VQA → Visual Dialog

Devi Parikh
People coloring a street on a college campus
It was a great event! It brought families out, and the whole community together.
Q. What are they coloring the street with?
A. Chalk
AI: What a nice picture! What event was this?
User: “Color College Avenue”. It was a lot of fun!
AI: I am sure it was! Do they do this every year?
User: I wish they would. I don’t think they’ve organized it again since 2012.
...
Aid visually-impaired users

FACEBOOK’S AI CAN CAPTION PHOTOS FOR THE BLIND ON ITS OWN
Peter just uploaded a picture from his vacation in Hawaii.

Great, is he at the beach?

No, on a mountain.

...
Aid ‘situationally-impaired’ analysts

Did anyone enter this room last week?

Yes, 127 instances logged on camera

Were any of them carrying a black bag?

...
Natural language instructions for robots

Is there smoke in any room around you?

Yes, in one room

Go there and look for people

...
Outline

Visual Question Answering

Visual Dialog
GuessWhat?! Visual object discovery through multi-modal dialogue

Harm de Vries  
University of Montreal  
mail@harmdevries.com

Florian Strub  
Univ. Lille, CNRS, Centrale Lille, Inria, UMR 9189 CRIStAL  
florian.strub@inria.fr

Sarah Chandar  
University of Montreal  
sarathcse2008@gmail.com

Olivier Pietquin  
DeepMind  
pietquin@google.com

Hugo Larochelle  
Twitter  
hlarochelle@twitter.com

Aaron Courville  
University of Montreal  
aaron.courville@gmail.com

Abstract

We introduce GuessWhat?!, a two-player guessing game as a testbed for research on the interplay of computer vision and dialogue systems. The goal of the game is to locate an unknown object in a rich image scene by asking a sequence of questions. Higher-level image understanding, like spatial reasoning and language grounding, is required to solve the proposed task. Our key contribution is the collection of a large-scale dataset consisting of 150K human-played games with a total of 800K visual question-answer pairs on 66K images. We explain our design decisions in collecting the dataset and introduce the oracle and questioner tasks that are associated with the two players of the game. We prototyped deep learning models to establish initial baselines of the introduced tasks.

<table>
<thead>
<tr>
<th>Questioner</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is it a vase?</td>
<td>Yes</td>
</tr>
<tr>
<td>Is it partially visible?</td>
<td>No</td>
</tr>
<tr>
<td>Is it in the left corner?</td>
<td>No</td>
</tr>
<tr>
<td>Is it the turquoise and purple one?</td>
<td>Yes</td>
</tr>
</tbody>
</table>
End-to-end optimization of goal-driven and visually grounded dialogue systems

Florian Strub
florian.strub@inria.fr
Univ. Lille, CNRS, Centrale Lille, Inria,
UMR 9189 - CRIStAL, F-59000 Lille, France

Harm de Vries
mail@harmdevries.com
University of Montreal

Jeremie Mary
jeremie.mary@univ-lille3.fr
Univ. Lille, CNRS, Centrale Lille, Inria,
UMR 9189 - CRIStAL, F-59000 Lille, France

Bilal Piot
piot@google.com
DeepMind

Aaron Courville
aaron.courville@gmail.com
University of Montreal

Olivier Pietquin
pietquin@google.com
DeepMind

Abstract
End-to-end design of dialogue systems has recently become a popular research topic thanks to powerful tools such as encoder-decoder architectures for sequence-to-sequence learning. Yet, most current approaches cast human-machine dialogue management as a supervised learning problem, aiming at predicting the next utterance of a participant given the full history of the dialogue. This vision is too simplistic to render the intrinsic planning problem inherent to dialogue as well as its grounded nature, making the context of a dialogue larger than the sole history. This is why only chit-chat and question answering tasks have been addressed so

Figure 1: Two example games of the GuessWhat?! dataset. The correct object is highlighted by a green mask.
Outline

Visual Question Answering

Visual Dialog
Visual Question Answering (VQA)
What is the mustache made of?
What is the mustache made of?
What is the mustache made of?

AI System

bananas
Visual Question Answering (VQA)

Ask any question about this image
Visual Question Answering (VQA)

Does it appear to be rainy?
Does this person have 20/20 vision?
Is this person expecting company?
What is just under the tree?
How many slices of pizza are there?
Is this a vegetarian pizza?
What color are her eyes?
What is the mustache made of?
VQA Dataset
VQA Dataset

>0.25 million images
254,721 images (COCO)
50,000 scenes
VQA Dataset

>0.25 million images

>0.76 million questions
Stump a smart robot! Ask a question about this image that a human can answer, but a smart robot probably can’t!

We have built a kitchen, beach
Ask a question
IMPORTANT: Try not to ask a question that can recognize the scene (e.g., “What can I make a big dinner in this kitchen?”) as it should not be able to answer.

Questions below:

- Do not repeat questions. Do not ask the same questions or the same questions with minor variations over and over again across images. Think of a new question each time specific to each image.

- Each question should be a single question. Do not ask questions that have multiple parts or multiple sub-questions in them.

- Do not ask generic questions that can be asked of many other images. Ask questions specific to each image.

Please ask a question about this image that a human can answer “if” looking at the image (and not otherwise), but would stump this smart robot:

Q1: Write your question here to stump this smart robot.
VQA Dataset

>0.25 million images

>0.76 million questions

~10 million answers

[Antol et al., ICCV 2015]
Papers using VQA

Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources

Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, Anthony Dick
School of Computer Science, The University of Adelaide
{qi.wu01, p.wang, chunhua.shen, anton.vandenhengel, anthony.dick}@adelaide.edu.au

Compositional Memory for Visual Question Answering

Aiwen Jiang1,2, Fang Wang2, Fatih Porikli2, Yi Li+ 2,3
1Jiangxi Normal University  
2NICTA and ANU  
3Toyota Research Institute North America
aiwen.jiang@nicta.com.au  
{fang.wang, fatih.porikli}@nicta.com.au  
yi.li@tema.toyota.com

Deep Compositional Question Answering with Neural Module Networks

Jacob Andreas  Marcus Rohrbach  Trevor Darrell  Dan Klein
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
{jda, rohrbach, trevor, klein}@cs, eecs, eecs, cs.berkeley.edu

Simple Baseline for Visual Question Answering

Bolei Zhou1, Yuandong Tian3, Sainbayar Sukhbaatar3, Arthur Szlam2, and Rob Fergus2
1Massachusetts Institute of Technology  
2Facebook AI Research

Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering

Huijuan Xu  
UMass Lowell  
hxu@cs.uml.edu
Kate Saenko  
UMass Lowell  
saenko@cs.uml.edu

Where To Look: Focus Regions for Visual Question Answering

Kevin J. Shih, Saurabh Singh, and Derek Hoiem
University of Illinois at Urbana-Champaign
{kjshih2, ss1, dhoiem}@illinois.edu

Stacked Attention Networks for Image Question Answering

Zichao Yang1, Xiaodong He2, Jianfeng Gao2, Li Deng2, Alex Smola1
1Carnegie Mellon University, 2Microsoft Research, Redmond, WA 98052, USA
py@cs.cmu.edu, {xiaohhe, jfgao, deng}@microsoft.com, alex@smola.org

ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering

Kan Chen  
University of Southern California  
kanchen@usc.edu
Haoyuan Gao  
Baidu Research - IDL  
gaohaoyuan@baidu.com
Weixi Xu  
Baidu Research - IDL  
weixi.xu@baidu.com
Jiang Wang  
Baidu Research - IDL  
wang.jiang@baidu.com
Liang-Chieh Chen  
UCLA  
lochen@cs.ucla.edu
Ram Nevatia  
University of Southern California  
evatia@usc.edu

29
VQA Challenge @ CVPR16

Recent progress in computer vision and natural language processing has demonstrated that lower-level tasks are much closer to being solved. We believe that the time is ripe to pursue higher-level tasks, one of which is Visual Question Answering (VQA), where the goal is to he
VQA Challenge @ CVPR16

<table>
<thead>
<tr>
<th>Team</th>
<th>Overall</th>
<th>Yes/No</th>
<th>Number</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC Berkeley &amp; Sony[^14]</td>
<td>68.47</td>
<td>83.24</td>
<td>39.47</td>
<td>58</td>
</tr>
<tr>
<td>Never Labs[^10]</td>
<td>64.79</td>
<td>63.31</td>
<td>38.7</td>
<td>54.62</td>
</tr>
<tr>
<td>DLAIT[^9]</td>
<td>63.68</td>
<td>83.25</td>
<td>40.07</td>
<td>52.09</td>
</tr>
<tr>
<td>snu-naverlab[^29]</td>
<td>63.18</td>
<td>83.16</td>
<td>39.14</td>
<td>51.33</td>
</tr>
<tr>
<td>POSTECH[^11]</td>
<td>63.17</td>
<td>81.67</td>
<td>38.16</td>
<td>52.79</td>
</tr>
</tbody>
</table>

Winning entry (MCB)

- Open-ended: 66%
- Multiple-choice: 70%
The Power of Language Priors
A giraffe is standing in grass next to a tree

The Power of Language Priors

Slide credit: Yash Goyal and Peng Zhang
Is there a clock ... ?
‘yes’ 98%

Is the man wearing glasses ... ?
‘yes’ 94%

Are the lights on ... ?
‘yes’ 85%

Do you see a ... ?
‘yes’ 87%
Is the man standing ... ?
‘no’ 69%

What sport is ... ?
‘tennis’ 41%

How many ... ?
‘2’ 39%

What animal is ... ?
‘dog’ 35%
Balancing the VQA dataset

Select an image for which answer to the question
What game is this?
is NOT tennis

SHOW INSTRUCTIONS

NOT POSSIBLE
PREVIOUS
NEXT
Balancing the VQA dataset

Is the TV on?

yes

no
Balancing the VQA dataset

How many pets are present?

