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

CVPR 2022

Team: CogModal Group

Authors: Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu

Speaker: Ting-Chih Chen

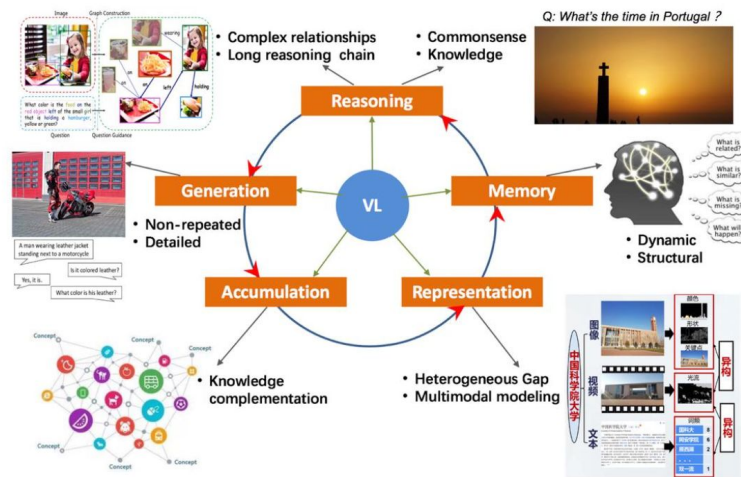


Content

- Motivation
- Recent works
- Model
- Experiments
- Summary & Future Work

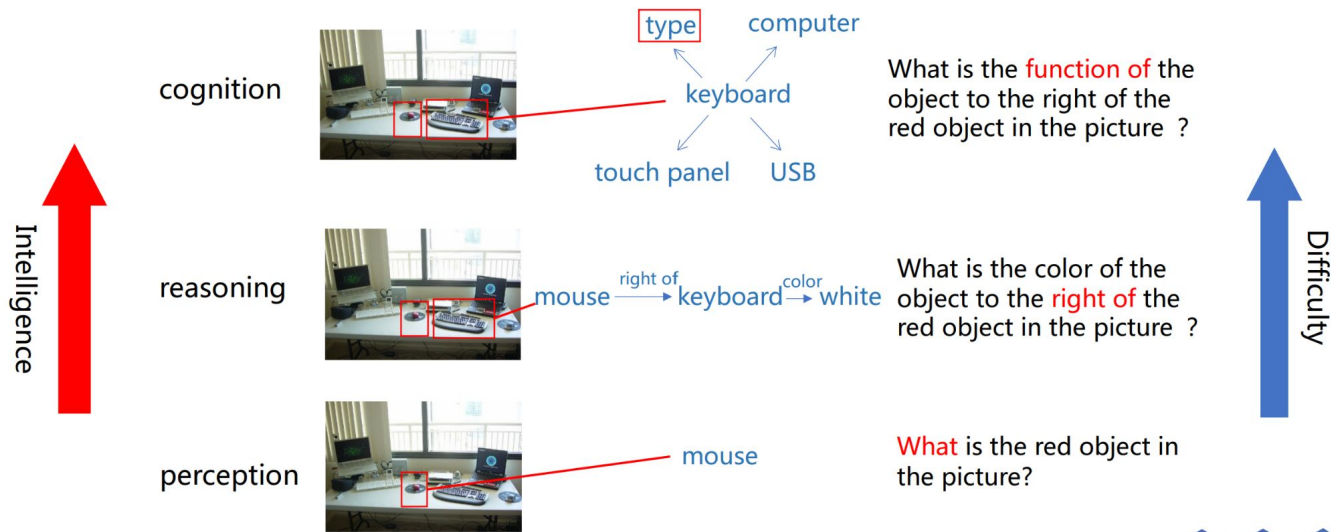
Motivation

- Knowledge-based VQA requires the ability of association **external knowledge**
- Limitation: existing solutions is that capture relevant knowledge from **text-only** knowledge bases
They lack multimodal knowledge for visual understanding



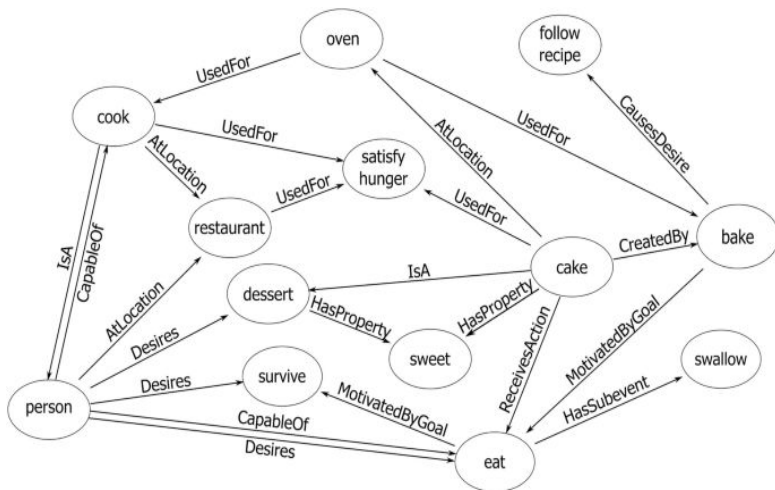
VQA evolution

- VQA evolves from **perception** to **reasoning** and then to **cognition**, requiring a gradually increase of intelligence



Recent works

- Structured KG: ConceptNet and DBpedia



About: [Michael Jordan](#)

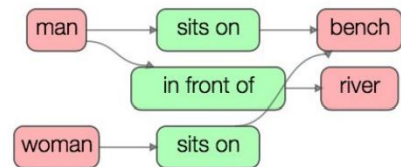
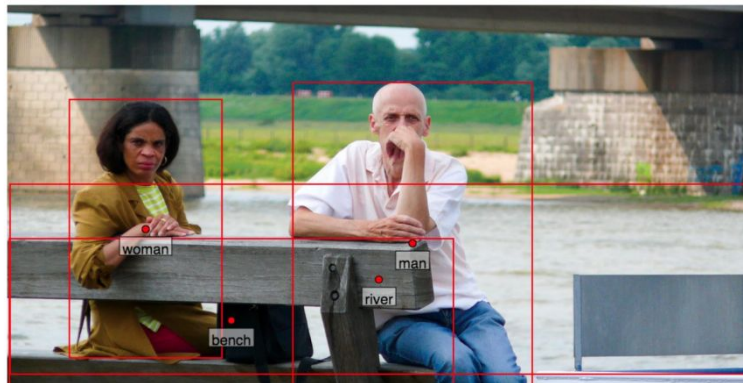
An Entity of Type: [species](#) from Named Graph: <http://dbpedia.org>, within Data Space: [dbpedia.org](#)

Michael Jeffrey Jordan OLY (born February 17, 1963), also known by his initials MJ, is an American businessman and former professional basketball player. His biography on the official NBA website states: "By acclamation, Michael Jordan is the greatest basketball player of all time." He played fifteen seasons in the National Basketball Association (NBA), winning six NBA championships with the Chicago Bulls. Jordan is the principal owner and chairman of the Charlotte Hornets of the NBA and of 23XI Racing in the NASCAR Cup Series. He was integral in popularizing the NBA around the world in the 1980s and 1990s, becoming a global cultural icon in the process.



