Visual Dialog

Abhishek Das[1], Satwik Kottur[2], Khushi Gupta[2]*, Avi Singh[3]*, Deshraj Yadav[4], José M.F. Moura[2], Devi Parikh[1], Dhruv Batra[1]
[1]Georgia Institute of Technology, [2]Carnegie Mellon University, [3]UC Berkeley, [4]Virginia Tech
[1]{abhshkdz, parikh, dbatra}@gatech.edu [2]{skottur, khushig, moura}@andrew.cmu.edu
[3]avisingh@cs.berkeley.edu [4]deshraj@vt.edu
visualdialog.org

Abstract

We introduce the task of Visual Dialog, which requires an AI agent to hold a meaningful dialog with humans in natural, conversational language about visual content. Specifically, given an image, a dialog history, and a question about the image, the agent has to ground the question in image, infer context from history, and answer the question accurately. Visual Dialog is disentangled enough from a specific downstream task so as to serve as a general test of machine intelligence, while being grounded in vision enough to allow objective evaluation of individual responses and benchmark progress. We develop a novel two-person chat data-collection protocol to curate a large-scale Visual Dialog dataset (VisDial). VisDial contains 1 dialog (10 question-answer pairs) on ∼140k images from the COCO dataset, with a total of ∼1.4M dialog question-answer pairs.

We introduce a family of neural encoder-decoder models for Visual Dialog with 3 encoders (Late Fusion, Hierarchical Recurrent Encoder and Memory Network) and 2 decoders (generative and discriminative), which outperform a number of sophisticated baselines. We propose a retrieval-based evaluation protocol for Visual Dialog where the AI agent is asked to sort a set of candidate answers and evaluated on metrics such as mean-reciprocal-rank of human response. We quantify gap between machine and human performance on the Visual Dialog task via human studies. Our dataset, code, and trained models will be released publicly at visualdialog.org. Putting it all together, we demonstrate the first 'visual chatbot'!

1. Introduction

We are witnessing unprecedented advances in computer vision (CV) and artificial intelligence (AI) – from 'low-level' AI tasks such as image classification [17], scene recognition [57], object detection [29] – to 'high-level' AI tasks

Figure 1: We introduce a new AI task – Visual Dialog, where an AI agent must hold a dialog with a human about visual content. We introduce a large-scale dataset (VisDial), an evaluation protocol, and novel encoder-decoder models for this task.

such as learning to play Atari video games [36] and Go [49], answering reading comprehension questions by understanding short stories [18, 59], and even answering questions about images [4, 34, 43, 64] and videos [51, 52]!

What lies next for AI? We believe that the next generation of visual intelligence systems will need to posses the ability to hold a meaningful dialog with humans in natural language about visual content. Applications include:

- Aiding visually impaired users in understanding their surroundings [5] or social media content [60] (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?'*),
- Interacting with an AI assistant (Human: *'Alexa – can*

*Work done while KG and AS were interns at Virginia Tech.

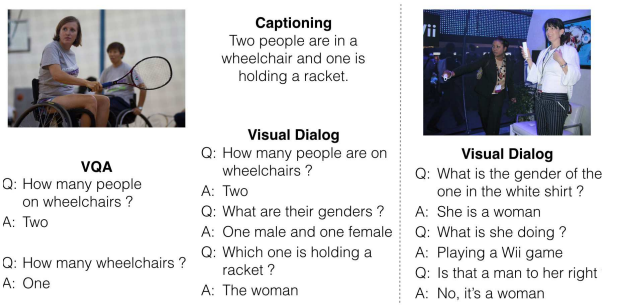


Figure 2: Differences between image captioning, Visual Question Answering (VQA) and Visual Dialog. Two (partial) dialogs are shown from our VisDial dataset, which is curated from a live chat between two Amazon Mechanical Turk workers (Sec. 3).

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 missions) where the operator may be 'situationally blind' and operating via language [35] (Human: *'Is there smoke in any room around you?'*, AI: *'Yes, in one room'*, Human: *'Go there and look for people'*).

Despite rapid progress at the intersection of vision and language – in particular, in image captioning and visual question answering (VQA) – it is clear that we are far from this grand goal of an AI agent that can 'see' and 'communicate'. In captioning, the human-machine interaction consists of the machine simply *talking at* the human (*'Two people are in a wheelchair and one is holding a racket'*), with no dialog or input from the human. While VQA takes a significant step towards human-machine interaction, it still represents only *a single round of a dialog* – unlike in human conversations, there is no scope for follow-up questions, no memory in the system of previous questions asked by the user nor consistency with respect to previous answers provided by the system (Q: *'How many people on wheelchairs?'*, A: *'Two'*; Q: *'How many wheelchairs?'*, A: *'One'*).

As a step towards conversational visual AI, we introduce a novel task – **Visual Dialog** – along with a large-scale dataset, an evaluation protocol, and novel deep models.

Task Definition. The concrete task in Visual Dialog is the following – given an image I, a history of a dialog consisting of a sequence of question-answer pairs (Q1: *'How many people are in wheelchairs?'*, A1: *'Two'*, Q2: *'What are their genders?'*, A2: *'One male and one female'*), and a natural language follow-up question (Q3: *'Which one is holding a racket?'*), the task for the machine is to answer the question in free-form natural language (A3: *'The woman'*). This task is the visual analogue of the Turing Test.

Consider the Visual Dialog examples in Fig. 2. The question *'What is the gender of the one in the white shirt?'* requires the machine to selectively focus and direct attention to a relevant region. *'What is she doing?'* requires co-reference resolution (whom does the pronoun 'she' refer to?), *'Is that a man to her right?'* further requires the machine to have visual memory (which object in the image were we talking about?). Such systems also need to be consistent with their outputs – *'How many people are in wheelchairs?'*, *'Two'*, *'What are their genders?'*, *'One male and one female'* – note that the number of genders being specified should add up to two. Such difficulties make the problem a highly interesting and challenging one.