2

1
Balancing the VQA dataset

What sign is this?

handicap

one way
Balancing the VQA dataset

Where is the child sitting?

fridge

arms
Balancing the VQA dataset

What is the cat doing on the rug?

sleeping

sitting
Balancing the VQA dataset

What color are the pants?

orange

brown
VQA v2.0

• More balanced than VQA v1.0
  – Entropy of answers increases by 56%

• Bigger than VQA v2.0
  – ~1.8 times image-question pairs
Benchmarking SOTA VQA models

• SOTA VQA models
  – Drop in performance by 7-8%
  – Gain 1-2% back when re-trained on balanced dataset

• By answer types
  – Biggest drop in performance in yes/no (10-12%)
  – Biggest improvement gained by re-training in yes/no (3-4%) and number (2-3%)
## Trends

<table>
<thead>
<tr>
<th></th>
<th>By Answer Type</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes/No</td>
<td>Number</td>
</tr>
<tr>
<td>UC Berkeley &amp; Sony[^14]</td>
<td>83.79</td>
<td>38.9</td>
</tr>
<tr>
<td>Naver Labs[^10]</td>
<td>83.78</td>
<td>37.67</td>
</tr>
<tr>
<td>DLAIT[^5]</td>
<td>83.65</td>
<td>39.18</td>
</tr>
<tr>
<td>snubi-naverlabs[^25]</td>
<td>83.64</td>
<td>38.43</td>
</tr>
</tbody>
</table>

**Trend Analysis:**
- **0.15%** increase in Yes/No across all datasets.
- **1.51%** increase in Number values.
- **7.03%** increase in Other categories.
- **3.5%** decrease in Overall scores.
VQA v2.0

2\textsuperscript{nd} VQA Challenge @ CVPR17!
Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering (CVPR 2017)

Yash Goyal (Virginia Tech)
Tejas Khot (Virginia Tech)
Doug Summers-Stay (Army Research Lab)
Dhruv Batra (Georgia Tech / FAIR)
Devi Parikh (Georgia Tech / FAIR)
(Another) problem with existing setup

Train

Q: What color is the dog?
A: White

Training Prior

- white
- red
- blue
- green
- yellow

Slide credit: Aishwarya Agrawal
(Another) problem with existing setup

**Train**

**Q:** What color is the dog?

**A:** White

**Test**

**Q:** What color is the dog?

**A:** Black

Slide credit: Aishwarya Agrawal
(Another) problem with existing setup

Train
Q: What color is the dog?
A: White

Test
Q: What color is the dog?
A: Black

Prediction: White
(Another) problem with existing setup

**Train**
- **Q:** Is the person wearing shorts?
- **A:** No

**Test**
- **Q:** Is the person wearing shorts?
- **A:** Yes

Prediction: No

Slide credit: Aishwarya Agrawal
(Another) problem with existing setup

- Similar priors in train and test
- Memorization does not hurt as much
- Problematic for benchmarking progress
Meet VQA-CP!

• Visual Question Answering under Changing Priors

• A new split of the VQA v1.0 dataset (Antol et al., ICCV 2015)
## Performance of VQA models on VQA-CP

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>Overall</th>
<th>Yes/No</th>
<th>Number</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>d-LSTM Q + norm I</td>
<td>VQA</td>
<td>54.23</td>
<td>79.81</td>
<td>33.26</td>
<td>40.35</td>
</tr>
<tr>
<td>(Antol et al. ICCV15)</td>
<td>VQA-CP</td>
<td>23.51</td>
<td>34.53</td>
<td>11.40</td>
<td>17.42</td>
</tr>
<tr>
<td>NMN</td>
<td>VQA</td>
<td>54.83</td>
<td>80.39</td>
<td>33.45</td>
<td>41.07</td>
</tr>
<tr>
<td>(Andreas et al. CVPR16)</td>
<td>VQA-CP</td>
<td>29.64</td>
<td>38.85</td>
<td>11.23</td>
<td>27.88</td>
</tr>
<tr>
<td>SAN</td>
<td>VQA</td>
<td>55.86</td>
<td>78.54</td>
<td>33.46</td>
<td>44.51</td>
</tr>
<tr>
<td>(Yang et al. CVPR16)</td>
<td>VQA-CP</td>
<td>26.88</td>
<td>35.34</td>
<td>11.34</td>
<td>24.70</td>
</tr>
<tr>
<td>MCB</td>
<td>VQA</td>
<td>60.97</td>
<td>81.62</td>
<td>34.56</td>
<td>52.16</td>
</tr>
<tr>
<td>(Fukui et al. EMNLP16)</td>
<td>VQA-CP</td>
<td>34.39</td>
<td>37.96</td>
<td>11.80</td>
<td>39.90</td>
</tr>
</tbody>
</table>

- 31% drop
- 25% drop
- 29% drop
- 27% drop
Grounded-VQA (GVQA)

Image (I) → VGG

Question (Q) → Extractor → Q_{main}

Answer Cluster Predictor (ACP)

Answer Predictor (AP)

Concept Extractor (CE)

Visual Concept Classifier (VCC)

Visual Verifier (VV)

Slide credit: Aishwarya Agrawal
Aishwarya Agrawal  
(Virginia Tech)

Dhruv Batra  
(Georgia Tech / FAIR)

Devi Parikh  
(Georgia Tech / FAIR)

Ani Kembhavi  
(AI2)
C-VQA: Compositional VQA

Training

Q: What color is the **plate**?
A: Green

Q: What color are **stop lights**?
A: Red

Testing

Q: What color is the **stop light**?
A: Green

Q: What is the color of the **plate**?
A: Red
Aishwarya Agrawal  
(Virginia Tech)

Dhruv Batra  
(Georgia Tech / FAIR)

Devi Parikh  
(Georgia Tech / FAIR)
Outline

Visual Question Answering

Visual Dialog
A man and a woman are holding umbrellas
A man and a woman are holding umbrellas.

What color is his umbrella?
A man and a woman are holding umbrellas

What color is his umbrella?
A man and a woman are holding umbrellas

What color is his umbrella?
A man and a woman are holding umbrellas

What color is his umbrella?
A man and a woman are holding umbrellas

His umbrella is black

What color is his umbrella?
A man and a woman are holding umbrellas

His umbrella is black

What color is his umbrella?

What about hers?
A man and a woman are holding umbrellas.

His umbrella is black.

What color is his umbrella?

What about hers?
Visual Dialog

A man and a woman are holding umbrellas

His **umbrella** is black

What color is his **umbrella**?

What about **hers**?
A man and a woman are holding umbrellas

What color is his umbrella?

His umbrella is black

What about hers?
A man and a woman are holding umbrellas

His umbrella is black

Hers is multi-colored

What color is his umbrella?

What about hers?
A man and a woman are holding umbrellas.

His umbrella is black.

Hers is multi-colored.

What color is his umbrella?

What about hers?

How many other people are in the image?
A man and a woman are holding umbrellas

His umbrella is black

Hers is multi-colored

What color is his umbrella?

What about hers?

How many other people are in the image?
A man and a woman are holding umbrellas

What color is his umbrella?

His umbrella is black

What about hers?

Hers is multi-colored

How many other people are in the image?

I think 3. They are occluded
A man and a woman are holding umbrellas

What color is his umbrella?

His umbrella is black

What about hers?

Hers is multi-colored

How many other people are in the image?

I think 3. They are occluded

How many are men?
A man and a woman are holding umbrellas

His umbrella is black

Hers is multi-colored

I think 3. They are occluded

What color is his umbrella?

What about hers?

How many other people are in the image?

How many are men?
Visual Dialog: Task

• Given
  – Image I
  – History of human dialog
    \((Q_1, A_1), (Q_2, A_2), \ldots, (Q_{t-1}, A_{t-1})\)
  – Follow-up Question \(Q_t\)

• Task
  – Produce free-form natural language answer \(A_t\)
Visual Dialog: Evaluation Protocol

• Given
  – Image I
  – History of human dialog \((Q_1, A_1), (Q_2, A_2), \ldots, (Q_{t-1}, A_{t-1})\)
  – Follow-up Question \(Q_t\)
  – 100 Answer Options
    • 50 answers from NN questions
    • 30 popular answers
    • 20 random answers

• Evaluation Task
  – Rank the list of 100 options

• Accuracy/Error
  – mean-rank-of-GT, mean-reciprocal-rank

Question: Do people look happy?

