第四讲 英文学术论文之写作思路
——相关工作和方法

于静 副研究员

中国科学院信息工程研究所

课程主页：https://mmlab-iie.github.io/course/

2022.07 @ Bilibili

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学
University of Chinese Academy of Sciences

一篇论文的组成——相关工作

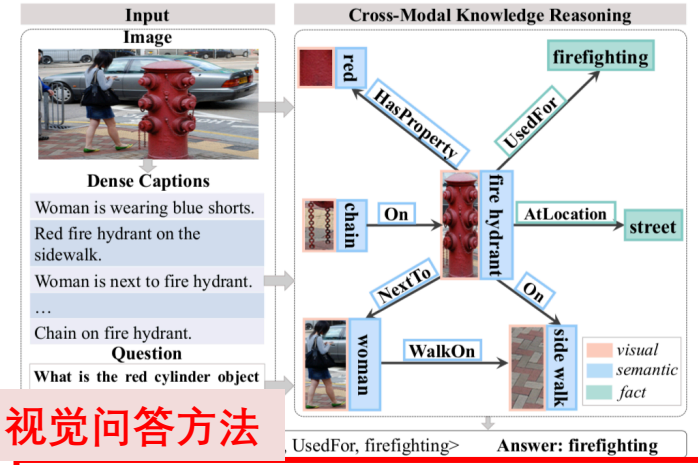
基本要求

- 包括理解本文的所有主题
- 包括问题相关的所有工作
- 从不同维度划分主题
- 同一主题方法归类
- 总结问题
- 引出本研究的区别和贡献

不要所有看过的论文！

不要罗列写上的论文！

一篇论文的组成——相关工作



视觉问答方法

Visual Question Answering. The typical solutions for VQA are based on the CNN-RNN architecture [Malinowski *et al.*, 2015] and leverage global visual features to represent image, which may introduce noisy information. Various attention mechanisms [Yang *et al.*, 2016; Lu *et al.*, 2016; Anderson *et al.*, 2018] have been exploited to highlight visual objects that are relevant to the question. However, they treat objects independently and ignore their informative relationships. [Battaglia *et al.*, 2018] demonstrates that human's ability of combinatorial generalization highly depends on the mechanisms for reasoning over relationships. Consistent with such proposal, there is an emerging trend to represent the image by graph structure to depict objects and relationships in VQA and other vision-language tasks [Hu *et al.*, 2019b; Wang *et al.*, 2019a; Li *et al.*, 2019b]. As an extension, [Jiang *et al.*, 2020] exploits natural language to enrich the graph-based visual representations. However, it solely captures the semantics in natural language by LSTM, which lacking of fine-grained correlations with the visual information. To go one step further, we depict an image by multiple layers of graphs from visual, semantic and factual perspectives to collect fine-grained evidence from different modalities.

基于知识的视觉问答方法

Fact-based Visual Question Answering. Human can easily combine visual observation with external knowledge for answering questions, which remains challenging for algorithms. [Wang *et al.*, 2018] introduces a fact-based VQA task, which provides a knowledge base of facts and associates each question with a supporting-fact. Recent works based on FVQA generally select one entity from fact graph as the answer and falls into two categories: query-mapping based methods and learning based methods. [Wang *et al.*, 2017] reduces the question to one of the available query templates and this limits the types of questions that can be asked. [Wang *et al.*, 2018] automatically classifies and maps the question to a query which does not suffer the above constraint. Among both methods, however, visual information are used to extract facts but not introduced during the reasoning process. [Narasimhan *et al.*, 2018] applies GCN on the fact graph where each node is represented by the fixed form of image-question-entity embedding. However, the visual information is wholly provided which may introduce redundant information for prediction. In this paper, we decipt an image by multi-layer graphs and perform cross-modal heterogeneous graph reasoning on them to capture complementary evidence from different layers that most relevant to the question.

