Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog

Zhe Gan[1], **Yu Cheng**[1], **Ahmed El Kholy**[1], **Linjie Li**[1], **Jingjing Liu**[1], **Jianfeng Gao**[2]
[1]Microsoft Dynamics 365 AI Research, [2]Microsoft Research
{zhe.gan, yu.cheng, ahmed.eikholy, lindsey.li, jingjl, jfgao}@microsoft.com

Abstract

This paper presents a new model for visual dialog, Recurrent Dual Attention Network (ReDAN), using multi-step reasoning to answer a series of questions about an image. In each question-answering turn of a dialog, ReDAN infers the answer progressively through multiple reasoning steps. In each step of the reasoning process, the semantic representation of the question is updated based on the image and the previous dialog history, and the recurrently-refined representation is used for further reasoning in the subsequent step. On the VisDial v1.0 dataset, the proposed ReDAN model achieves a new state-of-the-art of 64.47% NDCG score. Visualization on the reasoning process further demonstrates that ReDAN can locate context-relevant visual and textual clues via iterative refinement, which can lead to the correct answer step-by-step.

1 Introduction

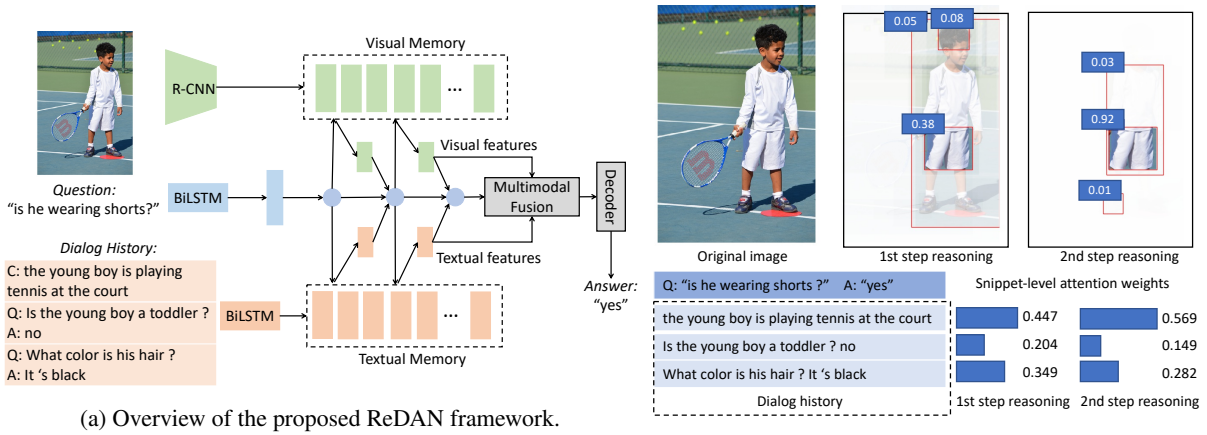
There has been a recent surge of interest in developing neural network models capable of understanding both visual information and natural language, with applications ranging from image captioning (Fang et al., 2015; Vinyals et al., 2015; Xu et al., 2015) to visual question answering (VQA) (Antol et al., 2015; Fukui et al., 2016; Anderson et al., 2018). Unlike VQA, where the model can answer a single question about an image, a visual dialog system (Das et al., 2017a; De Vries et al., 2017; Das et al., 2017b) is designed to answer a series of questions regarding an image, which requires a comprehensive understanding of both the image and previous dialog history.

Most previous work on visual dialog rely on attention mechanisms (Bahdanau et al., 2015; Xu et al., 2015) to identify specific regions of the image and dialog-history snippets that are relevant to the question. These attention models measure the relevance between the query and the attended image, as well as the dialog context. To generate an answer, either a discriminative decoder is used for ranking answer candidates, or a generative decoder is trained for synthesizing an answer (Das et al., 2017a; Lu et al., 2017). Though promising results have been reported, these models often fail to provide accurate answers, especially in cases where answers are confined to particular image regions or dialog-history snippets.

One hypothesis for the cause of failure is the inherent limitation of single-step reasoning approach. Intuitively, after taking a first glimpse of the image and the dialog history, readers often revisit specific sub-areas of both image and text to obtain a better understanding of the multimodal context. Inspired by this, we propose a Recurrent Dual Attention Network (ReDAN) that exploits multi-step reasoning for visual dialog.