Recent works

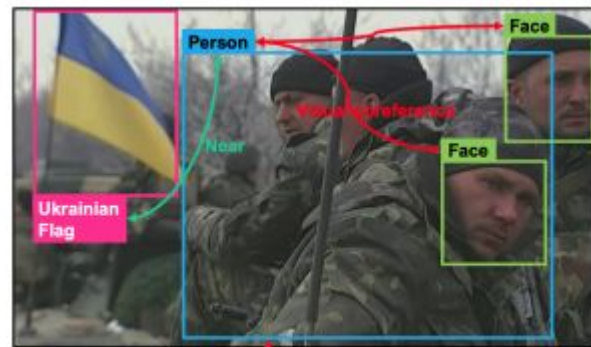
- Unstructured/ semi-structured KG: Wikipedia and Visual Genome
- Disadvantages:
 - Required knowledge by human annotations
 - KGs lack visual information to assist cross-modal understanding
 - The information is limited to the definite facts
 - KGs are difficult to represent high-order prediction



A man and a woman sit on a park bench along a river.

Recent works

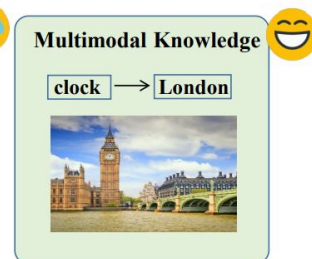
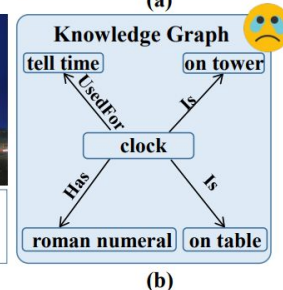
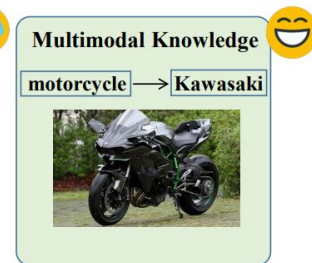
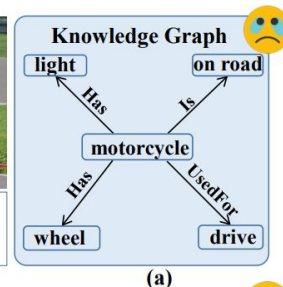
- Multimodal KG aims to correlate visual content with textual facts to form the augmented knowledge graph
- Disadvantages:
 - This kind of Multimodal KG still fails to model the high-order complex relationships



“... They put **troops** on the boarder, what for? ...”

Our goal

- How to **represent** the multimodal knowledge?
- How to **accumulate** the multimodal knowledge in the VQA scenarios?
- How to maintain the advantages of traditional knowledge graph in **explainable reasoning**?



Solution

- MuKEA represents multimodal knowledge unit by **explicit triplet** and **implicit relation**
- Explicit triplet: the visual objects referred by the question are embedded in the **head entity** and the embedding of the fact answer is kept in the **tail entity**
- Implicit relation: the **relationship** between head entity and tail entity

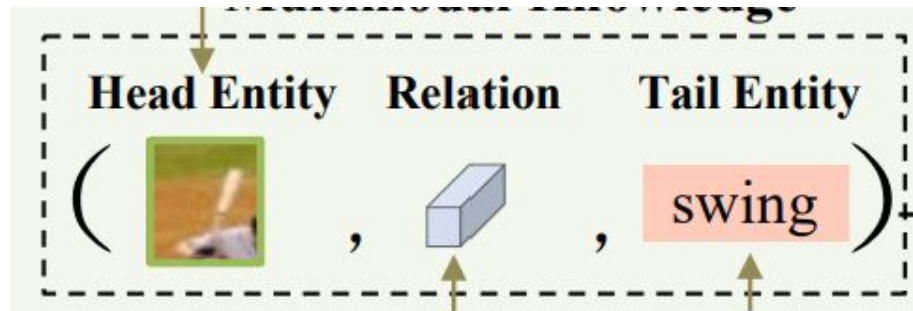
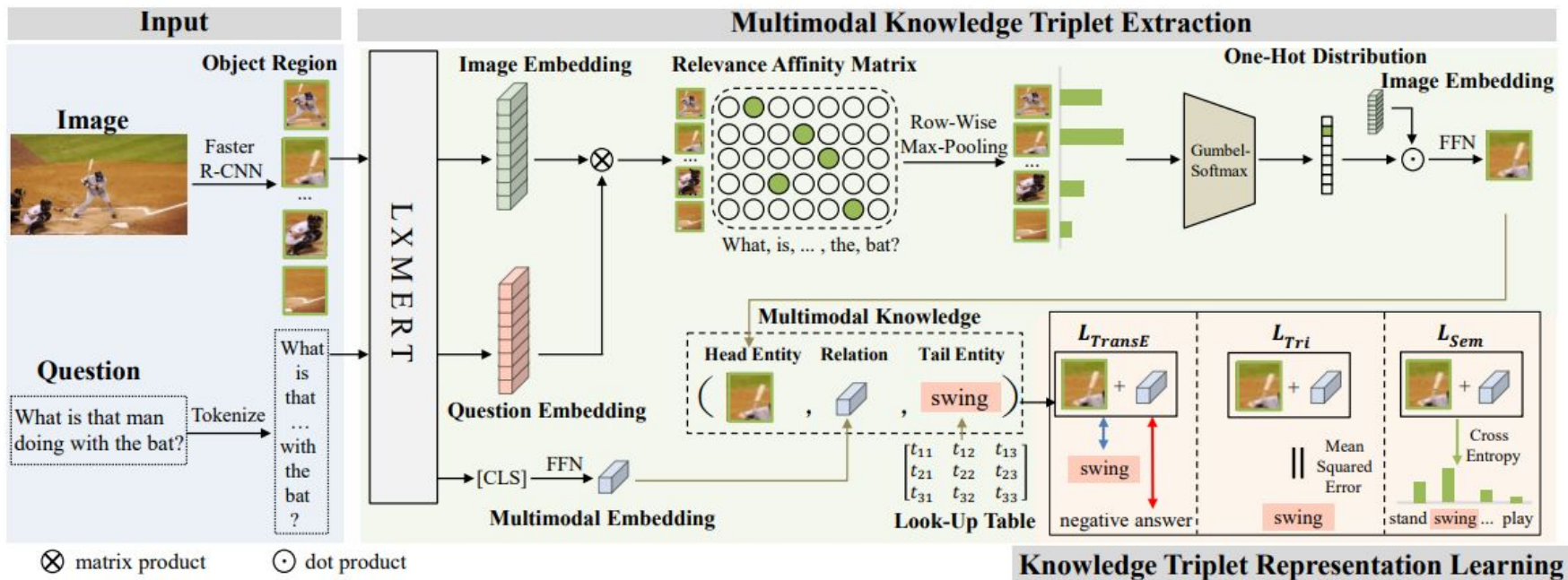


image resource: <https://arxiv.org/pdf/2203.09138.pdf>

MuKEA



MuKEA - multimodal knowledge triplet extraction

- Triplet format: (h,r,t)
 - h is visual content in the image focused by the question
 - t is a representation of the answer given the question-image pair
 - r is the implicit relationship between h and t