Why do we talk to machines? Prior work in language-only (non-visual) dialog can be arranged on a spectrum with the following two end-points:
goal-driven dialog (*e.g.* booking a flight for a user) $\longleftrightarrow$ goal-free dialog (or casual 'chit-chat' with chatbots).
The two ends have vastly differing purposes and conflicting evaluation criteria. Goal-driven dialog is typically evaluated on task-completion rate (how frequently was the user able to book their flight) or time to task completion [11, 38] – clearly, the shorter the dialog the better. In contrast, for chit-chat, the longer the user engagement and interaction, the better. For instance, the goal of the 2017 $2.5 Million Amazon Alexa Prize is to "create a socialbot that converses coherently and engagingly with humans on popular topics for 20 minutes."

We believe our instantiation of Visual Dialog hits a sweet spot on this spectrum. It is *disentangled enough* from a specific downstream task so as to serve as a general test of machine intelligence, while being *grounded enough* in vision to allow objective evaluation of individual responses and benchmark progress. The former discourages task-engineered bots for 'slot filling' [25] and the latter discourages bots that put on a personality to avoid answering questions while keeping the user engaged [58].

Contributions. We make the following contributions:

- We propose a new AI task: Visual Dialog, where a machine must hold dialog with a human about visual content.
- We develop a novel two-person chat data-collection protocol to curate a large-scale Visual Dialog dataset (VisDial). Upon completion[1], VisDial will contain 1 dialog each (with 10 question-answer pairs) on ~140k images from the COCO dataset [27], for a total of ~1.4M dialog question-answer pairs. When compared to VQA [4], VisDial studies a significantly richer task (dialog), overcomes a 'visual priming bias' in VQA (in VisDial, the questioner does not see the image), contains free-form longer answers, and is *an order of magnitude* larger.
- We introduce a family of neural encoder-decoder models

[1] VisDial data on COCO-`train` (~83k images) and COCO-`val` (~40k images) is already available for download at `https://visualdialog.org`. Since dialog history contains the ground-truth caption, we will not be collecting dialog data on COCO-`test`. Instead, we will collect dialog data on 20k extra images from COCO distribution (which will be provided to us by the COCO team) for our `test` set.

for Visual Dialog with 3 novel encoders
  - Late Fusion: that embeds the image, history, and question into vector spaces separately and performs a 'late fusion' of these into a joint embedding.
  - Hierarchical Recurrent Encoder: that contains a dialog-level Recurrent Neural Network (RNN) sitting on top of a question-answer (QA)-level recurrent block. In each QA-level recurrent block, we also include an attention-over-history mechanism to choose and attend to the round of the history relevant to the current question.
  - Memory Network: that treats each previous QA pair as a 'fact' in its memory bank and learns to 'poll' the stored facts and the image to develop a context vector.

  We train all these encoders with 2 decoders (generative and discriminative) – all settings outperform a number of sophisticated baselines, including our adaption of state-of-the-art VQA models to VisDial.
- We propose a retrieval-based evaluation protocol for Visual Dialog where the AI agent is asked to sort a list of candidate answers and evaluated on metrics such as mean-reciprocal-rank of the human response.
- We conduct studies to quantify human performance on this task.
- Putting it all together, on the project page we demonstrate the first visual chatbot!

2. Related Work

Vision and Language. A number of problems at the intersection of vision and language have recently gained prominence – image captioning [12, 13, 23, 56], video/movie description [45, 53, 54], text-to-image coreference/grounding [8, 19, 24, 39, 41, 44], visual storytelling [2, 20], and of course, visual question answering (VQA) [2, 4, 9, 14, 16, 32–34, 43, 62]. However, all of these involve (at most) a single-shot natural language interaction – there is no dialog. Concurrent with our work, two recent works [10, 37] have also begun studying this problem of visually-grounded dialog.

Visual Turing Test. Closely related to our work is that of Geman *et al*. [15], who proposed a fairly restrictive 'Visual Turing Test' – a system that asks templated, binary questions. In comparison, 1) our dataset has *free-form, open-ended* natural language questions collected via two subjects chatting on Amazon Mechanical Turk (AMT), resulting in a more realistic and diverse dataset (see Fig. 5). 2) The dataset in [15] only contains street scenes, while our dataset has considerably more variety since it uses images from COCO [27]. Moreover, our dataset is *two orders of magnitude larger* – 2,591 images in [15] vs ∼140k images, 10 question-answer pairs per image, total of ∼1.4M QA pairs.

Text-based Question Answering. Our work is related to text-based question answering or 'reading comprehension' tasks studied in the NLP community. Some recent large-scale datasets in this domain include the 30M Factoid Question-Answer corpus [46], 100K SimpleQuestions dataset [6], DeepMind Q&A dataset [18], the 20 artificial tasks in the bAbI dataset [59], and the SQuAD dataset for reading comprehension [40]. VisDial can be viewed as a *fusion* of reading comprehension and VQA. In VisDial, the machine must comprehend the history of the past dialog and then understand the image to answer the question. By design, the answer to any question in VisDial is not present in the past dialog – if it were, the question would not be asked. The history of the dialog *contextualizes* the question – the question *'what else is she holding?'* requires a machine to comprehend the history to realize who the question is talking about and what has been excluded, and then understand the image to answer the question.