Figure 1a provides an overview of the model architecture of ReDAN. First, a set of visual and textual memories are created to store image features and dialog context, respectively. In each step, a semantic representation of the question is used to attend to both memories, in order to obtain a question-aware image representation and question-aware dialog representation, both of which subsequently contribute to updating the question representation via a recurrent neural network. Later reasoning steps typically provide a sharper attention distribution than earlier steps, aiming at narrowing down the regions most relevant to the answer. Finally, after several iterations of reasoning steps, the refined question vector and the garnered visual/textual clues are fused to obtain a final multimodal context vector, which is fed to the decoder for answer generation. This multi-step reasoning process is performed in each turn of the dialog.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6463–6474
Florence, Italy, July 28 - August 2, 2019.



(a) Overview of the proposed ReDAN framework.

(b) An example of multi-step reasoning in ReDAN.

Figure 1: Model architecture and visualization of the learned multi-step reasoning strategies. In the first step, ReDAN first focuses on all relevant objects in the image (*e.g.*, "*boy*", "*shorts*"), and all relevant facts in the dialog history (*e.g.*, "*young boy*", "*playing tennis*", "*black hair*"). In the second step, the model narrows down to more context-relevant regions and dialog context (*i.e.*, the attention maps become sharper) which lead to the final correct answer ("*yes*"). The numbers in the bounding boxes and in the histograms are the attention weights of the corresponding objects or dialog history snippets.

Figure 1b provides an illustration of the iterative reasoning process. In the current dialog turn for the question "*is he wearing shorts?*", in the initial reasoning step, the system needs to draw knowledge from previous dialog history to know who "*he*" refers to (*i.e.*, "*the young boy*"), as well as interpreting the image to rule out objects irrelevant to the question (*i.e.*, *"net", "racket" and "court"*). After this, the system conducts a second round of reasoning to pinpoint the image region (*i.e.*, "*shorts*", whose attention weight increases from 0.38 to 0.92 from the 1st step to the 2nd step) and the dialog-history snippet (*i.e.*, "*playing tennis at the court*", whose attention weight increased from 0.447 to 0.569), which are most indicative of the correct answer ("*yes*").

The main contributions of this paper are threefold. (i) We propose a ReDAN framework that supports multi-step reasoning for visual dialog. (ii) We introduce a simple rank aggregation method to combine the ranking results of discriminative and generative models to further boost the performance. (iii) Comprehensive evaluation and visualization analysis demonstrate the effectiveness of our model in inferring answers progressively through iterative reasoning steps. Our proposed model achieves a new state-of-the-art of 64.47% NDCG score on the VisDial v1.0 dataset.

2 Related Work

Visual Dialog The visual dialog task was recently proposed by Das et al. (2017a) and De Vries et al. (2017). Specifically, Das et al. (2017a) released the VisDial dataset, which contains free-form natural language questions and answers. And De Vries et al. (2017) introduced the GuessWhat?! dataset, where the dialogs provided are more goal-oriented and aimed at object discovery within an image, through a series of yes/no questions between two dialog agents.

For the VisDial task, a typical system follows the encoder-decoder framework proposed in Sutskever et al. (2014). Different encoder models have been explored in previous studies, including late fusion, hierarchical recurrent network, memory network (all three proposed in Das et al. (2017a)), early answer fusion (Jain et al., 2018), history-conditional image attention (Lu et al., 2017), and sequential co-attention (Wu et al., 2018). The decoder model usually falls into two categories: (i) generative decoder to synthesize the answer with a Recurrent Neural Network (RNN) (Das et al., 2017a); and (ii) discriminative decoder to rank answer candidates via a softmax-based cross-entropy loss (Das et al., 2017a) or a ranking-based multi-class N-pair loss (Lu et al., 2017).

