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

image resource: https://mmlab-iie.github.io/
VQA evolution

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

image resource: https://mmlab-iie.github.io/uploads/MSRA_CVPR_2022_Pre-MuKEA.pdf
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.

Image resource: https://arxiv.org/pdf/1602.07332.pdf
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
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**?

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

![Diagram: Multimodal Knowledge Representation](https://arxiv.org/pdf/2203.09138.pdf)
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 \(t\)
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

image resource: https://arxiv.org/pdf/2203.09138.pdf
MuKEA - head entity extraction

Head entity extraction

- What is that man doing with the bat?

- Faster R-CNN

- Image Embedding

- Multimodal Knowledge Triplet Extraction

- One-Hot Distribution

- Image Embedding

- FFN

- head entity

\[ A = (W_1Q)^T(W_2V) \]

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

\[ a_i = \frac{\exp\left(\left(\log(a_i^{v-q}) + g_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\left(\log(a_j^{v-q}) + g_j\right)/\tau\right)} \]

\[ h = FFN\left(\sum_{i=1}^{K} a_i v_i\right) \]
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.

\[
A = (W_1Q)^T(W_2V) \tag{1}
\]

\[
a_i^{v,q} = \max_j A_{i,j} \tag{2}
\]

\[
\alpha_i = \frac{\exp((\log(a_i^{v,q}) + g_i)/\tau)}{\sum_{j=1}^{K} \exp((\log(a_j^{v,q}) + g_j)/\tau)} \tag{3}
\]

\[
h = \text{FFN}\left(\sum_{i=1}^{K} \alpha_i v_i\right) \tag{4}
\]
MuKEA - relation extraction

MuKEA extracts the cross-modal representation from the [CLS] token, and feed it into a FFN layer to obtain the relation embedding.
MuKEA - tail entity extraction

Head entity extraction

Relation extraction

Tail entity extraction

Head entity

Relation

Tail entity

Learning from scratch

Relation

$r = FFN([CLS])$

Head entity

$A = (W_1Q)^T(W_2V)$

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

$\alpha_i = \frac{\exp\left((\log(a_i^{v,q}) + g_i)/\tau\right)}{\sum_{j=1}^{K} \exp((\log(a_j^{v,q}) + g_j)/\tau)}$

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

image resource: https://mmlab-iie.github.io/uploads/MSRA_CVPR_2022_Pre-MuKEA.pdf
MuKEA - training stage

image resource: https://arxiv.org/pdf/2203.09138.pdf
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.)

\[
\mathcal{L}_{\text{TransE}} = \sum_{t^+ \in A^+} \sum_{t^- \in A^-} [\gamma + d(h+r, t^+) - d(h+r, t^-)]_+ \\
\mathcal{L}_{\text{Tri}} = \text{MSE}(h+r, t^+) \\
P(t^+) = \text{softmax}((T)^T (h + r)) \\
\mathcal{L}_{\text{Sem}} = -\log(P(t^+)) \\
\mathcal{L} = \mathcal{L}_{\text{TransE}} + \mathcal{L}_{\text{Tri}} + \mathcal{L}_{\text{Sem}}
\]
Knowledge accumulation and prediction

- MuKEA uses a two-stage 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

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

![Image](https://arxiv.org/pdf/2012.11014.pdf)
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.

image resource: https://arxiv.org/pdf/2012.11014.pdf
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 Table](https://arxiv.org/pdf/2012.11014.pdf)

- **Example Knowledge**
  - ConceptNet:
    - (color, used for, drink tea)
    - (play, used for, soccer)
  - DBpedia:
    - (fossil, is a, animal)
    - (maker, is a, country)
  - VisualGenome:
    - (train, is near, building)
    - (cafe, is on, road)

Image resource: https://arxiv.org/pdf/2012.11014.pdf
KRISP - Reasoning with symbolic knowledge

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

image resource: https://arxiv.org/pdf/1703.06103.pdf
KRISP - Integrating implicit & explicit knowledge

image resource: https://arxiv.org/pdf/2012.11014.pdf
Experiments

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

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

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

| Method                                      | Accuracy |
|---------------------------------------------|----------|
| MuKEA (full model)                         | 42.59    |
| Ablation of Loss Function                  |          |
| 2. w/o $\mathcal{L}_{\text{Tri}}$          | 41.35    |
| 3. w/o $\mathcal{L}_{\text{Sem}}$          | 42.06    |
| 4. w/o $\mathcal{L}_{\text{Tri}}$ & $\mathcal{L}_{\text{Sem}}$ | 40.84    |
| 5. w/o $\mathcal{L}_{\text{TransE}}$       | 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

image resource: https://mmlab-iie.github.io/uploads/MSRA_CVPR_2022_Pre-MuKEA.pdf
Experiment Analysis - Knowledge complementary

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

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

```
| Method       | Failure subset | Accuracy |
|--------------|----------------|----------|
| MuKEA        | 40.09          | 42.59    |
| MUTAN + AN*  | 40.06          | 25.43    |
| MuKEA + (MUTAN + AN*) | 40.46  | 35.39    |
| MuKEA + (MUTAN + AN*) oracle |        | 43.64    |
| Mucko*       | 26.45          | 27.17    |
| MuKEA + Mucko* | 27.68    | 35.97    |
| MuKEA + Mucko* oracle |        | 44.84    |
| KRISP*       | 27.68          | 32.02    |
| MuKEA + KRISP* | 27.68    | 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     |
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

| VQA 2.0 samples | OK-VQA samples | Knowledge after fine-tuning | Test samples |
|-----------------|----------------|-----------------------------|--------------|
| ![Train](image1) | ![Train](image2) | ![Knowledge](image3) | ![Test](image4) |
| Q: Where is the train?  
A: on tracks | Q: What kind of train is this?  
A: transportation |  | Q: Which animal in the picture has a neck that evolved to reach food?  
MuKEA: giraffe ✓ |
| ![Giraffe](image5) | ![Giraffe](image6) | ![Knowledge](image7) |  |
| Q: What type of animal is in the picture?  
A: giraffe | Q: What evolutionary advantage does the neck of a giraffe give it?  
A: reach food |  | Q: What is the function of the object on tracks?  
MuKEA: transportation ✓ |

Image source: https://mmlab-iie.github.io/uploads/MSRA_CVPR_2022_Pre-MuKEA.pdf
Qualitative analysis

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

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

- **Q:** What device is pictured?  
  - **Ground Truth:** remote

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

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

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

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

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

- **Q:** What kind of plane is this?  
  - **Ground Truth:** jet

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

- **Q:** Why is this dangerous?  
  - **Ground Truth:** 100 feet

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