Conversational Modeling and Chatbots. Visual Dialog is the visual analogue of text-based dialog and conversation modeling. While some of the earliest developed chatbots were rule-based [58], end-to-end learning based approaches are now being actively explored [7, 11, 22, 26, 47, 48, 55]. A recent large-scale conversation dataset is the Ubuntu Dialogue Corpus [30], which contains about 500K dialogs extracted from the Ubuntu channel on Internet Relay Chat (IRC). Liu *et al*. [28] perform a study of problems in existing evaluation protocols for free-form dialog. One important difference between free-form textual dialog and VisDial is that in VisDial, the two participants are not symmetric – one person (the 'questioner') asks questions about an image *that they do not see*; the other person (the 'answerer') sees the image and only answers the questions (in otherwise unconstrained text, but no counter-questions allowed). This role assignment gives a sense of purpose to the interaction (why are we talking? To help the questioner build a mental model of the image), and allows objective evaluation of individual responses.

3. The Visual Dialog Dataset (VisDial)

We now describe our VisDial dataset. We begin by describing the chat interface and data-collection process on AMT, analyze the dataset, then discuss the evaluation protocol.

Consistent with previous data collection efforts, we collect visual dialog data on images from the Common Objects in Context (COCO) [27] dataset, which contains multiple objects in everyday scenes. The visual complexity of these images allows for engaging and diverse conversations.

Live Chat Interface. Good data for this task should include dialogs that have (1) temporal continuity, (2) grounding in the image, and (3) mimic natural 'conversational' exchanges. To elicit such responses, we paired 2 workers on AMT to chat with each other in real-time (Fig. 3). Each worker was assigned a specific role. One worker (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

(a) What the 'questioner' sees. (b) What the 'answerer' sees. (c) Example dialog from our VisDial dataset.

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.

image to 'imagine the scene better'. The second worker (the 'answerer') sees the image and caption. Their task is to answer questions asked by their chat partner. Unlike VQA [4], answers are not restricted to be short or concise, instead workers are encouraged to reply as naturally and 'conversationally' as possible. Fig. 3c shows an example dialog.

This process is an unconstrained 'live' chat, with the only exception that the questioner must wait to receive an answer before posting the next question. The workers are allowed to end the conversation after 20 messages are exchanged (10 pairs of questions and answers). Further details about our final interface can be found in the supplement.

We also piloted a different setup where the questioner saw a highly blurred version of the image, instead of the caption. The conversations seeded with blurred images resulted in questions that were essentially 'blob recognition' – *'What is the pink patch at the bottom right?'*. For our full-scale data-collection, we decided to seed with just the captions since it resulted in more 'natural' questions and more closely modeled the real-world applications discussed in Section 1 where no visual signal is available to the human.

Building a 2-person chat on AMT. Despite the popularity of AMT as a data collection platform in computer vision, our setup had to design for and overcome some unique challenges – the key issue being that AMT is simply not designed for multi-user Human Intelligence Tasks (HITs). Hosting a live two-person chat on AMT meant that none of the Amazon tools could be used and we developed our own backend messaging and data-storage infrastructure based on Redis messaging queues and Node.js. To support data quality, we ensured that a worker could not chat with themselves (using say, two different browser tabs) by maintaining a pool of worker IDs paired. To minimize wait time for one worker while the second was being searched for, we ensured that there was always a significant pool of available HITs. If one of the workers abandoned a HIT (or was disconnected) midway, automatic conditions in the code kicked in asking the remaining worker to either continue asking questions or providing facts (captions) about the image (depending on their role) till 10 messages were sent by them. Workers who completed the task in this way were fully compensated, but our backend discarded this data and automatically launched a new HIT on this image so a real two-person conversation could be recorded. Our entire data-collection infrastructure (front-end UI, chat interface, backend storage and messaging system, error handling protocols) is publicly available[2].

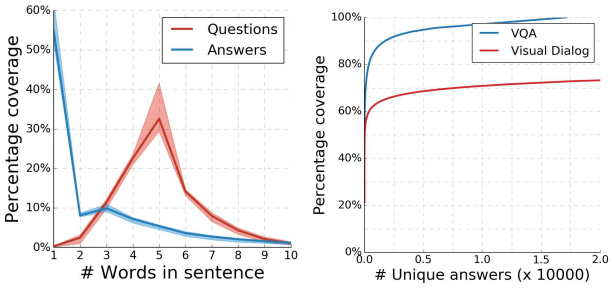


Figure 4: Distribution of lengths for questions and answers (left); and percent coverage of unique answers over all answers from the train dataset (right), compared to VQA. For a given coverage, VisDial has more unique answers indicating greater answer diversity.

4. VisDial Dataset Analysis

We now analyze the v0.9 subset of our VisDial dataset – it contains 1 dialog (10 question-answer pairs) on ∼123k images from COCO-trainval, a total of 1,232,870 QA pairs.

4.1. Analyzing VisDial Questions

Visual Priming Bias. One key difference between VisDial and previous image question-answering datasets (VQA [4], Visual 7W [63], Baidu mQA [14]) is the lack of a 'visual priming bias' in VisDial. Specifically, in all previous datasets, subjects saw an image while asking questions about it. As analyzed in [2, 16, 62], this leads to a particular bias in the questions – people only ask *'Is there a clock-tower in the picture?'* on pictures actually containing clock towers. This allows language-only models to perform remarkably well on VQA and results in an inflated sense of

[2] https://github.com/batra-mlp-lab/visdial-amt-chat

progress [16, 62]. As one particularly perverse example – for questions in the VQA dataset starting with *'Do you see a ... '*, blindly answering *'yes'* without reading the rest of the question or looking at the associated image results in an average VQA accuracy of 87%! In VisDial, questioners *do not* see the image. As a result, this bias is reduced.

