Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

Presented by:
Ali Mohammad Mehr | Amir Refaee | Ignacio Iturralde | Matt Dietrich
Outline

1) Motivation
2) Problem Statement + Contributions
3) Related Work
4) Methods and Models
5) Experiments
6) Discussion and Future Work
Motivation

1) What is visual dialog?
2) Why is it important?
Motivation

1) What is visual dialog?
2) Why is it important?

Source: https://visualdialog.org/
Motivation

1) What is visual dialog?
2) Why is it important?

Source: https://visualdialog.org/
Motivation

1) What is visual dialog?
2) Why is it important?

Source: https://visualdialog.org/
Motivation

1) What is visual dialog?
2) Why is it important?

Source: https://visualdialog.org/
Motivation

1) What is visual dialog?
2) Why is it important?

Applications
- Assist visually impaired users
- Analyze surveillance data
- Interact naturally with AI assistants (incl. robots)

Source: https://visualdialog.org/
Problem Statement

Cooperative image guessing game

Two zebras are walking around their pen at the zoo.

Q1: Any people in the shot?
A1: No, there aren’t any.

Q10: Are they facing each other?
A10: They aren’t.

I think we were talking about this image!
Problem Statement

Cooperative image guessing game

Questioner
- Sees only a caption, image pool
- Asks questions, guesses image
Problem Statement

Cooperative image guessing game

Questioner
- Sees only a caption, image pool
- Asks questions, guesses image

Answerer
- Sees the image
- Answers questions

I think we were talking about this image!
Problem Statement

Cooperative image guessing game

**Questioner**
- Sees only a caption, image pool
- Asks questions, guesses image

**Answerer**
- Sees the image
- Answers questions

**Reward** based on error/distance metric of prediction to ground truth
Problem Statement

Cooperative image guessing game

Questioner
- Sees only a caption, image pool
- Asks questions, guesses image

Answerer
- Sees the image
- Answers questions

Reward based on error/distance metric of prediction to ground truth

Reinforcement Learning!
Problem Statement

Challenges:

Two zebras are walking around their pen at the zoo.

Q1: Any people in the shot?
A1: No, there aren't any.

Q10: Are they facing each other?
A10: They aren't.

I think we were talking about this image!
Problem Statement

Challenges:
- Q-BOT: Interpret language, identify possible images, ask discerning questions
Problem Statement

Challenges:
- Q-BOT: Interpret language, identify possible images, ask discerning questions
- A-BOT: Model of understanding, answer with precision and concision

Two zebra are walking around their pen at the zoo.

Q1: Any people in the shot?  
A1: No, there aren’t any.

Q10: Are they facing each other?  
A10: They aren’t.

I think we were talking about this image!
Problem Statement

Challenges:
- Q-BOT: Interpret language, identify possible images, ask discerning questions
- A-BOT: Model of understanding, answer with precision and concision

Importance of Language:
- Interpretability
- Prevent cheating
Contributions

First instance of \textit{goal-driven training} for visual question answering and dialog agents
Contributions

First instance of **goal-driven training** for visual question answering and dialog agents

Experimental results:
1) Automatic emergence of grounded language + communication protocol
2) RL fine-tuned bots > supervised bots
Related Work

Vision and Language
- Visual Dialog [Das et al., 2017]
- GuessWhat?! Visual object discovery through multi-modal dialogue [de Vries et al., 2017]
Related Work

Vision and Language
- Visual Dialog [Das et al., 2017]
- GuessWhat?! Visual object discovery through multi-modal dialogue [de Vries et al., 2017]

Demo: http://demo.visualdialog.org/
Related Work

Vision and Language
- Visual Dialog [Das et al., 2017]
- GuessWhat?! Visual object discovery through multi-modal dialogue [de Vries et al., 2017]

Demo: http://demo.visualdialog.org/

**Questioner**
- Is it a vase? Yes
- Is it partially visible? No
- Is it in the left corner? No
- Is it the turquoise and purple one? Yes
Related Work