GT: Not really
• Yes they do
• I can't tell
• Not facing me
• Yes they look happy
• Yes I can only see 1 of their faces but she looks happy
• Not really but not unhappy either
VisDial Dataset
Live Two-Person Chat on Amazon Mechanical Turk
VisDial Dataset
Live Two-Person Chat on Amazon Mechanical Turk
Code for the chat interface used to collect the VisDial dataset on AMT [http://visualdialog.org/]

VisDial AMT Chat

Source for the two-person chat interface used to collect the VisDial dataset [arxiv.org/abs/1611.08669] on Amazon Mechanical Turk.
VisDial v0.9 Stats

>120k images (from COCO)

1 dialog/image

10 question-answer rounds/dialog

Total of >1.2 Million dialog QA pairs
VisDial Dataset

Code for the real-time chat interface used to collect the VisDial dataset on Amazon Mechanical Turk

VisDial v0.9

Training set (235M)
82,783 images

Validation set (108M)
40,504 images

Readme

- v0.9 Training is from COCO Training and v0.9 Validation set is from COCO Validation
- Numbers (in papers, etc.) should be reported on v0.9 val

Format

```json
[
  {
    'data': {
      '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',
        ...
      ]
    }
  }
]```
Models for Visual Dialog

Encoder
1. Late Fusion
2. Hierarchical Recurrent Encoder
3. Memory Network

Decoder
1. Generative
   - During training, maximizes LL of human response
   - For evaluation, ranks options by LL scores
2. Discriminative
   - Learn to rank 100 options
Visual Dialog Model #3

Memory Network Encoder

Slide credit: Abhishek Das
Do you think the woman is with him?

Question $Q_t$
Visual Dialog Model #3

Image 1

Do you think the woman is with him?

Question $Q_t$

<table>
<thead>
<tr>
<th>The man is riding his bicycle on the sidewalk.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the man wearing a helmet? No he does not have a helmet on.</td>
</tr>
<tr>
<td>How old is the man? He looks around 40 years old.</td>
</tr>
<tr>
<td>What color is his bike? It has black wheels and handlebars. I can’t see the body of the bike that well.</td>
</tr>
<tr>
<td>Is anyone else riding a bike? No he’s the only one.</td>
</tr>
<tr>
<td>Are there any people nearby? Yes there’s a woman walking behind him.</td>
</tr>
</tbody>
</table>

$t$ rounds of history

$\{(\text{Caption}, (Q_t, A_t), \ldots, (Q_{t-1}, A_{t-1}))\}$

Memory Network Encoder

Slide credit: Abhishek Das
Visual Dialog Model #3

Image 1

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. |
| How old is the man? He looks around 40 years old. |
| What color is his bike? It has black wheels and handlebars. I can’t see the body of the bike that well. |
| Is anyone else riding a bike? No he’s the only one. |
| Are there any people nearby? Yes there’s a woman walking behind him. |

$t$ rounds of history

$\{(\text{Caption}), (Q_t, A_t), \ldots, (Q_{t-1}, A_{t-1})\}$

Memory Network Encoder

Slide credit: Abhishek Das
Visual Dialog Model #3

Image 1
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.
How old is the man? He looks around 40 years old.
What color is his bike? It has black wheels and handlebars. I can't see the body of the bike that well.
Is anyone else riding a bike? No he's the only one.
Are there any people nearby? Yes there's a woman walking behind him.

$t$ rounds of history
{(Caption), ($Q_t, A_t$), ..., ($Q_{t-1}, A_{t-1}$)}

Memory Network Encoder

Slide credit: Abhishek Das
Visual Dialog Model #3

Memory Network Encoder

Slide credit: Abhishek Das
Visual Dialog Model #3

Memory Network Encoder

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.
- How old is the man? He looks around 40 years old.
- What color is his bike? It has black wheels and handlebars. I can't see the body of the bike that well.
- Is anyone else riding a bike? No he's the only one.
- Are there any people nearby? Yes there's a woman walking behind him.

$t$ rounds of history

$\{(\text{Caption}), (Q_t, A_t), \ldots, (Q_{t-1}, A_{t-1})\}$

Slide credit: Abhishek Das
Visual Dialog Model #3

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.
How old is the man? He looks around 40 years old.
What color is his bike? It has black wheels and handlebars. I can't see the body of the bike that well.
Is anyone else riding a bike? No he's the only one.
Are there any people nearby? Yes there's a woman walking behind him.

Memory Network Encoder

$t$ rounds of history

{(Caption), $(Q_t, A_t), ..., (Q_{t-1}, A_{t-1})$}
<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>0.3735</td>
<td>23.55</td>
<td>48.52</td>
<td>53.23</td>
<td>26.50</td>
</tr>
<tr>
<td>NN-Q</td>
<td>0.4570</td>
<td>35.93</td>
<td>54.07</td>
<td>60.26</td>
<td>18.93</td>
</tr>
<tr>
<td>NN-QI</td>
<td>0.4274</td>
<td>33.13</td>
<td>50.83</td>
<td>58.69</td>
<td>19.62</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>0.5048</td>
<td>39.78</td>
<td>60.58</td>
<td>66.33</td>
<td>17.89</td>
</tr>
<tr>
<td>LF-QH-G</td>
<td>0.5055</td>
<td>39.73</td>
<td>60.86</td>
<td>66.68</td>
<td>17.78</td>
</tr>
<tr>
<td>LF-QI-G</td>
<td>0.5204</td>
<td>42.04</td>
<td>61.65</td>
<td>67.66</td>
<td>16.84</td>
</tr>
<tr>
<td>LF-QIH-G</td>
<td>0.5199</td>
<td>41.83</td>
<td>61.78</td>
<td>67.59</td>
<td>17.07</td>
</tr>
<tr>
<td>HRE-QH-G</td>
<td>0.5102</td>
<td>40.15</td>
<td>61.59</td>
<td>67.36</td>
<td>17.47</td>
</tr>
<tr>
<td>HRE-QIH-G</td>
<td>0.5237</td>
<td>42.29</td>
<td>62.18</td>
<td>67.92</td>
<td>17.07</td>
</tr>
<tr>
<td>HREA-QIH-G</td>
<td>0.5242</td>
<td>42.28</td>
<td>62.33</td>
<td>68.17</td>
<td>16.79</td>
</tr>
<tr>
<td>MN-QH-G</td>
<td>0.5115</td>
<td>40.42</td>
<td>61.57</td>
<td>67.44</td>
<td>17.74</td>
</tr>
<tr>
<td>MN-QIH-G</td>
<td>0.5259</td>
<td>42.29</td>
<td>62.85</td>
<td>68.88</td>
<td>17.06</td>
</tr>
<tr>
<td><strong>Discriminative</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LF-Q-D</td>
<td>0.5508</td>
<td>41.24</td>
<td>70.45</td>
<td>79.83</td>
<td>7.08</td>
</tr>
<tr>
<td>LF-QH-D</td>
<td>0.5578</td>
<td>41.75</td>
<td>71.45</td>
<td>80.94</td>
<td>6.74</td>
</tr>
<tr>
<td>LF-QI-D</td>
<td>0.5759</td>
<td>43.33</td>
<td>74.27</td>
<td>83.68</td>
<td>5.87</td>
</tr>
<tr>
<td>LF-QIH-D</td>
<td>0.5807</td>
<td>43.82</td>
<td>74.68</td>
<td>84.07</td>
<td>5.78</td>
</tr>
<tr>
<td>HRE-QH-D</td>
<td>0.5695</td>
<td>42.70</td>
<td>73.25</td>
<td>82.97</td>
<td>6.11</td>
</tr>
<tr>
<td>HRE-QIH-D</td>
<td>0.5846</td>
<td>44.67</td>
<td>74.50</td>
<td>84.22</td>
<td>5.72</td>
</tr>
<tr>
<td>HREA-QIH-D</td>
<td>0.5868</td>
<td>44.82</td>
<td>74.81</td>
<td>84.36</td>
<td>5.66</td>
</tr>
<tr>
<td>MN-QH-D</td>
<td>0.5849</td>
<td>44.03</td>
<td>75.26</td>
<td>84.49</td>
<td>5.68</td>
</tr>
<tr>
<td>MN-QIH-D</td>
<td>0.5965</td>
<td>45.55</td>
<td>76.22</td>
<td>85.37</td>
<td>5.46</td>
</tr>
<tr>
<td><strong>VQA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAN1-QI-D</td>
<td>0.5764</td>
<td>43.44</td>
<td>74.26</td>
<td>83.72</td>
<td>5.88</td>
</tr>
<tr>
<td>HieCoAtt-QI-D</td>
<td>0.5788</td>
<td>43.51</td>
<td>74.49</td>
<td>83.96</td>
<td>5.84</td>
</tr>
</tbody>
</table>
Results

• Memory Network (generally) performs best
  
  – 0.53 MRR / ~17 mean rank (Generative)

  – 0.60 MRR / ~5.5 mean rank (Discriminative)

Slide credit: Abhishek Das

torch  computer-vision  natural-language-processing  deep-learning

3 commits  1 branch  0 releases  1 contributor

Branch: master  New pull request

abhshkdz committed on GitHub Updates demo link

Latest commit 5b97b5b on Apr 19

data  Changes type to `int` for lengths; Refs #1  a month ago

decoders  Initial commit  a month ago

encoders  Initial commit  a month ago

model_utils  Initial commit  a month ago

scripts  Initial commit  a month ago

vis  Initial commit  a month ago

.gitignore  Initial commit  a month ago

README.md  Updates demo link  a month ago

dataloader.lua  Initial commit  a month ago

evaluate.lua  Initial commit  a month ago
Hi, I am a Visual Chatbot, capable of answering a sequence of questions about images. Please upload an image and fire away!