Distributions. Fig. 4a shows the distribution of question lengths in VisDial – we see that most questions range from four to ten words. Fig. 5 shows 'sunbursts' visualizing the distribution of questions (based on the first four words) in VisDial *vs*. VQA. While there are a lot of similarities, some differences immediately jump out. There are more binary questions[3] in VisDial as compared to VQA – the most frequent first question-word in VisDial is 'is' *vs*. 'what' in VQA. A detailed comparison of the statistics of VisDial *vs*. other datasets is available in Table 1 in the supplement.

Finally, there is a stylistic difference in the questions that is difficult to capture with the simple statistics above. In VQA, subjects saw the image and were asked to stump a smart robot. Thus, most queries involve specific details, often about the background (*'What program is being utilized in the background on the computer?'*). In VisDial, questioners did not see the original image and were asking questions to build a mental model of the scene. Thus, the questions tend to be open-ended, and often follow a pattern:

- Generally starting with the entities in the caption:
  'An elephant walking away from a pool in an exhibit', *'Is there only 1 elephant?'*,
- digging deeper into their parts or attributes:
  'Is it full grown?', *'Is it facing the camera?'*,
- asking about the scene category or the picture setting:
  'Is this indoors or outdoors?', *'Is this a zoo?'*,
- the weather:
  'Is it snowing?', *'Is it sunny?'*,
- simply exploring the scene:
  'Are there people?', *'Is there shelter for elephant?'*,
- and asking follow-up questions about the new visual entities discovered from these explorations:
  'There's a blue fence in background, like an enclosure', *'Is the enclosure inside or outside?'*.

4.2. Analyzing VisDial Answers

Answer Lengths. Fig. 4a shows the distribution of answer lengths. Unlike previous datasets, answers in VisDial are longer and more descriptive – mean-length 2.9 words (VisDial) vs 1.1 (VQA), 2.0 (Visual 7W), 2.8 (Visual Madlibs).

Fig. 4b shows the cumulative coverage of all answers (y-axis) by the most frequent answers (x-axis). The difference between VisDial and VQA is stark – the top-1000 answers in VQA cover ~83% of all answers, while in VisDial that figure is only ~63%. There is a significant heavy tail in VisDial – most long strings are unique, and thus the coverage curve in Fig. 4b becomes a straight line with slope 1. In total, there are 337,527 unique answers in VisDial v0.9.

Answer Types. Since the answers in VisDial are longer strings, we can visualize their distribution based on the starting few words (Fig. 5c). An interesting category of answers emerges – *'I think so'*, *'I can't tell'*, or *'I can't see'* – expressing doubt, uncertainty, or lack of information. This is a consequence of the questioner not being able to see the image – they are asking contextually relevant questions, but not all questions may be answerable with certainty from that image. We believe this is rich data for building more human-like AI that refuses to answer questions it doesn't have enough information to answer. See [42] for a related, but complementary effort on question relevance in VQA.

Binary Questions vs Binary Answers. In VQA, binary questions are simply those with 'yes', 'no', 'maybe' as answers [4]. In VisDial, we must distinguish between binary questions and binary answers. Binary questions are those starting in 'Do', 'Did', 'Have', 'Has', 'Is', 'Are', 'Was', 'Were', 'Can', 'Could'. Answers to such questions can (1) contain only 'yes' or 'no', (2) begin with 'yes', 'no', and contain additional information or clarification, (3) involve ambiguity (*'It's hard to see'*, *'Maybe'*), or (4) answer the question without explicitly saying 'yes' or 'no' (Q: *'Is there any type of design or pattern on the cloth?'*, A: *'There are circles and lines on the cloth'*). We call answers that contain 'yes' or 'no' as binary answers – 149,367 and 76,346 answers in subsets (1) and (2) from above respectively. Binary answers in VQA are biased towards 'yes' [4, 62] – 61.40% of yes/no answers are 'yes'. In VisDial, the trend is reversed. Only 46.96% are 'yes' for all yes/no responses. This is understandable since workers did not see the image, and were more likely to end up with negative responses.

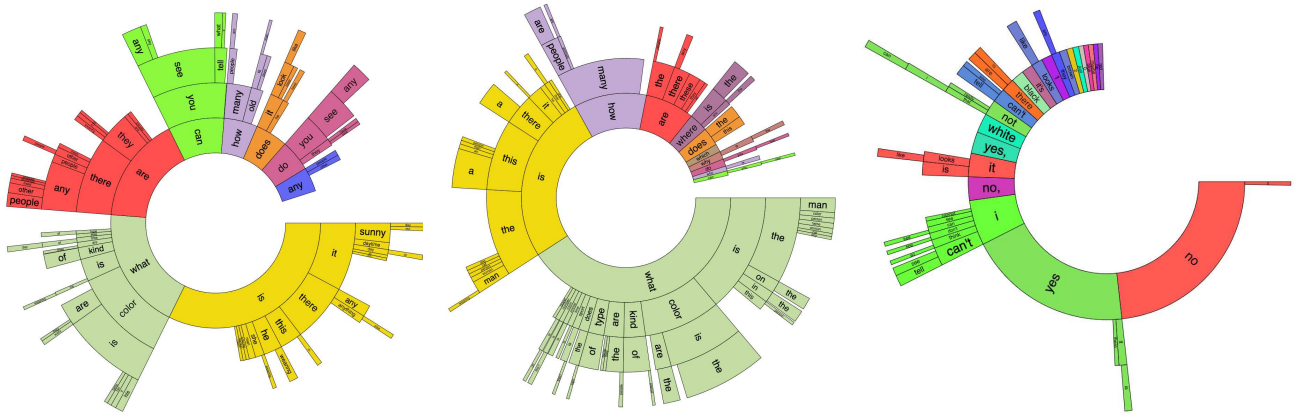
4.3. Analyzing VisDial Dialog

In Section 4.1, we discussed a typical flow of dialog in VisDial. We analyze two quantitative statistics here.