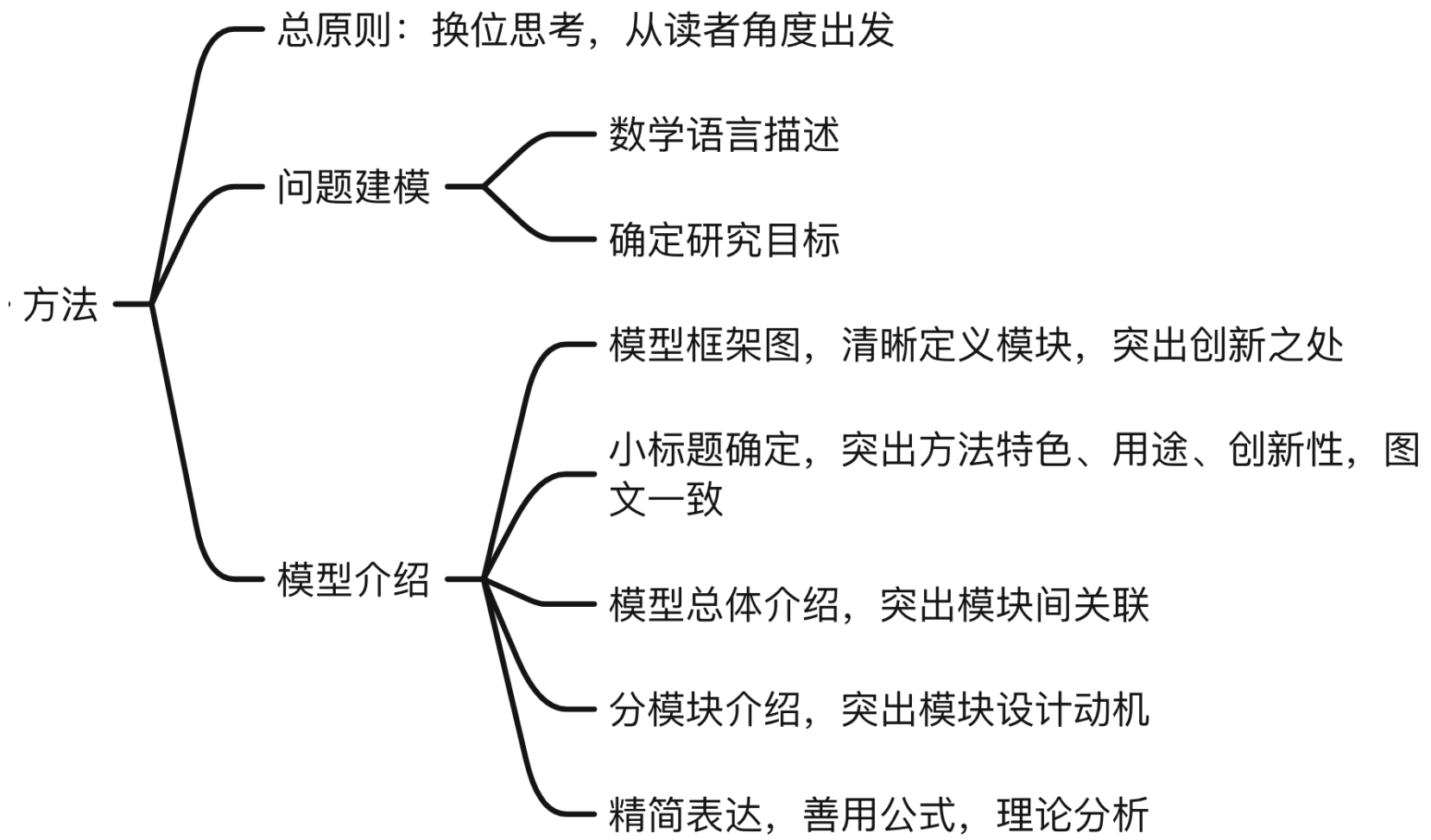
异构图神经网络方法

Heterogeneous Graph Neural Networks. Graph neu- …ntum in the last few …d with homogeneous graphs, heterogeneous graphs are more common in the real world. [Schlichtkrull *et al.*, 2018] generalizes graph convolutional network (GCN) to handle different relationships between entities in a knowledge base, where each edge with distinct relationships is encoded independently. [Wang *et al.*, 2019b; Hu *et al.*, 2019a] propose heterogeneous graph attention networks with dual-level attention mechanism. All of these methods model different types of nodes and edges on a unified graph. In contrast, the heterogeneous graph in this work contains multiple layers of subgraphs and each layer consists of nodes and edges coming from different modalities. For this specific constrain, we propose the intra-modal and cross-modal graph convolutions for reasoning over such multi-modal heterogeneous graphs.

