Out of the Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering

Medhini Narasimhan, Svetlana Lazebnik, Alexander Schwing

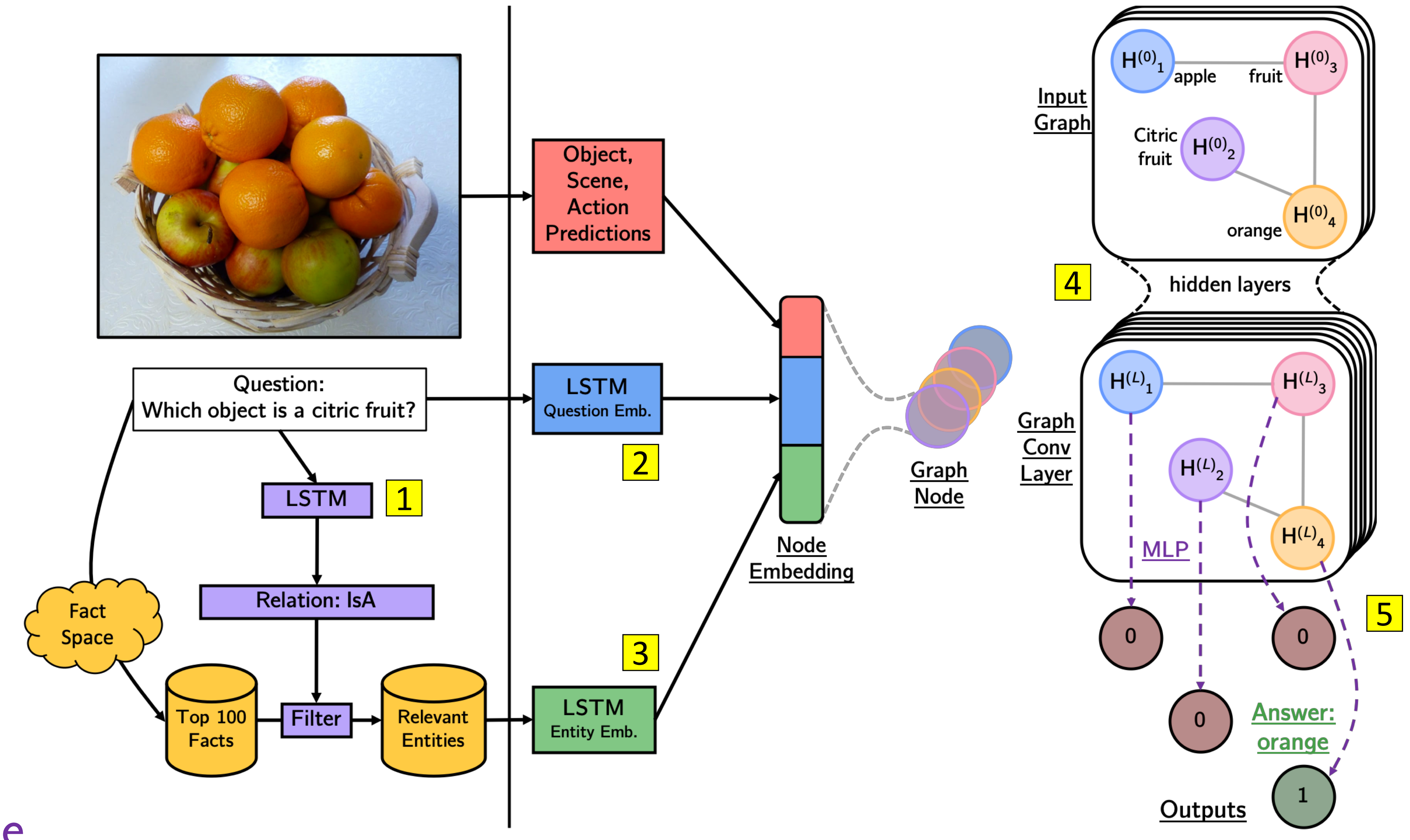
ILLINOIS

Overview

- **Objective:** To answer open ended questions about an image using facts from an external knowledge base.
- We use the **FVQA Dataset** containing image – question pairs and the corresponding **FVQA Knowledge Base** of facts. **[1]**
- We develop a model that **reasons using message passing** across multiple **relevant facts** before arriving at an answer.