Results

Q: Is the bottle open?
A: Yes
Q: Red or white wine?
A: Red
Q: Is there only 1 glass?
A: Yes
Q: Any food?
A: No
Q: What color is the table?
A: Brown
Q: Is it in a restaurant?
A: No
Q: Is it in the kitchen?
A: Yes
Q: Is the bottle open?
A: Yes
Q: Red or white wine?
A: Red
Q: Is there only 1 glass?
A: Yes
Q: Any food?
A: No
Q: What color is the table?
A: Brown
Q: Is it in a restaurant?
A: No
Q: Is it in the kitchen?
A: Yes

Q: What color is the building?
A: Brown
Q: What color is the clock?
A: Black and white
Q: Is it a digital or analog clock?
A: Analog
Q: What color are the numbers?
A: Black
Q: Are all of the numbers visible?
A: Yes
Q: Is it an outdoor view?
A: Yes
Q: Are there any people?
A: No
Q: Can you see other buildings?
A: No
Q: What color is the cat?
A: Orange and white
Q: Is the cat on the floor?
A: Yes
Q: What is the cat on?
A: I can’t tell
Q: Can you see anything in the mirror?
A: No, just the cat and the cat
Q: Can you see a ceiling light?
A: No
Q: What color are the cat’s eyes?
A: I can’t see the eyes
Q: What color is the cat’s nose?
A: Orange
Q: What color is the cat?
A: Orange and white
Q: Is the cat on the floor?
A: Yes
Q: What is the cat on?
A: I can’t tell
Q: Can you see anything in the mirror?
A: No, just the cat and the cat
Q: Can you see a ceiling light?
A: No
Q: What color are the cat’s eyes?
A: I can’t see the eyes
Q: What color is the cat’s nose?
A: Orange
Results

Q: What color is the cat?
A: Orange and white
Q: Is the cat on the floor?
A: Yes
Q: What is the cat on?
A: I can’t tell
Q: Can you see anything in the mirror?
A: No, just the cat and the cat
Q: Can you see a ceiling light?
A: No
Q: What color are the cat’s eyes?
A: I can’t see the eyes
Q: What color is the cat’s nose?
A: Orange

Slide credit: Abhishek Das
Q: What color is the vase?
A: White

Q: Can you tell what kind of flowers?
A: I can’t tell

Q: Is there any water in the vase?
A: Yes

Q: How much water is there?
A: A small amount

Q: What is the vase on?
A: Can’t tell

Q: Does this look like someone’s home?
A: Can’t tell
Q: What color is the vase?
A: White

Q: Can you tell what kind of flowers?
A: I can’t tell

Q: Is there any water in the vase?
A: Yes

Q: How much water is there?
A: A small amount

Q: What is the vase on?
A: Can’t tell

Q: Does this look like someone’s home?
A: Can’t tell

Slide credit: Abhishek Das
Visual Dialog (CVPR 2017)

Abhishek Das (Georgia Tech)
Satwik Kottur (CMU)
Avi Singh (UC Berkeley)
Deshraj Yadav (Virginia Tech)
Khushi Gupta (CMU)
José Moura (CMU)
Devi Parikh (Georgia Tech / FAIR)
Dhruv Batra (Georgia Tech / FAIR)
Q1: What is the man wearing?
A1: Green shirt and pants

Q2: What is he doing?
A2: He's on a motorcycle.

Q3: How old is the man?
A3: Maybe in his 40s

**Caption:** A man and a dog on a motorcycle.

- **Quantitative:**
  - Ground truth response scores higher more often

- **Qualitative:**
  - Responses are more informative
  - Responses are longer
  - Responses are more diverse

---

**Diagram Details:**

- **HCIAE Encoder**
- **LSTM**
- **Gumbel Sampler**
- **Deep metric learning**

- $a_t^{gt}$: Ground truth answer
- $a_N$: Negative answer N
- $e_t$: Encoder feature
- $f()$: Embedding function

---

**Slide credit:** Jiasen Lu
Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model (arXiv)

Jiasen Lu (Virginia Tech)
Jianwei Yang (Georgia Tech)

Anitha Kannan (Facebook AI Research)
Dhruv Batra (Georgia Tech / FAIR)
Devi Parikh (Georgia Tech / FAIR)
Open directions

• Improve dialog agents via self-talk
  – No additional human intervention
  – Are these agents better at human-bot interaction?

• Domain adaptation via self-talk
  – No need to collect a new dataset for each domain

• Dialog rollouts, future prediction, theory of mind, ...
Conclusion

- Natural progression in Vision+Language
  - Captioning $\rightarrow$ VQA $\rightarrow$ Visual Dialog

- VQA: Elevating the role of image understanding
  - Balancing
  - Changing priors
  - Compositional

- Visual Dialog
  - New AI task
  - Challenges: Memory, history, reasoning over time
  - VisDial dataset
    - Live 2-person Chat on AMT
    - 120k COCO images, 1 dialog/image, $\sim$1.2 Million dialog QA pairs
  - Visual Dialog Models (Neural Encoder-Decoders)
    - Late Fusion, Hierarchical Recurrent Encoder, Memory Network
Thank you.
Visual Dialog: Towards AI agents that can see, talk, and act

Dhruv Batra
Cooperative Visual Dialog Agents

Emergence of Grounded Dialog

Negotiation Dialog Agents
Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

[ICCV ‘17]

Abhishek Das* (Georgia Tech)
Satwik Kottur* (CMU)
José Moura (CMU)
Stefan Lee (Virginia Tech)
Dhruv Batra (Georgia Tech)
Visual Dialog: Task

• Given
  – Image I
  – History of human dialog
    \( (Q_1, A_1), (Q_2, A_2), \ldots, (Q_{t-1}, A_{t-1}) \)
  – Follow-up Question \( Q_t \)

• Task
  – Produce free-form natural language answer \( A_t \)
Problems

• No goal
  – Why are we talking?

• Agent not in control
  – Artificially injected at every round into a human conversation
  – Can’t steer conversation
  – Doesn’t get to see its errors during training

• Learning equivalent utterances
  – Many ways of answering the same question that should be treated equally, but aren’t
  – Is log-likelihood of human response really a good metric?
Image Guessing Game
Image Guessing Game

Q-Bot asks questions
Image Guessing Game

Q-Bot is blindfolded
Image Guessing Game

A-Bot answers questions
Image Guessing Game

A-Bot sees an image
Image Guessing Game
Image Guessing Game

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

Q2: Any other animal?
A2: No, just zebras.

Q3: Are they facing each other?
A3: They aren't.
Image Guessing Game

Q-BOT: I think we were talking about this image!

A-BOT: They aren't.
RL for Cooperative Dialog Agents

• Agents: (Q-bot, A-bot)

• Environment: Image

• Action:
  – Q-bot: question (symbol sequence) \( q_t \)
  – A-bot: answer (symbol sequence) \( a_t \)
  – Q-bot: image regression \( y_t \in \mathbb{R}^{4096} \)

• State
  – Q-bot: \( s_t^Q = [c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}] \)
  – A-bot: \( s_t^A = [I, c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}, q_t] \)

Any people in the shot?
No, there aren’t any.
RL for Cooperative Dialog Agents

- **Action:**
  - Q-bot: question (symbol sequence)
  - A-bot: answer (symbol sequence)
  - Q-bot: image regression

- **State**
  - Q-bot: \( s^Q_t = [c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}] \)
  - A-bot: \( s^A_t = [I, c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}, q_t] \)

\( q_t \) Any people in the shot? \( a_t \) No, there aren’t any.
\( \hat{y}_t \in \mathbb{R}^{4096} \)
RL for Cooperative Dialog Agents

- **Action:**
  - Q-bot: question (symbol sequence)
  - A-bot: answer (symbol sequence)
  - Q-bot: image regression

- **State**
  - Q-bot: $s^Q_t = [c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}]$
  - A-bot: $s^A_t = [I, c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}, q_t]$

- **Policy**
  - Q-bot: $\pi_Q(q_t|S^Q_{t-1})$
  - A-bot: $\pi_A(a_t|S^A_{t-1})$

- **Reward**
  $$r_t\left(\begin{array}{c}
  s^Q_t \\
  q_t, a_t, y_t
  \end{array}\right) = \ell(\hat{y}_{t-1}, y^{gt}) - \ell(\hat{y}_t, y^{gt})$$

\[\hat{y}_t \in \mathbb{R}^{4096}\]
Policy Networks

\[ \pi_Q(q_t | S_{t-1}^Q) \]

\[ \pi_A(a_t | S_{t-1}^A) \]

Q-Bot

A-Bot

Any people in the shot?

Q_t

No, there aren't any.

A_t

Rounds of Dialog

Q-BOT

A-BOT

(C) Dhruv Batra

Slide Credit: Abhishek Das
Policy Networks

Q-Bot

\[ \pi_Q(q_t | S_{t-1}^Q) \]

A-Bot

\[ \pi_A(a_t | S_{t-1}^A) \]

Slide Credit: Abhishek Das
Policy Networks

\[ \pi_Q(q_t | S^Q_{t-1}) \]

Q-Bot

\[ \pi_A(a_t | S^A_{t-1}) \]

A-Bot

Any people in the shot?

No, there aren’t any.

\[ [0.1, -2, 0, ..., 0.57] \]

Reward Function

Slide Credit: Abhishek Das

(C) Dhruv Batra
Policy Networks

