Dialog Systems and Visual Dialog

Sayyed Nezhadi
CSC2539
Feb 2017
What is a Dialog System?

- A **dialog system** is a machine (computer system) with the goal of conversing with human with a logical structure.
- The communication with machine can be done through **text, speech, gesture** and so on.
- A **Natural Dialog System** is a form of dialog system that tries to improve usability and user satisfaction by imitating human behaviour. (Berg, 2014)
- **Turing test**: a machine's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human.
Types of Dialog System

- **Goal-oriented agents**: it needs to understand the user input and complete a related task with a clear goal within a limited number of dialog turns.
  - **Finite-State**: Restaurant reservation, airline booking, ...
  - **Active Ontalogy/Frame Based**: Personal assistant, Siri, Alexa, Google Now
- **Chatbots**: general conversation with a wide scope
  - Chit-chatting
  - Entertainment
  - **Examples**: ELISA, ALICE, APRRY, ...
Finite-State Dialog

- A series of questions to be answered by user
- Full control of the conversation by the system
- Ignoring any unrelated answers
- Simple to build and good for simple tasks
- Only one information at a time
- Very practical but not a natural dialog

From: Dan Jurafsky slides
Active Ontology/Frame Based

- More natural conversation with mixed-initiative (Conversation initiative shifts between the user and the system)
- User can ask multiple questions or give multiple information in one sentence
- Using **Frame and Slots**: once all mandatory slots in a frame are filled, it will generate query to a knowledge base or external systems.
- Using **Natural Language Understanding** to extract slots from sentences (ML can be used).

Some texts from: Dan Jurafsky slides
Active Ontology/Frame Based - continued

- Voice Synthesis
- Voice Recognition
- Dialog Management
- Language Understanding
- Action Selection
- Session Context
- Knowledge Base

Clarifying Question
Text Input
Best Outcome
Missing Slots
Semantic Interpretation
Inferred User Input
Complete?
Yes
No

Voice Output
Voice Input
Voice Output

Based on a figure from Jerome Bellegarda
Example: Amazon Alexa

Amazon Echo

Amazon Alexa service

Custom Skill Service

External Systems

DB

Amazon Echo App Market

Register Skill

Sample Utterances

Skills

Intent Schema

Slots and Slot Types

Amazon Developer Portal
Example: Amazon Alexa

• **Skills** are voice enabled apps
• For every **Intent** we define as many as possible **sample utterances**
• Sample utterances can have **slots** in them
• Slots are categorized by **slot types**
• There are built-in intents to **start** or **stop** a skill or ask for help.

---

**Intent Schema:**
```
{   "intent": "airportInfoIntent",   "slots": [{      "name": "AIRPORTCODE",      "type": "FAACODES"    }]}
```

**Sample Utterances:**
- `airportInfoIntent {AIRPORTCODE}`
- `airportInfoIntent airport into {AIRPORTCODE}`
- `airportInfoIntent flight delay {AIRPORTCODE}`
- `airportInfoIntent info {AIRPORTCODE}`
- `airportInfoIntent flight status {AIRPORTCODE}`
- `airportInfoIntent airport {AIRPORTCODE}`
- `airportInfoIntent flight info {AIRPORTCODE}`
- ...

---

Slot Type “FAACODES” : AAC, AAF, AAH AAI, ...
General Chatbots

**Rule-based:**
- Based on pattern matching (AIML, ChatScript, Regex, ...)
- Using mental models
- Three different memory:
  - Utterance
  - Session
  - Global

**Corpus-based:**
- Large corpus data
- Deep neural networks
- Information retrieval (mine conversations and retrieve similar responses)

Sample Patterns (ELIZA)

```
<pattern>HELLO</pattern>
<random>
<li>How do you do. Please state your problem..</li>
<li>Hi. What seems to be your problem?.</li>
</random>

<pattern>YOU ARE *</pattern>
<random>
<li>What makes you think I am <star/>. </li>
<li>Does it please you to believe I am <star/>. </li>
<li>Do you sometimes wish you were <star/>. </li>
<li>Perhaps you would like to be <star/>.. </li>
</random>
```

Some texts from: Dan Jurafsky slides
A Neural Conversation Model (O. Vinyals, Q.V. Le 2015)

