1. Introduction

Visual Question Answering (VQA) aims to understand the natural language questions and related images to infer the correct answers. Due to the limited knowledge in the regular corpus during model training, some questions could not be answered. This requires the model be able to use external knowledge to infer the correct answer, called Knowledge-based Visual Question Answering (KB-VQA). For complex questions, it needs to capture evidence in the knowledge base and infer the correct path (Cui et al., 2023). To accomplish such a challenging task, the model needs to link questions, images, and entities in the knowledge base and embeds them in a multi-dimensional vector space to restore their complex semantic structures, when triples of linked entities are not enough to answer the question, multi-hop reasoning is needed. Finally the model decodes reasoning result to get the answer.

Most previous works (Zhang et al., 2022; Heo et al., 2022; Adjali et al., 2023) exist with neglected problems. Many real-world graphs, such as social networks and internet, tend to have scale-free and hierarchical structure (Chami et al., 2019). When these graphs are embedded in the Euclidean space, the distance of border nodes shrink as the graph is extended, which may lead to very close distance of leaf nodes belonging to different subtrees (Figure 1a). In other words, it is not intuitive that nodes with very different semantics may be very close to each other in the Euclidean vector space. Bourgain's theorem (Linial et al., 1995) shows that Euclidean space is unable to obtain comparably low distortion for trees, even using an unbounded number of dimensions. And after calculation, we find that most of the knowledge graphs in KBVQA also have a hierarchical structure. However, due to the natural defects of the Euclidean space, the embedding structure of knowledge graph in Euclidean space suffers from a large distortion, which limits the performance of multi-hop reasoning on knowledge evidence.



(a) Knowledge embedding in euclidean space. As the tree-like graph embedded in Euclidean space spreads, the edge nodes get closer to each other, leading to semantic distortion of the structure.

(b) Knowledge embedding in hyperbolic space. Each line segment in a Poincaré disc has the same length, e.g. a and b. Thus the relative positions of the nodes can be properly modeled.

Figure 1: Examples of graph embedding in euclidean and hyperbolic space.

In contrast, due to certain superior properties, hyperbolic space allows the embedding of graph structure with low distortion (Chami et al., 2020).

* Corresponding author

LREC-COLING 2024, pages 8505–8515
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

For example, 1) in the mapping from Euclidean space to Poincaré model, the angles between embedded vectors of knowledge are identical. That is, the mapping is conformal. 2)When the crowded structure modeled in the euclidean space is transferred to the hyperbolic space, due to the distortion of the space, it is able to maintain a reasonable relative position (Figure 1b). 3) In hyperbolic space, the volume grows exponentially with radius, while in Euclidean space it grows polynomially. This means that the hyperbolic space can embed more information than Euclidean space when they have same dimensions. These tree-like properties of the hyperbolic space are key features exploited for the knowledge graphs embedding of KBVQA.

Further, for complex multi-hop knowledge reasoning questions, the model requires multiple stages to iteratively perform the reasoning process. Therefore, different spatial patterns are needed to match the reasoning process at different stages of the reasoning. In addition, hyperbolic space is a non-Euclidean space with negative constant curvature, and the curvature determines the distribution structure of the space. For reasoning over complex questions, a single curvature cannot accurately portray the optimal spatial structure required for the knowledge graphs at different stages of reasoning. Therefore, it is necessary to use multiple adaptive curvatures to fit out different spatial structures that are best suited for each reasoning stage, respectively.

Therefore, we propose a multi-hop reasoning framework embedded in hyperbolic space, named HyperMR. It can utilize hypergraph (Kim et al., 2020; Heo et al., 2022) to encode higher-order semantics of complex structures in hyperbolic space with low distortion, and perform multi-hop knowledge reasoning. Specifically, we transform the input features in Euclidean space into hyperbolic embeddings to restore the complex hierarchical structure of the free-scale knowledge graphs. Then the question hypergraph and knowledge hypergraph are constructed in the hyperbolic space, and the higher-order semantics and relations of multi-hop fact knowledge are captured by the hyperedges. After that, we construct a hypergraph reasoning network based on the transformer encoder layer. Due to the different spatial structures required at different stages of question reasoning, and the different curvatures determine the degree of warping of the space. We map different layers of the hypergraph reasoning network into multiple hyperbolic spaces, adjusting the structural properties of each space separately with multiple trainable curvatures, achieving low-distortion, fine-grained reasoning. In summary, our contributions are as follows:

- We propose a hyperbolic hypergraph multi-hop reasoning method, named HyperMR, which exploits the high fidelity of hyperbolic geometry to model and reason about hierarchical knowledge evidence with low distortion, effectively improving model performance.
- We map the different stages of reasoning into multiple adjustable hyperbolic spaces, to obtain the optimal spatial structure that best suits the different stages of reasoning, achieving a low-distortion, fine-grained reasoning process.
- Extensive experiments on three mainstream datasets fully demonstrate the advanced multi-hop reasoning capability of HyperMR for strong-hierarchy data and the rationality of multi-space reasoning.