Q-Bot

\( \pi_Q(q_t | S^Q_{t-1}) \)

A-Bot

\( \pi_A(a_t | S^A_{t-1}) \)

Slide Credit: Abhishek Das
Policy Networks

Q-Bot

\[ \pi_Q(q_t | S_{t-1}^Q) \]

Any people in the shot?

QBOT

- History Encoder
- Fact Embedding
- Feature Regression Network
- Question Decoder
- Reward Function

Qt

\[ \pi_A(a_t | S_{t-1}^A) \]

\[ F_t^Q \]

\[ y_t \]

\[ [0.1, -2, 0, ..., 0.57] \]

\[ S_t^Q \]

\[ S_{t-1}^Q \]

\[ \alpha_t \]

\[ \alpha_t \]

\[ \hat{y}_t \]

\[ S_t^A \]

\[ S_{t-1}^A \]

\[ y \]

\[ F_{t-1}^A \]

\[ A-BOT \]

\[ A-BOT \]

(C) Dhruv Batra

Slide Credit: Abhishek Das
Policy Networks

Q-Bot

\[ \pi_Q(q_t | S^Q_{t-1}) \]

A-Bot

\[ \pi_A(a_t | S^A_{t-1}) \]

Q-BOT

- Question Decoder
- Fact Embedding
- Feature Regression Network

Any people in the shot?

[0.1, -2, 0, ..., 0.57]

Reward Function

Any

LSTM

\( S^Q_{t-1} \)

\( S^Q_t \)

\( \hat{y}_t \)

History Encoder

Answer Decoder

Fact Embedding

Question Encoder

History Encoder

No, there aren’t any

\( q_t \)

\( a_t \)

Slide Credit: Abhishek Das

(C) Dhruv Batra
Policy Networks

Q-Bot

\[ \pi_Q(q_t | S_{t-1}^Q) \]

A-Bot

\[ \pi_A(a_t | S_{t-1}^A) \]

Rounds of Dialog

Question Decoder

History Encoder

Fact Embedding

Feature Regression Network

Reward Function

[0.1, -2, 0, ..., 0.57]

LSTM

Any

People

Any people in the shot?

Q_t

\( y \)

\( F_t \)

\( \hat{y}_t \)

\( S_t \)

\( S_{t-1} \)

\( S_{t-1}^Q \)

\( S_t^Q \)

\( S_t^A \)

\( S_{t-1}^A \)

(C) Dhruv Batra

Slide Credit: Abhishek Das
Policy Networks

Q-Bot

\[ \pi_Q(q_t | S^Q_{t-1}) \]

[Diagram of the Q-Bot model with LSTM layers and question decoder]

Slide Credit: Abhishek Das
Policy Networks

Q-Bot

\[ \pi_Q(q_t | S^Q_{t-1}) \]

A-Bot

\[ \pi_A(a_t | S^A_{t-1}) \]

Rounds of Dialog

Q-BOT

History Encoder

Fact Embedding

Feature Regression Network

Question Decoder

Reward Function

[0.1, -2, 0, ..., 0.57]

Any people in the shot?

No, there aren’t any

\( q_t \)

\( a_t \)

Slide Credit: Abhishek Das

(C) Dhruv Batra
Policy Networks

Q-Bot
\[ \pi_Q(q_t | S_{t-1}^Q) \]

A-Bot
\[ \pi_A(a_t | S_{t-1}^A) \]

Slide Credit: Abhishek Das

(C) Dhruv Batra
Policy Networks

Q-Bot
\[ \pi_Q(q_t | S_{t-1}^Q) \]

A-Bot
\[ \pi_A(a_t | S_{t-1}^A) \]

VGG-16

Slide Credit: Abhishek Das
Policy Networks

\[ \pi_Q(q_t | S_{t-1}^Q) \]

\[ \pi_A(a_t | S_{t-1}^A) \]

Q-Bot

A-Bot

Slide Credit: Abhishek Das
Policy Networks

Q-Bot
\( \pi_Q(q_t | S_{t-1}^Q) \)

A-Bot
\( \pi_A(a_t | S_{t-1}^A) \)

Slide Credit: Abhishek Das
Two zebras are walking around their pen at the zoo.
Two zebras are walking around their pen at the zoo.

Is this zoo?
Yes

Two zebras are walking around their pen at the zoo.
Policy Networks

Q-Bot

π_Q(q_t|S_{t-1}^Q)

A-Bot

π_A(a_t|S_{t-1}^A)

Fact Embedding

LSTM

How many zebra?
Two

Is this zoo?
Yes

Two zebras are walking around their pen at the zoo.

Rounds of Dialog

Q-BOT

F^Q_t

History Encoder

Fact Embedding

Feature Regression Network

Reward Function

A-BOT

F^A_t

Question Encoder

Answer Decoder

Fact Embedding

History Encoder

Slide Credit: Abhishek Das

(C) Dhruv Batra

Policy Networks

Two zebras are walking around their pen at the zoo.

Is this zoo?
Yes

How many zebra?
Two
Policy Networks

Question-Answering System

Q-Bot

History Encoder

\( \pi_Q(q_t | s_{t-1}) \)

Policy Networks

Fact Embedding

A-Bot

\( \pi_A(a_t | s_{t-1}^A) \)

Two zebra are walking around their pen at the zoo. Is this zoo? Yes. How many zebra? Two. Two zebra are walking around their pen at the zoo.
Policy Networks

Two zebras are walking around their pen at the zoo. Is this zoo? Yes. How many zebra? Two. Two zebras are walking around their pen at the zoo.

(C) Dhruv Batra

Slide Credit: Abhishek Das
Two zebras are walking around their pen at the zoo. Is this zoo? Yes. How many zebras? Two.
Policy Networks

Q-Bot
\[ \pi_Q(q_t|S^Q_{t-1}) \]

A-Bot
\[ \pi_A(a_t|S^A_{t-1}) \]

Slide Credit: Abhishek Das
Policy Networks

Q-Bot
\[ \pi_Q(q_t | S^Q_{t-1}) \]

A-Bot
\[ \pi_A(a_t | S^A_{t-1}) \]

Slide Credit: Abhishek Das
Policy Networks

\[ \pi_Q(q_t | S_{t-1}^{Q}) \]

Q-Bot

\[ \pi_A(a_t | S_{t-1}^{A}) \]

A-Bot
Policy Networks

Q-Bot

$$\pi_Q(q_t | S^Q_{t-1})$$

A-Bot

$$\pi_A(a_t | S^A_{t-1})$$

Rounds of Dialog

Q-BOT

A-BOT

Question Decoder

Fact Embedding

History Encoder

Feature Regression Network

Reward Function

$$r_t \left( s^Q_t, (q_t, a_t, y_t) \right) = \ell \left( \hat{y}_{t-1}, y^{gt}_{t-1} \right) - \ell \left( \hat{y}_t, y^{gt}_t \right)$$

[0.1, -2, 0, ..., 0.57]
Policy Gradients

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

REINFORCE Gradients

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

\[ = \sum_{q_t, a_t} \pi_Q(q_t | s^Q_{t-1}) \nabla_{\theta_Q} \log \pi_Q(q_t | s^Q_{t-1}) \pi_A(a_t | s^A_t) r_t(\cdot) \]

\[ = \mathbb{E}_{\pi_Q, \pi_A} \left[ r_t(\cdot) \nabla_{\theta_Q} \log \pi_Q(q_t | s^Q_{t-1}) \right] \]
Turing Test

How old is this man?
Does he have short hair or long hair?
What color is his hair?
Is he wearing glasses?
Does he have facial hair?
How is he dressed?
Is he larger man or is he skinny?
Is this picture taken outside?
Can you tell if it’s daytime?
Can you tell what room he is standing in?

He looks like late teens
Medium
Brown
No
No
He’s white dress shirt with red tie ... school uniform
He’s skinny
Inside
I’m not sure
He’s in hallway

A man making the live long and prosper sign from star trek.

Where is man located?
What is man wearing?
Are there any other people in picture?
What color is table?
Is there anything else on table?
What are other people doing?
Are there any windows?
What color are walls?
What time of day do you think it is?
What is man wearing?

Looks like classroom of some sort
Black t shirt and jeans
Yes, quite few people in background
Picture is black and white, but it’s wood table
Not that I can see
They are standing in front of him, but I don’t ...
Not that I can see
I can’t see walls
I can’t tell
Black t shirt and jeans

(C) Dhruv Batra
### Human-Human dialog [4]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A couple of people in the snow on skis.</td>
<td><strong>How old are men?</strong></td>
<td><strong>Are people male or female?</strong></td>
<td><strong>Are people male or female?</strong></td>
</tr>
<tr>
<td></td>
<td>They look to be 30's</td>
<td>Male or female?</td>
<td>I think I am male, can’t see other 1...</td>
</tr>
<tr>
<td></td>
<td><strong>Are they in restaurant?</strong></td>
<td>I can’t tell, in close up</td>
<td><strong>What are they wearing?</strong></td>
</tr>
<tr>
<td></td>
<td>They look to be relaxed</td>
<td><strong>Is it snowing?</strong></td>
<td>They are all wearing snow pants and jackets</td>
</tr>
<tr>
<td></td>
<td><strong>Are they each holding a slice of pizza in their hands and giving them thumbs up with other?</strong></td>
<td>Yes</td>
<td><strong>Is it snowing?</strong></td>
</tr>
<tr>
<td></td>
<td>One holding slice, other holding box...</td>
<td><strong>I can’t tell, is it snowing?</strong></td>
<td>It does not appear to be snowing at the moment</td>
</tr>
<tr>
<td></td>
<td><strong>What kind of pizza do they have?</strong></td>
<td>Cheese</td>
<td><strong>Can you see any trees?</strong></td>
</tr>
<tr>
<td></td>
<td>Does box have pizza company name on it?</td>
<td>No</td>
<td>Yes, there are lots of trees in the background</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td><strong>Are there any toppings on pizza?</strong></td>
<td>No, I can not see any trees in the background</td>
</tr>
<tr>
<td></td>
<td><strong>What color are their shorts?</strong></td>
<td>Black</td>
<td><strong>Can you see any other people around?</strong></td>
</tr>
<tr>
<td></td>
<td>Black with red letters</td>
<td><strong>Are there any other people visible?</strong></td>
<td>Yes, there are lots of leaves in the background</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td><strong>Is it sky or night?</strong></td>
