Visual Dialog: A Survey

Krishna Garg
University of Illinois at Chicago
kgarg8@uic.edu

Natalie Parde
University of Illinois at Chicago
parde@uic.edu

September 13, 2021

Abstract

The *visual dialog* task is a relatively new research problem that extends the horizon for research at the intersection of computer vision, natural language processing, and deep learning. It seeks to enable autonomous agents to hold natural, free-form, goal-free, human-like conversations about visual imagery with other machines or human agents. The task was proposed to overcome the limitations of Visual Question Answering (VQA) systems, which can answer isolated questions but cannot make use of dialog history or prolonged visual context across multiple utterances. Doing so naturally introduces a variety of challenges. This survey paper provides an overview of the finer details of the visual dialog task and those associated challenges, and explains the various algorithms developed thus far to make progress towards solving them. It is our hope that aggregating this information in a convenient, concise survey will spur rapid advancements in this exciting area of research.

1 Introduction

Visual dialog represents an exciting frontier for artificial intelligence, offering the opportunity to blend techniques drawn from the more-traditional task of Visual Question Answering (VQA) with recent research in the decades-old[1] subfield of dialog systems. The task gained focus from researchers all over the world following work by Das et al. [4], in which they introduced the VisDial dataset. VisDial contains real-world conversations spanning more than a million images, and Das et al. used this dataset to create the first modern visual chatbot. The availability of such a massive collection of data in this domain has fueled researchers to (a) develop more realistic chatbots capable of situated conversation, (b) accelerate research in both computer vision and natural language processing, and (c) move the needle forward in the quest to solve the Visual Turing Test.[2]

Prior and similar to visual dialog is the task of *Visual Question Answering (VQA)* [6], but there are notable differences between the two. VQA systems engage the chatbot to answer a single question about the image, whereas visual dialog systems are much more involved. Visual dialog requires a series of questions about an image, each of which can only be answered after considering both visual context and dialog history. Furthermore, a key objective in visual dialog is to mimic human-human conversations. As such, generated answers are often longer and characterized by more descriptive language. In contrast, VQA answers tend to be short and blunt, often by design.

The potential real-world applications for visual dialog are wide-reaching. Some intriguing use cases include facilitating personal assistance for sight-impaired individuals, and assisting in surveillance tasks for potentially dangerous situations (e.g., emergencies, criminal activity, or security monitoring). However, although recently text-only chatbots have been in high supply and there has been significant progress in relevant computer vision areas such as image captioning, VQA, object detection, action recog-

[1] Early work dates to the 1960s [1] and 70s [2], with SHRDLU [3] being the first to incorporate a visual element.

[2] The goal of the artificial agent in this test is to fool the evaluator into thinking it is human based on its ability to identify objects, attributes, and relationships in an image [5].

nition, visual abstraction, reading comprehension, visual storytelling, and others, there has been less focus on visual dialog. It is our hope that this survey article stimulates further research in this area by providing a useful reference to the topic as well as underscoring promising areas for further exploration.

In this survey paper, we describe the datasets that have been employed for research in visual dialog in Section 2. We detail the task and the methods used to evaluate it in Section 3, and summarize the algorithms and models developed to date in Section 4. We compare the results of those approaches and discuss their shortcomings in Section 5. Finally, we present our conclusions and recommendations for future work in Section 6.

2 Datasets

Research in visual dialog has thus far been primarily driven by two datasets: *GuessWhat?!* [7], and *VisDial* [4]. Data included in the *Guesswhat?!* [7] corpus was collected in the context of a two-party cooperative game. In the game, two human players (a *questioner* and an *oracle*) viewed the same image and had access to a shared text-based interaction interface. The oracle selected an object from the image, and the questioner attempted to identify which object was selected by asking questions to the oracle, who responded with affirmative or negative answers. More than 150,000 games were played during the data collection period, yielding more than 800,000 question-answer pairs. However, all answers were limited to one of three options (*yes*, *no*, or *N/A*), limiting the potential transferability of the collected data to real-world scenarios.