Question: Which object in the image is more similar to a tiger?
Fact: (Cat, RelatedTo, Tiger)
Answer: Cat

Learning Knowledge Base Retrieval



Inference

1. Retrieval of Relevant Facts

- Fact consists of (visual concept, relation, attribute), e.g., *(Orange, IsA, Fruit)*
- 100 relevant facts retrieved based on GloVe similarity of the fact with the question and visual concepts in image
- One relation out of 13 possible is obtained from the question by using an LSTM 1, proposed in **[2]**
- Top 100 facts further reduced by filtering according to the predicted relation, e.g., *IsA*
- **Entity Embedding.** Each entity, *(visual concept, attribute)* in the fact is embedded using an LSTM 3

2. Question and Visual Concept Embedding

- Question: Embedding of dimension 100 learned using an LSTM 2
- Visual Concepts: Objects, scenes, and actions detected using pre-trained models

3. Node Embedding and Graph Construction

- The visual concept, question, and entity embeddings are concatenated to form an embedding of a node
- The nodes of the graph are connected based on the relations connecting the entities

4. Answer Prediction from the Graph

- A 2-layer graph convolution network (GCN) performs a joint assessment of the nodes in the graph
- Each hidden layer of the GCN is a non-linear function given by,

$$H^{(l)} = f(H^{(l-1)}, A) = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l-1)}W^{(l-1)}) \quad \forall l \in \{1, \ldots, L\}$$

- The output of the GCN is passed through an MLP which predicts the answer

Learning

1. Relation Prediction

- The LSTM 1 is trained using ground truth question-relation pairs and standard cross-entropy loss

2. Answer Prediction

- The answer predictor's parameters consist of the question and entity embedding, the layers of the GCN and MLP
- The LSTMs 2 and 3, the GCN 4, and the MLP 5 are trained end-to-end using the ground truth answer and binary cross-entropy loss

Quantitative and Qualitative Results

Method	Accuracy	
	@1	@3
FVQA [1]	56.91	64.65
FVQA Ensemble [1]	58.76	—
STTF [2]	62.20	75.60
Ours (1 layer GCN)	**57.89**	**65.14**
Ours (3 layer GCN)	**60.78**	**68.65**
Ours (2 layer GCN)	**69.35**	**80.25**
Human	77.99	-

Answer Prediction Results

Method	Synonyms (in FVQA)	Synonyms (Generated)	Homographs (in FVQA)
FVQA [1]	78	61	66.3
STTF [2]	91.6	89	79.4
Ours	**95.38**	**91**	**81.16**