方法写完后可动笔！

一篇论文的组成——方法

基本要求（最容易部分，可以先写）



一篇论文的组成——方法

CCF—A

- 问题-方法-实验，相互呼应
  - 动机：有理有据，足够具体
  - 方法：针对问题设计，每一步设计目标明确
    - 根据重点，重新组织方法介绍思路
    - 标题和图突出创新性和重点，相互呼应
    - 每一步方法设计都有理可依
  - 实验：针对方法逐一证明，针对动机逐一分析

CCF—C

- 问题-方法-实验，各为其说
  - 动机：大家都在研究，所以我研究
  - 方法：*step1->step2->step3*
  - 实验：达到了*SOTA*，缺乏分析

一篇论文的组成——方法

IJCAI 2020

Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering

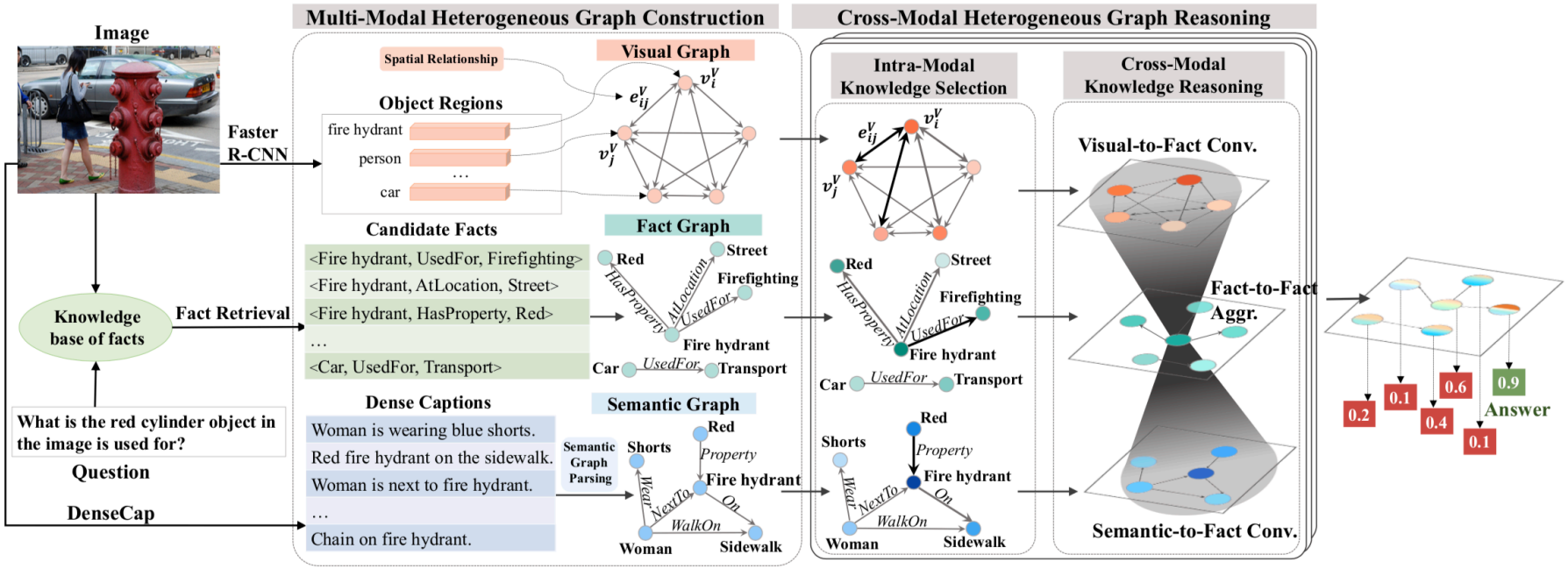
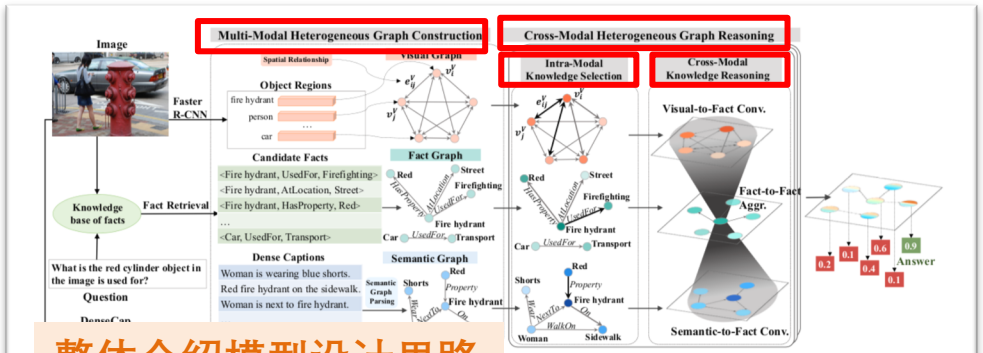


Figure 2: An overview of our model. The model contains two modules: Multi-modal Heterogeneous Graph Construction aims to depict an image by multiple layers of graphs and Cross-modal Hetegeneous Graph Reasoning supports intra-modal and cross-modal evidence selection.

