Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

Maria Lubeznov
Weining Hu
Popularity on Virtual Assistant
Motivation – Trend in Combining Vision and Language

1. Aiding visually impaired users in understanding their surroundings or social media content
2. Allowing medical personnel to better interpret medical scans
3. Helping AR/VR applications where a user could chat in natural language and work with a virtual companion who is seeing what they are seeing based on a visual common ground
Motivation – Trend in Combining Vision and Language

1. Aiding visually impaired users in understanding their surroundings or social media content
2. Allowing medical personnel to better interpret medical scans
3. Helping AR/VR applications where a user could chat in natural language and work with a virtual companion who is seeing what they are seeing based on a visual common ground
Image captioning vs VQA vs Visual dialog

Inputs
- Dialog history
- Image + caption
- Question

Output
- Answer

VQA
Q: How many people on wheelchairs?
A: Two

Q: How many wheelchairs?
A: One

Captioning
Two people are in a wheelchair and one is holding a racket.

Visual Dialog
Q: What is the gender of the one in the white shirt?
A: She is a woman

Q: What is she doing?
A: Playing a Wii game

Q: Is that a man to her right?
A: No, it's a woman
Motivation2 - They are not perfect

Vision+language recipe -> DATASETS (Collected from Amazon mechanical turk)

Treat dialog as static supervised learning -> Training (Ground truth answer provided)

These lead to two bad consequence!
Visual Dialog: Problem #1

Q: How many people on wheelchairs?
A: Two
Q: What gender are the people in the wheelchairs?
A: One is female, one is male
Q: Which one is holding the racket?
A: The female

Model can’t steer conversation and doesn’t get to see the future consequences of its utterances during training.
Visual Dialog: Problem #2

Q: How’s the weather?

Ground truth: Sunny

- Clear
- Can’t see the sky
- Looks warm
- It’s not raining
- I can’t see the sky, but I see shadows, so probably sunny

... 

Evaluation infeasible for utterances outside the dataset
Problem formulation

A-BOT
- Picks an image from dataset
- Answers question about the image

Q-BOT
- Asks questions
- Update its mental model of unseen image
- Makes prediction for image feature vector

RL reward for both agents
- how close the prediction to true image
Visual dialog dataset - VisDial

**VisDial**

**Training set** (235M)
82,783 images

**Validation set** (108M)
40,504 images

'questions': [
    'does it have a doorknob',
    'do you see a fence around the bear',
    ...
],

'answers': [
    'no, there is just green field in foreground',
    'countryside house',
    ...
],

'dialogs': [
    {
      'image_id': <COCO image id>,
      'caption': <image caption from COCO>,
      'dialog': [
        {
          'question': <index of question in `data.questions` list>,
          'answer': <index of answer in `data.answers` list>,
          'answer_options': <100 candidate answer indices from data.answers>,
          'gt_index': <index of `answer` in `answer_options`>
        },
        ...
        (10 rounds of dialog)
Reinforcement Learning for Dialog Agents

Diagram:

- State $S_t$
- Reward $R_t$
- Action $A_t$
- Environment

Flow:

1. Agent
2. Environment
3. State transition $S_{t+1}$
4. Reward update $R_{t+1}$
Reinforcement Learning for Dialog Agents

**Agent:** Cooperative A-Bot and Q-bot

**State:**

\[ s^A_t = [I, c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}, q_t] \quad s^Q_t = [c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}] \]

**Action:** question and answer pair

\[ (q_t, a_t) \]

**Reward:**

\[ r_t \left( s^Q_t, (q_t, a_t, y_t) \right) = \ell \left( \hat{y}_{t-1}, y^{gt}_{t-1} \right) - \ell \left( \hat{y}_t, y^{gt}_t \right) \]
Policy networks for Q-Bot and A-Bot
Policy network - Q-BOT

Fact encoder - LSTM $(q_t, a_t) \rightarrow \tilde{F}_t^Q \in \mathbb{R}^{512}$

State encoder - LSTM $(F_1^Q, \ldots, F_t^Q) \rightarrow S_t^Q$

Question decoder - LSTM which uses $S_{t-1}^Q$ as initial hidden state and generates $q_t$ by sequentially sampling words.