VisDial, proposed in the same year [4] and since updated multiple times (algorithms discussed in this survey article all use either *VisDial v0.9* or *VisDial v1.0*), contains fixed-length dialogs of question-answer pairs for images. The curators of the dataset [4] used COCO[3] images, and developed a two-person chat interface[4] on Amazon Mechanical Turk to crowdsource conversations regarding the images. Unlike *Guesswhat?!* and the single-turn *VQA*,[5] the *questioner* was blinded to the image itself and instead just saw the caption as a basis for constructing a question. The *answerer*, in turn, was encouraged to give a descriptive answer. The conversation ended after ten questions. To ensure data quality, unfinished conversations were not included in the dataset and the infrastructure was designed such that the workers could not chat with themselves using another separate browser tab. *VisDial v1.0*, the most recent iteration, contains 10 question-answer pairs for each of 123,287 images extracted from MS-COCO [8] and Flickr. Additional statistics regarding both *GuessWhat?!* and *VisDial* are provided in Table 1.

Dataset	**QA Pairs/Img.**	**Train**	**Val.**	**Test**
GuessWhat?!	Variable	108,696	23,292	23,292
VisDial v0.9	10	82,783	40,504	8000
VisDial v1.0	10 (Train/Val.) 1 (Test)	123,287	2064	8000

Table 1: Descriptive statistics for each major visual dialog dataset. **Train**, **Val.**, and **Test** refer to the number of images in each of those data splits.

2.1 Dataset Analysis:

VQA is mentioned a lot in this subsection, but not described at all in the previous subsection. In contrast, GuessWhat?! is described in detail in the previous subsection, but not analyzed at all in this subsection. I'd recommend adding (1) a description of VQA in the previous subsection, noting that you are describing it to more easily distinguish it from VisDial and GuessWhat?!, and (2) some comparisons between GuessWhat?! and VisDial in this subsection.

Guesswhat?! contains 155,280 dialogues (in total, 821,889 question-answer pairs). The dialogues focus on 66,537 unique images subsampled from COCO (filtered such that all images are greater than 500 px^2 and all objects in the image are between 3 to 20 px^2). The dialogues are shorter than those in *VisDial* (2.3 dialogues/image averaging 5.2 questions/dialogue, relative to 10 ques-

[3] COCO [8] is a large-scale object detection, segmentation and captioning dataset. More details can be found at http://cocodataset.org/.

[4] Source code publicly available at https://github.com/batra-mlp-lab/visdial-amt-chat.

[5] VQA, proposed in [9], contains open-ended questions about the images in COCO dataset. More details can be found at https://visualqa.org.

tions/dialogue and 1 dialogue/image in *VisDial*). Since the answers are limited to *yes, no, N/A* in this dataset, the most frequent words in the dataset are related to objects and entities in the image followed by spatial attributes (e.g., left/right) or visual features (e.g., color or size).

MS-COCO dataset was used even for Visual Question Answering task but it has a plenty of differences with the Visdial dataset. The Visdial dataset has more descriptive and longer answers than the VQA dataset. The mean-length of the answers in VisDial is 2.9 words compared to only 1.1 words in the VQA dataset. There are more unique answers in the VisDial dataset. The top 1000 answers span only 63% of the answers in VisDial, as against 83% in the VQA dataset. Definitely, the quality of the questions is much improved over the VQA dataset.

Another significant difference which VisDial has, unlike *Guesswhat?!* and other related datasets like VQA, Visual7W, VisualGenome, FM-IQA, DAQUAR, COCO-QA dataset, is that the questioner was blind-folded. Intuition behind this type of data collection was to reduce the visual priming bias while asking questions. When the image is visible, the questioner tends to follow a pattern while asking questions like ”do you see a particular object in the image” and thus overall the dataset becomes biased since a wild guess ”yes”, even without seeing the image, would give higher accuracy. VisDial v0.9 is much better in this regard since it has roughly 47-53% distribution for yes/no answers, as against roughly 61-39% distribution for yes/no answers in VQA dataset. The distribution for yes-no-N/A in *Guesswhat?!* is roughly 45-53-2%.