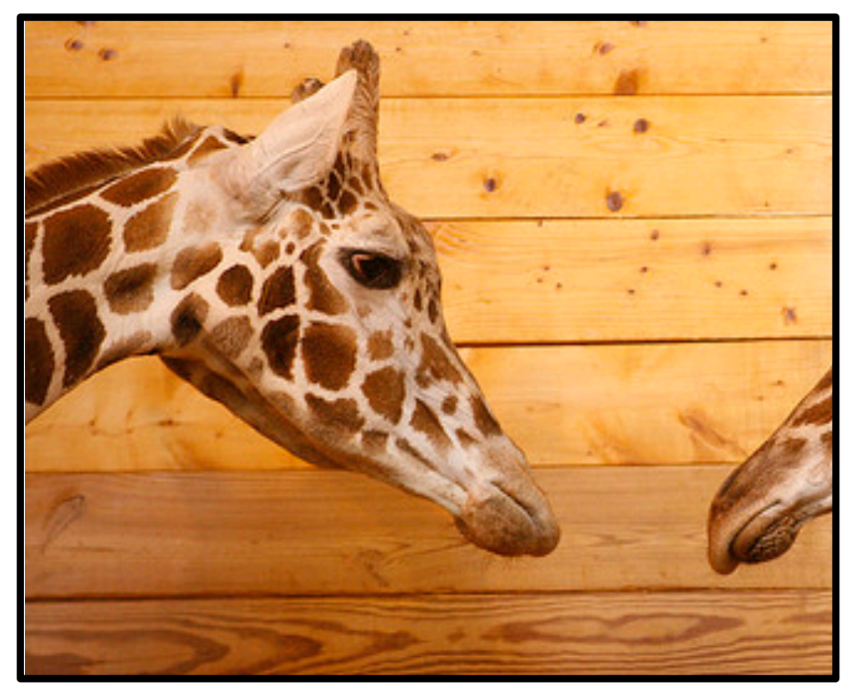
Answer Prediction Accuracy on Question-Fact pairs with Synonyms and Homographs

Correctly Answered Questions

Question: Which vehicle shown here can float?
Pred. Relation: CapableOf
Pred. Visual Concept: Boat (object)
Supporting Fact:
(Boat, CapableOf, Sailing)
Pred./GT Answer: Boat

Question: What is the place in this image used for?
Pred. Relation: UsedFor
Pred. Visual Concept: Kitchen (scene)
Supporting Fact:
(Kitchen, UsedFor, Cooking)
Pred./GT Answer: Kitchen

Question: What does the animal in the image like to chase?
Pred. Relation: CapableOf
Pred. Visual Concept: Cat (object)
Supporting Fact:
(Cat, CapableOf, Hunting mice)
Pred./GT Answer: Cat



Question: What is the plant-eating animal shown here?
Pred. Relation: Category
Pred. Visual Concept: Giraffe (object)
Supporting Fact:
(Giraffe, Category, Herbivore)
Pred./GT Answer: Giraffe

Question: What is the area in the image used for?
Pred. Relation: UsedFor
Pred. Visual Concept: Field (Scene)
Supporting Fact:
(Field, UsedFor, Grazing Animals)
Pred./GT Answer: Grazing Animals

Question: What in this image is made by baking?
Pred. Relation: Category
Pred. Visual Concept: Donut (object)
Supporting Fact:
(Donut, Category, Cooking)
Pred./GT Answer: Donut

