科研与英文学术论文写作

第七讲 英文学术论文之写作思路 ——模型框架图绘制概述

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

从论文图示与论文写作的关系说起

论文引言

科学问题 是否探索本质？为何重要？

写作逻辑 背景-现状-问题-方法如何环环相扣？

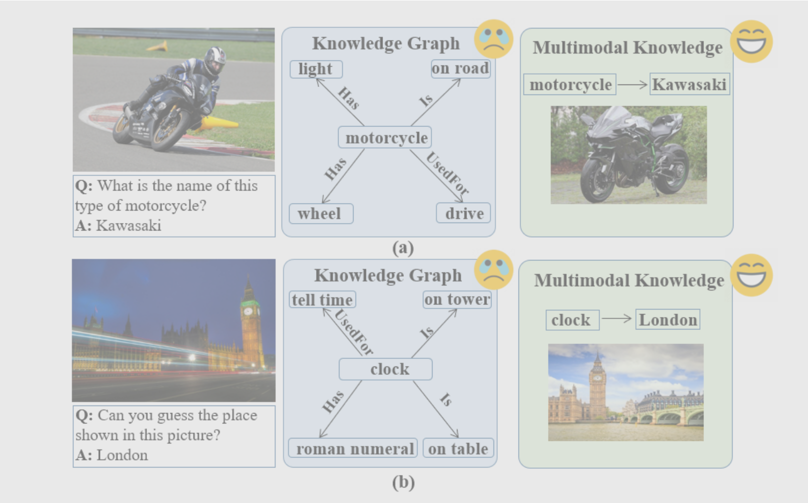
表达方式 如何抓住重点？如何深入浅出？

论文模型

解决方法 是否专注核心挑战？如何解决？

写作逻辑 方法如何分解？各自解决哪些问题？

表达方式 如何准确清晰？如何减少返工？



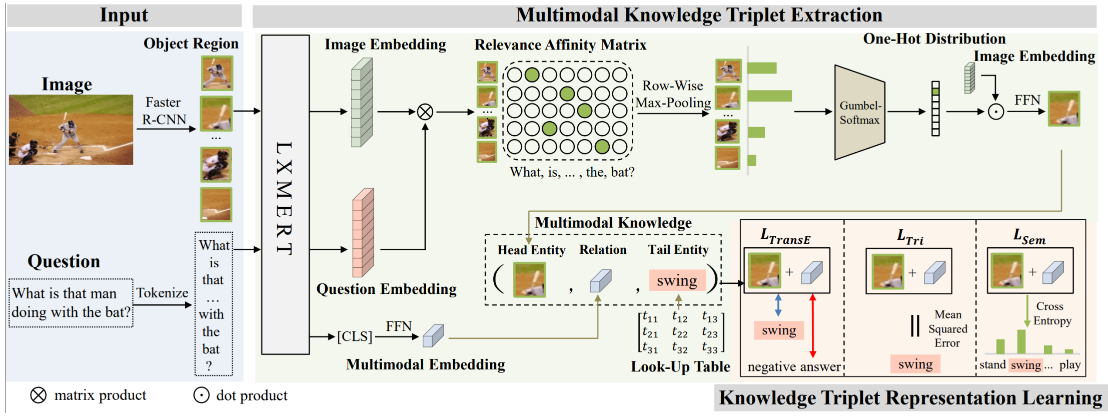
Q: What is the name of this type of motorcycle?
A: Kawasaki

(a)

Q: Can you guess the place shown in this picture?
A: London

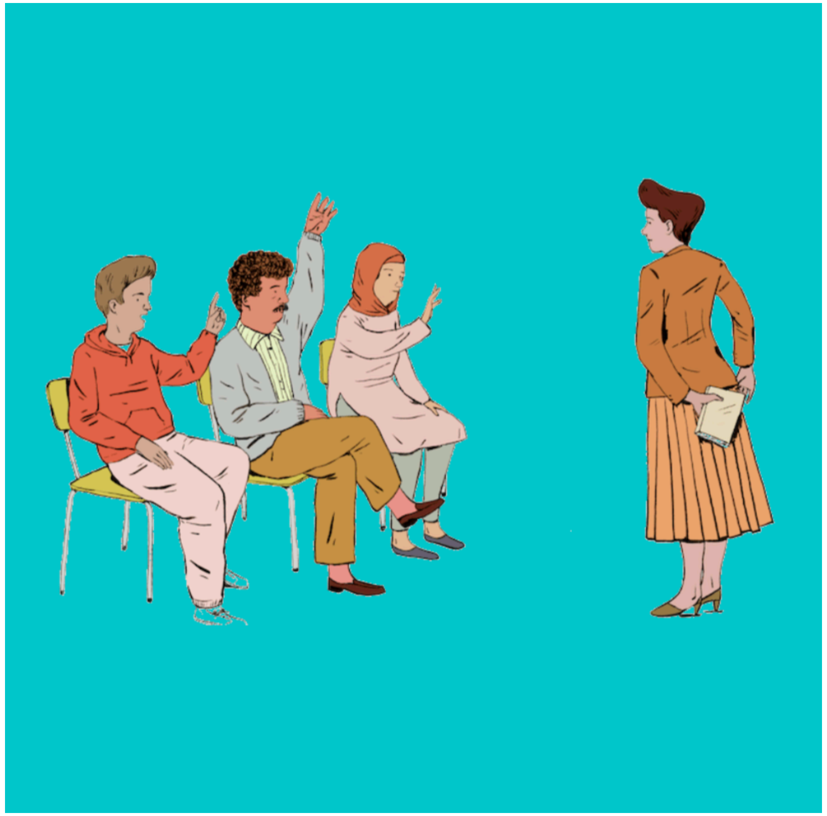
(b)

研究动机图



模型框架图

让我们一起来复盘



示例一：计算机视觉 | 新思路 | 小模型

MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-based Visual Question Answering

