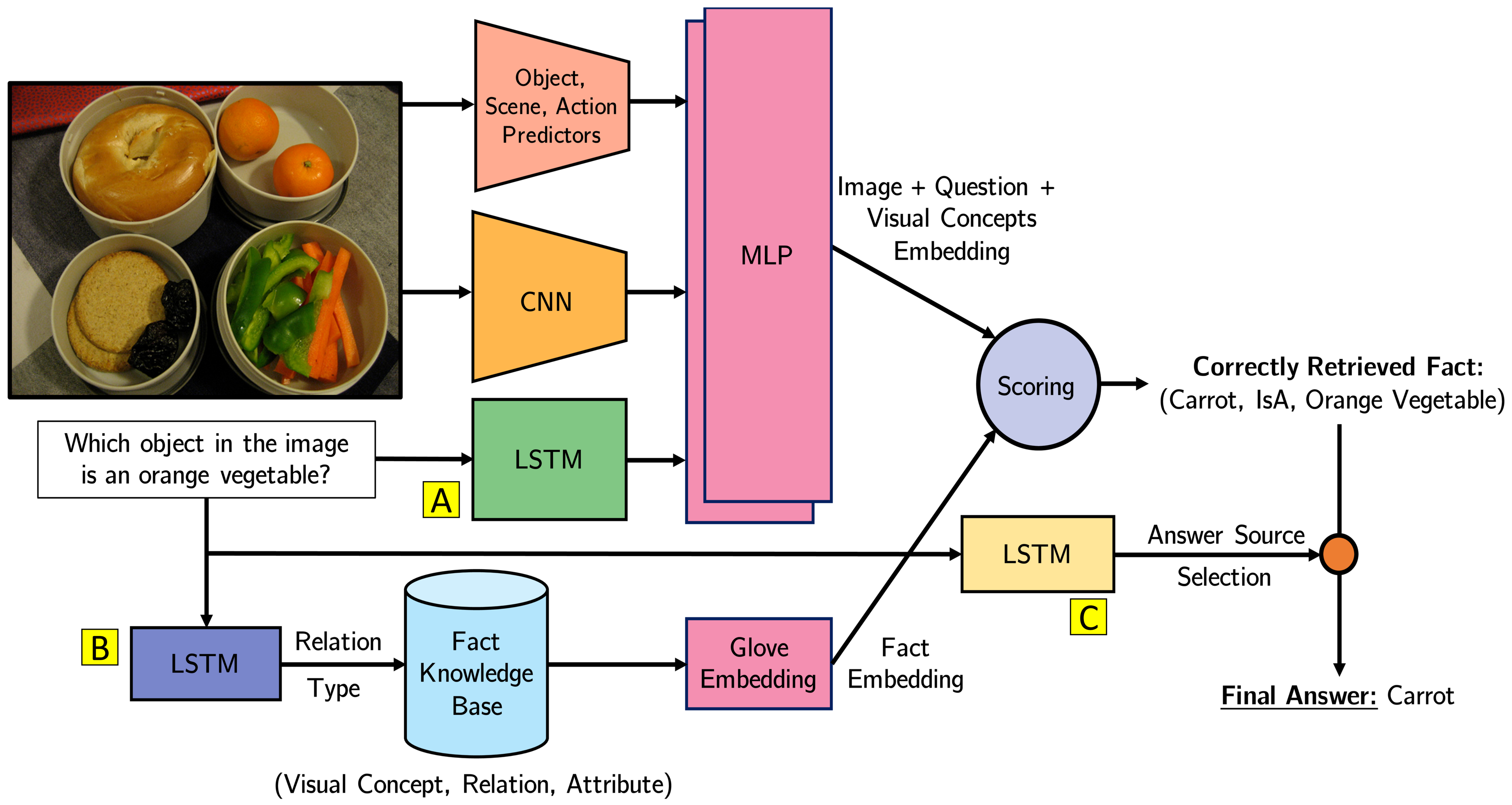
Straight to the Facts: Learning Knowledge Base Retrieval for Factual Visual Question Answering

Medhini Narasimhan, Alexander Schwing

Introduction

- **Motivation.** To answer open ended questions about an image using facts from an external knowledge base while **handling synonyms and homographs**. Answering a question correctly involves retrieving the right supporting fact and extracting the answer from it.
- **FVQA Dataset.** 2,190 images, 5,286 questions, and 4,126 unique facts corresponding to the questions.
- **FVQA Knowledge Base.** 193,449 facts, constructed by extracting the top visual concepts for all the images in the dataset and querying for those concepts in the three knowledge bases - WebChild, ConceptNet, and DBPedia.