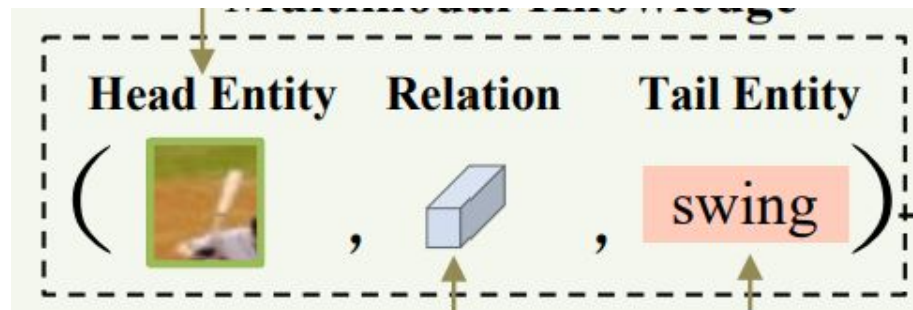
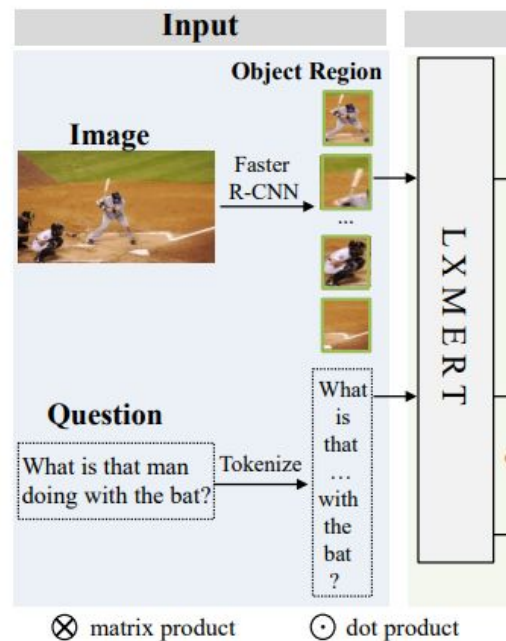


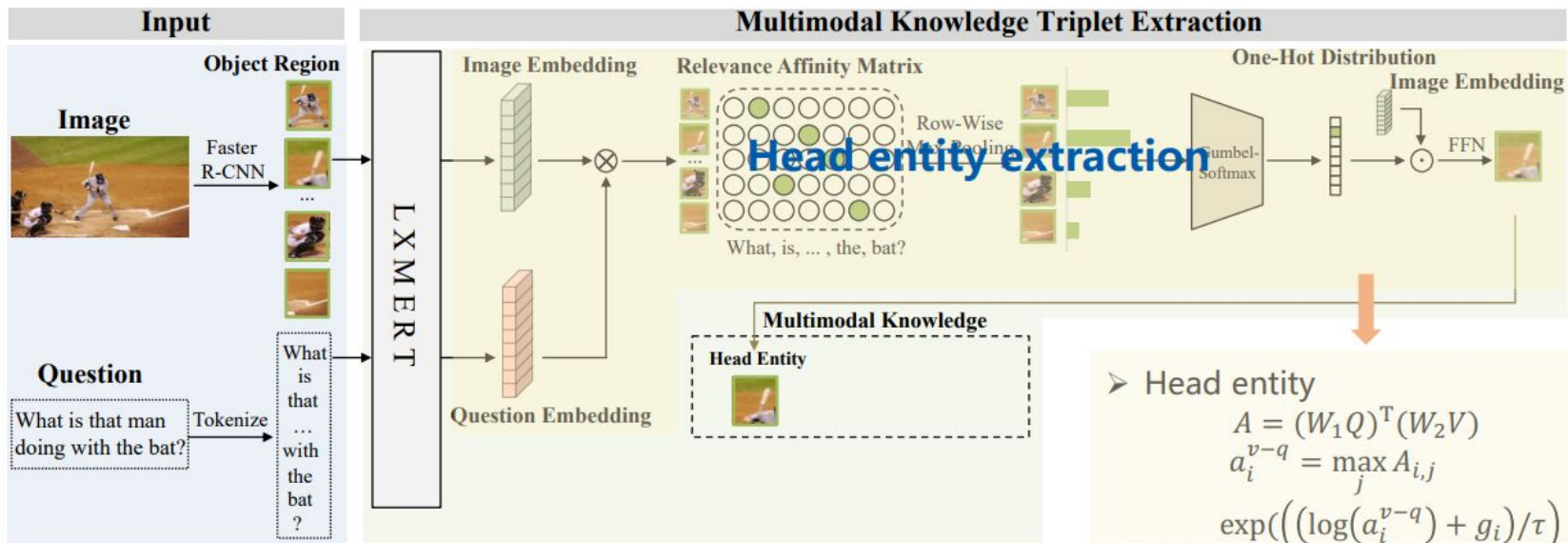
image resource: <https://arxiv.org/pdf/2203.09138.pdf>

MuKEA - image & question encoding

- MuKEA applies **Faster R-CNN** to detect a set of objects in image
- MuKEA tokenize the question using **WordPiece**
- MuKEA encodes the question and image for triple extraction with **pretrained LXMERT** to obtain the visual embeddings and the token embeddings



MuKEA - head entity extraction



➤ Head entity

$$A = (W_1 Q)^T (W_2 V)$$

$$a_i^{v-q} = \max_j A_{i,j}$$

$$a_i = \frac{\exp((\log(a_i^{v-q}) + g_i)/\tau)}{\sum_{j=1}^K \exp((\log(a_i^{v-q}) + g_i)/\tau)}$$

$$h = \text{FFN}\left(\sum_{i=1}^K a_i v_i\right)$$

MuKEA - head entity extraction

- Equ1. is to evaluate the relevance by computing the question-guided object-question relevance affinity matrix
- Equ2. is to evaluate the most relevance of each object to the question
- Equ3. is to obtain the approximate one-hot categorical distribution