Coreference in dialog. Since language in VisDial is the result of a sequential conversation, it naturally contains pronouns – 'he', 'she', 'his', 'her', 'it', 'their', 'they', 'this', 'that', 'those', *etc*. In total, 38% of questions, 19% of answers, and *nearly all* (98%) dialogs contain at least one pronoun, thus confirming that a machine will need to overcome coreference ambiguities to be successful on this task. We find that pronoun usage is low in the first round (as expected) and then picks up in frequency. A fine-grained per-round analysis is available in the supplement.

Temporal Continuity in Dialog Topics. It is natural for conversational dialog data to have continuity in the 'topics' being discussed. We have already discussed qualitative

[3] Questions starting in 'Do', 'Did', 'Have', 'Has', 'Is', 'Are', 'Was', 'Were', 'Can', 'Could'.



(a) VisDial Questions (b) VQA Questions (c) VisDial Answers

Figure 5: Distribution of first n-grams for (left to right) VisDial questions, VQA questions and VisDial answers. Word ordering starts towards the center and radiates outwards, and arc length is proportional to number of questions containing the word.

differences in VisDial questions *vs*. VQA. In order to quantify the differences, we performed a human study where we manually annotated question 'topics' for 40 images (a total of 400 questions), chosen randomly from the `val` set. The topic annotations were based on human judgement with a consensus of 4 annotators, with topics such as: asking about a particular object (*'What is the man doing?'*) , scene (*'Is it outdoors or indoors?'*), weather (*"Is the weather sunny?'*), the image (*'Is it a color image?'*), and exploration (*'Is there anything else?"*). We performed similar topic annotation for questions from VQA for the same set of 40 images, and compared topic continuity in questions. Across 10 rounds, VisDial question have 4.55 ± 0.17 topics on average, confirming that these are not independent questions. Recall that VisDial has 10 questions per image as opposed to 3 for VQA. Therefore, for a fair comparison, we compute average number of topics in VisDial over all subsets of 3 successive questions. For 500 bootstrap samples of batch size 40, VisDial has 2.14 ± 0.05 topics while VQA has 2.53 ± 0.09. Lower mean suggests there is more continuity in VisDial because questions do not change topics as often.

4.4. VisDial Evaluation Protocol

One fundamental challenge in dialog systems is evaluation. Similar to the state of affairs in captioning and machine translation, it is an open problem to automatically evaluate the quality of free-form answers. Existing metrics such as BLEU, METEOR, ROUGE are known to correlate poorly with human judgement in evaluating dialog responses [28].

Instead of evaluating on a downstream task [7] or holistically evaluating the entire conversation (as in goal-free chitchat [3]), we evaluate *individual responses* at each round ($t = 1, 2, \ldots, 10$) in a retrieval or multiple-choice setup.

Specifically, at test time, a VisDial system is given an image I, the 'ground-truth' dialog history (including the image caption) $C, (Q_1, A_1), \ldots, (Q_{t-1}, A_{t-1})$, the question Q_t, *and* a list of $N = 100$ candidate answers, and asked to return a sorting of the candidate answers. The model is evaluated on retrieval metrics – (1) rank of human response (lower is better), (2) recall@k, *i.e*. existence of the human response in top-k ranked responses, and (3) mean reciprocal rank (MRR) of the human response (higher is better).

The evaluation protocol is compatible with both discriminative models (that simply score the input candidates, *e.g*. via a softmax over the options, and cannot generate new answers), and generative models (that generate an answer string, *e.g*. via Recurrent Neural Networks) by ranking the candidates by the model's log-likelihood scores.

Candidate Answers. We generate a candidate set of correct and incorrect answers from four sets:
Correct: The ground-truth human response to the question.
Plausible: Answers to 50 most similar questions. Similar questions are those that start with similar tri-grams and mention similar semantic concepts in the rest of the question. To capture this, all questions are embedded into a vector space by concatenating the GloVe embeddings of the first three words with the averaged GloVe embeddings of the remaining words in the questions. Euclidean distances are used to compute neighbors. Since these neighboring questions were asked on different images, their answers serve as 'hard negatives'.
Popular: The 30 most popular answers from the dataset – *e.g*. 'yes', 'no', '2', '1', 'white', '3', 'grey', 'gray', '4', 'yes it is'. The inclusion of popular answers forces the machine to pick between likely *a priori* responses and plausible responses for the question, thus increasing the task difficulty.
Random: The remaining are answers to random questions in the dataset. To generate 100 candidates, we first find the union of the correct, plausible, and popular answers, and include random answers until a unique set of 100 is found.

5. Neural Visual Dialog Models

In this section, we develop a number of neural Visual Dialog answerer models. Recall that the model is given as input – an image I, the 'ground-truth' dialog history (including the image caption) $H = (\underbrace{C}_{H_0}, \underbrace{(Q_1, A_1)}_{H_1}, \ldots, \underbrace{(Q_{t-1}, A_{t-1})}_{H_{t-1}})$, the question Q_t, and a list of 100 candidate answers $\mathcal{A}_t = \{A_t^{(1)}, \ldots, A_t^{(100)}\}$ – and asked to return a sorting of $\mathcal{A}_t$.

At a high level, all our models follow the encoder-decoder framework, *i.e.* factorize into two parts – (1) an *encoder* that converts the input (I, H, Q_t) into a vector space, and (2) a *decoder* that converts the embedded vector into an output. We describe choices for each component next and present experiments with all encoder-decoder combinations.