CVPR 2022

Paper：https://arxiv.org/abs/2203.09138

Code：https://github.com/AndersonStra/MuKEA

Yang Ding , **Jing Yu*** , Bang Liu , Yue Hu, Mingxin Cui , Qi Wu

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

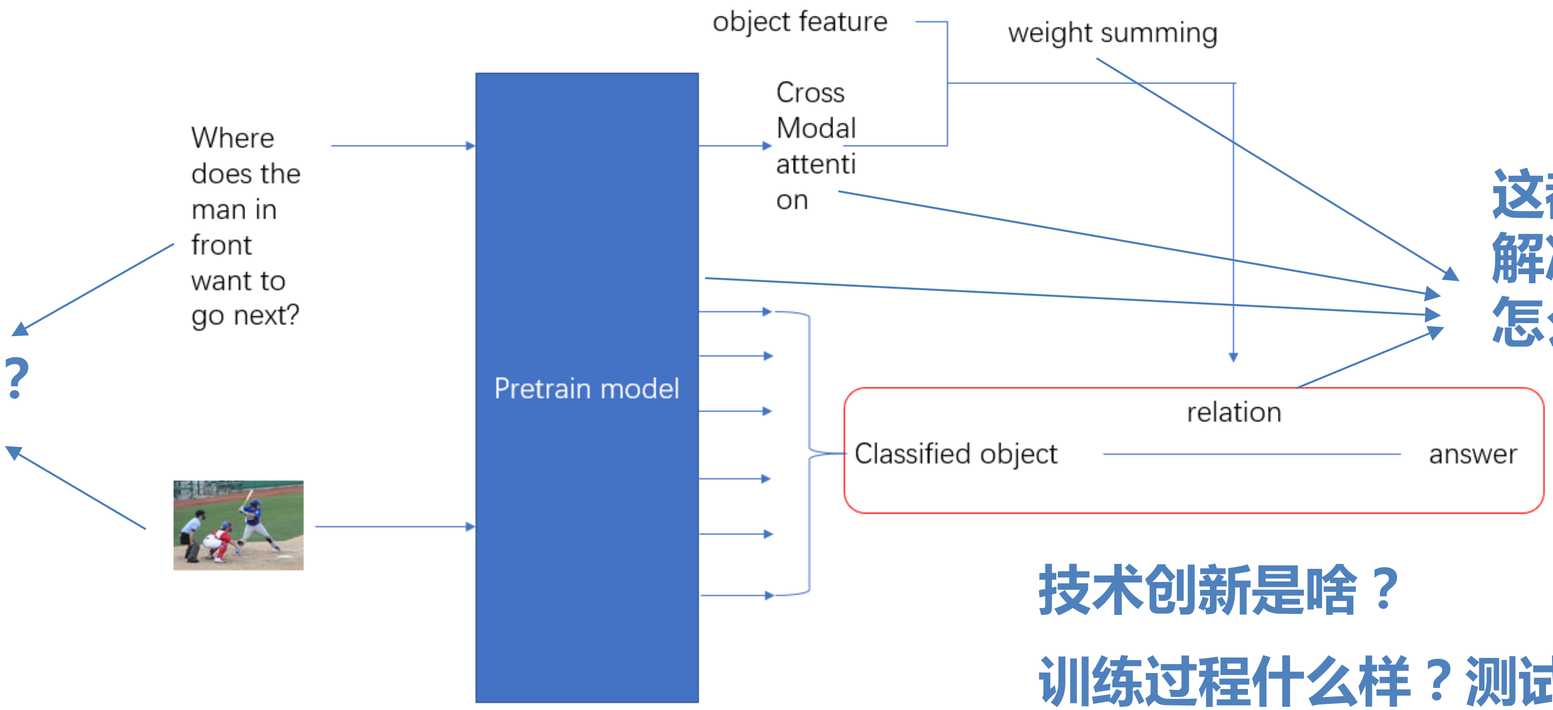
Université de Montréal

Mila

THE UNIVERSITY of ADELAIDE

模型框架图 V1

最初版本：直观的记录模型输入、输出



为什么这么小？

这都是啥？
解决了啥问题？
怎么解决的？

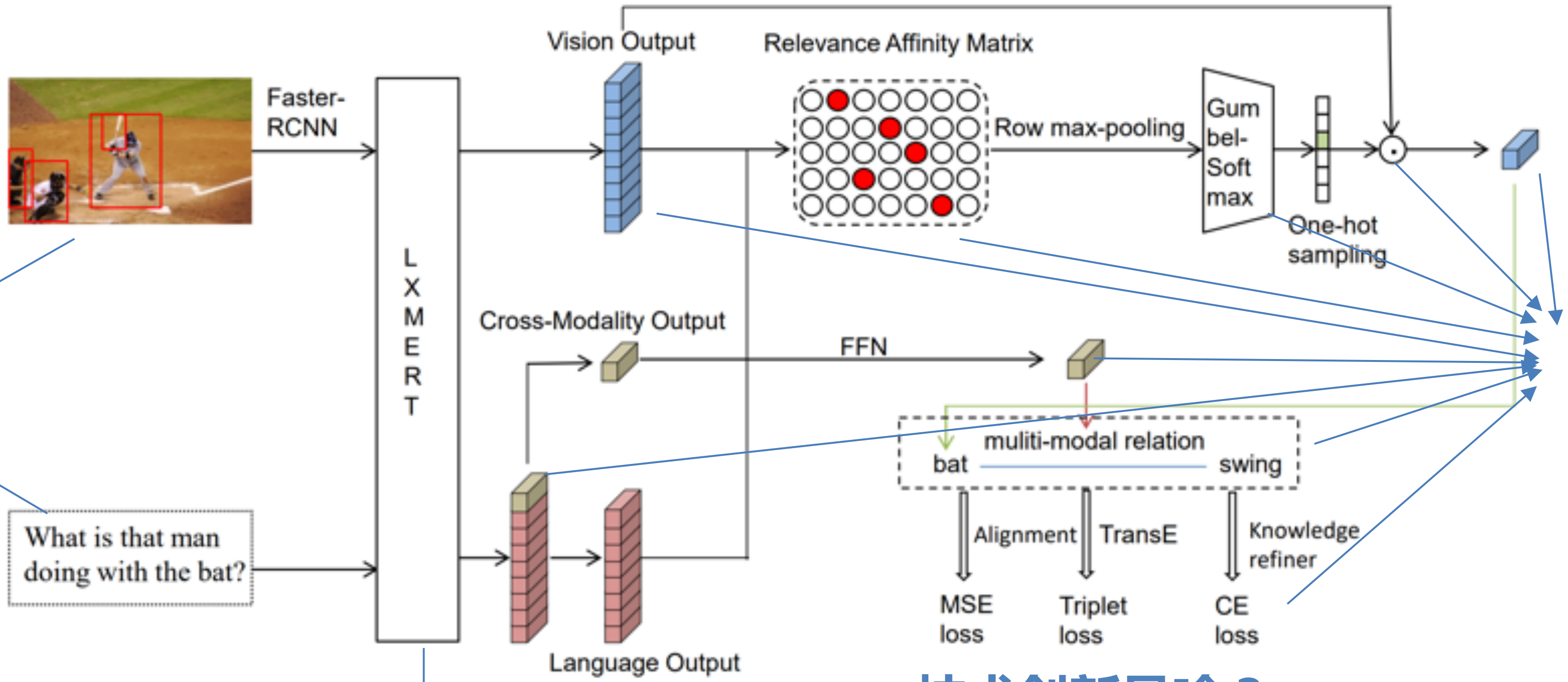