<td>Yes, I can not see any other people around</td>
</tr>
<tr>
<td></td>
<td><strong>What color are they wearing?</strong></td>
<td>White</td>
<td><strong>Can you see any buildings in the background?</strong></td>
</tr>
<tr>
<td>Two men giving thumbs up holding large slices of pizza.</td>
<td><strong>How old is man?</strong></td>
<td><strong>Are there any trees?</strong></td>
<td>No, I do not see any buildings in the background</td>
</tr>
<tr>
<td></td>
<td>In his teens</td>
<td><strong>Are there any trees?</strong></td>
<td><strong>Can you see any buildings in the background?</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Is he wearing helmet?</strong></td>
<td>Yes</td>
<td>No, I do not see any buildings in the background</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td><strong>Are they inside or outside?</strong></td>
<td><strong>Can you see any trees?</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Is he wearing any protective gear?</strong></td>
<td>No</td>
<td>Yes, there are lots of trees in the background</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td><strong>Is it daytime?</strong></td>
<td>No, I do not see any buildings in the background</td>
</tr>
<tr>
<td></td>
<td><strong>What color is his hair?</strong></td>
<td>Dirty blonde or dark brown</td>
<td><strong>Can you see any buildings in the background?</strong></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td><strong>Can you see any trees?</strong></td>
<td>No, I do not see any buildings in the background</td>
</tr>
<tr>
<td></td>
<td><strong>What color is his shirt?</strong></td>
<td>Black</td>
<td><strong>Can you see any trees?</strong></td>
</tr>
<tr>
<td></td>
<td>Black with red letters</td>
<td><strong>Are there any other people visible?</strong></td>
<td>Yes, there are lots of leaves in the background</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td><strong>Is it sky or night?</strong></td>
<td><strong>Can you see any buildings in the background?</strong></td>
</tr>
<tr>
<td></td>
<td><strong>What color are their shorts?</strong></td>
<td>Black</td>
<td>No, I do not see any buildings in the background</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td><strong>Can you see any trees?</strong></td>
<td><strong>Can you see any buildings in the background?</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Are there any other people visible?</strong></td>
<td>No</td>
<td>No, I do not see any buildings in the background</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td><strong>Is it sky or night?</strong></td>
<td><strong>Can you see any buildings in the background?</strong></td>
</tr>
<tr>
<td></td>
<td><strong>What color are their shorts?</strong></td>
<td>Black</td>
<td>No, I do not see any buildings in the background</td>
</tr>
<tr>
<td></td>
<td>I can’t see them</td>
<td><strong>What color is their shirt?</strong></td>
<td><strong>Can you see any buildings in the background?</strong></td>
</tr>
<tr>
<td>A man riding a skateboard down the side of a ramp.</td>
<td><strong>How old is man?</strong></td>
<td><strong>What color is his shirt?</strong></td>
<td>No, I do not see any buildings in the background</td>
</tr>
<tr>
<td></td>
<td>I can’t see his face, I don’t know</td>
<td>White</td>
<td><strong>Can you see any buildings in the background?</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Is he wearing helmet?</strong></td>
<td>Yes</td>
<td><strong>Can you see any trees?</strong></td>
</tr>
<tr>
<td></td>
<td>No</td>
<td><strong>Is it sky or night?</strong></td>
<td>Yes, it is sunny</td>
</tr>
<tr>
<td></td>
<td><strong>Is he wearing any protective gear?</strong></td>
<td>No</td>
<td><strong>What color is their shirt?</strong></td>
</tr>
<tr>
<td></td>
<td>No</td>
<td><strong>Is it sky or night?</strong></td>
<td><strong>What color is their shirt?</strong></td>
</tr>
<tr>
<td></td>
<td><strong>What color is his hair?</strong></td>
<td>Dirty blonde or dark brown</td>
<td><strong>What color is their shirt?</strong></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td><strong>Is it sky or night?</strong></td>
<td><strong>What color is their shirt?</strong></td>
</tr>
<tr>
<td></td>
<td><strong>What color is his shirt?</strong></td>
<td>Black with red letters</td>
<td><strong>What color is their shirt?</strong></td>
</tr>
<tr>
<td></td>
<td>Black with red letters</td>
<td><strong>Is it sky or night?</strong></td>
<td><strong>What color is their shirt?</strong></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td><strong>Is it sky or night?</strong></td>
<td><strong>What color is their shirt?</strong></td>
</tr>
<tr>
<td></td>
<td><strong>What color are their shorts?</strong></td>
<td>Black</td>
<td><strong>What color is their shirt?</strong></td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td><strong>Is it sky or night?</strong></td>
<td><strong>What color is their shirt?</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Are there any other people visible?</strong></td>
<td>No</td>
<td><strong>What color is their shirt?</strong></td>
</tr>
<tr>
<td></td>
<td>No</td>
<td><strong>Is it sky or night?</strong></td>
<td><strong>What color is their shirt?</strong></td>
</tr>
<tr>
<td></td>
<td><strong>What color are their shorts?</strong></td>
<td>Black</td>
<td><strong>What color is their shirt?</strong></td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td><strong>Is it sky or night?</strong></td>
<td><strong>What color is their shirt?</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Is this at skate park?</strong></td>
<td>Yes</td>
<td><strong>What color is their shirt?</strong></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td><strong>Is it sky or night?</strong></td>
<td><strong>What color is their shirt?</strong></td>
</tr>
<tr>
<td>An airplane sitting on an asphalt landing strip.</td>
<td><strong>What color is plane?</strong></td>
<td><strong>What color is plane?</strong></td>
<td><strong>What color is airplane?</strong></td>
</tr>
<tr>
<td></td>
<td>White and blue</td>
<td>White</td>
<td>White with red and blue stripes</td>
</tr>
<tr>
<td></td>
<td><strong>Is there writing on plane?</strong></td>
<td>Yes it says southwest</td>
<td>Yes, it is too far away</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td><strong>Time of day is?</strong></td>
<td><strong>What color is airplane?</strong></td>
</tr>
<tr>
<td></td>
<td><strong>What color is ground?</strong></td>
<td>Grey tarmac</td>
<td><strong>It is daytime</strong></td>
</tr>
<tr>
<td></td>
<td>Grey</td>
<td><strong>Can you see any people?</strong></td>
<td>No, it is too far away</td>
</tr>
<tr>
<td></td>
<td><strong>Any people?</strong></td>
<td>No people visible</td>
<td><strong>What time of day is it?</strong></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td><strong>Can you see any people?</strong></td>
<td><strong>It is daytime</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Can you see airport?</strong></td>
<td>Yes</td>
<td><strong>Can you see any people?</strong></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td><strong>Can you see any buildings?</strong></td>
<td>Yes, in background</td>
</tr>
<tr>
<td></td>
<td><strong>What color is it?</strong></td>
<td>Grey</td>
<td><strong>Can you see any people?</strong></td>
</tr>
<tr>
<td></td>
<td>Grey</td>
<td><strong>Are there clouds in sky?</strong></td>
<td>Yes, few</td>
</tr>
<tr>
<td></td>
<td><strong>Is there sunshine?</strong></td>
<td>No it’s kind of overcast</td>
<td><strong>Does it look like it’s going to rain?</strong></td>
</tr>
<tr>
<td></td>
<td>No</td>
<td><strong>Are there clouds in sky?</strong></td>
<td>Yes, it does</td>
</tr>
<tr>
<td></td>
<td><strong>Any other planes?</strong></td>
<td>No other planes</td>
<td><strong>Does it look like it’s going to rain?</strong></td>
</tr>
<tr>
<td></td>
<td>No</td>
<td><strong>What does it say?</strong></td>
<td>Yes, it does</td>
</tr>
<tr>