VisDial v0.9 also performs better than VQA in terms of temporal continuity, which means that the entire dialog has fewer topics and every topic has more questions, mimicking the high continuity in human conversations. The two datasets were compared for 40 images and the temporal continuity metric [lower the better] turned out to be 2.14±0.05 topics for VisDial and 2.53±0.09 topics for VQA for three successive questions.

However, we feel that VisDial dataset is still far from mimicking free-form human dialogs for the following reasons. The primary reason being the bot being unaware of the real world news, it can’t identify any celebrity from the image or it can’t identify any monuments from the image etc. In real life, many people can immediately recognize that the President is in White House just seeing the room where he is meeting people or addressing the press/ public. But the bot cant give all of that the information to the end user when questioned. Also, in human conversations there are questions which are related to the history or personal life of the personality which a dumb bot would never be able to answer. Again to give an instance, if the user already knows the celebrity shown in the image, the user naturally tends to ask whether the celebrity is having the same ”Top knot” haircut or whether he is accompanied by his/ her spouse or dog etc.

Secondly, the present dataset surely doesn’t account for the images with special effects like focus/ blur etc., images with different color scales [colors sometimes contribute a lot to the image understanding], if the image is skewed or is cartoon image etc. Thirdly, the bot can’t pose counter-arguments but just answers the questions. Finally, the test samples contain dialogs of just ten questions long. Even the best accuracy on the dataset will not be sufficient to say that the model will generalize well for longer conversations.

3 Task & Evaluation Methods

3.1 Task of Generative model

Given image I, dialog history $(Q_1, A_1), (Q_2, A_2), ..., (Q_{t-1}, A_{t-1})$, follow-up question Q_t, predict answer A_t.

3.2 Task of Discriminative model

Visdial has a pool of candidate answers in addition to the above items. Goal of the discriminative model is to sort this pool of candidate answers, with topmost answer being the closest to the ground truth.

The pool of candidate answers consists of a ground-truth answer and 99 negative samples. This set of 99 answers consists of 50 plausible answers, 30 popular answers, 19 random answers. 1) Plausible answers are the answers to questions starting with similar tri-grams and having the similar semantic concepts for the remaining words in the question. 2) Popular answers are the most frequently output responses to the questions in the entire dataset. 3) Random answers justify their category name.

Idea of the inclusion of 99 negative samples is to fool the discriminative model to output a wrong answer. A robust discriminative model should be able to rank the ground truth higher in the returned sorting.

3.3 Evaluation Methods

Since the goal is to evaluate an open-ended free-form answer, metrics like BLEU, METEOR, ROGUE etc. are not suitable for the task [10]. Therefore, the metrics like Recall@k, Mean Reciprocal Rank (MRR) and Mean Rank are used for the evaluation for Visdial v0.9. A new metric Normalized Discounted Cumulative Gain [] (NDCG) was introduced for the evaluation of performance of Visdial v1.0 where val and test datasets were human annotated. NDCG is an efficient metric to rank a list of answers when ideal sorting of the answers is available. It works by assigning a more discounting factor for the answers with less relevancy down the list. NDCG over top K-ranked answers (where K is the number of answers marked as correct by at least one of the four human annotators) ”is invariant to the order of options with identical relevance and to the order of options outside of the top K.” [11]

4 Approaches

The common approach across all the models is to use a combination of encoders and decoders. Encoder combines the three modalities (image, question, history) into a common vector space and decoder is then used to either output the correct answer or return the sorted candidate answers, depending upon whether the model is generative or discriminative.