Learning Knowledge Base Retrieval



Inference

1. **Image, Question, and Visual Concept Embedding**
   - **Image:** Low-level fc7 features extracted from a ResNet-152 model pre-trained on ImageNet
   - **Question:** Embedding of dimension 100 learned using LSTM A
   - **Visual Concepts:** Objects, scenes, and actions detected using pre-trained models
   - **Fusion:** Image, question, and visual concept features are combined using an MLP to form a 200d vector
2. **Fact Embedding**
   - Fact consists of (visual concept, relation, attribute), e.g., *(Orange, IsA, Fruit)*
   - One relation out of 13 possible is obtained from the question by using an LSTM B
   - Fact space reduced by filtering according to the predicted relation, e.g., *IsA*
   - Fact is encoded using 100d GloVe embeddings
3. **Scoring the facts**
   - Facts are scored by computing the cosine distance between the output of the MLP and the fact embeddings
   - Fact with highest score is chosen
4. **Answer from fact**
   - The answer is either the visual concept or the attribute within the chosen fact
   - Answer source is predicted from the question using an LSTM C

Learning

1. **Predicting the Relation and Answer Source**
   - The LSTM B is trained using ground truth question-relation pairs and standard cross-entropy loss
   - The LSTM C is trained using ground truth question-answer source pairs and binary cross-entropy loss
2. **Scoring the facts**
   - The score function is trained in multiple time steps by mining hard negatives in each step. Every iteration consists of the ground truth fact and 99 negatives
   - The parameters are learned using a classical margin loss that assigns the highest score to the image-question-ground truth fact embedding
   - The LSTM A, the MLP, and the score function are trained end-to-end

Quantitative Results

Method	Relation Prediction		Answer Source Prediction	
	Accuracy @1	Accuracy @3	Accuracy @1	Accuracy @3
FVQA	64.94	82.42	—	—
Ours	**75.40**	**91.97**	**97.30**	**100.00**

Method	Synonyms (FVQA)	Synonyms (Ours)	Homographs (FVQA)
FVQA	78	61	66.3
Ours	**91.6**	**89**	**79.4**

Method	Fact Prediction		Answer Prediction	
	Accuracy @1	Accuracy @3	Accuracy @1	Accuracy @3
FVQA	38.76	42.96	56.91	64.65
FVQA Ensemble	—	—	58.76	—
Ours – Q + I	28.98	32.34	26.68	30.27
Ours – Q + I + VC	62.30	74.90	60.30	73.10
Ours – Q + VC	**64.50**	**75.20**	**62.20**	**75.60**

Qualitative Results

Correctly Answered Questions

Question: What is a bookshelf used for?

Predicted Relation: UsedFor
Predicted Supporting Fact: (Bookshelf, UsedFor, Carrying Books)
Predicted Answer Source: KB

Predicted Answer: Carrying books
GT Answer: Carrying books

Question: What object in this image is capable of flying?

Predicted Relation: CapableOf
Predicted Supporting Fact: (Frisbee, CapableOf, Flying)
Predicted Answer Source: Image

Predicted Answer: Frisbee
GT Answer: Frisbee

Question: Which property does the place in the image have?

Predicted Relation: HasProperty
Predicted Supporting Fact: (Beach, HasProperty, Sandy)
Predicted Answer Source: KB

Predicted Answer: Sandy
GT Answer: Sandy

Question: What kind of sport do people usually practice in this place?

Predicted Relation: AtLocation
Predicted Supporting Fact: (Skiing, AtLocation, Ski-slope)
Predicted Answer Source: KB

Predicted Answer: Skiing
GT Answer: Skiing

Question: Which object in the image is used to make a cake?

Predicted Relation: UsedFor
Predicted Supporting Fact: (Oven, UsedFor, Baking)
Predicted Answer Source: Image

Predicted Answer: Oven
GT Answer: Oven

Question(Original): Which object in this image is related to sailing?