- Sequence to sequence (Seq2seq) mapping using recurrent neural networks (reads the input sequence one token at a time, and predicts the output sequence, also one token at a time)

- During training, the true output sequence is given to the model

- The model is trained to maximize the cross entropy of the correct sequence given its context

- During inference, the true output sequence is not observed, simply feed the predicted output token as input to predict the next output (“greedy” inference) or use beam search, and feed several candidates at the previous step to the next step. Selection based on the probability of the sequence
A Neural Conversation Model – sample output

- Trained on **IT Helpdesk Troubleshooting dataset** (Training: 30M tokens, Validation: 3M tokens)
- Trained a **single layer LSTM with 1024 memory cells** using stochastic gradient descent with gradient clipping. The vocabulary consists of the most common **20K words**, which includes **special tokens** indicating turn taking and actor.

**Describe your problem:** I forgot my password
**Machine:** hi, this is <NAME> from helpdesk connect, what can i help you with today?
**Human:** hi, i forgot my password
**Machine:** ok great
**Human:** can you tell me how i can recover it?
**Machine:** i’ll need to verify who it is at that machine
**Machine:** can we do a hangout
**Machine:** ?

**Human:** yes
**Machine:** <anonymized hangout instructions and verifications>
**Human:** cool, i am good now
**Machine:** how’s this looking good?
**Machine:** i’ll go ahead and close this chat.
**Machine:** take care. :)
**Human:** bye
**Machine:** bye
Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day

by James Vincent | @jvincent | Mar 24, 2016, 6:43 am EDT
End-To-End Dialog Systems Using Generative Hierarchical Neural Network Models (I.V. Serban et. al. 2016)

• Three RNN modules
  • Encoder RNN (encoding each sub-sequence into a vector)
  • Context RNN (encodes all previous sub-sequences into a vector)
  • Decoder RNN (generates the next sub-sequence)

* The randomness injected by the variable z corresponds to higher-level decisions, like topic or sentiment of the sentence.
Table 1: Wins, losses and ties (in %) of the VHRED model against the baselines based on the human study on Twitter (mean preferences ± 90% confidence intervals)

<table>
<thead>
<tr>
<th>Opponent</th>
<th>Short Contexts</th>
<th>Long Contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wins</td>
<td>Losses</td>
</tr>
<tr>
<td>VHRED vs LSTM</td>
<td>32.3 ± 2.4</td>
<td>42.5 ± 2.6</td>
</tr>
<tr>
<td>VHRED vs HRED</td>
<td>42.0 ± 2.8</td>
<td>31.9 ± 2.6</td>
</tr>
<tr>
<td>VHRED vs TF-IDF</td>
<td>51.6 ± 3.3</td>
<td>17.9 ± 2.5</td>
</tr>
</tbody>
</table>

Table 2: Twitter examples for the neural network models. The \( \rightarrow \) token indicates a change of turn.