技术创新是啥？

训练过程什么样？测试过程什么样？

主要问题：过程不明确，模块边界不清晰，如何解决没体现！

模型框架图 V2

初步刻画了模型的完整实现过程



这是啥？

这都是啥？？
解决了啥问题？
怎么解决的？

技术创新是啥？

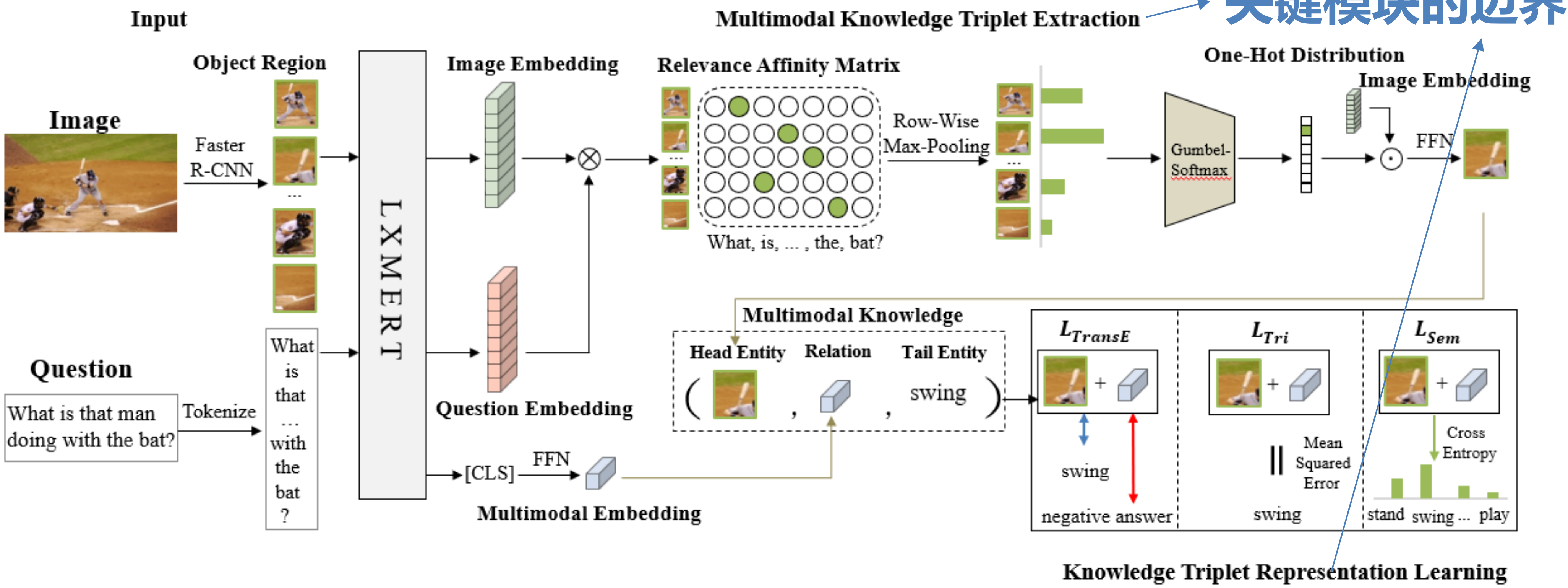
横版？竖版？

训练过程什么样？测试过程什么样？

主要问题：模块边界不清晰，如何解决没体现！

模型框架图 V3

明确定义所有过程目标、变量名称，细化每个阶段实现，避免线条交叉



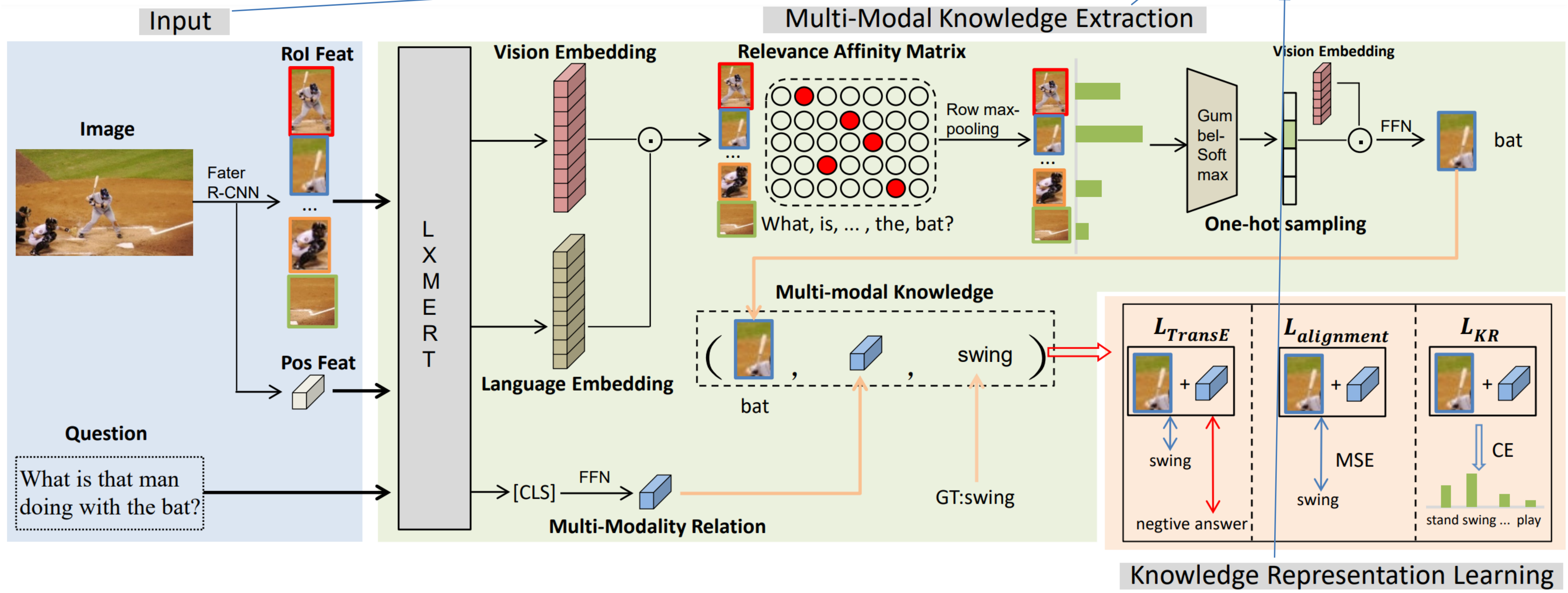
关键模块的边界在哪里？

主要问题：模块边界不清晰，算法逻辑不明确！

模型框架图 V4

每个模块用不同颜色区分，模块标题突出

排版不整齐！标题不明显！



主要问题：不直观！布局问题！