Decoders: We use two types of decoders:

- **Generative** (LSTM) decoder: where the encoded vector is set as the initial state of the Long Short-Term Memory (LSTM) RNN language model. During training, we maximize the log-likelihood of the ground truth answer sequence given its corresponding encoded representation (trained end-to-end). To evaluate, we use the model's log-likelihood scores and rank candidate answers.
  Note that this decoder does not need to score options during training. As a result, such models do not exploit the biases in option creation and typically underperform models that do [21], but it is debatable whether exploiting such biases is really indicative of progress. Moreover, generative decoders are more practical in that they can actually be deployed in realistic applications.
- **Discriminative** (softmax) decoder: computes dot product similarity between the input encoding and an LSTM encoding of each of the answer options. These dot products are fed into a softmax to compute the posterior probability over the options. During training, we maximize the log-liklihood of the correct option. During evaluation, options are simply ranked based on their posterior probabilities.

Encoders: We develop 3 different encoders (listed below) that convert inputs (I, H, Q_t) into a joint representation. In all cases, we represent I via the $\ell 2$-normalized activations from the penultimate layer of VGG-16 [50]. For each encoder E, we experiment with all possible ablated versions: $E(Q_t), E(Q_t, I), E(Q_t, H), E(Q_t, I, H)$ (for some encoders, not all combinations are 'valid'; details below).

- **Late Fusion (LF) Encoder**: In this encoder, we treat H as a long string with the entire history $(H_0, \ldots, H_{t-1})$ concatenated. Q_t and H are separately encoded with 2 different LSTMs, and 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)**: In this encoder, we capture the intuition that there is a hierarchical nature to our problem – each question Q_t is a sequence of words that need to be embedded, and the dialog as a whole is a sequence of question-answer pairs (Q_t, A_t). Thus, similar to [48], as shown in Fig. 6, we propose an HRE model that contains a dialog-RNN sitting on top of a recurrent block (R_t). The recurrent block R_t embeds the question and image jointly via an LSTM (early fusion), embeds each round of the history H_t, and passes a concatenation of these to the dialog-RNN above it. The dialog-RNN produces both an encoding for this round (E_t in Fig. 6) and a dialog context to pass onto the next round. We also add an attention-over-history ('Attention' in Fig. 6) mechanism allowing the recurrent block R_t to choose and attend to the round of the history relevant to the current question. This attention mechanism consists of a softmax over previous rounds $(0, 1, \ldots, t-1)$ computed from the history and question+image encoding.

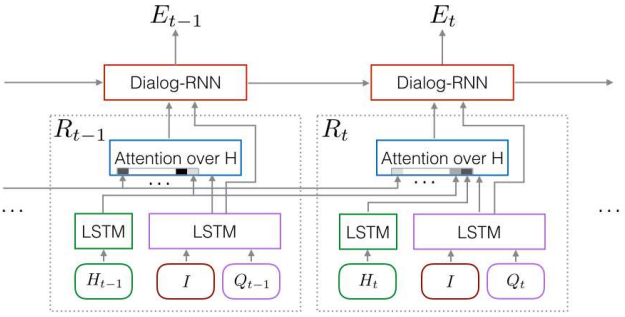


Figure 6: Architecture of HRE encoder with attention. At the current round R_t, the model has the capability to choose and attend to relevant history from previous rounds, based on the current question. This attention-over-history feeds into a dialog-RNN along with question to generate joint representation E_t for the decoder.

- **Memory Network (MN) Encoder**: We develop a MN encoder that maintains each previous question and answer as a 'fact' in its memory bank and learns to refer to the stored facts and image to answer the question. Specifically, we encode Q_t with an LSTM to get a 512-d vector, encode each previous round of history $(H_0, \ldots, H_{t-1})$ with another LSTM to get a $t \times 512$ matrix. We 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. Convex combination of history vectors using these attention probabilities gives us the 'context vector', which is passed through an fc-layer and added to the question vectorto construct the MN encoding. In the language of Memory Network [7], this is a '1-hop' encoding.

We use a '[encoder]-[input]-[decoder]' convention to refer to model-input combinations. For example, 'LF-QI-D' has a Late Fusion encoder with question+image inputs (no history), and a discriminative decoder. Implementation details about the models can be found in the supplement.

6. Experiments

Splits. VisDial v0.9 contains 83k dialogs on COCO-`train` and 40k on COCO-`val` images. We split the 83k into 80k for training, 3k for validation, and use the 40k as test.

Data preprocessing, hyperparameters and training details are included in the supplement.

Baselines We compare to a number of baselines: **Answer Prior:** Answer options to a test question are encoded with an LSTM and scored by a linear classifier. This captures ranking by frequency of answers in our training set without resolving to exact string matching. **NN-Q:** Given a test question, we find k nearest neighbor questions (in GloVe space) from train, and score answer options by their mean-similarity with these k answers. **NN-QI:** First, we find K nearest neighbor questions for a test question. Then, we find a subset of size k based on image feature similarity. Finally, we rank options by their mean-similarity to answers to these k questions. We use $k = 20, K = 100$.

Finally, we adapt several (near) state-of-art VQA models (SAN [61], HieCoAtt [32]) to Visual Dialog. Since VQA is posed as classification, we 'chop' the final VQA-answer softmax from these models, feed these activations to our discriminative decoder (Section 5), and train end-to-end on VisDial. Note that our LF-QI-D model is similar to that in [31]. Altogether, these form fairly sophisticated baselines.

Results. Tab. 1 shows the results for our proposed models and baselines on VisDial v0.9 (evaluated on 40k from COCO-`val`).