Vision and Language
- Visual Dialog [Das et al., 2017]
- GuessWhat?! Visual object discovery through multi-modal dialogue [de Vries et al., 2017]
Related Work

Vision and Language
- Visual Dialog [Das et al., 2017]
- GuessWhat?! Visual object discovery through multi-modal dialogue [de Vries et al., 2017]

20 Questions and Lewis Signaling Game
- Convention: A philosophical study [Lewis, 2008]

Supervised learning
Related Work

Vision and Language
- Visual Dialog [Das et al., 2017]
- GuessWhat?! Visual object discovery through multi-modal dialogue [de Vries et al., 2017]

20 Questions and Lewis Signaling Game

\[ s_1: \quad t_i \rightarrow m_1 \quad t_s \rightarrow m_2 \]
\[ s_2: \quad t_i \rightarrow m_1 \quad t_s \rightarrow m_2 \]
\[ s_3: \quad t_i \rightarrow m_1 \quad t_s \rightarrow m_2 \]
\[ s_4: \quad t_i \rightarrow m_1 \quad t_s \rightarrow m_2 \]

\[ r_1: \quad m_1 \rightarrow a_l \quad m_2 \rightarrow a_s \]
\[ r_2: \quad m_1 \rightarrow a_l \quad m_2 \rightarrow a_s \]
\[ r_3: \quad m_1 \rightarrow a_l \quad m_2 \rightarrow a_s \]
\[ r_4: \quad m_1 \rightarrow a_l \quad m_2 \rightarrow a_s \]

Figure 1.1: Lewis’s original example: the sexton’s and Revere’s admissible contingency plans.
Related Work

Vision and Language
- Visual Dialog [Das et al., 2017]
- GuessWhat?! Visual object discovery through multi-modal dialogue [de Vries et al., 2017]

20 Questions and Lewis Signaling Game
- Convention: A philosophical study [Lewis, 2008]

Supervised learning

Passive receiver, one-shot signaling
Related Work

Vision and Language
- Visual Dialog [Das et al., 2017]
- GuessWhat?! Visual object discovery through multi-modal dialogue [de Vries et al., 2017]

20 Questions and Lewis Signaling Game
- Convention: A philosophical study [Lewis, 2008]

Text-only or Classical Dialog
- Deep Reinforcement Learning for Dialogue Generation [Li et al., 2016]
- Adversarial Learning for Neural Dialogue Generation [Li et al., 2017]

Supervised learning
Passive receiver, one-shot signaling
Related Work

Vision and Language
- Visual Dialog [Das et al., 2017]
- GuessWhat?! Visual object discovery through multi-modal dialogue [de Vries et al., 2017]

20 Questions and Lewis Signaling Game
- Convention: A philosophical study [Lewis, 2008]

Text-only or Classical Dialog
- Deep Reinforcement Learning for Dialogue Generation [Li et al., 2016]
- Adversarial Learning for Neural Dialogue Generation [Li et al., 2017]
Related Work

Vision and Language
- Visual Dialog [Das et al., 2017]
- GuessWhat?! Visual object discovery through multi-modal dialogue [de Vries et al., 2017]

20 Questions and Lewis Signaling Game
- Convention: A philosophical study [Lewis, 2008]

Text-only or Classical Dialog
- Deep Reinforcement Learning for Dialogue Generation [Li et al., 2016]
- Adversarial Learning for Neural Dialogue Generation [Li et al., 2017]

Emergence of Language
- Learning to Communicate with Deep Multi-Agent Reinforcement Learning [Foerster et al., 2016]
- Emergence of Language with Multi-agent Games [Havrylov and Titov, 2017]
- Multi-Agent Cooperation and the Emergence of (Natural) Language [Lazaridou et al., 2017]
- Emergence of Grounded Compositional Language in Multi-Agent Populations [Mordatch and Abbeel, 2018]