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

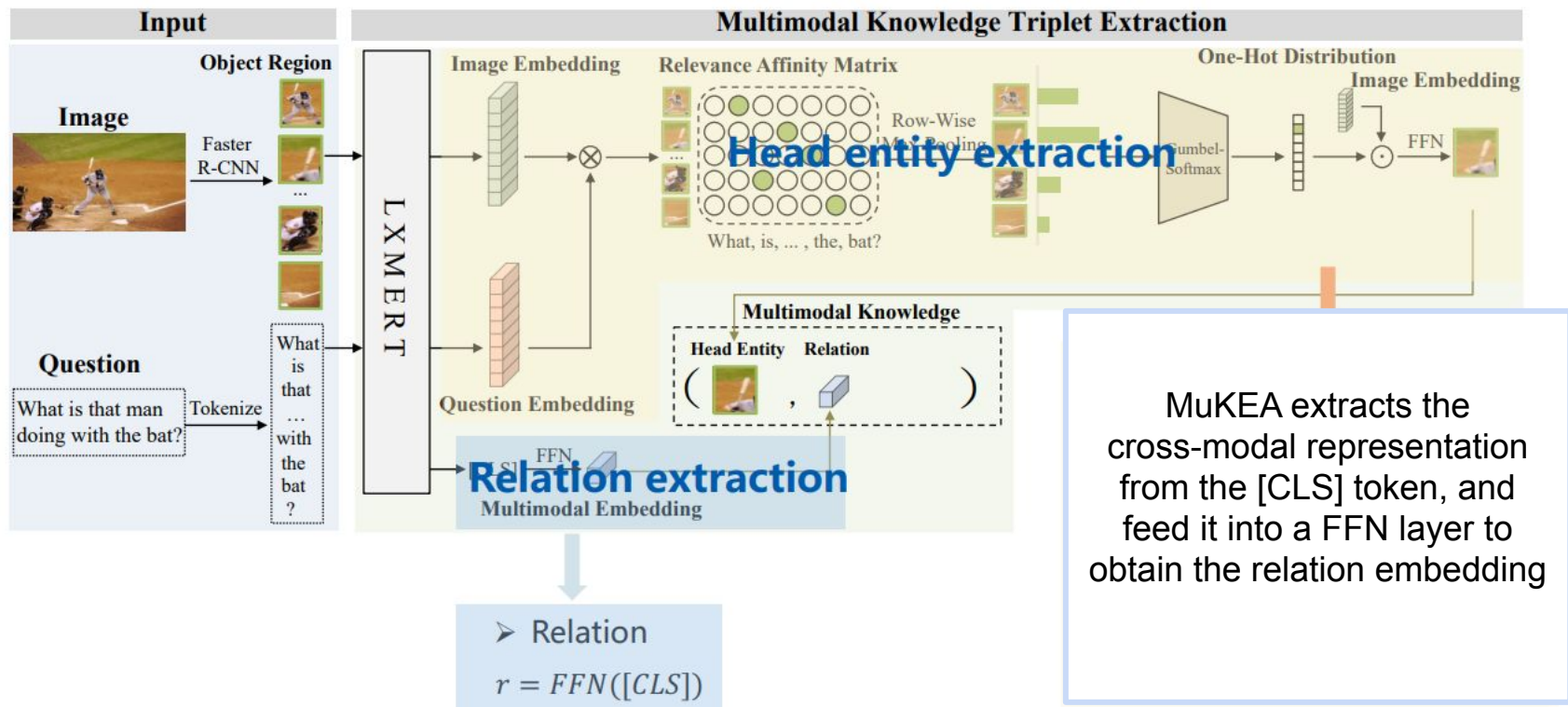
$$\mathbf{a}_i^{v-q} = \max_j \mathbf{A}_{i,j} \quad (2)$$

$$\alpha_i = \frac{\exp((\log(\mathbf{a}_i^{v-q}) + g_i)/\tau)}{\sum_{j=1}^K \exp((\log(\mathbf{a}_j^{v-q}) + g_j)/\tau)} \quad (3)$$

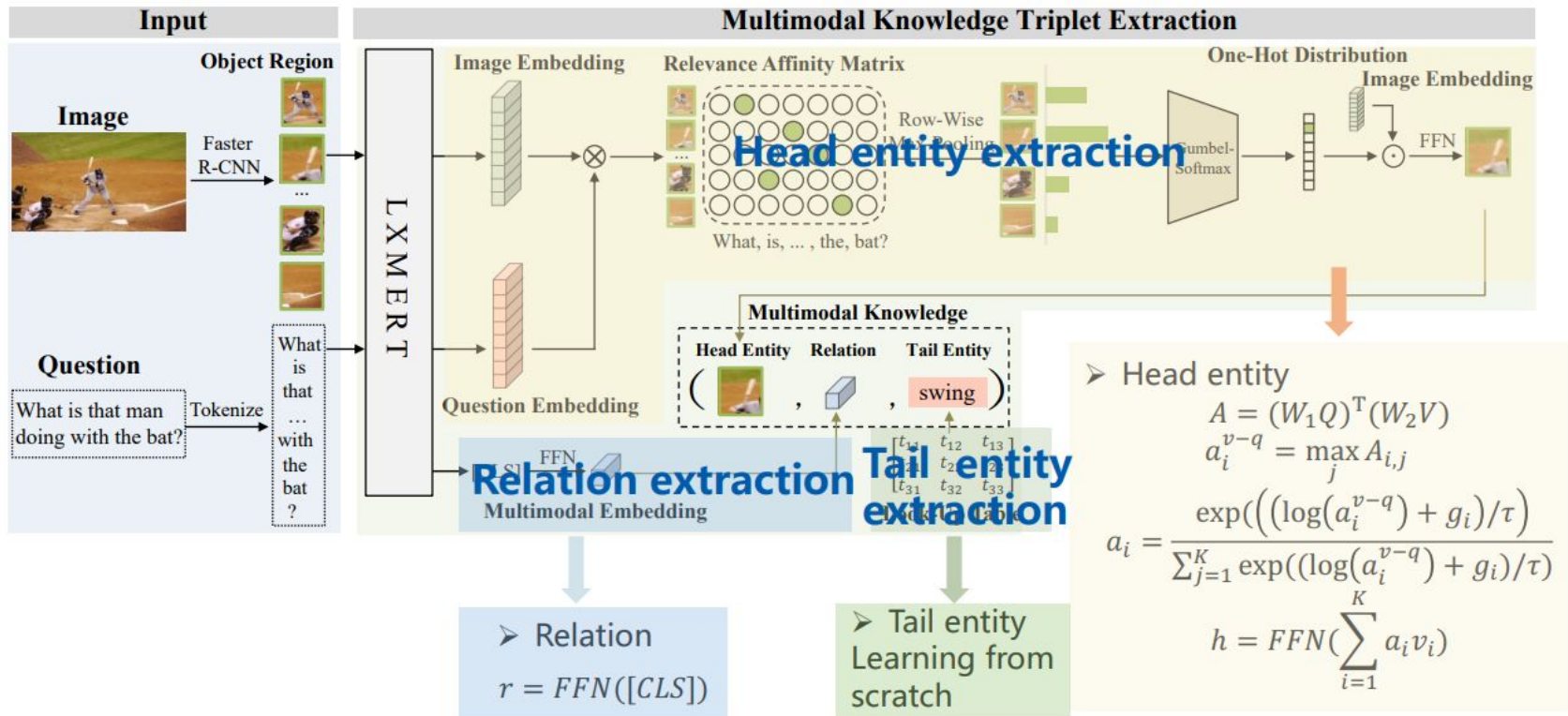
$$\mathbf{h} = \text{FFN}(\sum_{i=1}^K \alpha_i \mathbf{v}_i) \quad (4)$$