2. Related Work

2.1. Knowledge-based Visual Question Answering

The core of this task is the acquisition and integration of knowledge to answer questions that exceed the content of the images. Early explorations were dominated by the retrieval methods. Wang et al. (2017) parsed the input into a structured query and retrieved supporting knowledge from the fixed knowledge base to obtain the answer. After that, Kim et al. (2018) argued that the previous approaches ignored the interaction between commonsense knowledge and questions, so the answers were obtained by calculating the soft attention score between the retrieved knowledge and the question as the basis for matching. Yu et al. (2020); Zhu et al. (2021) constructed visual, factual, and semantic graphs, respectively, and performed joint reasoning on three graphs to find complementary evidence of multi-modality. Since the over-smoothing of graph convolution leads to limited long-distance message propagation in multi-hop reasoning, HAN (Kim et al., 2020) and Hypergraph Transformer (Heo et al., 2022) achieved efficient multi-hop reasoning by constructing hypergraphs. However, the hypergraph method uses triples as node units, focusing on semantic clustering and propagation of nodes while ignoring the accurate structure representation of the graph. In order to make the structure information of the graph work as it should, we use the low-distortion property of the hyperbolic space to model the knowledge graph.

2.2. Knowledge Graph Embeddings in Hyperbolic Space

Previous works (Huang et al., 2021; Wang et al., 2020) have conducted a lot of research on knowledge graphs embedded in Euclidean space and achieved a lot of outcomes, but most of them have problems such as too high embedding dimension or distorted embedding structure. Hyperbolic space has shown increasing interests due to its ability to model data with potential hierarchical structure. Nickel and Kiela (2017) proposed to learn a hierarchical representation of symbolic data through Poincaré embeddings. Chami et al. (2020) demonstrated that hyperbolic space are able to embed more information than Euclidean spaces in lower dimensions, and also explored the importance of curvature on spatial distributions. Further, by analyzing the effect of curvature for distortion of data with different hyperbolicity, Fu et al. (2021) demonstrated that adaptive fusion of hierarchical topology with feature information is necessary. By leveraging hyperbolic geometry for low-distortion modeling of hierarchical structure and applying learnable curvature, we accomplish accurate understanding and multi-hop reasoning about knowledge graphs.

2.3. Hyperbolic Geometry

Hyperbolic geometry is a necessary background for this work. Here we review the key points in hyperbolic geometry involved in our work. More details are described in (Chami et al., 2019) and (Ganea et al., 2018).

The hyperbolic space is the unique complete, simply connected Riemannian manifold as well as an isotropic space with constant negative sectional curvature. The curvature measures how far a geometric object deviates from the plane, where it determines the warping of hyperbolic space. The hyperbolic space degenerates to Euclidean space when the curvature is 0. There are several equivalent models that describe the hyperbolic space, such as Hyperboloid model and Poincaré model, etc. Here we focus on the study of multi-hop reasoning in the Poincaré ball. Assuming that the n-dimensional embedded Poincaré sphere has curvature $-c\,(c>0)$, its corresponding Riemannian manifold is: $\mathbb{H}^{n,c}=\left\{x\in\mathbb{H}^{n}\mid\|x\|^{2}<\frac{1}{c}\right\}$, where $\|\cdot\|$ is the L2 norm in the Euclidean space.

For any point $x_{\mathcal{H}}\in\mathbb{H}^{n,c}$ embedded in a hyperbolic space, there exists a tangent space $\mathcal{T}_{x}\mathbb{H}^{n,c}$, which is a local, first-order approximation of the hyperbolic manifold at x and the restriction of the Minkowski inner product to $\mathcal{T}_{x}\mathbb{H}^{n,c}$ is positive definite. Its geometric meaning is that it contains curves in all directions passing through point x in the manifold, which is locally approximated as Euclidean space at point x. This is very useful for performing various undefined operations in hyperbolic space. The method of mutual mapping between the hyperbolic space $\mathbb{H}^{n,c}$ and the tangent space at x is modeled in the Poincaré ball is given:

$$\exp_{x}^{c}(z)=z\oplus^{c}\left(\tanh\left(\sqrt{c}\frac{\lambda_{x}^{c}\|z\|}{2}\right)\frac{z}{\sqrt{c}\|z\|}\right),\qquad(1)$$

$$\log_{x}^{c}(y)=\frac{2}{\sqrt{c}\lambda_{x}^{c}}\tanh^{-1}\left(\sqrt{c}\|-x\oplus^{c}y\|\right)\frac{-x\oplus^{c}y}{\sqrt{c}\|-x\oplus^{c}y\|},\qquad(2)$$

where $x\in\mathcal{T}_{x}\mathbb{H}^{n,c},\ y\in\mathbb{H}^{n,c}$, the exponential mapping maps $\mathcal{T}_{x}\mathbb{H}^{n,c}$ to $\mathbb{H}^{n,c}$, the logarithmic mapping maps $\mathbb{H}^{n,c}$ to $\mathcal{T}_{x}\mathbb{H}^{n,c}$, $\lambda_{x}^{c}=2/(1-c\|x\|^{2})$is the conformal factor and the $\oplus^{c}$ refers to the Möbius addition (Ganea et al., 2018). Since the addition of two points in Poincaré ball may derive a point outside the ball, for this Möbius addition is a generalization of the addition operation in Euclidean space. It is defined as follows:

$$x\oplus^{c}y=\frac{\left(1+2c\langle x,y\rangle+c\|y\|^{2}\right)x+\left(1-c\|x\|^{2}\right)y}{1+2c\langle x,y\rangle+c^{2}\|x\|^{2}\|y\|^{2}},\qquad(3)$$

where $\langle x,y\rangle$ is the Euclidean inner product operation.

