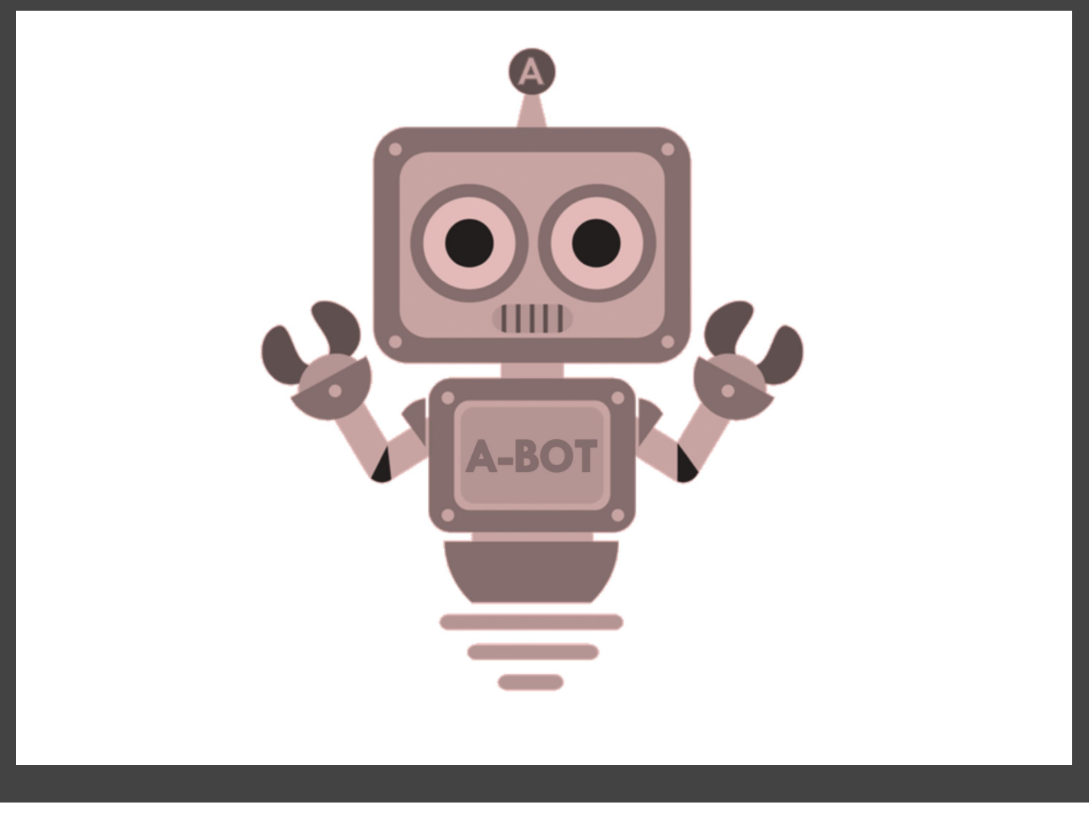
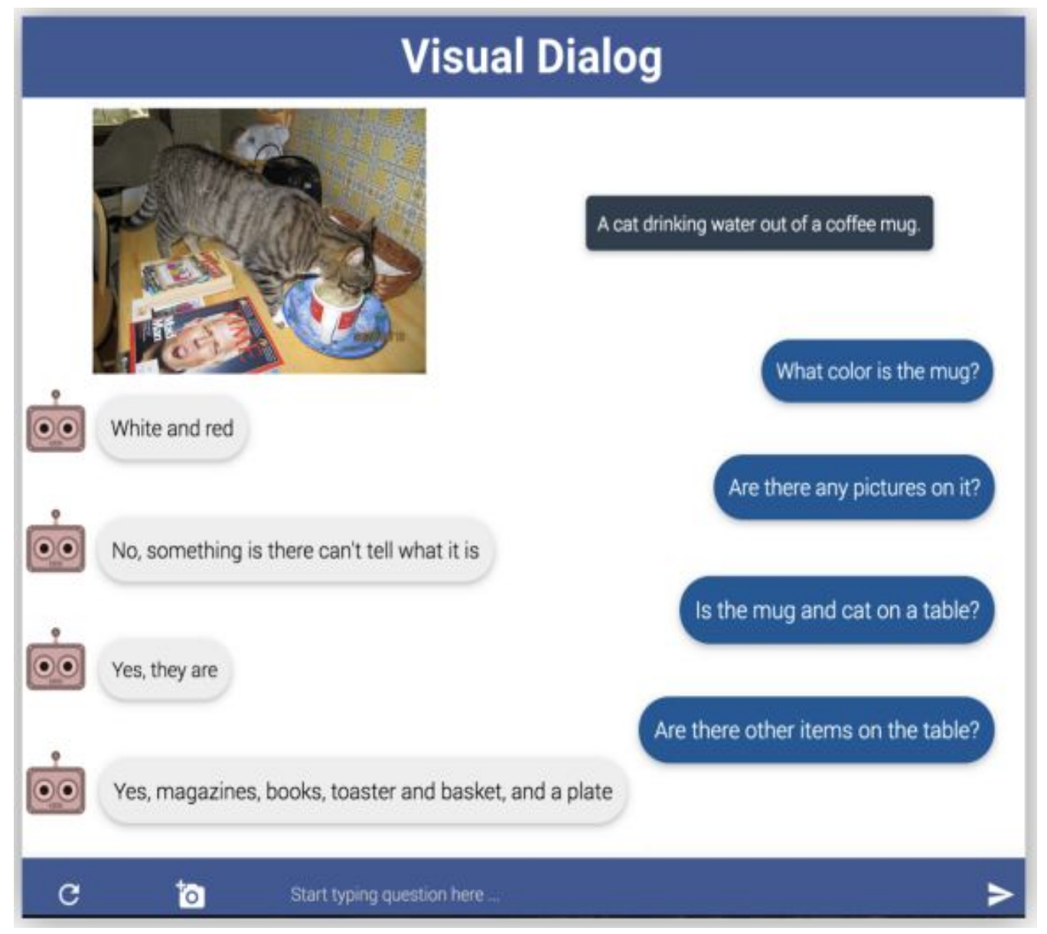