一篇论文的组成——方法



整体介绍模型设计思路

: Multi-modal Heterogeneous Graph Construction aims to depict an Reasoning supports intra-modal and cross-modal evidence selection.

where e_1 is a visual concept in the image, e_2 is an attribute or phrase and r represents the relationship between e_1 and e_2. The key is to choose a correct entity, *i.e.* either e_1 or e_2, from the supporting fact as the predicted answer. We first introduce a novel scheme of depicting an image by three layers of graphs, including the visual graph, semantic graph and fact graph respectively, imitating the understanding of various properties of an object and the relationships. Then we perform cross-modal heterogeneous graph reasoning that consists of two parts: *Intra-Modal Knowledge Selection* aims to choose question-oriented knowledge from each layer of graphs by intra-modal graph convolutions, and *Cross-Modal Knowledge Reasoning* adaptively selects complementary evidence across three layers of graphs by cross-modal graph convolutions. By stacking the above two processes multiple times, our model performs iterative reasoning across all the modalities and results in the optimal answer by jointly analyzing all the entities. Figure 2 gives detailed illustration of our model.

Semantic Graph Construction. In addition to visual information, high-level abstraction of the objects and relationships by natural language provides essential semantic information. Such abstraction is indispensable to associate the visual objects in the image with the concepts mentioned in both questions and facts. In our work, we leverage dense captions [Johnson *et al.*, 2016] to extract a set of local-level semantics in an image, ranging from the properties of a single object (color, shape, emotion, *etc.*) to the relationships between objects (action, spatial positions, comparison, *etc.*). We decipt an image by D dense captions, denoted as $Z = \{z_i\}_{i=1}^{D}$, where z_i is a natural language description about a local region in the image. Instead of using monolithic embeddings to represent the captions, we exploit to model them by a graph-based semantic representation, denoted as $\mathcal{G}^S = (\mathcal{V}^S, \mathcal{E}^S)$, which is constructed by a semantic graph parsing model [Anderson *et al.*, 2016]. The node $v_i^S \in \mathcal{V}^S$ represents the name or attribute of an object extracted from the captions while the edge $e_{ij}^S \in \mathcal{E}^S$ represents the relationship between v_i^S and

3.1 Multi-Modal Graph Construction

Visual Graph Construction. Since most of the que in FVQA grounded in the visual objects and their re ships, we construct a fully-connected visual graph to sent such evidence at appearance level. Given an image use Faster-RCNN [Ren *et al.*, 2017] to identify a set of $\mathcal{O} = \{o_i\}_{i=1}^{K}$ (K = 36), where each object o_i is associated with a visual feature vector $\boldsymbol{v}_i \in \mathbb{R}^{d_v}$ (d_v = 2048), a spatial feature vector $\boldsymbol{b}_i \in \mathbb{R}^{d_b}$ (d_b = 4) and a corresponding label. Specifically, $\boldsymbol{b}_i = [x_i, y_i, w_i, h_i]$, where (x_i, y_i), h_i and w_i

- 标题突出创新点和过程
- 表达逻辑一致
- 图文一致

base of facts following a scored based approach proposed in [Narasimhan *et al.*, 2018]. We compute the cosine similarity of the embeddings of every word in the fact with the words in cepts detected in the assign a similarity ased on the similar- retained, denoted as f_{100}. A relation type classifier is trained additionally to further filter the retrieved facts. Specifically, we feed the last hidden state of LSTM to an MLP layer to predict the relation type $\hat{r}_i$ of a question. We retain the facts among f_{100} only if their relationships agree with $\hat{r}_i$, *i.e.* $f_{rel} = f \in$

每一个过程首先介绍背后的动机和目标

node set and each node v_i^V corresponds to a detected object o_i. The feature of node v_i^V is represented by $\boldsymbol{v}_i^V$. Each edge $e_{ij}^V \in \mathcal{E}^V$ denotes the relative spatial relationships between two objects. We encode the edge feature by a 5-dimensional vector, *i.e.* $\boldsymbol{r}_{ij}^V = [\frac{x_j - x_i}{w_i}, \frac{y_j - y_i}{h_i}, \frac{w_j}{w_i}, \frac{h_j}{h_i}, \frac{w_j h_j}{w_i h_i}]$.