A few key takeaways – 1) As expected, all learning based models significantly outperform non-learning baselines. 2) All discriminative models significantly outperform generative models, which as we discussed is expected since discriminative models can tune to the biases in the answer options. 3) Our best generative and discriminative models are MN-QIH-G with 0.526 MRR, and MN-QIH-D with 0.597 MRR. 4) We observe that naively incorporating history doesn't help much (LF-Q *vs*. LF-QH and LF-QI *vs*. LF-QIH) or can even hurt a little (LF-QI-G *vs*. LF-QIH-G). However, models that better encode history (MN/HRE) perform better than corresponding LF models with/without history (*e.g*. LF-Q-D vs. MN-QH-D). 5) Models looking at I ({LF,MN,HRE }-QIH) outperform corresponding blind models (without I).

Human Studies. We conduct studies on AMT to quantitatively evaluate human performance on this task for all combinations of {with image, without image}×{with history, without history}. We find that without image, humans perform better when they have access to dialog history. As expected, this gap narrows down when they have access to the image. Complete details can be found in supplement.

	Model	MRR	R@1	R@5	R@10	Mean
Baseline	Answer prior	0.3735	23.55	48.52	53.23	26.50
	NN-Q	0.4570	35.93	54.07	60.26	18.93
	NN-QI	0.4274	33.13	50.83	58.69	19.62
Generative	LF-Q-G	0.5048	39.78	60.58	66.33	17.89
	LF-QH-G	0.5055	39.73	60.86	66.68	17.78
	LF-QI-G	0.5204	42.04	61.65	67.66	16.84
	LF-QIH-G	0.5199	41.83	61.78	67.59	17.07
	HRE-QH-G	0.5102	40.15	61.59	67.36	17.47
	HRE-QIH-G	0.5237	**42.29**	62.18	67.92	17.07
	HREA-QIH-G	0.5242	42.28	62.33	68.17	**16.79**
	MN-QH-G	0.5115	40.42	61.57	67.44	17.74
	MN-QIH-G	**0.5259**	**42.29**	**62.85**	**68.88**	17.06
Discriminative	LF-Q-D	0.5508	41.24	70.45	79.83	7.08
	LF-QH-D	0.5578	41.75	71.45	80.94	6.74
	LF-QI-D	0.5759	43.33	74.27	83.68	5.87
	LF-QIH-D	0.5807	43.82	74.68	84.07	5.78
	HRE-QH-D	0.5695	42.70	73.25	82.97	6.11
	HRE-QIH-D	0.5846	44.67	74.50	84.22	5.72
	HREA-QIH-D	0.5868	44.82	74.81	84.36	5.66
	MN-QH-D	0.5849	44.03	75.26	84.49	5.68
	MN-QIH-D	**0.5965**	**45.55**	**76.22**	**85.37**	**5.46**
VQA	SAN1-QI-D	0.5764	43.44	74.26	83.72	5.88
	HieCoAtt-QI-D	0.5788	43.51	74.49	83.96	5.84

Table 1: Performance of methods on VisDial v0.9, measured by mean reciprocal rank (MRR), recall@k and mean rank. Higher is better for MRR and recall@k, while lower is better for mean rank. Performance on VisDial v0.5 is included in the supplement.

7. Conclusions

To summarize, we introduce a new AI task – Visual Dialog, where an AI agent must hold a dialog with a human about visual content. We develop a novel two-person chat data-collection protocol to curate a large-scale dataset (VisDial), propose retrieval-based evaluation protocol, and develop a family of encoder-decoder models for Visual Dialog. We quantify human performance on this task via human studies. Our results indicate that there is significant scope for improvement, and we believe this task can serve as a testbed for measuring progress towards visual intelligence.

Acknowledgements. We thank Harsh Agrawal, Jiasen Lu for help with AMT data collection; Xiao Lin, Latha Pemula for model discussions; Marco Baroni, Antoine Bordes, Mike Lewis, Marc'Aurelio Ranzato for helpful discussions. We are grateful to the developers of Torch [1] for building an excellent framework. This work was funded in part by NSF CAREER awards to DB and DP, ONR YIP awards to DP and DB, ONR Grant N00014-14-1-0679 to DB, a Sloan Fellowship to DP, ARO YIP awards to DB and DP, an Allen Distinguished Investigator award to DP from the Paul G. Allen Family Foundation, ICTAS Junior Faculty awards to DB and DP, Google Faculty Research Awards to DP and DB, Amazon Academic Research Awards to DP and DB, AWS in Education Research grant to DB, and NVIDIA GPU donations to DB. SK was supported by ONR Grant N00014-12-1-0903. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government, or any sponsor.

References

[1] Torch. http://torch.ch/. 8

[2] H. Agrawal, A. Chandrasekaran, D. Batra, D. Parikh, and M. Bansal. Sort story: Sorting jumbled images and captions into stories. In *EMNLP*, 2016. 3, 4

[3] Amazon. Alexa. http://alexa.amazon.com/. 6

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 1, 2, 3, 4, 5

[5] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. VizWiz: Nearly Real-time Answers to Visual Questions. In *UIST*, 2010. 1

[6] A. Bordes, N. Usunier, S. Chopra, and J. Weston. Large-scale Simple Question Answering with Memory Networks. *arXiv preprint arXiv:1506.02075*, 2015. 3

[7] A. Bordes and J. Weston. Learning End-to-End Goal-Oriented Dialog. *arXiv preprint arXiv:1605.07683*, 2016. 3, 6, 7

[8] G. Christie, A. Laddha, A. Agrawal, S. Antol, Y. Goyal, K. Kochersberger, and D. Batra. Resolving language and vision ambiguities together: Joint segmentation and prepositional attachment resolution in captioned scenes. In *EMNLP*, 2016. 3

[9] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In *EMNLP*, 2016. 3

[10] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. C. Courville. GuessWhat?! Visual object discovery through multi-modal dialogue. In *CVPR*, 2017. 3