Most of the approaches for solving the Visual Dialog task are centered around proposing more efficient encoder modules, a bunch of them we will describe in the following section. Idea of the encoder module is to somehow combine the three modalities (image, question, dialog history) as efficiently as possible to form a feature vector. Most of the encoding strategies are based on incorporating the attention mechanisms. Attention mechanisms have recently gained a lot of focus in Visual Question Answering task [4] [12] [13] as well. Attention over image means we want to focus on particular regions of an image, which are most relevant to answering the question. Similarly, attention over history means we want to look up the most relevant questions that would make important contribution in answering the current question. For example, if there are multiple male persons in the image doing different activities and if the question asked is ”what is he doing”, attention over dialog history will actually attribute sense to the question as to which person is being talked about in the previous questions. Similarly, attention over the follow-up question can help to identify the objects and concepts the agent is asking about the image and accordingly we can ground those objects or concepts in the image while attempting to answer the question.

The first step of any encoder module is extraction of features from the image using a pre-trained CNN model. In the past, different architectures have been used like ResNet 101, VGG-19, VGG-16 [4] etc. for training the CNN models. Also, features from question and dialog history are extracted using LSTM-RNN model in most of the models. LSTM [14] models are especially helpful in capturing the long-range dependencies and this can be particularly useful when we want to answer a question based on the dialog history (which can go up to 10 question-answer pairs). Different encoders have different strategies of feature extraction from question and dialog history. Some use separate LSTM models for every question in the dialog history and some pass the entire history through a single LSTM. The latter approach is more common and has given better results.

Apart from improving the encoder modules, some models have used Adversarial Learning based approach for improving the accuracy of the models, while some other approaches have been centered around Reinforcement learning and some are just the combination of the two. Some other approaches have have focused on addressing the problem of resolving co-reference ambiguities inherent in the Visual Dialog task. Often the questions can be ambiguous, and the answers need to consider the previous utterances in the dialog history. The models for addressing the coreference resolution problem not just use attention mechanisms ... [need to complete at later point of time after reading 1-2 more papers]

Generative Adversarial Networks (GANs) have been extended to generate fake answers that are close to the ground truth, rather than generating fake images mimicking the real ones. GANs have two parts: Generator module (G) and Discriminator module (D). G tries to generate the answers as close to the ground truth as possible. D takes this answer as input and using some strategy or by computing a loss function, rewards G so that G can update its parameters and thus learn to generate more similar answers.

Reinforcement learning is also used in some of the models. Again, the setting remains similar to GAN-inspired models. Only difference is that in the former models, D is optimized for maximizing the reward so as to provide more informative feedback to G. [can also be improved later]

4.1 Visual Dialog [15]

4.1.1 Encoders

Authors propose four types of encoders.

Late Fusion (LF) Encoder: In this encoder, the image is passed through a pre-trained CNN, dialog history through an LSTM and the follow-up question through another LSTM to extract the features. The three feature vectors are then concatenated and further reduced to a 512-d vector space.

Hierarchical Recurrent Encoder (HRE): Two things are worth mentioning here. First, each question in the dialog history is passed through a separate LSTM unlike LF Encoder. Second, authors use LSTM models in two-level fashion. Lower-level (question-level) LSTM is used to capture the information of words of the follow-up question and the higher-level (dialog level) LSTM is used to capture the information of an utterance (pair of question-answer) in the entire dialog history.

Image features (I) are learnt using a CNN model (pre-trained on VGG-16). I and Q_t are then passed through an LSTM to give (I + Q_t) vector. This output is then concatenated with the output from LSTMs (H_{t-1}) used for encoding history. This is repeated for every question in the dataset and finally, all such outputs (I + Q_i where i = 1, , 10) are then passed through dialog-level LSTM to get the desired feature vector.

HRE-with Attention (HREA): This variant of HRE accounts for attending to a relevant utterance in the dialog history to answer a question. Before concatenating the output (I + Q_t) with (H_{t-1}), (I + Q_t) is passed through attention module to refer to a particular question(s) in the history.

Memory Network (MN) Encoder: The basic intuition behind this is to again use attention mechanism to refer to a relevant question in the history. Similar to HRE, the utterances are passed through separate LSTMs and the outputs can be seen as facts in the memory bank. Question (Q_t) is passed through another separate LSTM. Dot product is then computed between the question feature vector and each utterance in the history to compute attention-over-history probabilities. Finally, the resultant convex vector is passed through fc-layer of VGG-16 and then added to the (I + Q_t) feature vector to produce the desired embedding.