$f_{100} : r(f) \in \{\hat{r}_i\}$ ($\{\hat{r}_i\}$ contains top-3 predicted relationships in experiments). Then a fact graph $\mathcal{G}^F = (\mathcal{V}^F, \mathcal{E}^F)$ is built upon f_{rel} as the candidate facts can be naturally organized as graphical structure. Each node $v_i^F \in \mathcal{V}^F$ denotes an entity in f_{rel} and is represented by GloVe embedding of the entity, denoted as $\boldsymbol{v}_i^F$. Each edge $e_{ij}^F \in \mathcal{E}^F$ denotes the relationship between v_i^F and v_j^F and is represented by GloVe embedding $\boldsymbol{r}_{ij}$. The topological structure among facts can be effectively exploited by jointly considering all the entities in the fact graph.

3.2 Intra-Modal Knowledge Selection

Since each layer of graphs contains modality-specific knowledge relevant to the question, we first select valuable evidence independently from the visual graph, semantic graph and fact graph by **Visual-to-Visual Convolution**, **Semantic-to-Semantic Convolution** and **Fact-to-Fact Convolution** respectively. These three convolutions share the common operations but differ in their node and edge representations corresponding to the graph layers. Thus we omit the superscript of node representation $\boldsymbol{v}$ and edge representation $\boldsymbol{r}$ in the rest of this section. We first perform attention operations to highlight the nodes and edges that are most relevant to the question q and consequently update node representations via intra-modal graph convolution. This process mainly consists of the following three steps:

Question-guided Node Attention. We first evaluate the relevance of each node corresponding to the question by attention mechanism. The attention weight for v_i is computed as:

$$\alpha_i = \text{softmax}(\boldsymbol{w}_a^T \tanh(\mathbf{W}_1 \boldsymbol{v}_i + \mathbf{W}_2 \boldsymbol{q})) \quad (1)$$

where $\mathbf{W}_1, \mathbf{W}_2$ and $\boldsymbol{w}_a$ (as well as $\mathbf{W}_3, ..., \mathbf{W}_{11}, \boldsymbol{w}_b, \boldsymbol{w}_c$ mentioned below) are learned parameters. $\boldsymbol{q}$ is question embedding encoded by LSTM.

Question-guided Edge Attention. Under the guidance of question, we then evaluate the importance of edge e_{ji} constrained by the neighbor node v_j regarding to v_i as:

$$\beta_{ji} = \text{softmax}(\boldsymbol{w}_b^T \tanh(\mathbf{W}_3 \boldsymbol{v}'_j + \mathbf{W}_4 \boldsymbol{q}')) \quad (2)$$

where $\boldsymbol{v}'_j = \mathbf{W}_5[\boldsymbol{v}_j, \boldsymbol{r}_{ji}]$, $\boldsymbol{q}' = \mathbf{W}_6[\boldsymbol{v}_i, \boldsymbol{q}]$ and $[\cdot,\cdot]$ denotes concatenation operation.

Intra-Modal Graph Convolution. Given the node and edge attention weights learned in Eq. 1 and Eq. 2, the node representations of each layer of graphs are updated following]. We repre-

具体过程分小标题

$$\boldsymbol{m}_i = \sum_{j \in \mathcal{N}_i} \beta_{ji} \boldsymbol{v}'_j \quad (3)$$

$$\hat{\boldsymbol{v}}_i = \text{ReLU}(\mathbf{W}_7[\boldsymbol{m}_i, \alpha_i \boldsymbol{v}_i]) \quad (4)$$

where $\mathcal{N}_i$ is the neighborhood set of node v_i.

We conduct the above intra-modal knowledge selection on $\mathcal{G}^V$, $\mathcal{G}^S$ and $\mathcal{G}^F$ independently and obtain the updated node representations, denoted as $\{\hat{\boldsymbol{v}}_i^V\}_{i=1}^{\mathcal{N}^V}$, $\{\hat{\boldsymbol{v}}_i^S\}_{i=1}^{\mathcal{N}^S}$ and $\{\hat{\boldsymbol{v}}_i^F\}_{i=1}^{\mathcal{N}^F}$ accordingly.