[11] J. Dodge, A. Gane, X. Zhang, A. Bordes, S. Chopra, A. Miller, A. Szlam, and J. Weston. Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems. In *ICLR*, 2016. 2, 3

[12] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *CVPR*, 2015. 3

[13] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From Captions to Visual Concepts and Back. In *CVPR*, 2015. 3

[14] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. In *NIPS*, 2015. 3, 4

[15] D. Geman, S. Geman, N. Hallonquist, and L. Younes. A Visual Turing Test for Computer Vision Systems. In *PNAS*, 2014. 3

[16] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 3, 4, 5

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 1

[18] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *NIPS*, 2015. 1, 3

[19] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 3

[20] T.-H. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell. Visual storytelling. In *NAACL HLT*, 2016. 3

[21] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *ECCV*, 2016. 7

[22] A. Kannan, K. Kurach, S. Ravi, T. Kaufmann, A. Tomkins, B. Miklos, G. Corrado, L. Lukács, M. Ganea, P. Young, et al. Smart Reply: Automated Response Suggestion for Email. In *KDD*, 2016. 3

[23] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 3

[24] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *CVPR*, 2014. 3

[25] O. Lemon, K. Georgila, J. Henderson, and M. Stuttle. An ISU dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling in the TALK in-car system. In *EACL*, 2006. 2

[26] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky. Deep Reinforcement Learning for Dialogue Generation. In *EMNLP*, 2016. 3

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. DollÃ¡r, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 2, 3

[28] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *EMNLP*, 2016. 3, 6

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single Shot MultiBox Detector. In *ECCV*, 2016. 1

[30] R. Lowe, N. Pow, I. Serban, and J. Pineau. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *SIGDIAL*, 2015. 3

[31] J. Lu, X. Lin, D. Batra, and D. Parikh. Deeper LSTM and Normalized CNN Visual Question Answering model. https://github.com/VT-vision-lab/VQA_LSTM_CNN, 2015. 8

[32] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *NIPS*, 2016. 3, 8

[33] M. Malinowski and M. Fritz. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *NIPS*, 2014. 3

[34] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015. 1, 3

[35] H. Mei, M. Bansal, and M. R. Walter. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *AAAI*, 2016. 2

[36] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep rein-

forcement learning. *Nature*, 518(7540):529–533, 02 2015. 1

[37] N. Mostafazadeh, C. Brockett, B. Dolan, M. Galley, J. Gao, G. P. Spithourakis, and L. Vanderwende. Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation. *arXiv preprint arXiv:1701.08251*, 2017. 3

[38] T. Paek. Empirical methods for evaluating dialog systems. In *Proceedings of the workshop on Evaluation for Language and Dialogue Systems-Volume 9*, 2001. 2

[39] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 3

[40] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP*, 2016. 3

[41] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. Linking people with "their" names using coreference resolution. In *ECCV*, 2014. 3

[42] A. Ray, G. Christie, M. Bansal, D. Batra, and D. Parikh. Question Relevance in VQA: Identifying Non-Visual And False-Premise Questions. In *EMNLP*, 2016. 5

[43] M. Ren, R. Kiros, and R. Zemel. Exploring Models and Data for Image Question Answering. In *NIPS*, 2015. 1, 3

[44] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 3

[45] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. In *CVPR*, 2015. 3

[46] I. V. Serban, A. García-Durán, Ç. Gülçehre, S. Ahn, S. Chandar, A. C. Courville, and Y. Bengio. Generating Factoid Questions With Recurrent Neural Networks: The 30M Factoid Question-Answer Corpus. In *ACL*, 2016. 3

[47] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI*, 2016. 3

[48] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. *arXiv preprint arXiv:1605.06069*, 2016. 3, 7

[49] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. 1

[50] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 7

[51] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In *CVPR*, 2016. 1

[52] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S. C. Zhu. Joint Video and Text Parsing for Understanding Events and Answering Queries. *IEEE MultiMedia*, 2014. 1

[53] S. Venugopalan, M. Rohrbach, J. Donahue, R. J. Mooney, T. Darrell, and K. Saenko. Sequence to Sequence - Video to Text. In *ICCV*, 2015. 3

[54] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. J. Mooney, and K. Saenko. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. In *NAACL HLT*, 2015. 3

[55] O. Vinyals and Q. Le. A Neural Conversational Model. *arXiv preprint arXiv:1506.05869*, 2015. 3

[56] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 3

[57] L. Wang, S. Guo, W. Huang, Y. Xiong, and Y. Qiao. Knowledge Guided Disambiguation for Large-Scale Scene Classification with Multi-Resolution CNNs. *arXiv preprint arXiv:1610.01119*, 2016. 1

[58] J. Weizenbaum. ELIZA. http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm. 2, 3

[59] J. Weston, A. Bordes, S. Chopra, and T. Mikolov. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In *ICLR*, 2016. 1, 3

[60] S. Wu, H. Pique, and J. Wieland. Using Artificial Intelligence to Help Blind People 'See' Facebook. http://newsroom.fb.com/news/2016/04/using-artificial-intelligence-to-help-blind-people-see-facebook/, 2016. 1

[61] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola. Stacked Attention Networks for Image Question Answering. In *CVPR*, 2016. 8

[62] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh. Yin and Yang: Balancing and Answering Binary Visual Questions. In *CVPR*, 2016. 3, 4, 5

[63] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded Question Answering in Images. In *CVPR*, 2016. 4

[64] C. L. Zitnick, A. Agrawal, S. Antol, M. Mitchell, D. Batra, and D. Parikh. Measuring machine intelligence through visual question answering. *AI Magazine*, 2016. 1