<table>
<thead>
<tr>
<th>Context</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>jamaica blood!! &lt;url&gt; ( \rightarrow ) lol. yes [name]!. [name]! ( \rightarrow ) what’s poppin in toronto homie</td>
<td>VHRED: nothing much. just chillin. what’s good with you?</td>
</tr>
<tr>
<td></td>
<td>LSTM: something!!!!!!!!!!</td>
</tr>
<tr>
<td></td>
<td>HRED: something!!!!!!!</td>
</tr>
<tr>
<td>whhhhhyyyyy is it suddenly cold in tallahassee today?? this</td>
<td>VHRED: the sun is shining!!!</td>
</tr>
<tr>
<td>is florida! how can i dream of the beach when my fingers are</td>
<td>LSTM: i know!!!!!!!!!!</td>
</tr>
<tr>
<td>icye cold?? ( \rightarrow ) sittin on deck in sun looking at lake</td>
<td>HRED: something!!!!!!</td>
</tr>
<tr>
<td>travis in austin! enjoying day b4 birthday</td>
<td></td>
</tr>
<tr>
<td>reaches for [name]! ( \rightarrow ) - approaches him warily - hey there</td>
<td>VHRED: - looks up at you and smiles - you’re so cute. - kisses your cheek -</td>
</tr>
<tr>
<td>buddy. - swings him into the air - how you doin? ( \rightarrow ) tilts</td>
<td>LSTM: - smiles - i’m fine thank you. how are you?</td>
</tr>
<tr>
<td>his head and your wariness and reaches for your face [name].</td>
<td>HRED: - grins, looking at you as i look at you -</td>
</tr>
<tr>
<td>i think i want an ipad2. but i already have an ibook. want vs. need.</td>
<td>VHRED: that’s what i’m thinking. i’m going to buy a new one.</td>
</tr>
<tr>
<td>( \rightarrow ) if you have an ibook. then you deserve an ipad!</td>
<td>LSTM: i want one!!!!</td>
</tr>
<tr>
<td>ah echt? dan antwoord ik volgende keer via twitter als je me smst.</td>
<td>HRED: i don’t know what to do with it. i don’t know what to do.</td>
</tr>
<tr>
<td>( \rightarrow ) haha doe maar p. goed bezig hoor myrthe d ( \rightarrow</td>
<td>VHRED: haha, ik heb er geen zin in de something, maar ik heb er geen</td>
</tr>
<tr>
<td>goed bezig? die snap ik even niet p</td>
<td>something meer something</td>
</tr>
<tr>
<td></td>
<td>LSTM: haha something something something</td>
</tr>
<tr>
<td></td>
<td>HRED: something something something something something something something</td>
</tr>
</tbody>
</table>
End-to-End Goal-Oriented Dialog (A. Bordes et. al 2017)

• The workhorse of traditional dialog systems is **slot-filling**

• End-to-end dialog systems, usually based on neural networks, shown promising performance in **non goal-oriented** chit-chat settings, where they were trained to predict the next utterance in social media and forum threads

• Conducting **goal-oriented dialog** requires skills that go beyond language modeling, e.g., asking questions to **clearly define a user request**, querying Knowledge Bases (KBs), interpreting results from queries to **display options to users** or completing a transaction

• The paper shows: end-to-end dialog system based on Memory Networks can reach promising, **yet imperfect**, performance and learn to perform **non-trivial operations**
End-to-End Goal-Oriented Dialog

Goal-oriented dialog tasks:

• A user (in green) chats with a bot (in blue) to book a table at a restaurant. Models must predict bot utterances and API calls (in dark red). Task 1 tests the capacity of interpreting a request and asking the right questions to issue an API call.

• Task 2 checks the ability to modify an API call.

• Task 3 and 4 test the capacity of using outputs from an API call (in light red) to propose options (sorted by rating) and to provide extra-information.

• Task 5 combines everything.
Figure 1: (a): A single layer version of our model. (b): A three layer version of our model. In practice, we can constrain several of the embedding matrices to be the same (see Section 2.2).
## End-to-End Goal-Oriented Dialog - results