<td></td>
<td><strong>Any grass?</strong></td>
<td>No grass</td>
<td><strong>Does it look like it’s going to rain?</strong></td>
</tr>
<tr>
<td></td>
<td>No</td>
<td><strong>Is it rain?</strong></td>
<td>Yes, it does</td>
</tr>
<tr>
<td></td>
<td><strong>Any trees?</strong></td>
<td>No tree</td>
<td><strong>Does it look like it’s going to rain?</strong></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td><strong>Is it rain?</strong></td>
<td>Yes, it does</td>
</tr>
<tr>
<td></td>
<td><strong>Any writing on it?</strong></td>
<td>Any writing on plane</td>
<td><strong>Does it look like it’s going to rain?</strong></td>
</tr>
<tr>
<td></td>
<td>White</td>
<td><strong>Is it rain?</strong></td>
<td>Yes, it does</td>
</tr>
</tbody>
</table>

(C) Dhruv Batra
How old is man?
What is he wearing?
What color is his shirt?
what color are his pants?
what color is his shirt?
What color are his shoes?
What color are his shoes?
What color are his shoes?

SL Agents
I can’t see his face, I don’t know
shorts and t shirt
white
black
blue
Black
Black
Black
Black
White color is skateboard?
Is he wearing helmet?
Is he wearing knee pads?
Is it day or night?
Is it sunny?
Can you see any other people?
Is it day or night?
Is it sunny?
What color is skateboard?
Is he wearing helmet?

RL Agents
It is hard to tell but I think it’s black
Yes
No, no knee pads
It is daytime
It is bright out, but I see no shadows, so I can not
There are 2 other people in background
It is daytime
Yes, it is sunny
It is hard to tell but I think it’s black
Yes, he is wearing helmet
Image Guessing

Percentile vs. Dialog Round

RL-full-QAf
SL-pretrained
Concurrent Work

GuessWhat?! Visual object discovery through multi-modal dialogue

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, Aaron Courville

(Submitted on 23 Nov 2016 (v1), last revised 6 Feb 2017 (this version, v2))

We introduce GuessWhat?! a two-player guessing game as a testbed for research on the the game is to locate an unknown object in a rich image scene by asking a sequence of questions reasoning and language grounding, is required to solve the proposed task. Our key contribution is to create a dataset of 150K human-played games with a total of 800K visual question-answer pairs on 66K images and introduce the oracle and questioner tasks that are associated with the two players of initial baselines of the introduced tasks.

End-to-end optimization of goal-driven and visually grounded dialogue systems

Florian Strub, Harm de Vries, Jeremie Mary, Bilal Piot, Aaron Courville, Olivier Pietquin

(Submitted on 15 Mar 2017)

End-to-end design of dialogue systems has recently become a popular research topic thanks to powerful tools such as encoder-decoder architectures for sequence-to-sequence learning. Yet, most current approaches cast human-machine dialogue management as a supervised learning problem, aiming at predicting the next utterance of a participant given the full history of the dialogue. This vision is too simplistic to render the intrinsic planning problem inherent to dialogue as well as its grounded nature, making the context of a dialogue larger than the sole history. This is why only chit-chat and question answering tasks have been addressed so far using end-to-end architectures. In this paper, we introduce a Deep Reinforcement Learning method to optimize visually grounded task-oriented dialogues, based on the policy gradient algorithm. This approach is tested on a dataset of 120K dialogues collected through Mechanical Turk and provides encouraging results at solving both the problem of generating natural dialogues and the task of discovering a specific object in a complex picture.

(C) Dhruv Batra
Outline

Cooperative Visual Dialog Agents

Emergence of Grounded Dialog

Negotiation Dialog Agents
Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog

[EMNLP ‘17]

Satwik Kottur*
(CMU)

José Moura
(CMU)

Stefan Lee
(Virginia Tech)

Dhruv Batra
(Georgia Tech)
Toy World

- Sanity check

- Simple, synthetic world
  - Instances - (shape, color, style)
  - Total of $4^3(64)$ instances

- Example instances:
Task & Talk

• Task (G)
  • Inquire pair of attributes
  • (color, shape), (shape, color)

• Talk
  • Single token per round
  • Two rounds

• Q-bot guesses a pair
  • Reward : +1 / -1
  • Prediction order matters!

Instance
(purple, square, filled)

Q: Y
A: 2

Q: Z
A: 3

Guess: (purple, square)

Get reward!
Emergence of Grounded Dialog

T: (style, color)
P: (solid, green)

X
color?

3
green

Z
style?

4
solid

T: (style, shape)
P: (filled, triangle)

Y
shape?

1
triangle

Z
style?

2
filled

(C) Dhruv Batra
Emergence of Grounded Dialog

- Compositional grounding
- Predict dialog for unseen instances

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Task (color, shape)</th>
<th>q1, q2</th>
</tr>
</thead>
<tbody>
<tr>
<td>color</td>
<td>(color, shape)</td>
<td>Y, X</td>
</tr>
<tr>
<td>shape</td>
<td>(shape, color)</td>
<td>Y, X</td>
</tr>
<tr>
<td>style</td>
<td>(shape, style)</td>
<td>Y, Z</td>
</tr>
<tr>
<td>color</td>
<td>(color, style)</td>
<td>Z, X</td>
</tr>
<tr>
<td>style</td>
<td>(style, color)</td>
<td>X, Z</td>
</tr>
</tbody>
</table>

(a) A-BOT
(b) Q-BOT

Q1: Y
Q2: X
Q1: 2
Q2: 2
## Summary of findings

<table>
<thead>
<tr>
<th>Setting</th>
<th>Vocabulary</th>
<th>Memory</th>
<th>Generalization</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$</td>
<td>V_Q</td>
<td>$</td>
<td>$</td>
</tr>
<tr>
<td>A. Over-complete</td>
<td>64</td>
<td>64</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. Attribute</td>
<td>3</td>
<td>12</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Minimal</td>
<td>3</td>
<td>4</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Deep Multi-Agent Communication

• NIPS ‘16

• Arxiv ‘17
  – Learning to play guess who? and inventing a grounded language as a consequence. Emilio Jorge, Mikael Kågebäck, and Emil Gustavsson.
  – Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. Serhii Havrylov and Ivan Titov.
Outline

Cooperative Visual Dialog Agents

Emergence of Grounded Dialog

Task (color, shape)
Q1: Y
Q2: X
Q1: 2
Q2: 2

Negotiation Dialog Agents

Q1: I'd like the ball and hats
Q2: I need the hats, you can have the ball
Q1: Ok, if I get both books?
Q2: Ok, deal

2 points each
1 point each
5 points each

0 points each
3 point each
1 points each
Deal or No Deal? End-to-End Learning for Negotiation Dialogues

[EMNLP ‘17]

Mike Lewis (FAIR)
Denis Yarats (FAIR)
Yann Dauphin (FAIR)
Devi Parikh (Georgia Tech)
Dhruv Batra (Georgia Tech)
Why Negotiation?

Adversarial  Negotiation  Cooperative

Slide Credit: Mike Lewis
Why Negotiation?

Negotiation useful when:

- Agents have different goals
- Not all can be achieved at once
- (all the time)
Why Negotiation?

• Both **linguistic** and **reasoning** problem

• *Interpret* multiple sentences, and *generate* new message

• Plan ahead, make proposals, counter-offers, bluffing, lying, compromising

Slide Credit: Mike Lewis
Both agents given *reward function*, can’t observe each other’s goals.

Both agents *independently* select agreement.

Agent 1 Goals -> Dialogue

Agent 2 Goals -> Dialogue

Dialogue until they *agree* on common action

Agent 1 Output -> Agent 1 Reward

Agent 2 Output -> Agent 2 Reward

If agents agree, they are given reward

Slide Credit: Mike Lewis
Object Division Task

Agents shown *same* set of objects but *different* values for each.

Asked to agree how to divide objects between them:

- 2 points each
- 1 point each
- 5 points each

Slide Credit: Mike Lewis
Multi-Issue Bargaining

I'd like the ball and hats

I need the hats, you can have the ball

Ok, if I get both books?

Ok, deal

Slide Credit: Mike Lewis
Data Collection on AMT

Divide these items between you and your partner.

Your partner sees the same items but with different values.
You get some items, and your partner will get the rest.

*If you get a great deal for you then we will pay a bonus!*