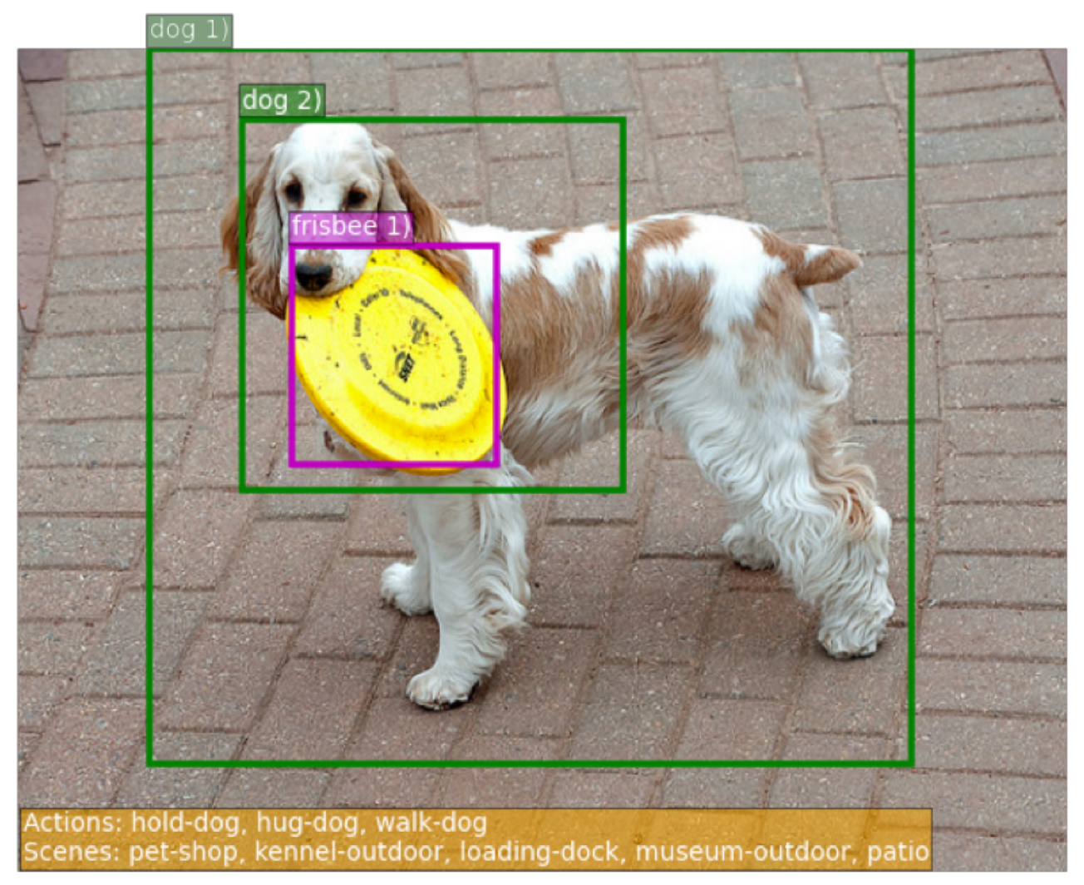
Question: What object in this image is spiky?
Pred. Relation: RelatedTo
Pred. Visual Concept: Pineapple (object)
Supporting Fact:
(Pineapple, RelatedTo, Spiky)
Pred./GT Answer: Pineapple

Question: Which object in this image is venomous?
Pred. Relation: HasProperty
Pred. Visual Concept: Snake (object)
Supporting Fact:
(Snake, HasProperty, Venomous)
Pred./GT Answer: Snake

Question: Which action shown here is faster than walking?
Pred. Relation: Comparative (faster)
Pred. Visual Concept: Cycling (action)
Supporting Fact:
(Cycling, Faster, Walking)
Pred./GT Answer: Cycling

Question: What is on the ground in this image?
Pred. Relation: AtLocation
Pred. Visual Concept: Beach (Scene)
Supporting Fact:
(Sand, AtLocation, Beach)
Pred./GT Answer: Sand

Visual Concepts Prediction


dog 1)
dog 2)
frisbee 1)
Actions: hold-dog, hug-dog, walk-dog
Scenes: pet-shop, kennel-outdoor, loading-dock, museum-outdoor, patio