<table>
<thead>
<tr>
<th>Task</th>
<th>Rule-based Systems</th>
<th>TF-IDF Match</th>
<th>Nearest Neighbor</th>
<th>Supervised Embeddings</th>
<th>Memory Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1: Issuing API calls</td>
<td>100 (100)</td>
<td>5.6 (0)</td>
<td>22.4 (0)</td>
<td>55.1 (0)</td>
<td>99.9 (99.6)</td>
</tr>
<tr>
<td>T2: Updating API calls</td>
<td>100 (100)</td>
<td>3.4 (0)</td>
<td>16.4 (0)</td>
<td>68.3 (0)</td>
<td>100 (100)</td>
</tr>
<tr>
<td>T3: Displaying options</td>
<td>100 (100)</td>
<td>8.0 (0)</td>
<td>8.0 (0)</td>
<td>58.8 (0)</td>
<td>74.9 (2.0)</td>
</tr>
<tr>
<td>T4: Providing information</td>
<td>100 (100)</td>
<td>9.5 (0)</td>
<td>17.8 (0)</td>
<td>28.6 (0)</td>
<td>59.5 (3.0)</td>
</tr>
<tr>
<td>T5: Full dialogs</td>
<td>100 (100)</td>
<td>4.6 (0)</td>
<td>8.1 (0)</td>
<td>57.1 (0)</td>
<td>96.1 (49.4)</td>
</tr>
<tr>
<td>T1(OOV): Issuing API calls</td>
<td>100 (100)</td>
<td>5.8 (0)</td>
<td>22.4 (0)</td>
<td>44.1 (0)</td>
<td>72.3 (0)</td>
</tr>
<tr>
<td>T2(OOV): Updating API calls</td>
<td>100 (100)</td>
<td>3.5 (0)</td>
<td>16.8 (0)</td>
<td>68.3 (0)</td>
<td>78.9 (0)</td>
</tr>
<tr>
<td>T3(OOV): Displaying options</td>
<td>100 (100)</td>
<td>8.3 (0)</td>
<td>8.3 (0)</td>
<td>58.8 (0)</td>
<td>74.4 (0)</td>
</tr>
<tr>
<td>T4(OOV): Providing inform.</td>
<td>100 (100)</td>
<td>9.8 (0)</td>
<td>17.2 (0)</td>
<td>28.6 (0)</td>
<td>57.6 (0)</td>
</tr>
<tr>
<td>T5(OOV): Full dialogs</td>
<td>100 (100)</td>
<td>4.6 (0)</td>
<td>9.0 (0)</td>
<td>48.4 (0)</td>
<td>65.5 (0)</td>
</tr>
<tr>
<td>T6: Dialog state tracking 2</td>
<td>33.3 (0)</td>
<td>1.6 (0)</td>
<td>1.6 (0)</td>
<td>21.9 (0)</td>
<td>41.1 (0)</td>
</tr>
<tr>
<td>Concierge(*)</td>
<td>n/a</td>
<td>1.1 (0.2)</td>
<td>n/a</td>
<td>13.4 (0.5)</td>
<td>16.7 (1.2)</td>
</tr>
</tbody>
</table>

*Data extracted from a real online concierge service performing restaurant booking*
Visual Dialog (A. Das et. al. 2016)

Computer Vision and Artificial Intelligence Trends:
• Image classification
• Scene recognition
• Object detection
• Learning to play video games
• Image and video QA

What’s Next?
• Visual Dialog: Ability to hold a meaningful dialog with humans in natural language about visual content
Aiding visually impaired users in understanding their surroundings or social media content

**AI:** ‘John just uploaded a picture from his vacation in Hawaii’, **Human:** ‘Great, is he at the beach?’, **AI:** ‘No, on a mountain’

Aiding analysts in making decisions based on large quantities of surveillance data

**Human:** ‘Did anyone enter this room last week?’, **AI:** ‘Yes, 27 instances logged on camera’, **Human:** ‘Were any of them carrying a black bag?’, **AI:** ‘No’

Interacting with an AI assistant

**Human:** ‘Alexa – can you see the baby in the baby monitor?’, **AI:** ‘Yes, I can’, **Human:** ‘Is he sleeping or playing?’

Robotics applications (e.g. search and rescue mission)

**Human:** ‘Is there smoke in any room around you?’, **AI:** ‘Yes, in one room’, **Human:** ‘Go there and look for people’
Visual Dialog vs. Dialog System

• **Visual Dialog Task** (visual analogue of the Turing Test): given an image I, a history of a dialog consisting of a sequence of question-answer pairs, and a natural language follow-up question, the task for the machine is to answer the question in free-form natural language.

• Visual Dialog is more specific than a general chatbot because the dialog is about a specific image.

• Visual Dialog is not geared toward a specific goal (similar to goal-driven dialog systems). Therefore slot-filling methods won’t work.
Visual Dialog vs. VQA

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

Visual Dialog
Q: How many people are on wheelchairs?
A: Two
Q: What are their genders?
A: One male and one female
Q: Which one is holding a racket?
A: The woman

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

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
Visual Dialog Dataset – Data Collection

• Good data for this task should include dialogs that have:
  • Temporal continuity
  • Grounding in the image
  • Mimic natural ‘conversational’ exchanges

• Collected visual dialog data on images from the Common Objects in Context (COCO) dataset, which contains multiple objects in everyday scenes.