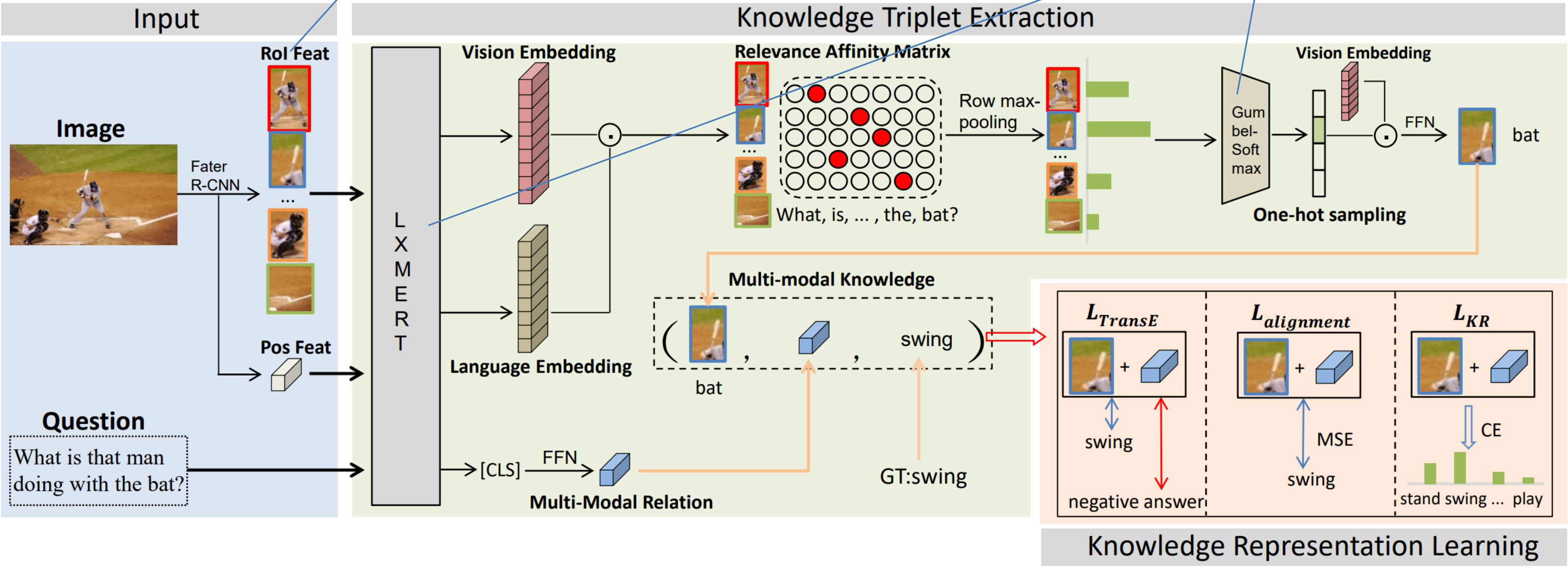
于静 中科院信息工程研究所 CogModal GROUP

模型框架图 V5

标题整体布局合理、一致

字体不一致

排版不规范



主要问题：看着乱！字体、排版问题！

于静 中科院信息工程研究所

CogModal GROUP

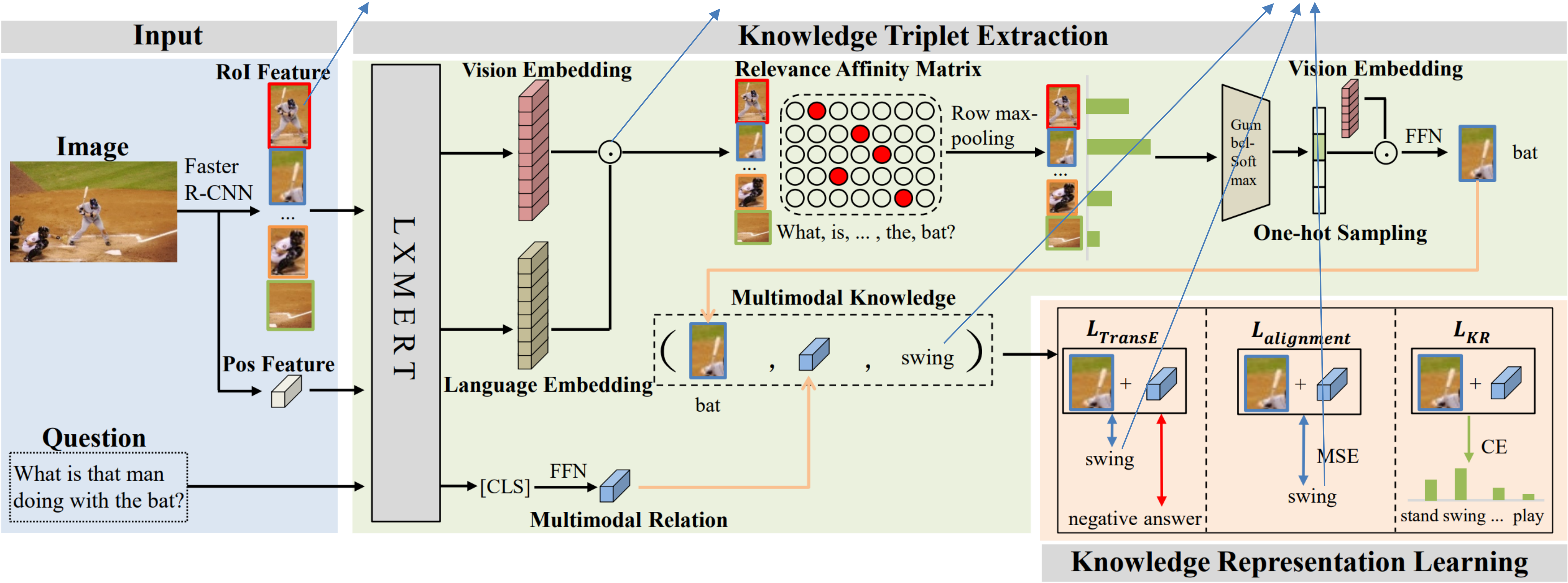
模型框架图 V6

字体、字号一致，排版符合常识

图片大小不统一

重要符号没有注释

同样的符号背景不一致



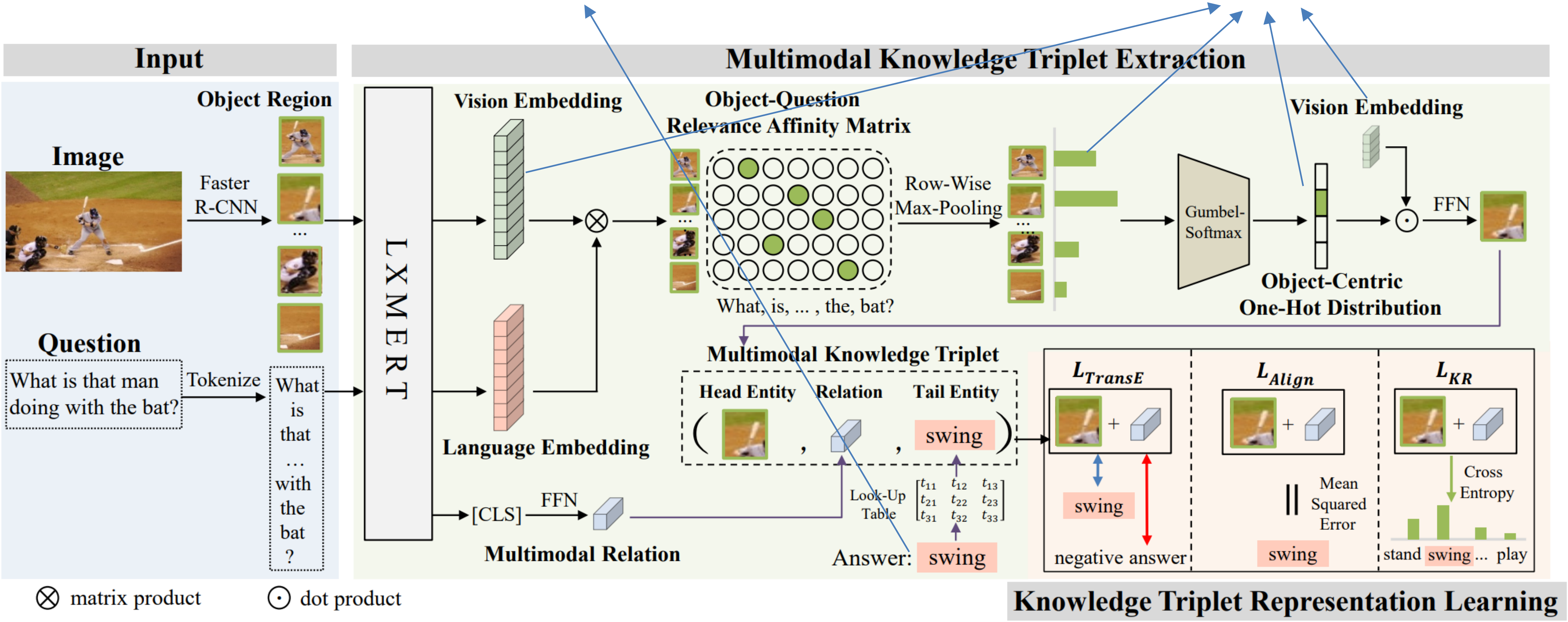
主要问题：颜色乱！文字多！线条多！不直观！

模型框架图 V7

相同内容表现形式一致，符号有注释

元素冗余

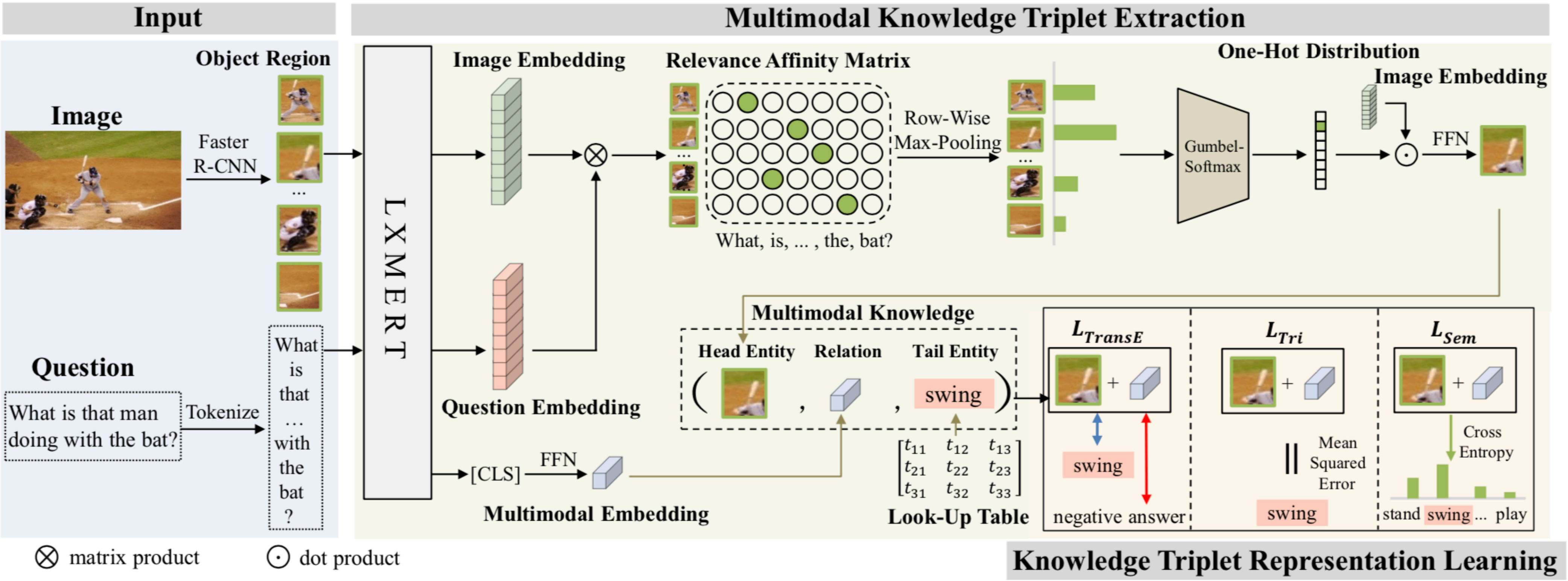
向量不统一