The path through the shortest distance between two points x,y embedded in non-Euclidean geometry is called the geodesic, which is a generalization of the straight-line distance in Euclidean geometry. This is very meaningful for node embedding and multi-hop reasoning in knowledge graphs. The distance in the Poincaré ball is defined as follows:

$$d_{c}(x,y)=\frac{2}{\sqrt{c}}\operatorname{arctanh}\left(\sqrt{c}\|-x\oplus^{c}y\|\right).\qquad(4)$$

3. Method

In this section we will introduce HyperMR, a framework for KBVQA multi-hop reasoning in hyperbolic space, which implements higher-order semantics reasoning about knowledge under low-distortion hierarchy in hyperbolic space. Due to the lack of pre-trained word embedding in hyperbolic space, like GloVe (Pennington et al., 2014), we map their embedded features in Euclidean space to hyperbolic space (Eq.5) after entity linking of questions, images and knowledge graphs. Then we construct hypergraph for knowledge and questions and use hyperbolic hypergraph reasoning network to perform higher-order semantics reasoning on hyperedges. Each stage of the reasoning process, consisting of an attention network, is mapped to a different hyperbolic space. Multiple trainable curvatures are used in spatial adjustments during embedding and reasoning to find the most appropriate spatial structure for the current stage.

3.1. Task Definition

We describe the notation of the KBVQA task without loss of generality. Given the questions $\mathcal{Q}$, the

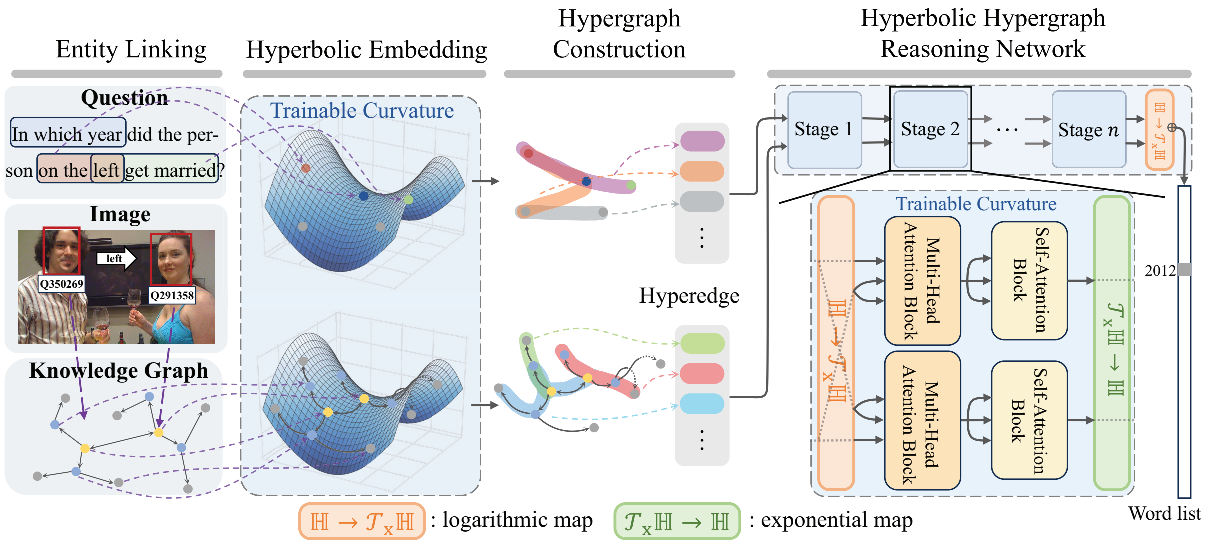

Entity Linking
Hyperbolic Embedding
Hypergraph Construction
Hyperbolic Hypergraph Reasoning Network
Question
In which year did the person on the left get married?
Image
left
Q350269
Q291358
Knowledge Graph
Trainable Curvature
Hyperedge
Stage 1
Stage 2
Stage n
Trainable Curvature
$\mathbb{H} \to \mathcal{T}_x\mathbb{H}$
Multi-Head Attention Block
Self-Attention Block
Multi-Head Attention Block
Self-Attention Block
$\mathcal{T}_x\mathbb{H} \to \mathbb{H}$
2012
Word list
$\mathbb{H} \to \mathcal{T}_x\mathbb{H}$: logarithmic map
$\mathcal{T}_x\mathbb{H} \to \mathbb{H}$: exponential map

Figure 2: The overview of HyperMR. First the question-image pairs are linked to entities of knowledge graphs and embeded into the hyperbolic space. After that hypergraphs are constructed for the questions and knowledge, and multi-hop reasoning is performed using Hyperbolic Hypergraph Reasoning Network, where each stage is mapped to a different hyperbolic space with different trainable curvatures. In the end, the joint representation is transferred to the tangent space to get the answer.

images $\mathcal{I}$ and the knowledge graphs $\mathcal{G} = (\mathcal{V}, \mathcal{R}, \mathcal{S})$, where $\mathcal{V}$, $\mathcal{R}$, $\mathcal{S}$ represent the set of entities, relations and triples, respectively. The triplet is defined as $\mathcal{S} = \{(h,\ r,\ t) \mid h, t \in \mathcal{V},\ r \in \mathcal{R}\} \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$, where $h,\ t,\ r$ stand for the head entities, the tail entities and the relations between them, respectively. The purpose of the task is to find an inference process $\varphi(\cdot)$, which reasons about the knowledge graphs based on questions and images to derive the probability distribution P_v of the set of nodes and thus the answer. $P_v = \varphi(\{v_i \in \mathcal{V} \mid \mathcal{Q}, \mathcal{I}, \mathcal{G}\}) \Rightarrow \text{Ans} \in \mathbb{D}^d$, where $\mathbb{D}^d$ is the list of words of size d.