• Free form, open-ended natural language questions collected via two workers chatting on Amazon Mechanical Turk (AMT) real-time
  • The ‘questioner’ sees only a single line of text describing an image (caption from COCO); the image remains hidden to the questioner.
  • Their task is to ask questions about this hidden image so as to ‘imagine the scene better’
  • The ‘answerer’ sees the image and the caption. Their task is to answer the questions asked by their chat partner.
Figure 3: Collecting visually-grounded dialog data on Amazon Mechanical Turk via a live chat interface where one person is assigned the role of ‘questioner’ and the second person is the ‘answerer’. We show the first two questions being collected via the interface as Turkers interact with each other in Fig. 3a and Fig. 3b. Remaining questions are shown in Fig. 3c.
Visual Dialog Dataset – Analysis

• **One** dialog (10 question-answer pairs) on **68k** images from COCO (58k train and 10k val), or a total of **680,000** QA pairs

• More **natural** conversation comparing to other image QA datasets because the questioner doesn’t see the image (**no visual priming bias**)

• Higher mean-length of answers (3.1 words) and less **binary answers** (e.g. ‘Yes’, ‘No’)

• **Coreference in dialog:** 38% of questions, 22% of answers, and nearly all (99%) dialogs contain at least one pronoun

• **Temporal Continuity in Dialog Topics:** based on human evaluation on smaples, across 10 rounds, VisDial question have **4:55 +/- 0:17** topics on average, confirming that these are not independent questions
Visual Dialog Dataset – Analysis

<table>
<thead>
<tr>
<th>Dataset</th>
<th># QA</th>
<th># Images</th>
<th>Q Length</th>
<th>A Length</th>
<th>A Length &gt; 2</th>
<th>Top-1000 A</th>
<th>Human Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAQUAR [31]</td>
<td>12,468</td>
<td>1,447</td>
<td>11.5 ± 2.4</td>
<td>1.2 ± 0.5</td>
<td>3.4%</td>
<td>96.4%</td>
<td>-</td>
</tr>
<tr>
<td>Visual Madlibs [60]</td>
<td>56,468</td>
<td>9,688</td>
<td>4.9 ± 2.4</td>
<td>2.8 ± 2.0</td>
<td>47.4%</td>
<td>57.9%</td>
<td>-</td>
</tr>
<tr>
<td>COCO-QA [41]</td>
<td>117,684</td>
<td>69,172</td>
<td>8.7 ± 2.7</td>
<td>1.0 ± 0.0</td>
<td>0.0%</td>
<td>100%</td>
<td>-</td>
</tr>
<tr>
<td>Baidu [12]</td>
<td>316,193</td>
<td>316,193</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VQA [4]</td>
<td>614,163</td>
<td>204,721</td>
<td>6.2 ± 2.0</td>
<td>1.1 ± 0.4</td>
<td>3.8%</td>
<td>82.7%</td>
<td>✓</td>
</tr>
<tr>
<td>Visual7W [62]</td>
<td>327,939</td>
<td>47,300</td>
<td>6.9 ± 2.4</td>
<td>2.0 ± 1.4</td>
<td>27.6%</td>
<td>63.5%</td>
<td>✓</td>
</tr>
<tr>
<td>VisDial (Ours)</td>
<td>680,000</td>
<td>68,000</td>
<td>4.64 ± 0</td>
<td>3.08 ± 0</td>
<td>42.2%</td>
<td>58.1%</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 2: Comparison of existing image question answering datasets with VisDial

(a) Questions vs. Answers
(b) Percentage coverage
Visual Dialog - Evaluation Protocol

• Evaluate individual responses independently at each round \( t = 1, 2, \ldots, 10 \) in a retrieval or multiple-choice setup

• The model is evaluated on retrieval metrics:
  • rank of human response
  • recall@k, i.e., existence of the human response in top-k ranked responses
  • mean reciprocal rank (MRR) of the human response

• **Candidate Answers:** ground-truth, answers to 50 similar questions, 30 most popular answers, 19 random answers

\[
I = \text{image} \\
H = \left( \frac{C}{H_0}, \left( \frac{Q_1, A_1}{H_1}, \ldots, \frac{Q_{t-1}, A_{t-1}}{H_{t-1}} \right) \right) \\
Q_t = \text{question}
\]

INPUT

OUTPUT

sorting of 100 candidate answers

\( \{ A_t^{(1)}, \ldots, A_t^{(100)} \} \)
Neural Visual Dialog Models

• Experimented with the encoder-decoder combinations

• **Encoders:** convert inputs \((I, H, Q_t)\) into a joint representation

• In all cases, we represent \(I\) via the l2-normalized activations from the penultimate layer of VGG-16