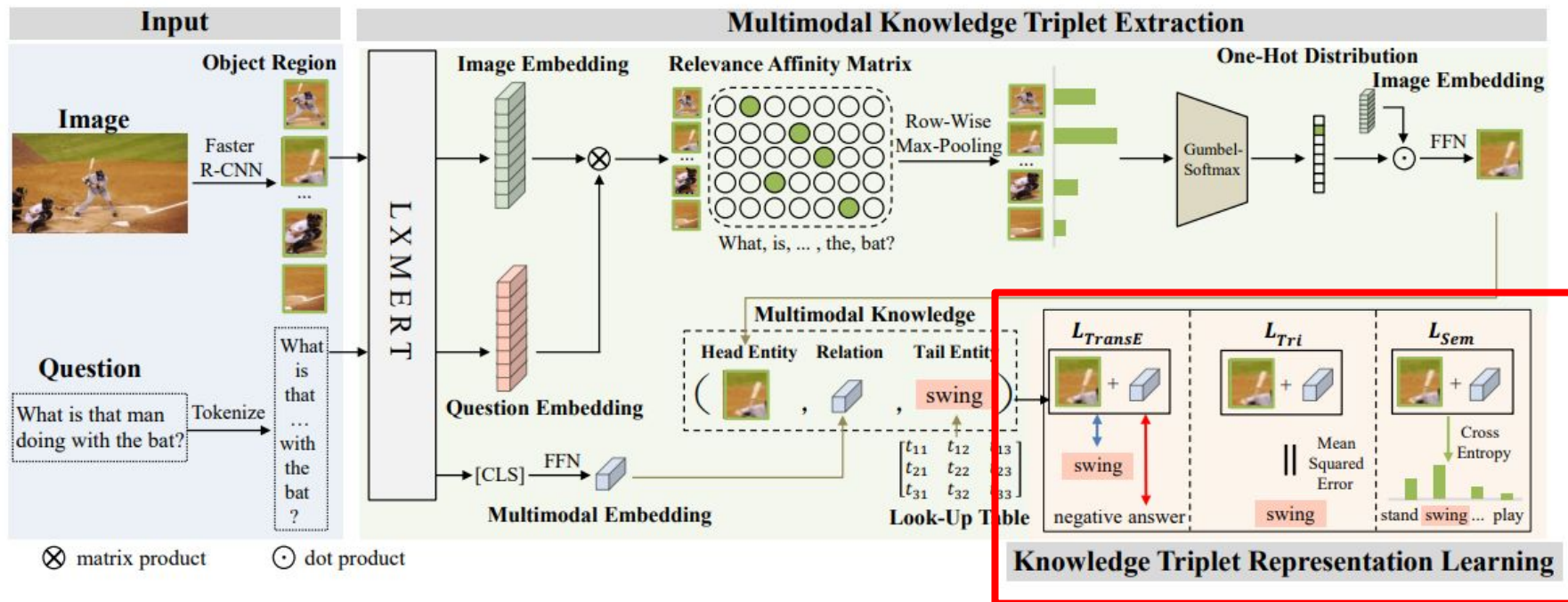
MuKEA - relation extraction



MuKEA - tail entity extraction



MuKEA - training stage



Knowledge triplet representation learning

- Three objective loss functions to learn the triplet representation to bridge the heterogeneous gap and the semantic gap
 - Preserve the embedding structure (equ5.) -> **issue**
 - Force the strict topological relation (equ6.)
 - Learn a common semantic space (equ7,8.)

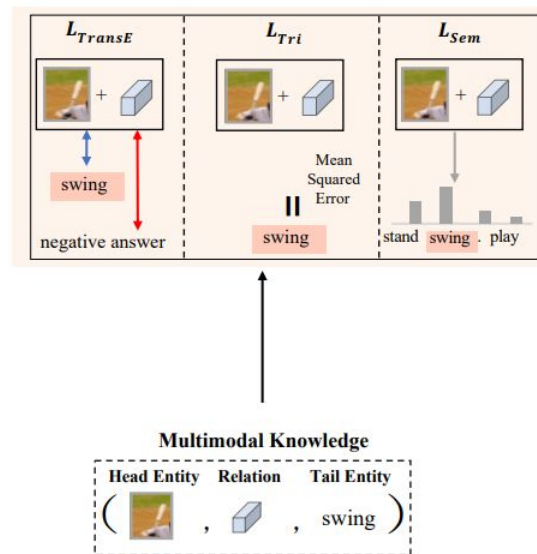
$$\mathcal{L}_{TransE} = \sum_{t^+ \in \mathcal{A}^+} \sum_{t^- \in \mathcal{A}^-} [\gamma + d(\mathbf{h} + \mathbf{r}, t^+) - d(\mathbf{h} + \mathbf{r}, t^-)]_+ \quad (5)$$

$$\mathcal{L}_{Tri} = \text{MSE}(\mathbf{h} + \mathbf{r}, t^+) \quad (6)$$

$$P(t^+) = \text{softmax}((\mathbf{T})^T(\mathbf{h} + \mathbf{r})) \quad (7)$$

$$\mathcal{L}_{Sem} = -\log(P(t^+)) \quad (8)$$

$$\mathcal{L} = \mathcal{L}_{TransE} + \mathcal{L}_{Tri} + \mathcal{L}_{Sem} \quad (9)$$



Knowledge accumulation and prediction

- MuKEA use two stages training strategy to accumulate multimodal knowledge
 - **Pre-training** on VQA 2.0 dataset
 - **Fine-tuning** on the downstream KB-VQA task to learn complex knowledge
- Inference:
 - MuKEA computes the distance between $h_{inf}+r_{inf}$ and each tail entity t_i in the look-up table T , and **select the tail entity with the minimum distance**

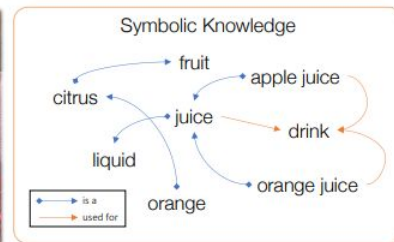
$$t_{inf} = \arg \min_{t_i \in T} d(h_{inf} + r_{inf}, t_i) \quad (10)$$

KRISP - Idea

- **Implicit knowledge** can be efficiently learned from models pre-trained on large-scale corpora
- **Explicit knowledge** can be learned from explicit and symbolic knowledge in knowledge base
- By integrating the two models, both implicit and explicit knowledge can be combined for reasoning.