3.2. Entity Linking

Visual targets (e.g., name, object, position, etc.) in images as well as key entities in questions are detected. These visual targets and entities are then matched to the knowledge entities in the knowledge graphs, constituting question-image-KG entity links.

3.3. Hyperbolic Embedding

After embedding the questions and the knowledge graphs linked by question-image pairs into the Euclidean space, we utilize the exponential mapping to map the features embedded in the Euclidean space to hyperbolic manifold in the Poincaré ball. Assume that $x^{(0,\mathcal{R})} \in \mathbb{R}^d$ is a d-dimensional vector embedded in Euclidean space. We map features from the tangent space $\mathcal{T}_0\mathbb{H}^{n,c}$ centered at the point $\left(0, \boldsymbol{x}^{0,\mathcal{R}}\right)$ to the hyperbolic manifold $\mathbb{H}^{n,c}$:

$$\begin{aligned} \boldsymbol{x}^{0,\mathcal{H}} &= \exp_0^c\left(\left(0, \boldsymbol{x}^{0,\mathcal{R}}\right)\right) \\ &= \left(0, \tanh\left(\sqrt{c}\,\|\boldsymbol{x}\|\right)\frac{\boldsymbol{x}}{\sqrt{c}\,\|\boldsymbol{x}\|}\right). \end{aligned} \tag{5}$$

3.4. Hypergraph Construction

Based on previous works (Heo et al., 2022; Kim et al., 2020), we construct the hypergraph in a hyperbolic space. Hypergraph is a graph representation based on higher-order semantics, which consists of a set of nodes and hyperedges, and a hyperedge can connect more than two nodes at the same time. It implements abstract clustering of graphs based on higher-order relationships of different nodes.

For the knowledge hypergraph, we utilize the traversal approach to construct the hypergraph, and preserve the higher-order semantics by using the triple as a basic unit, starting with the linked entity node and considering connecting all nodes linked to it via hyperedges. For the n-hop questions, the traversal connects n facts and merges them into a hyperedge.

For the question hypergraph, there exists an intuitive manner for constructing graphs. That is, the semantics in a sentence often consists of a sense-group of n-gram phrases, e.g., the question in Figure 2. Therefore, we take the sense-groups of the adjacent 3-gram in the sentence as node units, and traverse them to construct the hypergraph.

3.5. Hyperbolic Hypergraph Reasoning Network

In this part, we first performs the linear transform on the hyperbolic input, after which the features are mapped to the tangent space for multi-head attention reasoning, and then mapped back to the hyperbolic space. With the above process, dimension normalization and higher-order semantics reasoning for hypergraphs in hyperbolic space are achieved.

3.5.1. Poincaré Ball Linear Transform

After mapping the features from the Euclidean space to the hyperbolic space and constructing the hypergraph, the excessive dimension becomes a problem. Especially for the hypergraph construction, if taking the 3-gram phrase as a node of question hypergraph, the feature of a node is 900 dimensions. Previous work (Liu et al., 2019) proved that: excessive dimensions provide very limited improvement in the representation ability of hyperbolic manifolds.

Thus, according to the work of Chami et al. (2019), we perform a linear transformation on the embedding vectors in hyperbolic space using the weight matrix $W \in \mathbb{R}^{n' \times n}$ and the bias $b \in \mathbb{R}$, where the n' and n are the matrix dimensions after and before mapping. Specifically, matrix multiplication in the Poincaré ball is defined by Möbius scalar multiplication:

$$W \otimes^c \boldsymbol{x} = \exp_0^c \left(W \ \log_0^c \left(\boldsymbol{x} \right) \right). \tag{6}$$

For bias addition, b is defined as a vector located in the tangent space $\mathcal{T}_0 \mathbb{H}^{n,c}$, which is parallel transported to hyperbolic point x in the tangent space $\mathcal{T}_x \mathbb{H}^{n,c}$, and then mapped into the manifold.

$$\boldsymbol{x}^{\mathcal{H}} \oplus^c b = \exp_x^c \left(P_{o \to x}^c \left(b \right) \right), \tag{7}$$

where $P_{o \to x}^c \left(v \right) = \log_x^c \left(\boldsymbol{x} \oplus^c \exp_0^c \left(v \right) \right) = \frac{\lambda_o^c}{\lambda_x^c} v$ is the parallel transportation from $\mathcal{T}_0 \mathbb{H}^{n,c}$ to $\mathcal{T}_x \mathbb{H}^{n,c}$.