• For each encoder \(E\), we experiment with all possible ablated versions: \(E(Q_t), E(Q_t, I), E(Q_t, H), E(Q_t, I, H)\)

• **Decoders:** rank candidate answers based on the joint representation from encoders

• Generative (LSTM) and Discriminative (Softmax)
Visual Dialog - Encoders

• **Late Fusion (LF) Encoder:**
  - Treat H as a long string with the entire history \((H_0, ..., H_{t-1})\) concatenated.
  - \(Q_t\) and H are separately encoded with 2 different LSTMs
  - individual representations of participating inputs \((I, H, Q_t)\) are concatenated and linearly transformed to a desired size of joint representation.

• **Hierarchical Recurrent Encoder (HRE):**
  - Similar architecture as ‘Hierarchical Latent Variable Encoder-Decoder Model’

• **Memory Network (MN) Encoder:**
  - Encode \(Q_t\) with an LSTM to get a 512 -d vector
  - encode each previous round of history \((H_0, ..., H_{t-1})\) with another LSTM to get a \(t \times 512\) matrix.
  - Compute inner product of question vector with each history vector to get scores over previous rounds, which are fed to a softmax to get attention-over-history probabilities.
Visual Dialog – Late Fusion Encoder

Image I
Do you think the woman is with him?

Question $Q_t$

The man is riding his bicycle on the sidewalk. Is the man wearing a helmet? No he does not have a helmet on. ... Are there any people nearby? Yes there’s a woman walking behind him.

t rounds of history (concatenated)

A fully-connected layer and tanh nonlinearity
Each QA-pair in dialog history is independently encoded by another LSTM with shared weights.

The image-question representation, computed for every round from 1 through t, is concatenated with history representation from the previous round and constitutes a sequence of question-history vectors.
A fully-connected layer and tanh nonlinearity

No I don't think they are together
Answer $A_t$

compute attention over the $t$ facts by inner product
Visual Dialog - Decoders

• **Generative (LSTM):**
  - Encoded vector is set as the initial state of the LSTM language model
  - Maximizes the log-likelihood of the ground truth answer sequence given its corresponding encoded representation (trained end-to-end)
  - Uses the model’s loglikelihood scores and rank candidate answers

• **Discriminative (softmax):**
  - Computes dot product similarity between the input encoding and an LSTM encoding of each of the answer options
  - The dot products are fed into a softmax to compute the posterior probability over the options
  - Maximizes the loglikelihood of the correct options and options are simply ranked based on their posterior probabilities.
Visual Dialog – Generative Decoder

- Encoded vector is set as the initial state of the LSTM language model
- Maximizes the log-likelihood of the ground truth answer sequence given its corresponding encoded representation (trained end-to-end)
- Uses the model’s loglikelihood scores and rank candidate answers
Computes dot product similarity between the input encoding and an LSTM encoding of each of the answer options

The dot products are fed into a softmax to compute the posterior probability over the options

Maximizes the log likelihood of the correct options and options are simply ranked based on their posterior probabilities
### Visual Dialog – Experiment Results