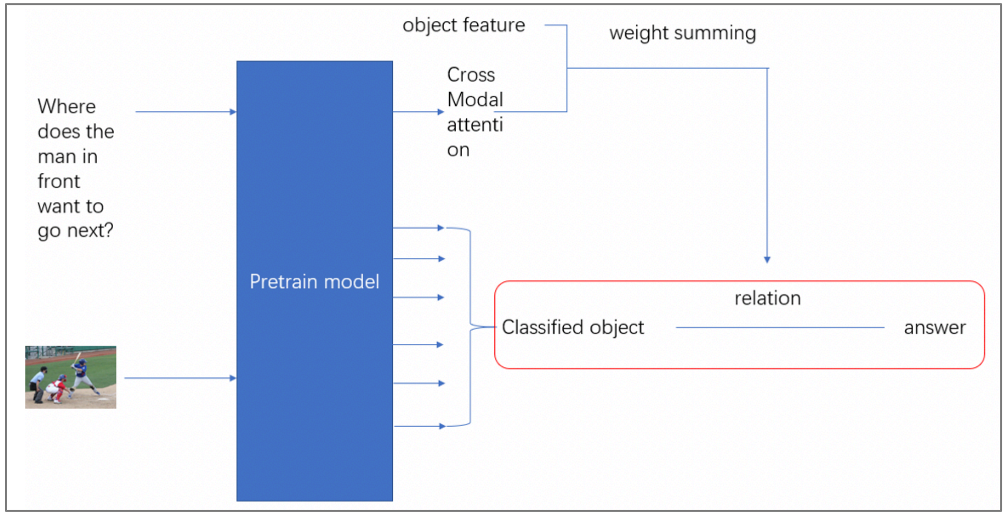
主要问题：相同内容前后不对应！信息冗余！

模型框架图 V8

最终版本！

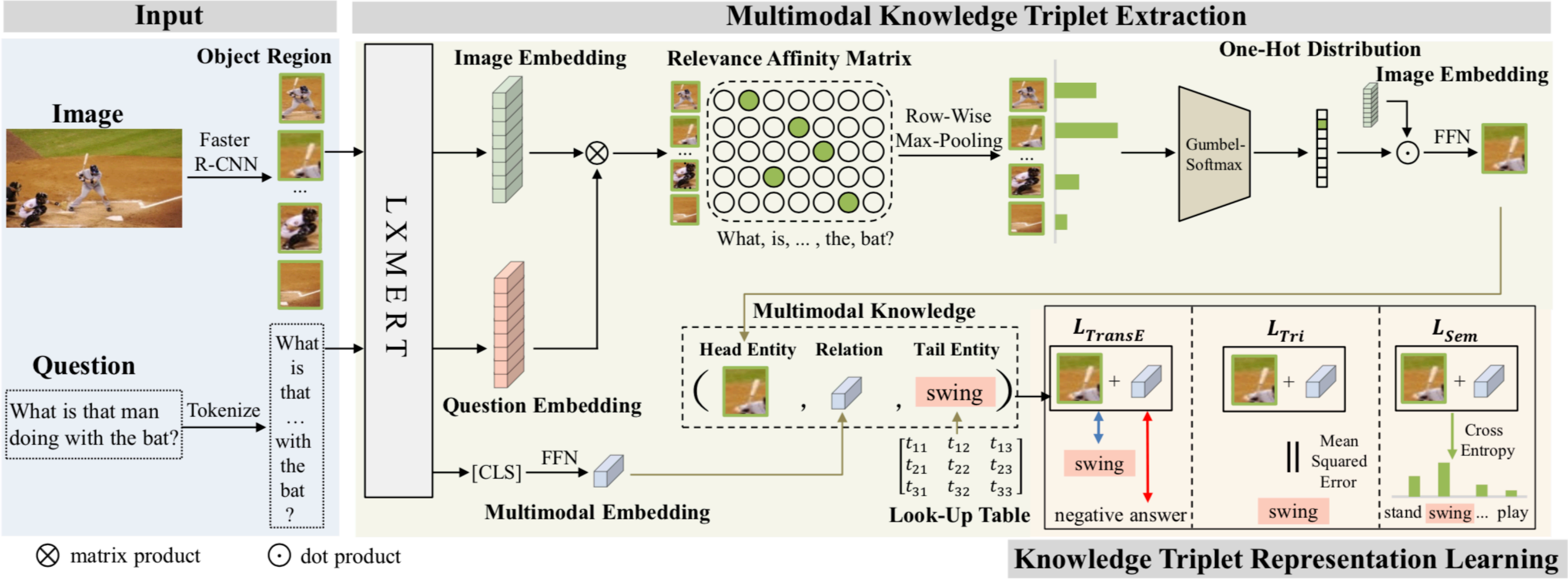


模型框架图 从0到1



10月
最初版本！

20+ 版本

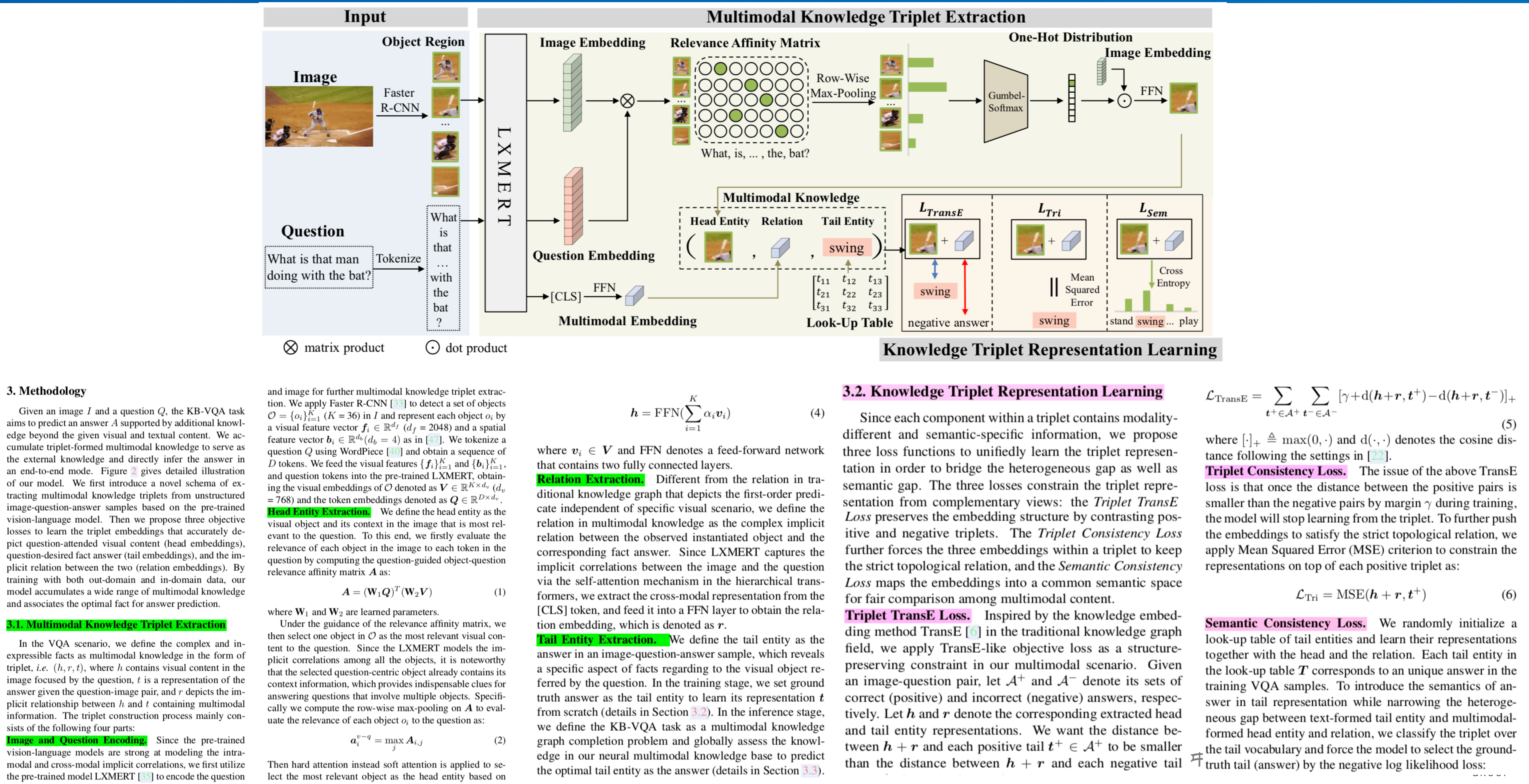


11月
最终版本！

于静 中科院信息工程研究所

CogModal GROUP

模型框架图 从0到1



3. Methodology

Given an image I and a question Q, the KB-VQA task aims to predict an answer A supported by additional knowledge beyond the given visual and textual content. We accumulate triplet-formed multimodal knowledge to serve as the external knowledge and directly infer the answer in an end-to-end mode. Figure 2 gives detailed illustration of our model. We first introduce a novel schema of extracting multimodal knowledge triplets from unstructured image-question-answer samples based on the pre-trained vision-language model. Then we propose three objective losses to learn the triplet embeddings that accurately depict question-attended visual content (head embeddings), question-desired fact answer (tail embeddings), and the implicit relation between the two (relation embeddings). By training with both out-domain and in-domain data, our model accumulates a wide range of multimodal knowledge and associates the optimal fact for answer prediction.