3.5.2. Multiple Hyperbolic Space Reasoning

Humans usually divide the reasoning process into several stages, such as dividing a huge question into several parallel sub-questions, or completing the reasoning process in a step-by-step progression. Inspired by this, we divide the reasoning process into different stages as well, taking advantage of the property that hyperbolic space can be warped and adjusted, so as to find the most suitable spatial structure for each reasoning stage. Final select the hyperedge that contains the correct answer. Algorithm 1 shows the inputs and outputs as well as the curvature assignments for each stage of the reasoning process:

Algorithm 1 Multiple hyperbolic space reasoning

```
Input: The value of nodes $NV$; A set of curvatures
  $C = \{c_1, c_2, ..., c_n\}$
Output: The feature of nodes $NF$
  initialization
  for $i \in [1, n]$ do
    if $i = 1$ then
      $NF_{i-1} = hyperbolicEmb(NV)^{c_i}$
    else
      $stage_i(NF_{i-1})_{in}^{ci-1} = stage_{i-1}(NF_{i-1})_{out}^{c_{i-1}}$
      $NF_i = stage_i(NF_{i-1})_{out}^{c_i}$
    end if
  end for
  return $NF_n$
```

In order to accomplish higher-order semantics reasoning for hypergraphs in hyperbolic space, the encoder layer of transformer (Vaswani et al., 2017) is used to construct multi-head attention block and self-attention block. In addition, we process the attention reasoning blocks using exponential and logarithmic mappings, allowing the operations of attention, normalization, and feed-forward network in the standard encoder layer to be implemented in hyperbolic space.

Let the knowledge hyperedges be E^k, and the question hyperedges be E^q. For two multi-head attention blocks, we set query Q, key K, and value V as respectively:

$$Q_k = E^k W_{Q_k}; K_q = E^q W_{K_q}; V_q = E^q W_{V_q},$$

$$Q_q = E^q W_{Q_q}; K_k = E^k W_{K_k}; V_k = E^k W_{V_k},$$

where $W_{(\cdot)}$ is the attention weight matrix.

Then scaling dot product attention $\mathrm{Att}\left(Q, K, V\right) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$ is performed by $\mathrm{Att}\left(Q_k, K_q, V_q\right)$ and $\mathrm{Att}\left(Q_q, K_k, V_k\right)$.

For the two self-attention blocks, we use similar approach as above, but where the query, key, value of the multi-head attention is set to the same value: $\mathrm{Att}\left(Q_q, K_q, V_q\right)$ and $\mathrm{Att}\left(Q_k, K_k, V_k\right)$.

3.6. Decoding

The representations of the knowledge hyperedges and the question hyperedges are updated by multiple hyperbolic space reasoning network, and aggregated into a joint vector representation j. We use the logarithmic mapping $\log_0^c(\cdot)$ to project j into the tangent space of the origin to obtain $j^{(0,\mathcal{R})} \in \mathcal{T}_0 \mathbb{H}^{n,c}$, after which we compute the dot-product similarity $p = j^{(0,\mathcal{R})} \mathbb{D}^T$ with the answer word list $\mathbb{D}^{d \times w}$,w being the word embedding dimension. Finally we use the cross-entropy between the calculated answer and ground-truth as the loss function.

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

Table 1: Statistics of three benchmark datasets: PathQuestion (PQ), PathQuestion-Large (PQL) and Knowledge-aware Visual Question Answering (KVQA).

Hyperbolicity δ=1.5	Original (ORG)			Paraphrased (PRP)			Mean
	1-hop	2-hop	3-hop	1-hop	2-hop	3-hop	
BLSTM	-	-	-	-	-	-	51.0
GCN (Kipf and Welling, 2016)	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
HyperTransformer (Heo et al., 2022)	88.1	90.2	91.0	87.8	90.5	90.7	89.7
DSAMR (Sun et al., 2024)	89.1	91.0	91.1	89.0	90.7	91.1	90.3
HyperMR	**90.7**	**91.9**	**91.9**	**90.7**	**91.7**	**91.7**	**91.4**

Table 2: The accuracy of advanced methods for various types in KVQA dataset. ORG and PRP denote the question type.1-hop, 2-hop, 3-hop represent the number of hops in the ground-truth path. $\delta = 1.5$ stands for the hyperbolicity of the KVQA dataset as 1.5.

4. Experiment

4.1. Experiment Setup

4.1.1. Datasets

In this work, we evaluate our method on the Knowledge-aware VQA dataset (KVQA) and the textual question answering datasets PathQuestion (PQ) and PathQuestion-Large (PQL), which focus on multi-hop reasoning ability. In addition, we compute Gromovs δ - hyperbolicity, a notion from group theory that measures how tree-like a graph is. Lower δ represents a graph that is more hierarchical, and when δ = 0 represents a tree. Hyperbolicity is used to aid in investigating the properties of knowledge representation in hyperbolic spaces.

1. **KVQA** (Shah et al., 2019) is the largest dataset for exploring VQA over knowledge graph. It consists of 183,007 question-answer pairs involving more than 39,414 named entities based on Wikidata (Vrandečić and Krötzsch, 2014) and 24,602 images from Wikipedia.

2. **PQ** and **PQL** (Zhou et al., 2018) datasets are two question-and-answer datasets that focus on multi-hop reasoning, including 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 knowledge of PQ and PQL has different structure and PQL contains more knowledge facts.

4.1.2. Baselines

This work is the first to conduct the KBVQA task in hyperbolic space. Besides, the text QA baselines in the PQ and PQL datasets is not in the VQA domain and only serves as a supporting proof. Therefore the baselines contain only the more advanced methods of VQA embedded in Euclidean space. For these KBVQA methods, we divide them into three categories: graph-based, memory-based and attention-based networks. In order to evaluate the pure reasoning ability of models regardless of entity linking performance, we give ground-truth named entities in the images for all methods.

Graph-based networks. Graph Convolutional Networks (GCN) (Kipf and Welling, 2016) and Gated Graph Neural Networks (GGNN) (Li et al., 2016) propagate information among neighbors by learning the question graph and knowledge graph, and aggregating nodes to complete the update of the graph representation.