4.2 Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model [12]

4.2.1 Motivation

Generative neural network models trained with maximum likelihood estimation tend to produce very generic response like *I dont know*, *I cant tell*. On the other hand, although discriminative models perform better than generative models [4], but they cant be used for having real conversations with the chatbot. Discriminative models just return a sorting of the pool of the candidate answers in decreasing order of their relevance. So, the motivation was to combine the practical usefulness of the generative models and the good performance of the discriminative models.

4.2.2 History-Conditioned Image Attentive Encoder (HCIAE)

This is slightly improved over the encoders proposed by Das et al. [4]. Not just image and question are used to attend to the history, furthermore the attended history and question are used to attend to the image. This helps to focus on relevant regions in the image while answering the question.

4.2.3 Algorithm

The model proposed by the authors is much similar to Generative Adversarial Networks (GANs) [16] with only slight variations. Generative part (G) tries to generate an answer closest to ground truth whereas discriminative part

(D) is used to assess the closeness of the answer and to give feedback to G.

Output from HCIAE encoder is fed to G to produce a distribution over candidate answers. An answer is then sampled from the distribution using Gumbel-Softmax [reference] sampler. This answer is then fed to D. Based on the input of 100 candidate answers including the ground truth answer, D pre-learns a function which tells the closeness of the generated answer to the ground-truth. The novelty of the paper also lies in using multi-class N-pair loss function [12] instead of multi-class logistic loss function. This loss function particularly prevents the correct but different answers from getting overly penalized. Finally, G updates its parameters based on the feedback from D.

4.3 Are you Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning [13]

4.3.1 Motivation

This is another model inspired by GANs and applied to multimodal features. Additionally, reinforcement learning is used to update G. The difference from the previous model is that D doesnt take candidate answers as inputs and borrows the attention weights from G. These attention weights can be seen as a form of reasoning for the answer and thus, help in producing human-like responses (which are longer and descriptive).

4.3.2 Sequential Co-attention Encoder

We have three modalities –image, question and the dialog history. Authors use two of the modalities at a time to co-attend to the third modality. The difference with the other encoders is that in this encoder, CNN is pre-trained on Very Deep Convolutional Networks [17] and there is a cycle of attention sequences (co-attention is used three times sequentially instead of just one or two).

4.3.3 Algorithm

The model has both generative (G) and discrimator (D) parts, similar to GANs. G takes in the encoded feature vector and passes it through LSTM (decoder) to generate an answer. D takes three inputs i.e. image feature vector, history feature vector and combined feature vector of follow up question and the generated answer. Output from D is the probability whether the dialog is human or not. This probability is used as reward to update G. D uses REINFORCE algorithm [18] to maximize the reward. Authors propose two more variants. First, they use Monte Carlo search to provide appropriate reward to the words of the generated answer, instead of just rewarding the entire sentence. Second, they use human feedback (teacher-forcing) [19][20] for updating G.

4.4 Stacked Co-Attention for VisDial1.0 [21]

4.4.1 Motivation

Although the prior works [4] [12] [13] have used attention models, but still they fail to answer questions related to fine-grained details in an image. Stacked attention models [22] work by first locating all the objects and concepts referred in the question and then pinpointing to the most relevant region in the image over multiple iterations. This type of multi-step reasoning helps resolve the problem of co-reference ambiguity. Also, authors use N-pair discriminative loss function which was shown to perform better than naive logistic loss function [12].

4.4.2 Encoder

This encoder is an improved version of Co-Attention Encoder [13]. It is like a stack of three co-attention encoders.

4.4.3 Algorithm

Image features are extracted in bottom-up fashion using Faster RCNN (ResNet 101) pretrained on COCO-dataset, in a similar fashion with [23]. The rest of the architecture is pretty much similar except for the new encoder and discriminative loss function.