3.1. Multimodal Knowledge Triplet Extraction

In the VQA scenario, we define the complex and inexpressible facts as multimodal knowledge in the form of triplet, *i.e.* (h, r, t), where h contains visual content in the image focused by the question, t is a representation of the answer given the question-image pair, and r depicts the implicit relationship between h and t containing multimodal information. The triplet construction process mainly consists of the following four parts:

Image and Question Encoding. Since the pre-trained vision-language models are strong at modeling the intra-modal and cross-modal implicit correlations, we first utilize the pre-trained model LXMERT [35] to encode the question and image for further multimodal knowledge triplet extraction. We apply Faster R-CNN [33] to detect a set of objects $\mathcal{O} = \{o_i\}_{i=1}^{K}$ (K = 36) in I and represent each object o_i by a visual feature vector $\boldsymbol{f}_i \in \mathbb{R}^{d_f}$ (d_f = 2048) and a spatial feature vector $\boldsymbol{b}_i \in \mathbb{R}^{d_b}(d_b = 4)$ as in [47]. We tokenize a question Q using WordPiece [40] and obtain a sequence of D tokens. We feed the visual features $\{\boldsymbol{f}_i\}_{i=1}^{K}$ and $\{\boldsymbol{b}_i\}_{i=1}^{K}$, and question tokens into the pre-trained LXMERT, obtaining the visual embeddings of $\mathcal{O}$ denoted as $\boldsymbol{V} \in \mathbb{R}^{K \times d_v}$ (d_v = 768) and the token embeddings denoted as $\boldsymbol{Q} \in \mathbb{R}^{D \times d_v}$.

Head Entity Extraction. We define the head entity as the visual object and its context in the image that is most relevant to the question. To this end, we firstly evaluate the relevance of each object in the image to each token in the question by computing the question-guided object-question relevance affinity matrix $\boldsymbol{A}$ as:

$$\boldsymbol{A} = (\mathbf{W}_1 \boldsymbol{Q})^T (\mathbf{W}_2 \boldsymbol{V}) \tag{1}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are learned parameters.

Under the guidance of the relevance affinity matrix, we then select one object in $\mathcal{O}$ as the most relevant visual content to the question. Since the LXMERT models the implicit correlations among all the objects, it is noteworthy that the selected question-centric object already contains its context information, which provides indispensable clues for answering questions that involve multiple objects. Specifically we compute the row-wise max-pooling on $\boldsymbol{A}$ to evaluate the relevance of each object o_i to the question as:

$$\boldsymbol{a}_i^{v-q} = \max_j \boldsymbol{A}_{i,j} \tag{2}$$

Then hard attention instead soft attention is applied to select the most relevant object as the head entity based on

$$\boldsymbol{h} = \mathrm{FFN}(\sum_{i=1}^{K} \alpha_i \boldsymbol{v}_i) \tag{4}$$

where $\boldsymbol{v}_i \in \boldsymbol{V}$ and FFN denotes a feed-forward network that contains two fully connected layers.

Relation Extraction. Different from the relation in traditional knowledge graph that depicts the first-order predicate independent of specific visual scenario, we define the relation in multimodal knowledge as the complex implicit relation between the observed instantiated object and the corresponding fact answer. Since LXMERT captures the implicit correlations between the image and the question via the self-attention mechanism in the hierarchical transformers, we extract the cross-modal representation from the [CLS] token, and feed it into a FFN layer to obtain the relation embedding, which is denoted as $\boldsymbol{r}$.

Tail Entity Extraction. We define the tail entity as the answer in an image-question-answer sample, which reveals a specific aspect of facts regarding to the visual object referred by the question. In the training stage, we set ground truth answer as the tail entity to learn its representation $\boldsymbol{t}$ from scratch (details in Section 3.2). In the inference stage, we define the KB-VQA task as a multimodal knowledge graph completion problem and globally assess the knowledge in our neural multimodal knowledge base to predict the optimal tail entity as the answer (details in Section 3.3).

3.2. Knowledge Triplet Representation Learning

Since each component within a triplet contains modality-different and semantic-specific information, we propose three loss functions to unifiedly learn the triplet representation in order to bridge the heterogeneous gap as well as semantic gap. The three losses constrain the triplet representation from complementary views: the *Triplet TransE Loss* preserves the embedding structure by contrasting positive and negative triplets. The *Triplet Consistency Loss* further forces the three embeddings within a triplet to keep the strict topological relation, and the *Semantic Consistency Loss* maps the embeddings into a common semantic space for fair comparison among multimodal content.

Triplet TransE Loss. Inspired by the knowledge embedding method TransE [6] in the traditional knowledge graph field, we apply TransE-like objective loss as a structure-preserving constraint in our multimodal scenario. Given an image-question pair, let $\mathcal{A}^+$ and $\mathcal{A}^-$ denote its sets of correct (positive) and incorrect (negative) answers, respectively. Let $\boldsymbol{h}$ and $\boldsymbol{r}$ denote the corresponding extracted head and tail entity representations. We want the distance between $\boldsymbol{h} + \boldsymbol{r}$ and each positive tail $\boldsymbol{t}^+ \in \mathcal{A}^+$ to be smaller than the distance between $\boldsymbol{h} + \boldsymbol{r}$ and each negative tail

$$\mathcal{L}_{\mathrm{TransE}} = \sum_{\boldsymbol{t}^+ \in \mathcal{A}^+} \sum_{\boldsymbol{t}^- \in \mathcal{A}^-} [\gamma + \mathrm{d}(\boldsymbol{h}+\boldsymbol{r}, \boldsymbol{t}^+) - \mathrm{d}(\boldsymbol{h}+\boldsymbol{r}, \boldsymbol{t}^-)]_+ \tag{5}$$

where $[\cdot]_+ \triangleq \max(0, \cdot)$ and $\mathrm{d}(\cdot, \cdot)$ denotes the cosine distance following the settings in [22].

Triplet Consistency Loss. The issue of the above TransE loss is that once the distance between the positive pairs is smaller than the negative pairs by margin γ during training, the model will stop learning from the triplet. To further push the embeddings to satisfy the strict topological relation, we apply Mean Squared Error (MSE) criterion to constrain the representations on top of each positive triplet as:

$$\mathcal{L}_{\mathrm{Tri}} = \mathrm{MSE}(\boldsymbol{h} + \boldsymbol{r}, \boldsymbol{t}^+) \tag{6}$$

Semantic Consistency Loss. We randomly initialize a look-up table of tail entities and learn their representations together with the head and the relation. Each tail entity in the look-up table $\boldsymbol{T}$ corresponds to an unique answer in the training VQA samples. To introduce the semantics of answer in tail representation while narrowing the heterogeneous gap between text-formed tail entity and multimodal-formed head entity and relation, we classify the triplet over the tail vocabulary and force the model to select the ground-truth tail (answer) by the negative log likelihood loss:

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