Memory-based networks. Memory Network (MemNN) (Sukhbaatar et al., 2015), a baseline in the early stage when memory-based approaches were still popular, embeds knowledge in slots and

Hyperbolicity δ	PathQuestion				PathQuestion-Large			
	PQ-2H	PQ-3H	PQ-M	Mean	PQL-2H	PQL-3H	PQL-M	Mean
	δ=2	δ=2	δ=2		δ=0	δ=1	-	
Seq2Seq (Sutskever et al., 2014)	89.9	77.0	-	-	71.9	64.7	-	-
MemNN (Sukhbaatar et al., 2015)	89.5	79.2	86.8	85.2	61.2	53.6	55.8	56.9
KV-MemNN (Miller et al., 2016)	91.5	79.4	85.2	85.4	70.5	63.4	68.6	67.5
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	73.4	71.8	65.7	66.9	68.1
IRN-weak (Zhou et al., 2018)	91.9	83.3	85.8	87.0	63.0	61.8	62.4	62.4
SRN (Qiu et al., 2020)	96.3	89.2	89.3	91.6	78.6	77.5	78.3	78.1
HyperTransformer (Heo et al., 2022)	96.4	90.3	89.5	92.1	90.5	77.9	94.5	87.6
DSAMR (Sun et al., 2024)	**98.4**	**91.1**	**91.7**	**93.7**	95.6	81.7	98.8	92.0
HyperMR	96.2	90.5	89.5	92.1	**96.3**	**83.7**	**98.9**	**93.0**

Table 3: Accuracy of advanced methods under PQ and PQL datasets. Hyperbolicity is also given for different question types. The methods in the top half are of the fully supervised type, and the bottom half including our methods are of the weakly supervised type.

computes the soft attention between each memory slot and the question to obtain a joint representation.

Attention-based networks. Bilinear Attention Networks (BAN) (Kim et al., 2018), Hypergraph Attention Networks (HAN) (Kim et al., 2020), Hypergraph Transformer (Heo et al., 2022) and Dual-Stream Attention (DSAMR) (Sun et al., 2024) consider interactions between knowledge and question based on co-attention mechanism. Among them, BAN is updated by computing the attention of the knowledge entity with the question, and both HAN and Hypergraph Transformer belong to hypergraph-based reasoning. The difference is that HAN performs the construction of hyperedges using randomly selected nodes as starting nodes, while Hypergraph Transformer uses the knowledge entities linked to the image-question pairs as the starting point for the construction of hyperedges. DSAMR adds the dual-stream attention module to Hypergraph Transformer, which effectively solves the problem of too much redundant information in the last layer of the attention network, and extracts the necessary knowledge for answer prediction.

4.1.3. Implementation Details

The question and knowledge graphs are first initialized in Euclidean space as a 300-dimensional vector by GloVe (Pennington et al., 2014). After mapping to the hyperbolic space, subsequent operations are performed with 256 dimensional hidden size. For person entities in KVQA, we use the well-known pre-trained models RetinaFace (Deng et al., 2020) for face detection and ArcFace (Deng et al., 2022) for face feature extraction. For PQ and PQL datasets, we divide the training set, validation set, and testing set according with 8:1:1. Adam (Kingma and Ba, 2015) is used as an optimizer for all learnable parameters. All experiments report an average accuracy of 5 trials.

4.2. Experiment Results

4.2.1. Results for KVQA

As shown in Table 2, our proposed method outperforms baselines in all settings, especially when the KVQA dataset accuracy is close to 1. Graph-based networks focus on information aggregation and transfer among nodes, memory-based networks and attention-based networks focus on attention interaction for knowledge and questions. They both ignore that knowledge has a hierarchical structure, which is crucial for knowledge understanding and multi-hop reasoning. The improvement we obtained is due to the low-distortion modeling of hierarchical knowledge graphs by hyperbolic geometry. The auxiliary experimental proof is described in detail in section 4.2.4.

4.2.2. Results for PQ and PQL

In order to evaluate the multi-hop reasoning ability of the proposed method, we additionally conducted experiments on the PQ and PQL datasets of textual QA task. The PQ and PQL datasets have three question types: 2-hop, 3-hop and M, to denote the number of hops in the ground-truth path, with M representing a mixture of 2-hop and 3-hop.

The results, as shown in Table 3, show that our method achieves great improvement in all question types on the PQL dataset. For PQL-2H and PQL-M, our method achieves an advanced score of 95.1% and 97.5%, respectively. For PQL-3H the lower precision is due to fewer QA pairs being included (2,625 in PQL-M compared to 1,031 in PQL-3H, Table 1), leading to insufficient training. The results

	Original (ORG)			Paraphrased (PRP)			Mean
	1-hop	2-hop	3-hop	1hop	2-hop	3-hop	
Single space w/ fixed curvature	90.2	91.2	91.5	90.1	91.2	91.4	90.9
Single space w/ trainable curvature	90.5	91.6	91.7	90.5	91.5	91.5	91.2
Multiple space w/ trainable curvature	**90.7**	**91.9**	**91.9**	**90.7**	**91.7**	**91.7**	**91.4**
HyperMR by hyperboloid	84.3	85.0	85.2	83.7	86.9	85.6	85.1

Table 4: Results for the KVQA dataset at different settings. The top three are modeled by a Poincaré model, and the bottom one is modeled by a hyperbolic model.