Reinforcement Learning (RL) was used in Das et al. (2017b); Chattopadhyay et al. (2017) to train two agents to play image guessing games. Lu et al. (2017) proposed a training schema to effectively transfer knowledge from a pre-trained discriminative model to a generative dialog model. Generative Adversarial Network (Goodfellow et al., 2014; Yu et al., 2017b; Li et al., 2017) was also

used in Wu et al. (2018) to generate answers indistinguishable from human-generated answers, and a conditional variational autoencoder (Kingma and Welling, 2014; Sohn et al., 2015) was developed in Massiceti et al. (2018) to promote answer diversity. There were also studies investigating visual coreference resolution, either via attention memory implicitly (Seo et al., 2017) or using a more explicit reasoning procedure (Kottur et al., 2018) based on neural module networks (Andreas et al., 2016). In addition to answering questions, question sequence generation is also investigated in Jain et al. (2018); Massiceti et al. (2018).

For the GuessWhat?! task, various methods (such as RL) have been proposed to improve the performance of dialog agents, measured by task completion rate as in goal-oriented dialog system (Strub et al., 2017; Shekhar et al., 2018; Strub et al., 2018; Lee et al., 2018; Zhang et al., 2018). Other related work includes image-grounded chitchat (Mostafazadeh et al., 2017), dialog-based image retrieval (Guo et al., 2018), and text-only conversational question answering (Reddy et al., 2018; Choi et al., 2018). A recent survey on neural approaches to dialog modeling can be found in Gao et al. (2018).

In this work, we focus on the VisDial task. Different from previous approaches to visual dialog, which all used a single-step reasoning strategy, we propose a novel multi-step reasoning framework that can boost the performance of visual dialog systems by inferring context-relevant information from the image and the dialog history iteratively.

Multi-step Reasoning The idea of multi-step reasoning has been explored in many tasks, including image classification (Mnih et al., 2014), text classification (Yu et al., 2017a), image generation (Gregor et al., 2015), language-based image editing (Chen et al., 2018), Visual Question Answering (VQA) (Yang et al., 2016; Nam et al., 2017; Hudson and Manning, 2018), and Machine Reading Comprehension (MRC) (Cui et al., 2017; Dhingra et al., 2017; Hill et al., 2016; Sordoni et al., 2016; Shen et al., 2017; Liu et al., 2018).

Specifically, Mnih et al. (2014) introduced an RNN for image classification, by selecting a sequence of regions adaptively and only processing the selected regions. Yu et al. (2017a) used an RNN for text classification, by learning to skip irrelevant information when reading the text input. A recurrent variational autoencoder termed DRAW was proposed in Gregor et al. (2015) for multi-step image generation. A recurrent attentive model for image editing was also proposed in Chen et al. (2018) to fuse image and language features via multiple steps.

For VQA, Stacked Attention Network (SAN) (Yang et al., 2016) was proposed to attend the question to relevant image regions via multiple attention layers. For MRC, ReasoNet (Shen et al., 2017) was developed to perform multi-step reasoning to infer the answer span based on a given passage and a question, where the number of steps can be dynamically determined via a termination gate.

Different from SAN for VQA (Yang et al., 2016) and ReasoNet for MRC (Shen et al., 2017), which reason over a single type of input (either image or text), our proposed ReDAN model incorporates multimodal context that encodes both visual information and textual dialog. This multimodal reasoning approach presents a mutual enhancement between image and text for a better understanding of both: on the one hand, the attended image regions can provide additional information for better dialog interpretation; on the other hand, the attended history snippets can be used for better image understanding (see the dotted red lines in Figure 2).

Concurrent Work We also include some concurrent work for visual dialog that has not been discussed above, including image-question-answer synergistic network (Guo et al., 2019), recursive visual attention (Niu et al., 2018), factor graph attention (Schwartz et al., 2019), dual attention network (Kang et al., 2019), graph neural network (Zheng et al., 2019), history-advantage sequence training (Yang et al., 2019), and weighted likelihood estimation (Zhang et al., 2019).

3 Recurrent Dual Attention Network

The visual dialog task (Das et al., 2017a) is formulated as follows: given a question Q_ℓ grounded in an image I, and previous dialog history (including the image caption C) $H_\ell = \{C, (Q_1, A_1), \cdots, (Q_{\ell-1}, A_{\ell-1})\}$ (ℓ is the current dialog turn) as additional context, the goal is to generate an answer by ranking a list of N candidate answers $\mathcal{A}_\ell = \{A_\ell^{(1)}, \ldots, A_\ell^{(N)}\}$.