4.5 Image-Question-Answer Synergistic Network for Visual Dialog [24]

4.5.1 Motivation

Other visual dialog models simply ignore the candidate pool of answers. Authors of this paper propose that even the candidate pool of answers can be used to attend to the image features and the dialog history.

4.5.2 Algorithm

Features for the image are extracted using Bottom-up and Top-down approach [23]. The model has two stages. 1) In the primary stage, the feature vectors of the image, the dialog history and the question are passed through the Co-attention Encoder. The candidate pool of answers A_t is ranked using the N-pair loss function, similar to [12]. The spotlight of this stage is that the top N-ranked answers are selected and then appended with their corresponding question to form a new candidate set B_t. The basic idea is that these N answers form hard samples, which means they are closest to the ground truth and other answers are called easy samples. According to the authors' analysis, nearly 90% of the answers constitute easy samples and are actually not relevant for the question. 2) In the synergistic stage, the feature vectors for B_t, the dialog history and the image are passed through the encoder and finally, the answers are re-ranked using a decoder.

Another interesting part of the paper was the fusion method used for combining the different features. Borrowing the idea from [25], multi-modal bilinear pooling is used which fuses the feature vectors from different modalities more efficiently.

4.6 Visual Reference Resolution using Attention Memory for Visual Dialog [26]

4.6.1 Motivation

Visual Dialog entails a problem of resolving co-reference ambiguities. Often the questions can be ambiguous, and the answers need to consider the previous utterances in the dialog history. In the following section, we will present some of the works done to address this challenge.

We have discussed some attention network models in the previous sections. Most of them focus on attending to some image region or to particular question in the dialog history and thus hope to implicitly resolve the problem of coreference ambiguity. But this paper introduces the novel idea of storing attention maps for every time step in the form of associative (attention map v/s key) memory and thus, for answering any ambiguous question, they consider the most relevant stored attention map.

4.6.2 Algorithm

Three different encoders are used for each component in the triplet (question q, history H, image I): Recurrent Neural Network (RNN) with LSTM units for q, Hierarchical Recurrent Neural Network (HRNN) for H and CNN to get a feature map for I. Next, they compute f_t^{att}, an attended image feature embedding conditioned on the context embedding c_t (a joint embedding space of question and history).

They propose three-step attention mechanism. First, tentative attention is calculated by computing similarity between c_t(the joint embedding space of question and history) and the image feature map vector. This is sufficient to answer the unambiguous questions. Second, retrieved attention is calculated by looking up the most relevant attention map in the attention memory. Third, both the attentions are stacked together and are fed to a convolution layer which is then flattened and fed to a fully connected softmax layer to generate the final attention map. The dynamic parameter prediction approach [27] is used to combine the two attentions where the parameters of the layer are learnt dynamically based on the current question.

Finally, the final attention map, its corresponding retrieved key, c_t and f_t^{att} are fused together and fed to a fully connected layer which finally generates the final encoding(e_t). The intended answer is generated by decoding e_t.

The authors justify their approach by evaluating the results on a self-proposed MNIST Dialog dataset which contains a lot of co-reference ambiguities. Their algorithm on this dataset produces 96.39% accuracy and 0.6210 MRR on VisDial v0.9 dataset. Additionally, they show the improvement in the performance of both the datasets by incorporating sequential dialog structure (where a learnable parameter accounts for more recent attentions to be referred again).

5 Results and Discussion

The results for the Visual Dialog task are summarized in the tables [2 3]. Table [2] shows the performance of the different models on the VisDial v0.9 validation dataset which contains around 40k images as mentioned in tables [**??** **??**]. Stacked Co-attention is the best performing model with Mean Reciprocal Rank (MRR) of 0.6398 and Mean Rank (MR) of 4.47 which implies that on an average, the ground truth answer was ranked at 4.47 among 100 answers. Also, from the results we can easily infer that improving the attention mechanism in the models makes a direct impact on the retrieval metrics. MN Encoder uses attention once, HCIAE uses attention twice, Co-Attention Encoder uses attention thrice in a cyclic fashion and Stacked Co-Attention Encoder further improves the attention using a stack of Co-Attention Encoders. Thus, we see performance in increasing order for MN, HCIAE, Co-Attention Encoder and Stacked Co-Attention Encoder.