	PathQuestion				PathQuestion-Large			
	PQ-2H	PQ-3H	PQ-M	Mean	PQL-2H	PQL-3H	PQL-M	Mean
Single space w/ fixed curvature	93.2	87.1	86.0	88.8	94.2	79.2	96.6	90.0
Single space w/ trainable curvature	94.2	87.7	90.1	90.1	95.1	81.3	97.5	91.3
Multiple space w/ trainable curvature	**96.2**	**90.5**	**90.3**	**92.3**	**96.3**	**83.7**	**98.9**	**93.0**
HyperMR by hyperboloid	84.8	81.4	82.1	82.8	94.4	78.9	95.4	89.6

Table 5: Results for PQ and PQL datasets at different settings. The top three are modeled by a Poincaré model, and the bottom one is modeled by a hyperbolic model.

of the poor performance on the PQ dataset are analyzed in detail in section 4.2.3.

4.2.3. Analyze for Hierarchical Data Representation in Hyperbolic Geometry

We measured the knowledge of KVQA, PQ and PQL dataset Gromovs δ-hyperbolicity, a notion that measures how tree-like a graph is. Where the δ corresponding to the knowledge of PQL-2H is 0, indicating that it is completely tree-structure. Corresponding to the experimental results in Table 2, the accuracy improvement of PQL-2H is also the most significant. The δ of the knowledge of PQL-3H and KVQA are 1 and 1.5, respectively, with a corresponding reduction in the accuracy enhancement. While the δ of the knowledge in PQ dataset is 2, hyperbolicity is further weakened and no longer has a clear hierarchical structure to be embedded in the hyperbolic space. In addition, compared to the PQL and KVQA datasets, the PQ dataset has much fewer knowledge entities (Table 1) , and the knowledge structure is simpler. That is, poorly hierarchical and simple structured graph data is not suitable for embedding in hyperbolic space. Even so, it still obtained comparable performance to HyperTransformer on the PQ dataset.

Besides modelling the hyperbolic space using the Poincaré ball model, we also tested the performance of the hyperbolic model, as shown in Tables 4 and 5. The hyperbolic model has a large gap compared to the Poincaré model in all question types across all datasets. Because the model was very unstable during the training phase, with large fluctuations in accuracy and loss, we believe that the hyperbolic model is not suitable for the data distribution in this work.

4.2.4. Effect of Multiple Hyperbolic Space

We find that single space with trainable curvature in Table 4 and 5 performs worse compared to HyperTransformer in Table 2 and 3. This is counterintuitive, since when the trainable curvature is adjusted to 0, the structure of the hyperbolic space flattens out and degenerates into Euclidean space. That means at worst it should get results comparable to a Euclidean space, not worse. We believe that using a single trainable curvature in a multi-layer network cannot accurately fit all the requirements of the training process. When the update weights of different layers in the backpropagation are not distributed evenly, e.g., one layer plays too large a role in the curvature adjustment process, resulting in the curvature not being adjusted to 0 (the curvature in the PQ dataset in Fig.3 is very close to 0), achieving worse results than the Euclidean space.

In contrast, multiple hyperbolic spaces with multiple adjustable curvatures achieve comparable performance to HyperTransformer. This demonstrates that mapping the different stages of reasoning into spaces with different forms achieves a more fine-grained training process, and also more accurately fits the spatial structural requirements of the different stages during reasoning.

5. Conclusion and Future Work

In this paper, we introduced HyperMR, a multi-hop reasoning framework over knowledge graph for KBVQA task. It exploits the expressiveness of hy-

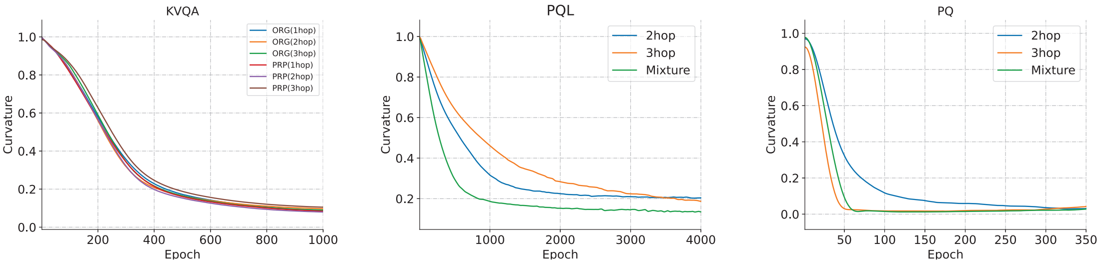


Figure 3: Convergence process of one trainable curvature in trebling hyperbolic spaces.

perbolic geometry for low-distortion modeling of hierarchical structure, thus performing multi-hop reasoning on complex knowledge evidence. In addition, In order to satisfy the spatial morphology requirements of different stages of the reasoning process, we map each reasoning stage into multiple hyperbolic spaces , achieving a more fine-grained reasoning process. We conducted extensive experiments on the KVQA, PQ and PQL datasets to demonstrate the effectiveness of HyperMR for strong-hierarchy knowledge graphs.

In future work, We plan to explore the relationship between different stages of reasoning and spatial morphology. How more "curved" or more "flat" spatial morphology contributes to the graph modeling and multi-hop reasoning. This will further reveal the properties of hyperbolic spaces in graph modeling.