Feature regression network - 1 FC layer $\hat{y}_t = f(S_t^Q)$

Parameters - $\theta_Q$ (LSTMs) and $\theta_f$ (FC layer)
Policy network - A-BOT

**Question encoder** - LSTM $q_t \rightarrow Q_t^A \in \mathbb{R}^{512}$

**Fact encoder** - same as in Qbot.

**State encoder** - LSTM $(y, Q_t^A, F_t^A), \ldots, (y, Q_t^A, F_{t-1}^A) \rightarrow S_t^A$

**Answer decoder** - LSTM which uses $S_t^A$ as initial hidden state and generates $a_t$ by sequentially sampling words

**Parameters** - (LSTMs) $\theta_A$
Training

1. Supervised pre training on VisDial
   a. Faster RL convergence
   b. To prevent the bots from inventing their own uninterpretable language

2. Fine tuning with RL - parameters are updated in response to experienced rewards.
Supervised training

BOTs are trained separately.

Q-BOT
● Is trained to generate the follow-up question by the questionnaire, given the caption and the QA history.

A-BOT
● Is trained to generate the response by the answerer, given the image, caption and the QA history.

CNN for image feature extraction - pre trained on ImageNet
Reinforcement fine tuning

Objective: maximize the expected reward, under agents’ policies, for each dialog round (time $t$).

$$\max (J(\theta_A, \theta_Q, \theta_g)) \text{ where } J(\theta_A, \theta_Q, \theta_g) = \mathbb{E}_{\pi_Q, \pi_A} \left[ r_t(s^Q_t, (q_t, a_t, y_t)) \right]$$

REINFORCE algorithm is applied to calculate the $\nabla_{\theta_Q} J$ and $\nabla_{\theta_A} J$.

Intuition:
- Informative QA round $\rightarrow$ positive reward $\rightarrow$ positive gradient (probability)
- Poor exchange $\rightarrow$ negative reward $\rightarrow$ negative gradient (probability)

$\theta_f$ is updated in a “supervised” way.
Synthetic example

64 images.

\[ V_A = \{1, 2, 3, 4\} \quad V_Q = \{X, Y, Z\} = \{\text{shape, color, style}\} \]

RL iteration = 10K episodes
### Results - supervised vs RL dialogues

#### Image + Caption

**Human-Human dialog [4]**

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>What are their genders?</td>
<td>1 man 1 woman</td>
</tr>
<tr>
<td>Are they both adults?</td>
<td>Yes</td>
</tr>
<tr>
<td>Do they wear goggles?</td>
<td>Looks like sunglasses</td>
</tr>
<tr>
<td>Do they have hats on?</td>
<td>Man does</td>
</tr>
<tr>
<td>Are there any other people?</td>
<td>No</td>
</tr>
<tr>
<td>What color is man’s hat?</td>
<td>Black</td>
</tr>
<tr>
<td>Is it snowing now?</td>
<td>No</td>
</tr>
<tr>
<td>What is woman wearing?</td>
<td>Blue jacket and black pants</td>
</tr>
<tr>
<td>Are they smiling?</td>
<td>Yes</td>
</tr>
<tr>
<td>Do you see trees?</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**K-Pretrained Q-BOT-A-BOT dialog**

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many people?</td>
<td>2</td>
</tr>
<tr>
<td>Male or female?</td>
<td>I can’t tell</td>
</tr>
<tr>
<td>What are they wearing?</td>
<td>I can’t tell, they are far away</td>
</tr>
<tr>
<td>What color are skis?</td>
<td>No</td>
</tr>
<tr>
<td>Are they wearing goggles?</td>
<td>Yes</td>
</tr>
<tr>
<td>Are they wearing goggles?</td>
<td>Yes</td>
</tr>
<tr>
<td>Are they wearing hats?</td>
<td>Yes</td>
</tr>
<tr>
<td>Are they wearing goggles?</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**RL-full-QAf Q-BOT-A-BOT dialog**