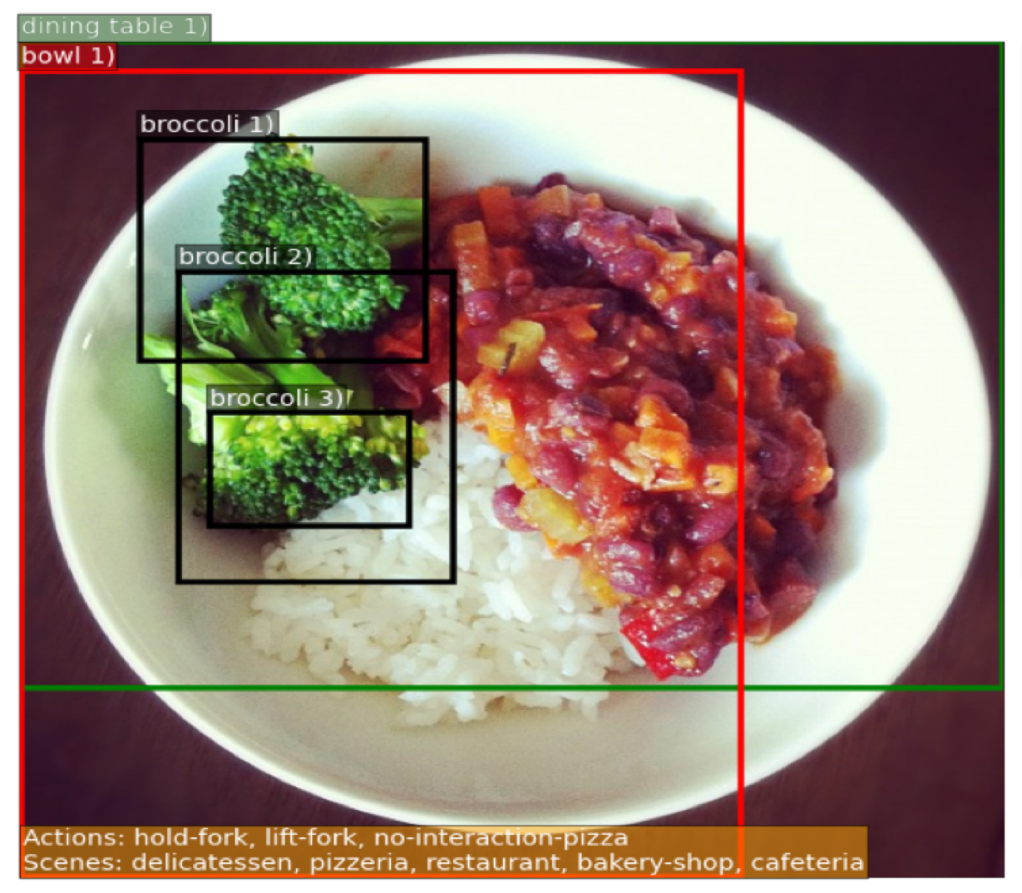
Question: What object in this image can fly?

Relevant Object: Frisbee

Predicted/GT Relation: CapableOf

Supporting Fact: (Frisbee, CapableOf, Flying)

Predicted/GT Answer: Frisbee


dining table 1)
bowl 1)
broccoli 1)
broccoli 2)
broccoli 3)
Actions: hold-fork, lift-fork, no-interaction-pizza
Scenes: delicatessen, pizzeria, restaurant, bakery-shop, cafeteria

Question: What are the greens shown in this image?

Relevant Object: Broccoli

Predicted/GT Relation: IsA

Supporting Fact: (Broccoli, IsA, Green Vegetable)

Predicted/GT Answer: Broccoli

Incorrectly Answered Questions

Question: What is the object that the picture is taken from used for?
Pred. Relation: UsedFor
GT Supporting Fact: (Airplane, UsedFor, Flying)
Pred. Answer: Printing pictures
GT Answer: Flying

Error: GT Fact not retrieved in Top–100.

Question: What object in this image is used to play polka music?
Pred. Relation: UsedFor
GT Relation: ReceivesAction
GT Supporting Fact:
(Accordion, ReceivesAction, Polka Music)
Pred. Answer: Guitar
GT Answer: Accordion

Error: Incorrect annotation / Wrong relation predicted.

Question: What object in this image is used for entering data?
Pred. Relation: UsedFor
GT Supporting Fact: (Keyboard, UsedFor, Data entry)
Pred. Answer: Laptop
GT Answer: Keyboard

Error: GCN predicted the wrong node.

References

[1] Wang P, Wu Q, Shen C, Dick A, van den Hengel A. Fvqa: Fact-based visual question answering. *IEEE TPAMI,* 2018.
[2] Narasimhan M, Schwing AG. Straight to the Facts: Learning Knowledge Base Retrieval for Factual Visual Question Answering. In *ECCV,* 2018.