Supervised learning
Passive receiver, one-shot signaling
Prescribed vs. adversarial learning
Cooperative Image Guessing Game - Agents

A questioner bot (Q-bot)
Primed with a 1-sentence description i.e. “Two zebras are walking around their pen at the zoo”
Does not see the image

An answerer bot (A-bot)
Sees the image
Sees the caption
Cooperative Image Guessing Game - Turn and Episode

Any people in the shot?
No, there aren’t any.
Are they facing each other?
They aren’t.
Cooperative Image Guessing Game - Objective

\( \hat{y} \) - vector embedding of the image
\( \hat{y}_{gt} \) - VGG-16 features
\( L(\hat{y}, \hat{y}_{gt}) \) – Euclidean distance
State-Action Space

Action
Discrete token vocabulary $V$ common between both agents, i.e. English tokens

State
Each agent has a different state due to information asymmetry
Q-Bot: state at round $t$ is the caption and dialog history so far

$$s_t^Q = [c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}]$$

A-Bot: state at round $t$ includes the image as well

$$s_t^A = [I, c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}, q_t]$$
Policy

Stochastic policies $\pi_Q(q_t|s_t^Q; \theta_Q)$ and $\pi_A(a_t|s_t^A; \theta_A)$ learned by two separate deep neural networks parametrized by $\theta_Q$ and $\theta_A$

Feature Regression network for Q-bot:

$$\hat{y}_t = f(s_t^Q, q_t, a_t; \theta_f) = f(s_{t+1}^Q; \theta_f)$$

Goal is to learn $\theta_Q$, $\theta_A$, and $\theta_f$
Environment and Reward

Image as the environment

Common reward for both agents:

\[
 r_t \left( s_t^Q, (q_t, a_t, y_t) \right) = \ell \left( \hat{y}_{t-1}, y^{gt}_t \right) - \ell \left( \hat{y}_t, y^{gt}_t \right)
\]

Total Reward:

\[
 \sum_{t=1}^{T} r_t \left( s_t^Q, (q_t, a_t, y_t) \right) = \ell \left( \hat{y}_0, y^{gt}_T \right) - \ell \left( \hat{y}_T, y^{gt}_T \right)
\]

overall improvement due to dialog
Policy Networks
**Q-Bot**

**Fact Encoder:** LSTM  
Final hidden state $F_t^Q \in \mathbb{R}^{512}$  
$(q_t, a_t) \rightarrow F_t^Q$

**State/History Encoder:** LSTM  
$(F_1^Q, \ldots, F_t^Q) \rightarrow S_t^Q$

**Question Decoder:** LSTM  
$S_{t-1}^Q \rightarrow q_t$

**Feature Regression Network**  
Fully connected layer  
$\hat{y} = f(S_t^Q)$

$\theta_f$: combined LSTM parameter

[Diagram and text related to Q-Bot's operations and structure]
A-Bot

Question Encoder: LSTM
Final hidden state $Q_t^A \in \mathbb{R}^{512}$
$q_t \rightarrow Q_t^A$

Fact Encoder: LSTM
Final hidden state $F_t^A \in \mathbb{R}^{512}$
$(q_t, a_t) \rightarrow F_t^A$

State/History Encoder: LSTM
$((y, Q_1^A, F_0^A), ..., (y, Q_t^A, F_{t-1}^A)) \rightarrow S_t^A$

Answer Decoder: LSTM
$S_t^A \rightarrow a_t$

$\theta_A$: combined LSTM parameters
Joint Training with Policy Gradients

Based on REINFORCE algorithm:

- Update policy parameters \((\theta_Q, \theta_A, \theta_f)\)
  - in response to experienced rewards
- The objective is to maximize the expected reward summed over all episodes

\[
\max_{\theta_A, \theta_Q, \theta_g} J(\theta_A, \theta_Q, \theta_g)
\]

\[
J(\theta_A, \theta_Q, \theta_g) = \mathbb{E}_{\pi_Q, \pi_A} \left[ \sum_{t=1}^{T} r_t(s_t^Q, (q_t, a_t, y_t)) \right]
\]
Joint Training with Policy Gradients

Based on REINFORCE algorithm:

- Update policy parameters \( (\theta_Q, \theta_A, \theta_f) \)
  - in response to experienced rewards
- The objective is to maximize the expected reward summed over all episodes

\[
J(\theta_A, \theta_Q, \theta_g) = \mathbb{E}_{\pi_Q, \pi_A} \left[ \sum_{t=1}^{T} r_t(s_t^Q, (q_t, a_t, y_t)) \right]
\]

- This is considering the entire dialog as a single RL episode
  - Does not differentiate between individual good or bad exchanges

\[
J(\theta_A, \theta_Q, \theta_g) = \mathbb{E}_{\pi_Q, \pi_A} \left[ r_t(s_t^Q, (q_t, a_t, y_t)) \right]
\]
Joint Training with Policy Gradients

Based on REINFORCE algorithm:

- Update policy parameters \((\theta_Q, \theta_A, \theta_f)\)
  - in response to experienced rewards
- The objective is to maximize the expected reward

\[
J(\theta_A, \theta_Q, \theta_g) = \mathbb{E}_{\pi_Q, \pi_A} \left[ r_t(s_t^Q, (q_t, a_t, y_t)) \right]
\]

\[
\nabla_{\theta_Q} J = \nabla_{\theta_Q} \left[ \mathbb{E}_{\pi_Q, \pi_A} \left[ r_t(\cdot) \right] \right]
\]

\[
= \nabla_{\theta_Q} \left[ \sum_{q_t, a_t} \pi_Q(q_t | s_{t-1}^Q) \pi_A(a_t | s_t^A) r_t(\cdot) \right]
\]

\[
= \sum_{q_t, a_t} \pi_Q(q_t | s_{t-1}^Q) \nabla_{\theta_Q} \log \pi_Q(q_t | s_{t-1}^Q) \pi_A(a_t | s_t^A) r_t(\cdot)
\]
Joint Training with Policy Gradients

Based on REINFORCE algorithm:

- Update policy parameters \((\theta_Q, \theta_A, \theta_f)\)
  - in response to experienced rewards
- The objective is to maximize the expected reward

\[
\nabla_{\theta_Q} J = \nabla_{\theta_Q} \left[ \mathbb{E}_{\pi_Q, \pi_A} [r_t(\cdot)] \right] \\
= \nabla_{\theta_Q} \left[ \sum_{q_t, a_t} \pi_Q(q_t|s_{t-1}^Q) \pi_A(a_t|s_t^A) r_t(\cdot) \right] \\
= \sum_{q_t, a_t} \pi_Q(q_t|s_{t-1}^Q) \nabla_{\theta_Q} \log \pi_Q(q_t|s_{t-1}^Q) \pi_A(a_t|s_t^A) r_t(\cdot) \\
= \mathbb{E}_{\pi_Q, \pi_A} \left[ r_t(\cdot) \nabla_{\theta_Q} \log \pi_Q(q_t|s_{t-1}^Q) \right]
\]
Joint Training with Policy Gradients

Based on REINFORCE algorithm:

- Update policy parameters \((\theta_Q, \theta_A, \theta_f)\)
  - in response to experienced rewards

\[
\nabla_{\theta_Q} J = \mathbb{E}_{\pi_Q, \pi_A} \left[ r_t (\cdot) \ \nabla_{\theta_Q} \log \pi_Q \left( q_t | s_{t-1} \right) \right]
\]