Table [3] summarizes the performance of different models on VisDial v1.0 dataset. Model using Stacked Co-Attention Encoder (discussed in section 4.4) was the runner-up for Visual Dialog Challenge 2018 [28] and the Synergistic model (discussed in section 4.5) was the winner in the same challenge. Synergistic model pushed the state-of-the art performance by +0.02 in terms of MRR, -0.68 in terms of MR and +1.41 in terms of NDCG. The model improved the Recall @ (1, 5, 10) metric by 1.65, 2.64 and 2.8 respectively. The higher performance of the Synergistic model can be attributed again to the improved attention mechanism in the model since even the top N-answers, returned in the primary stage, were used to attend the image regions and the dialog history in the synergistic stage.

Model	MRR	R@1	R@5	R@10	Mean
LF [4]	0.5807	43.82	74.68	84.07	5.78
HRE [4]	0.5846	44.67	74.50	84.22	5.72
MN [4]	0.5965	45.55	76.22	85.37	5.46
HCIAE [12]	0.6222	48.48	78.75	87.59	4.81
Co-Att [15]	0.6398	50.29	80.71	88.81	4.47
Stacked Co-Att [21]	0.6403	50.27	80.85	88.96	4.43
AMEM [26]	0.6227	48.53	78.66	87.43	4.86

Table 2: Results on VisDial 0.9 val (Discriminative Model) Mean Rank: lower the better, other metrics: higher the better

Model	MRR	R@1	R@5	R@10	Mean	NDCG
LF [4]	0.5542	40.95	72.45	82.83	5.95	0.4531
HRE [4]	0.5416	39.93	70.45	81.50	6.41	0.4546
MN [4]	0.5549	40.98	72.30	83.30	5.92	0.4750
Stacked Co-Att [21]	0.6144	47.65	78.13	87.88	4.65	56.47
Synergistic [24]	0.6342	49.30	80.77	90.68	3.97	57.88

Table 3: Results on VisDial 1.0 test-std (Discriminative Model) Mean Rank: lower the better, other metrics: higher the better

6 Conclusion and Future Work

Visual Dialog has recently come up as a new area of research and Artificial Intelligence powered by Visual Dialog task can take exciting future research directions with nice applications.

In the survey paper, we present an overview of some of the models that have been proposed for solving the Visual Dialog task. Finally, we summarize and compare the performance of the different models on VisDial v0.9 and VisDial v1.0 datasets. It is very evident from the results that attention has a vital role to play in improving the performance of models. Most of the models focus on this aspect.

However, we are highly optimistic that performance of the models can be improved a lot if we incorporate external knowledge bases into the model. Existing models are not capable of answering the questions which require common sense reasoning. For instance, if an image shows some students in a library and the question is *what are the people doing here*, visual chatbot is not capable of giving a human-like reasoning and thus ends up answering like *people are sitting around a table* instead of answer like *students might have gathered for collaboration on a project*. To the best of our knowledge, there is not much analysis done by the researchers as to how well the model is performing on various aspects like how well the model is able to co-reference resolution, how well the model is able to attend to the dialog history, how well the model is performing on reasoning questions, etc. In [29], the authors point out some unnoticed flaws in the overall task either due to over constrained evaluation metrics or due to some implicit biases in the dataset, their overly simplified model, which doesnt even consider visual scenes and dialog history, performs as par with the models proposed in the state-of-the-art for this task. Although VisDial dataset doesnt contain visual priming bias since the questioner is blind-folded, but this may introduce some other biases. Firstly, for a blind person to imagine mental model of even fine-grained details of the image by just ten questions is a big challenge. There is some work done by Das et al. related to this aspect as well in paper [15]. Secondly, the questioner asks the questions based on the caption he sees on AMT chat server but how detailed notion of an image the caption itself is capturing, that itself needs to be considered.