Which liquid here comes from a citrus fruit?



?

Orange juice

KRISP - Reasoning with implicit knowledge

- Question encoding: KRISP tokenize a question Q using **WordPiece** as in BERT. Then, KRISP embed them with the pre-trained **BERT** embeddings and append BERT's positional encoding.
- Visual features: KRISP uses bottom-up features collected from **Faster-RCNN**.



KRISP - Reasoning with symbolic knowledge

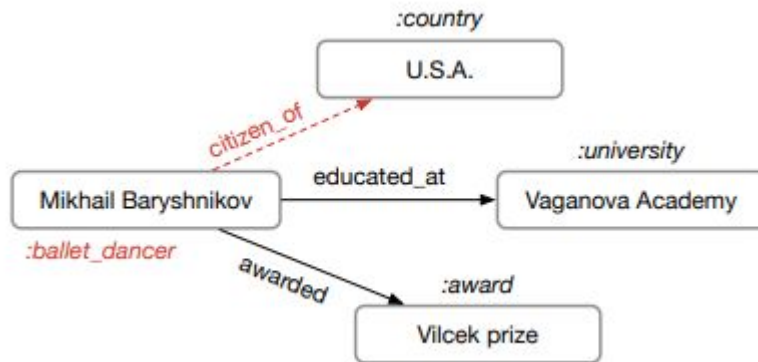
- Visual symbols:
 - KRISP uses the pre-trained visual recognition systems to get image features
 - Visual concepts: places, objects, parts of objects and attributes
 - Totally, KRISP has 4000 visual concepts
- Knowledge graph construction:
 - **Trivia knowledge:** facts about famous people, places or events
 - **Commonsense knowledge:** what are houses made of, what is a wheel part of
 - **Scientific knowledge:** what genus are dogs, what are different kinds of nutrients
 - **Situational knowledge:** where do cars tend to be located, what tends to be inside bowls

	Relation Types																					
	has part	is a	used for	has a	at location	has property	located near	instance of	related to	made of	part of	capable of	causes	is on	is in	has	is made of	is at	is part of	is near	is for	
hasPart KB	X																					
DBPedia	X	X																				
ConceptNet		X	X	X	X	X	X	X	X	X	X	X	X									
VisualGenome														X	X	X	X	X	X	X	X	X

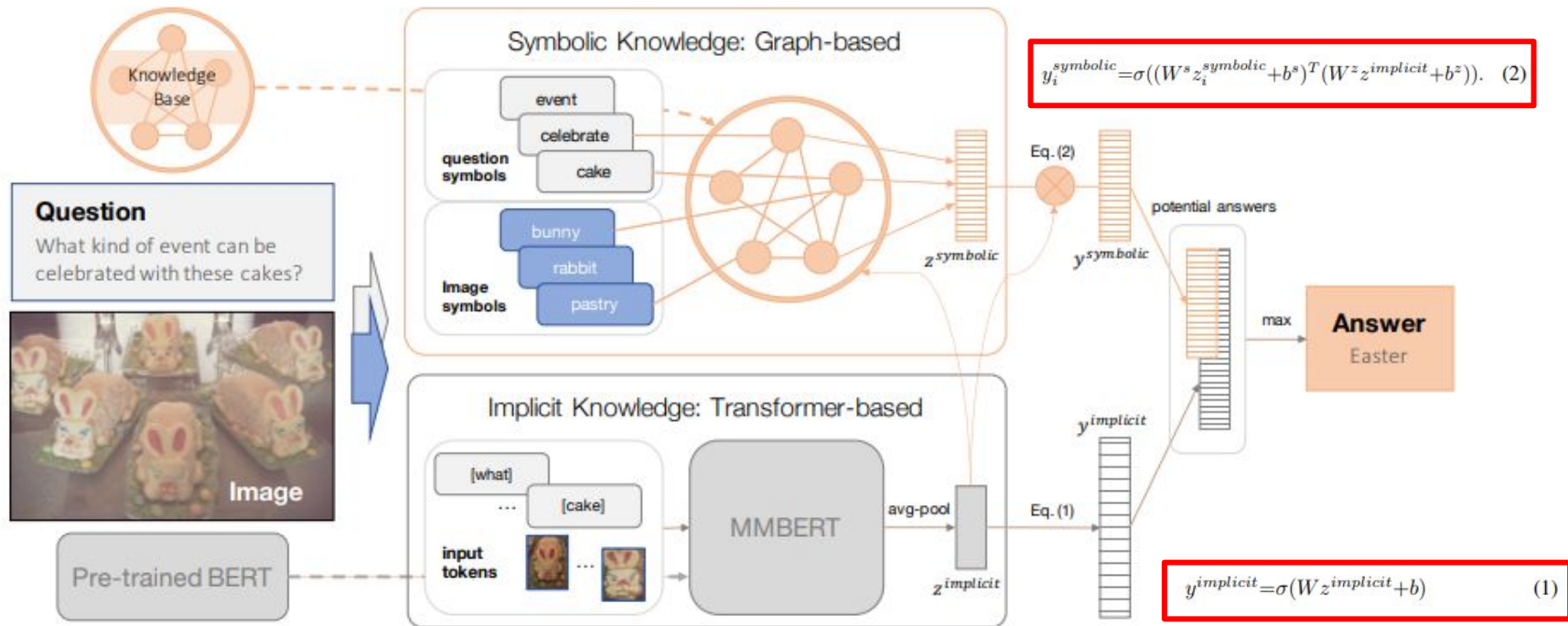
hasPart KB	DBPedia	ConceptNet	VisualGenome
(bear, has part, coat)	(poland, is a, country)	(saloon, used for, drink)	(tree, is near, building)
(wasp, has part, wing)	(mark, is a, currency)	(stream, at location, forest)	(car, is on, road)
(cnidarian, has part, cell)	(easyjet, is a, company)	(eye, used for, look)	(building, is made of, bricks)
(alfalfa plant, has part, leave)	(gerbera, is a, insect)	(tearoom, used for, drink tea)	(outlet, is on, wall)
(water, has part, water molecule)	(new era, is a, automobile)	(heifer, at location, barnyard)	(tracks, is for, train)
(human, has part, bone)	(brussels, has part, ixelles)	(quartz, is a, mineral)	(chair, is near, table)
(hare, has part, long ear)	(syrah, is a, grape)	(star, at location, galaxy)	(food, is in, bowl)
(fern, has part, spore)	(leona, is a, ship)	(hotel room, used for, sleep in)	(giraffe, has, spots)

KRISP - Reasoning with symbolic knowledge

- Graph network
- KRISP uses **Relational Graph Convolutional Network(RGCN)** as the base
- RGCN can natively supports having different calculations between nodes for different edge types and edge directions



KRISP - Integrating implicit & explicit knowledge



Experiments

- 2 datasets:
 - Outside Knowledge VQA (OK-VQA)
 - Knowledge-Routed VQA (KRVQA)