6. Ethics Statement

We ensure that our research was conducted in accordance with the relevant ethical guidelines and regulations. We obtained all necessary permissions and approvals from any involved institutions or organizations, and maintained the confidentiality and anonymity of any participants or sensitive data throughout the study. In addition, we made every effort to minimize any potential harm or negative consequences that could arise from our research.

7. Bibliographical References

Omar Adjali, Paul Grimal, Olivier Ferret, Sahar Ghannay, and Hervé Le Borgne. 2023. Explicit knowledge integration for knowledge-aware visual question answering about named entities. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, ICMR, page 29–38.

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.

Antoine Bordes, Jason Weston, and Nicolas Usunier. 2014b. Open question answering with weakly supervised embedding models. In *Machine Learning and Knowledge Discovery in Databases: European Conference. Proceedings, Part I 14*, PKDD, pages 165–180. Springer.

Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. 2020. Low-dimensional hyperbolic knowledge graph embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL, pages 6901–6914.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. 2019. Hyperbolic graph convolutional neural networks. In *Advances in neural information processing systems*, volume 32 of *NeurIPS*.

Hai Cui, Tao Peng, Ridong Han, Jiayu Han, and Lu Liu. 2023. Path-based multi-hop reasoning over knowledge graph for answering questions via adversarial reinforcement learning. *Knowledge-Based Systems*, 276:110760.

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 *International Conference on Learning Representations*, ICLR.

J Deng, J Guo, J Yang, N Xue, I Kotsia, and S Zafeiriou. 2022. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, CVPR, pages 5203–5212.

Xingcheng Fu, Jianxin Li, Jia Wu, Qingyun Sun, Cheng Ji, Senzhang Wang, Jiajun Tan, Hao Peng, and S Yu Philip. 2021. Ace-hgnn: adaptive curvature exploration hyperbolic graph neural network. In *2021 IEEE international conference on data mining, ICDM '21*, pages 111–120. IEEE.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, volume 31 of *NeurIPS*.

Yu-Jung Heo, Eun-Sol Kim, Woo Suk Choi, and Byoung-Tak Zhang. 2022. Hypergraph Transformer: Weakly-supervised multi-hop reasoning for knowledge-based visual question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL, pages 373–390.

Junjie Huang, Huawei Shen, Liang Hou, and Xueqi Cheng. 2021. Sdgnn: Learning node representation for signed directed networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35 of *AAAI*, pages 196–203.

Eun-Sol Kim, Woo Young Kang, Kyoung-Woon On, Yu-Jung Heo, and Byoung-Tak Zhang. 2020. Hypergraph attention networks for multimodal learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, CVPR, pages 14581–14590.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in neural information processing systems*, volume 31 of *NeurIPS*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, ICLR.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, ICLR.

Yujia Li, Richard Zemel, Marc Brockschmidt, and Daniel Tarlow. 2016. Gated graph sequence neural networks. In *4th International Conference on Learning Representations*, ICLR.

Nathan Linial, Eran London, and Yuri Rabinovich. 1995. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15:215–245.

Qi Liu, Maximilian Nickel, and Douwe Kiela. 2019. Hyperbolic graph neural networks. In *Advances in neural information processing systems*, volume 32 of *NeurIPS*.

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*, EMNLP, pages 1400–1409.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Advances in neural information processing systems*, volume 30 of *NIPS*.

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.

Yunqi Qiu, Yuanzhuo Wang, Xiaolong Jin, and Kun Zhang. 2020. Stepwise reasoning for multi-relation question answering over knowledge graph with weak supervision. In *Proceedings of the 13th international conference on web search and data mining*, WSDM, pages 474–482.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, volume 28 of *NIPS*.

Yanhan Sun, Zhenfang Zhu, Zicheng Zuo, Kefeng Li, Shuai Gong, and Jiangtao Qi. 2024. Dsamr: Dual-stream attention multi-hop reasoning for knowledge-based visual question answering. *Expert Systems with Applications*, 245:123092.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, volume 27 of *NIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, volume 30 of *NIPS*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Henge. 2017. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI, pages 1290–1296.

Yanbang Wang, Yen-Yu Chang, Yunyu Liu, Jure Leskovec, and Pan Li. 2020. Inductive representation learning in temporal networks via causal anonymous walks. In *International Conference on Learning Representations*, ICLR.

Jing Yu, Zihao Zhu, Yujing Wang, Weifeng Zhang, Yue Hu, and Jianlong Tan. 2020. Cross-modal knowledge reasoning for knowledge-based visual question answering. *Pattern Recognit*, 108:107563.

Jing Zhang, Xiaokang Zhang, Jifan Yu, Jian Tang, Jie Tang, Cuiping Li, and Hong Chen. 2022. Subgraph retrieval enhanced model for multi-hop knowledge base question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL, pages 5773–5784.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2021. Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, IJCAI, pages 1097–1103.

8. Language Resource References

Bollacker, Kurt and Evans, Colin and Paritosh, Praveen and Sturge, Tim and Taylor, Jamie. 2008. *Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge*.

Shah, Sanket and Mishra, Anand and Yadati, Naganand and Talukdar, Partha Pratim. 2019. *Kvqa: Knowledge-aware visual question answering*.

Vrandečić, Denny and Krötzsch, Markus. 2014. *Wikidata: A Free Collaborative Knowledgebase*. ACM.

Zhou, Mantong and Huang, Minlie and Zhu, Xiaoyan. 2018. *An Interpretable Reasoning Network for Multi-Relation Question Answering*. COLING.