3.3 Cross-Modal Knowledge Reasoning

To answer the question correctly, we fully consider the complementary evidence from visual, semantic and factual information. Since the answer comes from one entity in the fact graph, we gather complementary information from visual graph and semantic graph to fact graph by cross-modal convolutions, including *visual-to-fact convolution* and *semantic-to-fact convolution*. Finally, a *fact-to-fact aggregation* is performed on the fact graph to reason over all the entities and form a global decision.

Visual-to-Fact Convolution. For the entity v_i^F in fact graph, the attention value of each node v_j^V in the visual graph w.r.t. v_i^F is calculated under the guidance of question:

$$\gamma_{ji}^{V\text{-}F} = \text{softmax}(\boldsymbol{w}_c \tanh(\mathbf{W}_8 \hat{\boldsymbol{v}}_j^V + \mathbf{W}_9[\hat{\boldsymbol{v}}_i^F, \boldsymbol{q}])) \quad (5)$$

The complementary information $\boldsymbol{m}_i^{V\text{-}F}$ from visual graph for v_i^F is computed as:

$$\boldsymbol{m}_i^{V\text{-}F} = \sum_{j \in \mathcal{N}^V} \gamma_{ji}^{V\text{-}F} \hat{\boldsymbol{v}}_j^V \quad (6)$$

Semantic-to-Fact Convolution. The complementary information $\boldsymbol{m}_i^{S\text{-}F}$ from the semantic graph is computed in the same way as in Eq. 5 and Eq. 6.

Then we fuse the complementary knowledge for v_i^F from three layers of graphs via a gate operation:

$$\boldsymbol{gate}_i = \sigma(\mathbf{W}_{10}[\boldsymbol{m}_i^{V\text{-}F}, \boldsymbol{m}_i^{S\text{-}F}, \hat{\boldsymbol{v}}_i^F]) \quad (7)$$

$$\tilde{\boldsymbol{v}}_i^F = \mathbf{W}_{11}(\boldsymbol{gate}_i \circ [\boldsymbol{m}_i^{V\text{-}F}, \boldsymbol{m}_i^{S\text{-}F}, \hat{\boldsymbol{v}}_i^F]) \quad (8)$$

where σ is sigmoid function and "$\circ$" denotes element-wise product.

Fact-to-Fact Aggregation. Given a set of candidate entities in the fact graph $\mathcal{G}^F$, we aim to globally compare all the entities and select an optimal one as the answer. Now the representation of each entity in the fact graph gathers question-oriented information from three modalities. To jointly evaluate the possibility of each entity, we perform the attention-based graph convolutional network similar to Fact-to-Fact Convolution introduced in Section 3.2 to aggregate information in the fact graph and obtain the transformed entity representations.

We iteratively perform intra-modal knowledge selection and cross-modal knowledge reasoning in multiple steps to obtain the final entity representations. After T steps, each entity representation $\tilde{\boldsymbol{v}}_i^{F(T)}$ captures the structural information within T-hop neighborhood across three layers.

3.4 Learning

The concatenation of entity representation $\tilde{\boldsymbol{v}}_i^{F(T)}$ and question embedding $\boldsymbol{q}$ is passed to a binary classifier to predict its probability as the answer , *i.e.* $\hat{y}_i = p_\theta([\tilde{\boldsymbol{v}}_i^{F(T)}, \boldsymbol{q}])$. We apply the binary cross-entropy loss in the training process:

$$l_n = -\sum_{i \in \mathcal{N}^F} [a \cdot y_i \ln \hat{y}_i + b \cdot (1 - y_i) \ln(1 - \hat{y}_i)] \quad (9)$$

where y_i is the ground truth label for v_i^F and a, b represent loss function weights for positive and negative samples respectively. The entity with the largest probability is selected as the final answer.

注意：
动机？
其他方式？

欢迎大家在B站、知乎专栏、邮件留言交流！

于静

邮箱：yujing02@iie.ac.cn

课程主页：https://mmlab-iie.github.io/course/

研究组主页：https://mmlab-iie.github.io/

知乎专栏：https://www.zhihu.com/column/c_1284803871596797952

课程主页

研究组主页

知乎专栏

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学
University of Chinese Academy of Sciences