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are people male or female or both?</td>
<td>I think I is male, can’t see other 1 ...</td>
</tr>
<tr>
<td>What are they wearing?</td>
<td>They are all wearing snow pants and jackets</td>
</tr>
<tr>
<td>Is it snowing?</td>
<td>It does not appear to be snowing at moment</td>
</tr>
<tr>
<td>Can you see any trees?</td>
<td>Yes, there are lot of trees in background</td>
</tr>
<tr>
<td>Can you see any buildings in background?</td>
<td>No, I can not</td>
</tr>
<tr>
<td>Can you see any trees?</td>
<td>Yes, there are lot of trees in background</td>
</tr>
<tr>
<td>Can you see any buildings in background?</td>
<td>Yes, I do not see any buildings</td>
</tr>
<tr>
<td>Does it look like they are in park?</td>
<td>It does not appear to be</td>
</tr>
</tbody>
</table>

#### Image + Caption

**Human-Human dialog [4]**

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are there any people in picture?</td>
<td>No, there are no people</td>
</tr>
<tr>
<td>Is cheese white?</td>
<td>Yes, cheese is white</td>
</tr>
<tr>
<td>Are crackers white?</td>
<td>No, cracker is brownish</td>
</tr>
<tr>
<td>Are there lots of greens?</td>
<td>No, only 1 sprig</td>
</tr>
<tr>
<td>Was picture taken in kitchen?</td>
<td>I cannot tell, it is too zoomed in</td>
</tr>
<tr>
<td>How many crackers are there?</td>
<td>There is 1</td>
</tr>
<tr>
<td>Is there design on plate?</td>
<td>No, plate is plane white</td>
</tr>
<tr>
<td>Is there good lighting?</td>
<td>Yes, lighting is decent</td>
</tr>
<tr>
<td>Is plate on table?</td>
<td>I cannot tell I can only see plate as background</td>
</tr>
<tr>
<td>Does cracker look delicious?</td>
<td>Absolutely</td>
</tr>
</tbody>
</table>

**K-Pretrained Q-BOT-A-BOT dialog**

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are there any people?</td>
<td>No</td>
</tr>
<tr>
<td>What color is plate?</td>
<td>White</td>
</tr>
<tr>
<td>Is plate on table?</td>
<td>I can’t tell</td>
</tr>
<tr>
<td>What color is plate?</td>
<td>White</td>
</tr>
<tr>
<td>Is plate on table?</td>
<td>I can’t tell</td>
</tr>
<tr>
<td>Is there anything else on table?</td>
<td>Is plate large or small?</td>
</tr>
<tr>
<td>Is there anything else on table?</td>
<td>Is it on table?</td>
</tr>
<tr>
<td>Is there anything else on table?</td>
<td>What kind of donut?</td>
</tr>
<tr>
<td>Is there anything else on table?</td>
<td>What kind of chips?</td>
</tr>
<tr>
<td>Is there anything else on table?</td>
<td>Any people?</td>
</tr>
<tr>
<td>Is there anything else on table?</td>
<td>What is this inside or outside?</td>
</tr>
<tr>
<td>Is there anything else on table?</td>
<td>What color is table?</td>
</tr>
<tr>
<td>Is there anything else on table?</td>
<td>Any choice?</td>
</tr>
<tr>
<td>Is there anything else on table?</td>
<td>Is it large or small?</td>
</tr>
</tbody>
</table>