The best performing model that we have seen so far has MRR of 0.6364 and this is far from any real-world deployment system standards. We hope that future Visual Dialog Challenges invoke lot of enthusiastic participation from the researchers all over the world and we see the chatbot mimicking humans more closely in near future.

Acknowledgements

We heartily thank Ashima Garg for valuable contributions in the form of insightful suggestions and comments during the entire work.

References

[1] J. Weizenbaum, “Eliza—a computer program for the study of natural language communication between man and machine,” *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[2] D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd, “Gus, a frame-driven dialog system,” *Artificial intelligence*, vol. 8, no. 2, pp. 155–173, 1977.

[3] T. Winograd, “Procedures as a representation for data in a computer program for understanding natural language,” Massachusetts Institution of Technology, AI Technical Report 235, Tech. Rep., 1971.

[4] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, “Visual dialog,” *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2017.

[5] D. Geman, S. Geman, N. Hallonquist, and L. Younes, “Visual turing test for computer vision systems,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 12, pp. 3618–3623, 2015. [Online]. Available: https://www.pnas.org/content/112/12/3618

[6] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “Vqa: Visual question answering,” *Proceedings of the IEEE international conference on computer vision*, p. 24252433, 2015.

[7] H. D. Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. C. Courville, “Guesswhat?! visual object discovery through multi-modal dialogue,” *CVPR*, vol. 1, 2017.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick, “Microsoft coco: Common objects in context,” *ECCV*, 2014.

[9] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “Vqa: Visual question answering,” *ICCV*, 2015.

[10] 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," *EMNLP*, 2016.

[11] Y. Wang, L. Wang, Y. Li, D. He, W. Chen, and T.-Y. Liu, "A theoretical analysis of normalized discounted cumulative gain (ndcg) ranking measures," *In Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, 2013.

[12] . J. Lu, A. Kannan, J. Yang, D. Parikh, and D. Batra, "Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model," *arXiv preprint arXiv:1706.01554*, 2017.

[13] Q. Wu, P. Wang, C. Shen, I. Reid, and A. van den Hengel, "Are you talking to me? reasoned visual dialog generation through adversarial learning," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, p. 61066115, 2018.

[14] S.Hochreiter and J.Schmidhuber, "Longshort-term memory," *Neural computation*, vol. 9(8), p. 17351780, 1997.

[15] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra, "Learning cooperative visual dialog agents with deep reinforcement learning," *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, p. 26722680, 2014.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8(3-4), pp. 229–256, 1992.

[19] A. M. Lamb, A. G. A. P. GOYAL, Y. Zhang, S. Zhang, A. C. Courville, and Y. Bengio, "Professor forcing: A new algorithm for training recurrent networks," *In Advances In Neural Information Processing Systems*, p. 46014609, 2016.

[20] J. Li, W. Monroe, T. Shi, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generationarxiv preprint arxiv:1701.06547," *arXiv preprint arXiv:1701.06547*, 2017.

[21] Y. Tianhao, Z. Zheng-Jun, and Z. Hanwang, "Stacked co-attention for visdial 1.0," 2018.

[22] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," *In Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 21–29, 2016.

[23] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image caption and vqa," *CVPR*, 2018.

[24] D. Guo, C. Xu, and D. Tao, "Image-question-answer synergistic network for visual dialog," *arXiv preprint arXiv:1902.09774*, 2018.

[25] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bi- linear pooling with co-attention learning for visual question answering," *IEEE International Conference on Computer Vision (ICCV)*, p. 18391848, 2017.

[26] P. H. Seo, A. Lehrmann, B. Han, and L. Sigal, "Visual reference resolution using attention memory for visual dialog," *NIPS*, 2017.

[27] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," *CVPR*, 2016.

[28] "https://visualdialog.org/challenge/."

[29] N. S. D. Massiceti, P. Dokania and P. Torr, "Visual dialogue without vision or dialogue," *arXiv preprint arXiv: 1812.06417*.