Visual Conversational Agent

Vincent Kang | David Zeng | Serena Wang
11-777 Multimodal Machine learning

Motivation

- AI systems should be able to handle all kinds of complex human-machine interactions
- Need to understand visual elements in conversations



Visual Dialog

A cat drinking water out of a coffee mug.

What color is the mug?

White and red

Are there any pictures on it?

No, something is there can't tell what it is

Is the mug and cat on a table?

Yes, they are

Are there other items on the table?

Yes, magazines, books, toaster and basket, and a plate

Dataset

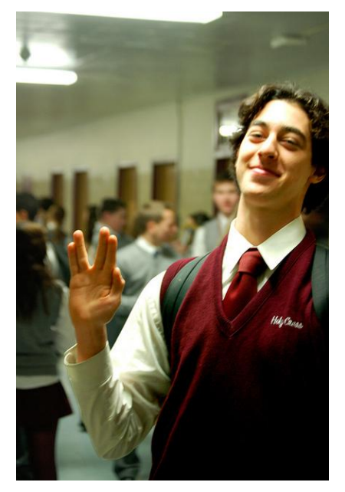
- When given an image, a history of question-answer pairs, a follow-up question, and 100 candidate answers, our model should return a ranking of the candidate answers
- Dataset contains 130,000 images and 10 rounds of dialog per image
- More yes/no questions in the VisDial dataset than in other VQA datasets, but fewer of those questions in VisDial have "yes" answers and some have ambiguous answers like "I don't know" or "I can't tell"

Challenges

- Typical VQA challenges

Q: How many toothbrushes showing?
Model: 3
GT: 2 (ranked 2)

Q: How many cell phones are there?
Model: 6
GT: At least 12 (ranked 15)



Q: How old is this man?
Model: 30s
GT: He looks like late teens. (ranked 54)

- Need to attend over previous rounds of dialog to get context from history

Q: Is there only 1 bird?
A: yes
Q: What color is it?
Model: white
GT: looks green and yellow

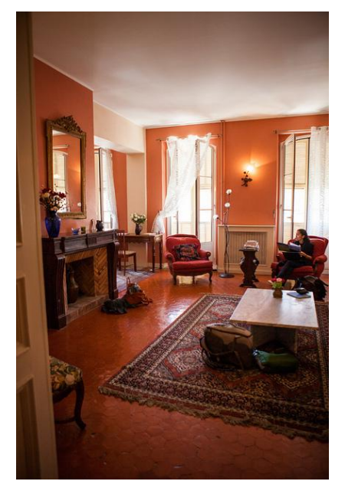
Q: Is there a lot of plants?
A: I only see 2
Q: What color are they?
Model: 1 is light red , 2 are dark red , 1 is white and the 1 is brown brick
GT: green

Q: Is it a wii remote he is holding?
A: Yes
Q: Is the other person holding a remote also?
Model: Yes
GT: No (ranked 2)

Proposed Models

Similarity Metrics

- There may be multiple correct answers within the dataset that are semantically similar
- Incorporate similarity as an evaluation metric and also into the loss function so model is penalized less for choosing answers that are similar in meaning to the ground truth answer.



Q: How old does the person look?
Model: I can’t tell
GT: She’s far away, I can’t tell. (ranked 59)
Cosine Sim: 0.92

- We can formalize the loss as a cross-entropy loss with soft labels:

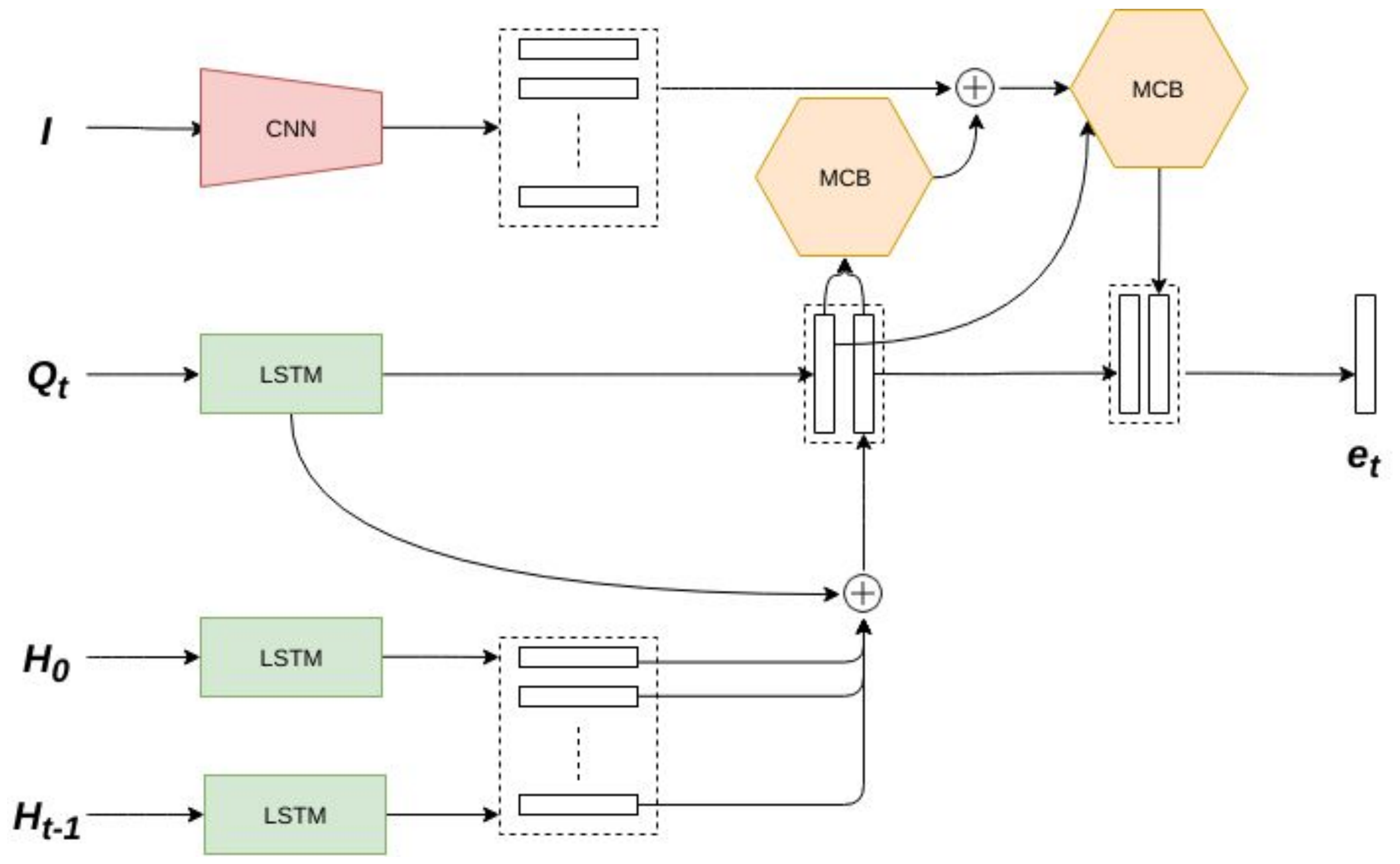
$$\mathcal{L}_\theta = -\sum_i p(A_t^i) \log Q(A_t^i | Q_t, I, H; \theta)$$

Where the soft labels p are given by

$$p(A_t^i) = \frac{e^{\lambda \cdot \mathrm{sim}(A_t^{gt}, A_t^i)}}{\sum_j e^{\lambda \cdot \mathrm{sim}(A_t^{gt}, A_t^j)}}$$

Early Fusion

- Early fusion models can significantly improve performance for VQA
- Existing models often concatenate unimodal representations or use late fusion encoders
- Replace the concatenations of unimodal representations with VQA early fusion techniques



Results and Methodology

Similarity Metrics

- We introduce the following metrics for evaluation: word mover’s distance (WMD), cosine similarity, BLEU-4, ROUGE_L
- For training we KL divergence as a loss with a cosine similarity metric based on word vectors trained from gLoVe’s word2vec model.

Early Fusion

- Added multimodal compact bilinear pooling (MCB) in two different ways
  1. Merge question and image features using MCB and use this to attend to history
  2. Merge question and history using MCB and use this to attend to image. Merge question with attended image with MCB and then combine with history.
- MCB did not seem to improve performance
  - MCB may add too much dimensionality to the input representation
  - May require longer training time or different hyperparameters than baseline models

Results

	MRR	WMD	Cosine Sim	BLEU	ROUGE
Late Fusion	0.5981	2.88	0.78	0.47	0.45
Late Fusion with similarity	0.5813	2.96	0.77	0.45	0.48
HCIAE-D-NP	0.6187	2.73	0.79	0.51	0.52
HCIAE-D-NP with MCB	0.6149	2.75	0.79	0.51	0.53