Predicted Relation: RelatedTo
Predicted Supporting Fact: (Boat, RelatedTo, Sail)
Predicted Answer Source: Image

Predicted Answer: Boat
GT Answer: Boat

Correctly Answered Synonymous Questions

Question(Original): Which object in the image is capable of floating on water?
Question(Synonymous): Which vehicle shown here can sail?

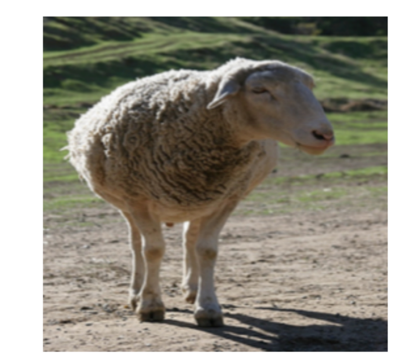
Predicted Relation: CapableOf
Predicted Supporting Fact: (Boats, CapableOf, Floating on water)
Predicted Answer Source: Image

Predicted Answer: Boat
GT Answer: Boat

Question(Original): Which object in this image is used to measure the passage of time?
Question(Synonymous): What in this image can tell time?

Predicted Relation: UsedFor
Predicted Supporting Fact: (Clock, UsedFor, measure the passage of time)
Predicted Answer Source: Image

Predicted Answer: Clock
GT Answer: Clock



Question(Original): Which object in this image is related to wool?
Question(Synonymous): Which object in this image is the source of a woolen sweater?

Predicted Relation: RelatedTo
Predicted Supporting Fact: (Sheep, RelatedTo, Wool)
Predicted Answer Source: Image

Predicted Answer: Sheep
GT Answer: Sheep

Question(Original): Which instrument in this image is common in jazz?
Question(Synonymous): Which musical instrument is shown here?

Predicted Relation: IsA
Predicted Supporting Fact: (Saxophone, IsA, Jazz instrument)
Predicted Answer Source: Image

Predicted Answer: Saxophone
GT Answer: Saxophone

Question(Original): Which object in this image is used for lighting?
Question(Synonymous): Which object in this image do you need in a dark room?

Predicted Relation: UsedFor
Predicted Supporting Fact: (Lamp, UsedFor, Lighting)
Predicted Answer Source: Image

Predicted Answer: Lamp
GT Answer: Lamp

Question(Original): What in this image is capable of hunting a mouse?
Question(Synonymous): What in this image preys on a mouse?

Predicted Relation: CapableOf
Predicted Supporting Fact: (Cat, CapableOf, Killing a mouse)
Predicted Answer Source: Image

Predicted Answer: Cat
GT Answer: Cat

Visual Concepts Prediction and Retrieved Facts

Question: Which object are you likely to find in a monkey's hand?

Objects Detected: Banana, Bowl, Cup, Bottle, Laptop, Keyboard, Dining Table. Book, Orange

Predicted Relation: AtLocation

Top-3 Retrieved Facts:
(Bananas, AtLocation, monkey's hand)
(Banana's, AtLocation, Grocery store)
(Cup, AtLocation, Kitchen)

Predicted Answer: Banana

Question: Which object in this image is considered to be a shelter?

Scenes Detected:
Alley, Residential Neighborhood, Street, House, Motel

Predicted Relation: IsA

Top-3 Retrieved Facts:
(House, IsA, Shelter)
(Car, IsA, Heavier Than Horse)
(Car, IsA, Motorvehicle)

Predicted Answer: House

Incorrectly Answered Questions

Question: What object in this image is round?

Predicted Relation: HasProperty
Predicted Supporting Fact: (Person, HasProperty, Alive)
GT Supporting Fact: (TennisBall, HasProperty, Round)

Predicted Answer Source: Image
GT Answer Source: Image

Predicted Answer: Person
GT Answer: TennisBall

Question: Which action is less strenuous than the action in the image?

Predicted Relation: Comparative
Predicted Supporting Fact: (Jumping, Comparative-more strenuous, Dressage)

Predicted Answer Source: Image
GT Answer Source: KB

Predicted Answer: Jumping
GT Answer: Dressage

Question: What sort of food can you see in the image?

Predicted Relation: IsA
GT Relation: Category

Predicted Supporting Fact: (Lemon, isA, Fruit)
GT Supporting Fact: (Fruits, Category, Food)

Predicted Answer Source: Image

Predicted Answer: Lemon
GT Answer: Fruits

Follow Up: Upcoming NIPS Paper

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