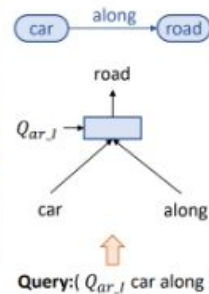
<p>Vehicles and Transportation</p>  <p>Q: What sort of vehicle uses this item? A: firetruck</p>	<p>Brands, Companies and Products</p>  <p>Q: When was the soft drink company shown first created? A: 1898</p>	<p>Objects, Material and Clothing</p>  <p>Q: What is the material used to make the vessels in this picture? A: cooper</p>	<p>Sports and Recreation</p>  <p>Q: What is the sports position of the man in the orange shirt? A: goalie</p>	<p>Cooking and Food</p>  <p>Q: What is the name of the object used to eat this food? A: chopsticks</p>
<p>Geography, History, Language and Culture</p>  <p>Q: What days might I most commonly go to this building? A: Sunday</p>	<p>People and Everyday Life</p>  <p>Q: Is this photo from the 50's or the 90's? A: 50's</p>	<p>Plants and Animals</p>  <p>Q: What phylum does this animal belong to? A: chordata, chordata</p>	<p>Science and Technology</p>  <p>Q: How many chromosomes do these creatures have? A: 23</p>	<p>Weather and Climate</p>  <p>Q: What is the warmest outdoor temperature at which this kind of weather can happen? A: 32 degrees</p>

image resource: <https://okvqa.allenai.org/>



Q: Tell me the object that the car is driving on?
A: road
KM-net: road
Facts-pred: (car, along, road)

Qtype: 1
First-order
KB-not-related

image resource: <https://www.sysu-hcp.net/resources/datasets/index.html>

Experiment Analysis - OK-VQA

- avoid cascading error
- MuKEA captures the question-centric and information-abstract multimodal knowledge

Method	Knowledge Resources	Accuracy
ArticleNet (AN) [24]	Wikipedia	5.28
Q-only [24]	—	14.93
BAN [14]	—	25.17
+AN [24]	Wikipedia	25.61
+ KG-AUG [16]	Wikipedia + ConceptNet	26.71
MUTAN [5]	—	26.41
+ AN [24]	Wikipedia	27.84
Mucko [46]	ConceptNet	29.20
GRUC [41]	ConceptNet	29.87
KM ⁴ [44]	multimodal knowledge from OK-VQA	31.32
ViLBERT [20]	—	31.35
LXMERT [34]	—	32.04
KRISP(w/o mm pre.) [23]	DBpedia + ConceptNet + VisualGenome + haspartKB	32.31
KRISP(w/ mm pre.) [23]	DBpedia + ConceptNet + VisualGenome + haspartKB	38.90
ConceptBert [9]	ConceptNet	33.66
Knowledge is Power [45]	YAGO3	39.24
MuKEA	multimodal knowledge from VQA 2.0 and OK-VQA	42.59



+3.35%

Experiment Analysis - KRVQA

- In the vision-only questions, MuKEA requires multimodal commonsense to **bridge the low-level visual content and high-level semantics**

Method	KB-not-related							KB-related					Overall
	one-step			two-step				one-step	two-step				
	0	1	2	3	4	5	6	2	3	4	5	6	
Q-type [7]	36.19	2.78	8.21	33.18	35.97	3.66	8.06	0.09	0.00	0.18	0.06	0.33	8.12
LSTM [7]	45.98	2.79	2.75	43.26	40.67	2.62	1.72	0.43	0.00	0.52	1.65	0.74	8.81
FiLM [29]	52.42	21.35	18.50	45.23	42.36	21.32	15.44	6.27	5.48	4.37	4.41	7.19	16.89
MFH [43]	43.74	28.28	27.49	38.71	36.48	20.77	21.01	12.97	5.10	6.05	5.02	14.38	19.55
UpDn [2]	56.42	29.89	28.63	49.69	43.87	24.71	21.28	11.07	8.16	7.09	5.37	13.97	21.85
MCAN [42]	49.60	27.67	25.76	39.69	37.92	21.22	18.63	12.28	9.35	9.22	5.23	13.34	20.52
+ knowledge retrieval [7]	51.32	27.14	25.69	41.23	38.86	23.25	21.15	13.59	9.84	9.24	5.51	13.89	21.30
MuKEA	59.12	44.88	37.36	52.47	48.08	35.63	31.61	17.62	6.14	9.85	6.22	18.28	27.38



+6.08%

Ablation study

Method	Accuracy
1. MuKEA (full model)	42.59
Ablation of Loss Function	
2. w/o \mathcal{L}_{Tri}	41.35
3. w/o \mathcal{L}_{Sem}	42.06
4. w/o \mathcal{L}_{Tri} & \mathcal{L}_{Sem}	40.84
5. w/o \mathcal{L}_{TransE}	24.50
Ablation of Triplet Representation	
6. head entity w/ soft-attention	40.67
7. relation w/ self-attention	40.79
8. tail entity w/ GloVe	41.42
Ablation of Triplet Structure	
9. w/o h	39.83
10. w/o r	39.40
Ablation of Knowledge Source	
11. w/o VQA 2.0 knowledge	36.35
12. w/o OK-VQA knowledge	27.20
Ablation of Pre-training Knowledge	
13. w/o LXMERT pre-training	33.52

● Confirm the complementary of each loss function.

● Assess the influence of triplet extraction methods.

● Prove the importance of triplet structure.

● Both basic knowledge and domain-specific knowledge are important.

● Influence of prior knowledge accumulated in the pre-trained LXMERT

Experiment Analysis - Knowledge complementary

- Multimodal knowledge and existing KB knowledge respectively deals with **different types** of open-ended questions.

Method	Failure subset		
	MUTAN + AN*	Mucko*	KRISP*
MuKEA	40.09	40.06	40.46

(a)

Method	Failure subset
	MuKEA
MUTAN + AN*	26.45
Mucko*	27.68
KRISP*	27.68

(b)

- Complementary benefits** of multimodal knowledge and existing knowledge bases

Method	Accuracy
MuKEA	42.59
MUTAN + AN*	25.43
MuKEA + (MUTAN + AN*)	35.39
MuKEA + (MUTAN + AN*) oracle	43.64
Mucko*	27.17
MuKEA + Mucko*	35.97
MuKEA + Mucko* oracle	44.84
KRISP*	32.02
MuKEA + KRISP*	37.75
MuKEA + KRISP* oracle	47.15