*If you often get low scores then your work may be rejected.*

<table>
<thead>
<tr>
<th>Items to Split between You and Partner</th>
<th>Value Each to You</th>
<th>Number You Get</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hat</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Basketball</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Fellow Turk connected. Please send a message!

Type Message Here:

Message

Send

No deal was agreed.
Dataset

- ~6k dialogs
- Average 6.6 turns/dialog
- Average 7.6 words/turn
- 80% agreed solutions
- 77% Pareto Optimal solutions
Baseline Model

Language model predicts both agent’s tokens

<write> Give me both books <read> ok deal

Input Encoder

Output Decoder

Attention over complete dialogue

Separate classifier for each output

Read input at each timestep

Slide Credit: Mike Lewis
SL-Pretraining

• Train to maximize likelihood of human-human dialogues

• Decode by sampling likely messages
SL-Pretraining

- Model knows nothing about task, just tries to imitate human actions
- Agrees too easily
- Can’t go beyond human strategies

Slide Credit: Mike Lewis
Goal-based RL-Finetuning

- Very sensitive to hyperparameters

- Generate dialogues using self-play

- Backpropagate reward using REINFORCE

- Interleave with supervised updates

Slide Credit: Mike Lewis
Dialog Rollouts: Goal-based Decoding

- Dialog rollouts use model to simulate remainder of conversation
- Average scores to estimate future reward
Intrinsic Evaluation

Supervised learning gives most “human like” dialog.
End-to-End Evaluation against SL negotiators

Slide Credit: Mike Lewis
End-to-End Evaluation against Turkers

Relative Score (all) vs. Relative Score (agreed)

% Agreed vs. % Pareto Optimal

Slide Credit: Mike Lewis
I can not make that deal. I need the ball and book, you can have the hats.

Can I have the hats and book?

I can not make that deal. I need the ball and book, you can have the hats.

I need the book and 2 hats.

No deal then.

No deal doesn’t work for me sorry.

How about I give you the book and I keep the rest.

Sorry, I want the book and one hat.

I need the book and hats.

Can I have the hats and book?

Model generates meaningful novel language.
That would work for me. I can take the ball and 1 hat
I would like the ball and two hats
I need the book and 3 hats
That would work for me. I can take the ball and 1 hat

Model can be deceptive to achieve its goals
Conclusion

- Negotiation is **useful** and **challenging**

- End-to-End approach trades cheaper data for difficult modelling

- Goal-based training and decoding improves over likelihood

- Model can generate meaningful language be be deceptive to achieve their goals
Outline

Cooperative Visual Dialog Agents

Emergence of Grounded Dialog

Negotiation Dialog Agents
Sneak Peek:

Inner Dialog:
Pragmatic Visual Dialog Agents that Rollout a Mental Model of their Interlocutors
Inner Dialog

A-ROBOT

Two zebra are walking around their pen at the zoo.

Any people in the shot?

No, there aren’t any.

O-ROBOT

What color is the grass?

It looks green.

How many zebra?

Two.

(C) Dhruv Batra
What next?

• So far
  – Vision + Language
    • Captioning $\rightarrow$ VQA $\rightarrow$ Visual Dialog

• Interacting with an intelligent agent
  – Perceive + Communicate + Act
  – Vision + Language + Reinforcement Learning
  – Ok Google – can you find my picture where I was wearing this red shirt? And order me a new one?
There is a cloud on the left side of the frame all the way at the top. The rest of the sky is clear.

What is located in the grass?

There is a tent on the right side sitting on the grass with a bear behind it on the left side.

What else is on the grass?

There is a little boy sitting with his legs straight on the right side of the tent.

Is there anything else?

Yes, there is a red grill to the right of the bear and a little girl sitting on the right side of the tent.

If the tent is on the right side there is no room for the kids? Is the tent on the right side of the picture?

I am so sorry the grill is on the left of the bear. Yes, the little boy is on the bottom right of the picture.

Where is the little girl?

He is in front of the tent. The tent is in the middle of the grass on the right side.

Is that all? Are we goo?

She is on the left side of the tent, she is crying with her legs straight.

Great! Are we good?

Yes, we are good.

— Fin. —
Agents in Virtual Environments

AI2 Thor

SUNCG dataset
What next?

• So far
  – Vision + Language
    • Captioning $\rightarrow$ VQA $\rightarrow$ Visual Dialog

• Interacting with an intelligent agent
  – Perceive + Communicate + Act
  – Vision + Language + Reinforcement Learning
  – Ok Google – can you find my picture where I was wearing this red shirt? And order me a new one?
What next?

• So far
  – Vision + Language
    • Captioning → VQA → Visual Dialog

• Interacting with an intelligent agent
  – Perceive + Communicate + Act
  – Vision + Language + Reinforcement Learning
  – Ok Google – can you find my picture where I was wearing this red shirt? And order me a new one?

• Teaching with natural language
  – ”No, not that shirt. *This* one.”
ParlAI: A Dialog Research Software Platform

Alexander H. Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, Jason Weston

(Submitted on 18 May 2017)

We introduce ParlAI (pronounced "par-lay"), an open-source software platform for dialog research implemented in Python, available at this http URL. Its goal is to provide a unified framework for training and testing of dialog models, including multitask training, and integration of Amazon Mechanical Turk for data collection, human evaluation, and online/reinforcement learning. Over 20 tasks are supported in the first release, including popular datasets such as SQuAD, bAbi tasks, MCTest, WikiQA, QACNN, QADailyMail, CBT, bAbi Dialog, Ubuntu, OpenSubtitles and VQA. Included are examples of training neural models with PyTorch and Lua Torch, including both batch and hogwild training of memory networks and attentive LSTMs.

ParlAI (pronounced “parlay”)
A Dialogue Dataset “Universe”

<table>
<thead>
<tr>
<th>QA datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>bAbi tasks</td>
</tr>
<tr>
<td>MCTest</td>
</tr>
<tr>
<td>SQuAD, NewsQA, MS MARCO</td>
</tr>
<tr>
<td>SimpleQuestions</td>
</tr>
<tr>
<td>WebQuestions, WikiQA</td>
</tr>
<tr>
<td>WikiMovies, MTurkWikiMovies</td>
</tr>
<tr>
<td>MovieDD (Movie-Recommendations)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dialogue Goal-Oriented</th>
</tr>
</thead>
<tbody>
<tr>
<td>bAbi Dialog tasks</td>
</tr>
<tr>
<td>Camrest</td>
</tr>
<tr>
<td>Dialog-based Language Learning</td>
</tr>
<tr>
<td>MovieDD (QA,Recs dialogue)</td>
</tr>
<tr>
<td>CommAI-env</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sentence Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>QACNN</td>
</tr>
<tr>
<td>QADailyMail</td>
</tr>
<tr>
<td>CBT</td>
</tr>
<tr>
<td>BookTest</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dialogue Chit-Chat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ubuntu multiple-choice</td>
</tr>
<tr>
<td>UbuntuGeneration</td>
</tr>
<tr>
<td>Movies SubReddit</td>
</tr>
<tr>
<td>Reddit</td>
</tr>
<tr>
<td>Twitter</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VQA/Visual Dialog</th>
</tr>
</thead>
<tbody>
<tr>
<td>VQA already in v1.0.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Add your own dataset!</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open source...</td>
</tr>
</tbody>
</table>
Computer Vision Lab

Devi Parikh  
Assistant Professor

Xiao Lin  
Ph.D. Student

Arjun Chandrasekaran  
Ph.D. Student

Ramakrishna Vedantam  
Ph.D. Student

Peng Zhang  
Ph.D. Student

Jiasen Lu  
Ph.D. Student

Ram Prasaath Selvaraju  
Ph.D. Student

Jianwei Yang  
Ph.D. Student

Arijit (Arren) Ray  
M.S. Student

Prithvijit Chattopadhyay  
Intern
Thanks!