**RL-full-QAf Q-BOT-A-BOT dialog**

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is plate large or small?</td>
<td>Large</td>
</tr>
<tr>
<td>Is it on table?</td>
<td>I can’t tell</td>
</tr>
<tr>
<td>What kind of donut?</td>
<td>Plain with sugar on it</td>
</tr>
<tr>
<td>What kind of chips?</td>
<td>Potato chips</td>
</tr>
<tr>
<td>Any people?</td>
<td>No</td>
</tr>
<tr>
<td>Is this inside or outside?</td>
<td>I can’t tell, it’s close up</td>
</tr>
<tr>
<td>What color is table?</td>
<td>White</td>
</tr>
<tr>
<td>Any choice?</td>
<td>No</td>
</tr>
<tr>
<td>What color is table?</td>
<td>I can’t tell</td>
</tr>
<tr>
<td>Is it large or small?</td>
<td>I can’t see plate</td>
</tr>
</tbody>
</table>
Evaluation

Dataset: test split for VisDial v0.5 (~9.5k images)
Input: Image + caption
Rounds: 10
Output: Sort the entire test set in ascending distance to this prediction

(a) Guessing Game Evaluation.

(b) Visual Dialog Answerer Evaluation.

<table>
<thead>
<tr>
<th>Model</th>
<th>MRR</th>
<th>R@5</th>
<th>R@10</th>
<th>Mean Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL-pretrained</td>
<td>0.436</td>
<td>53.41</td>
<td>60.09</td>
<td>21.83</td>
</tr>
<tr>
<td>Frozen-Q</td>
<td>0.428</td>
<td>53.12</td>
<td>60.19</td>
<td>21.52</td>
</tr>
<tr>
<td>Frozen-f</td>
<td>0.432</td>
<td>53.28</td>
<td>60.11</td>
<td>21.54</td>
</tr>
<tr>
<td>RL-full-QAf</td>
<td>0.428</td>
<td>53.08</td>
<td>60.22</td>
<td>21.54</td>
</tr>
<tr>
<td>Frozen-Q-multi</td>
<td><strong>0.437</strong></td>
<td><strong>53.67</strong></td>
<td><strong>60.48</strong></td>
<td><strong>21.13</strong></td>
</tr>
</tbody>
</table>
Evaluation

$\ell_2$ distance to ground truth image in fc7 space

Round 1: What kind of pizza is it? Cheese and maybe mushroom.

Round 2: Are they male or female? 1 is male, 1 is female.

Round 4: Are they indoors or outdoors? Outdoors.

Round 5: Is there anything else on plate? Yes, there are 2 other plates in background.

Round 6: Is there anything else on table? No.

$c$ Qualitative Retrieval Results.
Weakness of the evaluation

1. No evaluation on the response time
2. No evaluation on how the initial caption quality affect the later dialog
Strengths and weaknesses of the approach

Strengths
● Self talk → unlimited data. No need to collect very large datasets
● Evaluation is image guessing → no need to evaluate natural language generation.
● Method is agent driven - learning to deal with actions consequences

Weaknesses
● No explanation of why the agent becomes less sure after 2 dialog rounds.
● Authors mention that A-BOT responses are not enough “human like”.
Potential extensions

Our thoughts
● Address repetitive questions
● It may be a good idea to combine this method with some sort of CGAN from paper presented last week, to better emulate human dialogs.
● Evaluate the impact of different image captioning quality on final reward achievement

Follow up papers
● Evaluate if and how this method can be applied on human-machine interaction ([1]).
● Investigate how similar method can be applied on other tasks that require negotiation ([2])
References

[1] Evaluating Visual Conversational Agents via Cooperative Human-AI Games

[2] Deal or No Deal? End-to-End Learning for Negotiation Dialogues

[3] Demo

Appendix - technical details

- Supervised training for the first K rounds of dialog and transition to RL for the remaining $10 - K$ rounds
- In each epoch, K gradually annealed to 0.
- 15 epochs
- Adam optimizer with learning rate 0.001
- Gradients are clamped to [-5,5] to avoid explosion.