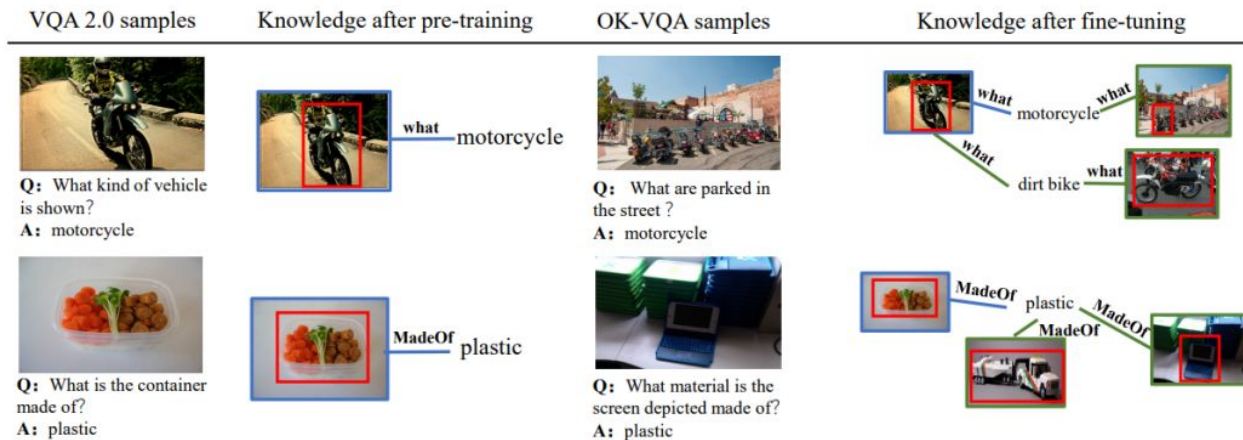
Experiment Analysis - Long-tail analysis

- This analysis is to prove the model's generalization ability on the rare answers while not overfitting on the 'head' ones
- **mAccuracy** is to fairly evaluate the performance on the long-tail distributed answers
- mAccuracy calculates the accuracy for each unique answer separately and average for all the answers

Method	Accuracy	mAccuracy
KRISP*	32.31	26.91
MuKEA	42.59	35.42

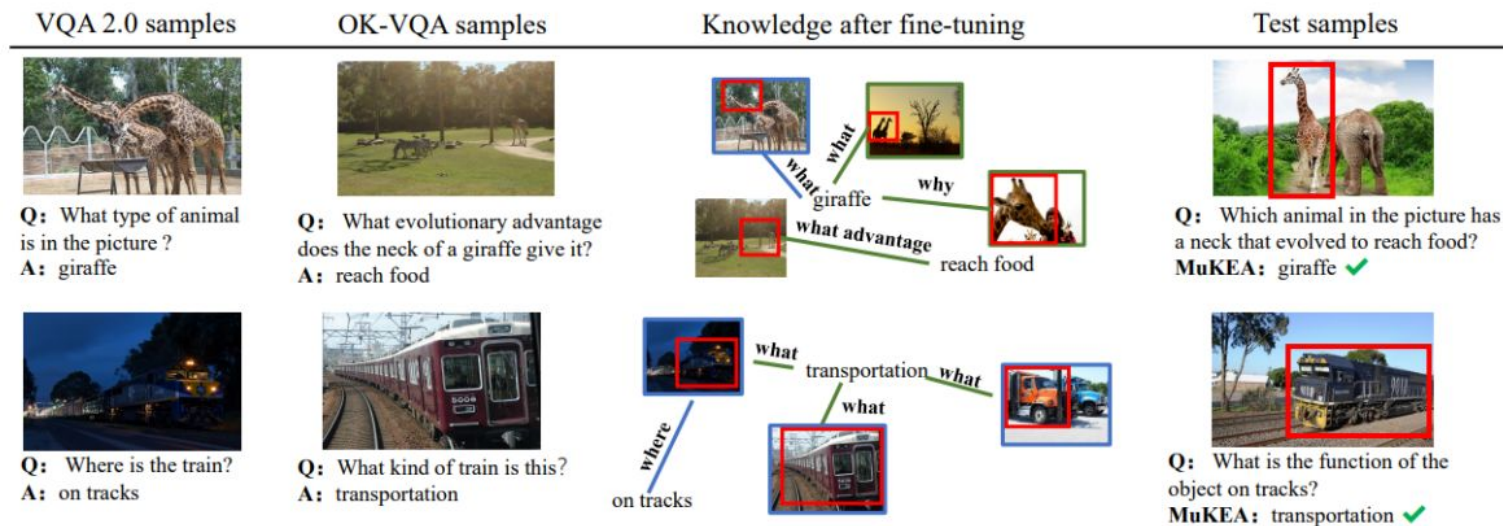
Experiment Analysis - Progressive Knowledge Accumulation

- This table shows how the basic visual knowledge in **VQA 2.0** helps to learn more complex knowledge in **OK-VQA**



Experiment Analysis - Zero-shot Analysis of Accumulated Multi-modal Knowledge

- MuKEA correlates **giraffe** with **evolution** through the manually constructed question




Qualitative analysis

- MuKEA captures **instantiated knowledge**
- MuKEA contains **multi-object involved complex knowledge**



Q: What electronic device is being featured in this photo?


KRISP: laptop ✗	MuKEA: remote ✓
Knowledge graph (screen, is on, laptop) (laptop, has, screen)	Multimodal knowledge (button,  , remote)



Q: What device is pictured?
Ground Truth: remote



Q: What kind of plane is this?


KRISP: biplane ✗	MuKEA: prop plane ✓
Knowledge graph (biplane, is a, airplane)	Multimodal knowledge (propeller,  , prop plane)



Q: What type of fuel does this plane use?
Ground Truth: jet



Q: What type of architecture is shown in these buildings?


KRISP: victorian ✗	MuKEA: gothic ✓
Knowledge graph (victorian, is a, comic)	Multimodal knowledge (city,  , gothic)



Q: What style of architecture is pictured in this photo?
Ground Truth: gothic



Q: Why is this dangerous?


KRISP: danger ✗	MuKEA: drown ✓
Knowledge graph (danger, has property, bad)	Multimodal knowledge (water,  , drown)



Q: What is the largest one of these natural occurrences ever recorded?
Ground Truth: 100 feet



Q: What style of oranges are in the stack?


KRISP: granny smith ✗	MuKEA: navel ✓
Knowledge graph (apple, capable of, granny smith)	Multimodal knowledge (orange,  , navel)



Q: What kind of orange is this?
Ground Truth: navel



Q: What is the name for a child of the species shown?

KRISP: herd ✗	MuKEA: calf ✓
Knowledge graph (sheep, is in, herd) (herd, has part, lamb)	Multimodal knowledge (cow,  , calf)



Q: The baby of this animal is called what?
Ground Truth: calf

Summary & Future Work

- Summary:
 - MuKEA focuses on **multimodal knowledge** instead of language knowledge for KB-VQA
 - Multimodal knowledge is represented by **explicit triplets** via three loss functions
 - A pre-training and fine-tuning strategy **accumulates multimodal knowledge** from basic to complex
- Future Work:
 - How to effectively **combine** multimodal knowledge with existing knowledge bases?
 - How to accumulate **generic** multimodal knowledge for vision-language tasks?

Thank you