- **Dialogs**: 50K train, 8K validation, 10K test

<table>
<thead>
<tr>
<th>Model</th>
<th>MRR</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Answer prior</td>
<td>0.311</td>
<td>19.85</td>
<td>39.14</td>
<td>44.28</td>
<td>31.56</td>
</tr>
<tr>
<td>NN-Q</td>
<td>0.392</td>
<td>30.54</td>
<td>46.99</td>
<td>49.98</td>
<td>30.88</td>
</tr>
<tr>
<td>NN-QI</td>
<td>0.385</td>
<td>29.71</td>
<td>46.57</td>
<td>49.86</td>
<td>30.90</td>
</tr>
<tr>
<td><strong>Generative</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LF-Q-G</td>
<td>0.403</td>
<td>29.74</td>
<td>50.10</td>
<td>56.32</td>
<td>24.06</td>
</tr>
<tr>
<td>LF-QH-G</td>
<td>0.425</td>
<td>32.49</td>
<td>51.56</td>
<td>57.80</td>
<td>23.11</td>
</tr>
<tr>
<td>LF-QI-G</td>
<td>0.437</td>
<td>34.06</td>
<td>52.50</td>
<td>58.89</td>
<td>22.31</td>
</tr>
<tr>
<td>LF-QIH-G</td>
<td>0.430</td>
<td>33.27</td>
<td>51.96</td>
<td>58.09</td>
<td>23.04</td>
</tr>
<tr>
<td>HRE-QH-G</td>
<td>0.430</td>
<td>32.84</td>
<td>52.36</td>
<td>58.64</td>
<td>22.59</td>
</tr>
<tr>
<td>HRE-QIH-G</td>
<td>0.442</td>
<td>34.37</td>
<td>53.40</td>
<td>59.74</td>
<td>21.75</td>
</tr>
<tr>
<td>HREA-QIH-G</td>
<td>0.442</td>
<td>34.47</td>
<td>53.43</td>
<td>59.73</td>
<td>21.83</td>
</tr>
<tr>
<td>MN-QH-G</td>
<td>0.434</td>
<td>33.12</td>
<td>53.14</td>
<td>59.61</td>
<td>22.14</td>
</tr>
<tr>
<td>MN-QIH-G</td>
<td><strong>0.443</strong></td>
<td><strong>34.62</strong></td>
<td><strong>53.74</strong></td>
<td><strong>60.18</strong></td>
<td><strong>21.69</strong></td>
</tr>
</tbody>
</table>

Higher Better | Lower Better

<table>
<thead>
<tr>
<th>Model</th>
<th>MRR</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>LF-Q-D</td>
<td>0.482</td>
<td>34.29</td>
<td>63.42</td>
<td>74.31</td>
<td>8.87</td>
</tr>
<tr>
<td>LF-QH-D</td>
<td>0.505</td>
<td>36.21</td>
<td>66.56</td>
<td>77.31</td>
<td>7.89</td>
</tr>
<tr>
<td>LF-QI-D</td>
<td>0.502</td>
<td>35.76</td>
<td>66.59</td>
<td>77.61</td>
<td>7.72</td>
</tr>
<tr>
<td>LF-QIH-D</td>
<td>0.511</td>
<td>36.72</td>
<td>67.46</td>
<td>78.30</td>
<td>7.63</td>
</tr>
<tr>
<td>HRE-QH-D</td>
<td>0.489</td>
<td>34.74</td>
<td>64.25</td>
<td>75.40</td>
<td>8.32</td>
</tr>
<tr>
<td>HRE-QIH-D</td>
<td>0.502</td>
<td>36.26</td>
<td>65.67</td>
<td>77.05</td>
<td>7.79</td>
</tr>
<tr>
<td>HREA-QIH-D</td>
<td>0.508</td>
<td>36.76</td>
<td>66.54</td>
<td>77.75</td>
<td>7.59</td>
</tr>
<tr>
<td>MN-OH-D</td>
<td>0.524</td>
<td>36.84</td>
<td>67.78</td>
<td>78.92</td>
<td>7.25</td>
</tr>
<tr>
<td>MN-QIH-D</td>
<td><strong>0.529</strong></td>
<td><strong>37.33</strong></td>
<td><strong>68.47</strong></td>
<td><strong>79.54</strong></td>
<td><strong>7.03</strong></td>
</tr>
</tbody>
</table>

**Discriminative**

**VQA**

<table>
<thead>
<tr>
<th>Model</th>
<th>Human Accuracies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-Q</td>
<td>0.441</td>
</tr>
<tr>
<td>Human-QH</td>
<td>0.485</td>
</tr>
<tr>
<td>Human-QI</td>
<td>0.619</td>
</tr>
<tr>
<td>Human-QIH</td>
<td>0.635</td>
</tr>
</tbody>
</table>
Visual Dialog - Conclusions

• Demonstrated the first visual chatbot.

• The results and analysis indicates that there is significant scope for improvement, the authors believe this task can serve as a testbed for measuring progress towards visual intelligence.

Potential Improvements:

• Using a model to generate responses rather than ranking candidate answers

• Include language features (e.g. part-of-speech) as the inputs

• Extend it to videos
THANKS