- Estimate the expectation with sample averages
  - Sample a question from Q-BOT
  - Sample its answer from A-BOT
  - Compute the scalar reward for this round
  - Multiply that scalar reward to gradient of log-probability of this exchange
  - Propagate backward to compute gradients w.r.t. all parameters \(\theta_Q, \theta_A\).
Emergence of Grounded Dialog

Challenges to succeed in the image guessing:
- Learning a common language
  - Understand the difference between words for color and words for poses.
- develop mappings between symbols and image representations
  - How it looks likes when someone is standing up in a picture.
- A-BOT needs to ground language in visual perception to answer questions
- Q-BOT must learn to predict plausible image representations
Emergence of Grounded Dialog

Challenges to succeed in the image guessing:
- Learning a common language
  - Understand the difference between words for color and words for poses.
- develop mappings between symbols and image representations
  - How it looks likes when someone is standing up in a picture.
- A-BOT needs to ground language in visual perception to answer questions
- Q-BOT must learn to predict plausible image representations

These challenges need to be handled in an end-to-end manner
- From a distant reward function

A sanity check is needed to see if it is really possible!
Emergence of Grounded Dialog

A simple setup:
- Images with 4 shapes, 4 colors, 4 styles
  - For a total of 64 unique images
- A-BOT has perfect perception
- Q-BOT is to deduce two attributes of image
  - In a particular order

Vocabulary:
- Vocabulary size is crucial
  - For a non-trivial ‘non-cheating’ behavior
- If for the A-BOT vocabulary $V_A, |V_A| \geq 64$
  - A-BOT conveys the entire image in
    - a single token
    - E.g. 1 = (red, square, filled)
- $V_A = \{1, 2, 3, 4\}$
- $V_Q = \{X, Y, Z\}$
Emergence of Grounded Dialog

Policy Learning:
- The state-action space is discrete and small
- Both bots are fully specified tables of Q-values
  - $Q: [\text{state}, \text{action}] \rightarrow \text{future reward estimate}$
- Learn the policies by Q-learning with Monte Carlo estimation over 10k episodes
  - Updates are done alternately where one bot is frozen while the other is updated
- Ensure enough exploration
  - by randomly choosing actions not aligned with the learned policy

Results:
- The two invent their own communication protocol
  - Q-BOT
    - X -> color, Y -> shape, Z -> style
  - A-BOT
    - 1 -> purple, 2 -> green, 3 -> blue, 4 -> red
    - 1 -> triangle, 2 -> square, 3 -> circle, 4 -> star
Experiments

‘Sanity Check’ Experiment

Model Experiments on VisDial*
- Supervised Learning pretrained model (no RL)
- Frozen-Q or -A: Fix Q- or A-bot to SL-pretrained train active agent (and regression network) with RL
- Freeze regression network and train both agents with RL
- Agents and Regression trained with RL (after SL-pretrain)

*VisDial is dataset: 680k QA-pairs (10 QA-pairs for each of 68k COCO images)
Experiment Evaluation

Guessing Game
- Image retrieval experiment based on test split of VisDial
- Agents presented with image + automatically generated caption
- Look at distance between Q-Bot representations and all images in test set

Emulating Human Dialogs
- Log-likelihood of A-Bot answer v. 100 candidate responses of VisDial

Human Study
- Human interpretability shows that interpretability of bots’ dialogs and image-discriminative language are both successful and best with the RL-full-QAf model

<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-pretrain</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>0.437</td>
<td>53.67</td>
<td>60.48</td>
<td>21.13</td>
</tr>
</tbody>
</table>

(a) Guessing Game Evaluation.
(b) Visual Dialog Answerer Evaluation.
Discussion and Future Work

Strengths:
- Use of RL makes less labeling necessary
- Simplicity of model’s parts to build a complex network

Weaknesses:
- Network forgetfulness e.g. asking the same question over and over again
- Network inconsistency e.g. different answers for same/similar questions
- Use of vector evaluation with Euclidean distance seems simplistic (?)
- Could try to incorporate attention for both the image and question/answer
Thank You!