Figure 2 provides an overview of the Recurrent Dual Attention Network (ReDAN). Specifically,

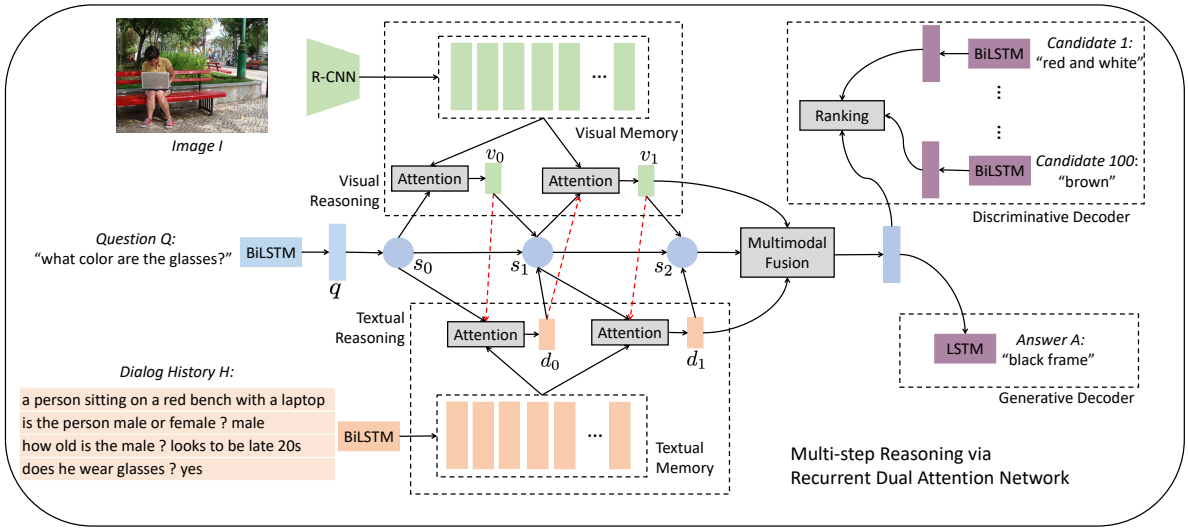


Figure 2: Model Architecture of Recurrent Dual Attention Network for visual dialog. Please see Sec. 3 for details.

ReDAN consists of three components: (*i*) *Memory Generation Module* (Sec. 3.1), which generates a set of visual and textual memories to provide grounding for reasoning; (*ii*) *Multi-step Reasoning Module* (Sec. 3.2), where recurrent dual attention is applied to jointly encode question, image and dialog history into a multimodal context vector for decoding; and (*iii*) *Answer Decoding Module* (Sec. 3.3), which derives the final answer for each question based on the multimodal context vector. The following sub-sections describe the details of these components.

3.1 Memory Generation Module

In this module, the image I and the dialog history H_ℓ are transformed into a set of memory vectors (visual and textual).

Visual Memory We use a pre-trained Faster R-CNN (Ren et al., 2015; Anderson et al., 2018) to extract image features, in order to enable attention on both object-level and salient region-level, each associated with a feature vector. Compared to image features extracted from VGG-Net (Simonyan and Zisserman, 2014) and ResNet (He et al., 2016), this type of features from Faster R-CNN has achieved state-of-the-art performance in both image captioning and VQA (Anderson et al., 2018; Teney et al., 2018) tasks. Specifically, the image features $\mathbf{F}_I$ for a raw image I are represented by:

$$\mathbf{F}_I = \text{R-CNN}(I) \in \mathbb{R}^{n_f \times M}\,, \tag{1}$$

where $M = 36$ is the number of detected objects in an image[1], and $n_f = 2048$ is the dimension of the feature vector. A single-layer perceptron is used to transform each feature into a new vector that has the same dimension as the query vector (described in Sec. 3.2):

$$\mathbf{M}_v = \tanh(\mathbf{W}_I \mathbf{F}_I) \in \mathbb{R}^{n_h \times M}\,, \tag{2}$$

where $\mathbf{W}_I \in \mathbb{R}^{n_h \times n_f}$. All the bias terms in this paper are omitted for simplicity. $\mathbf{M}_v$ is the visual memory, and its m-th column corresponds to the visual feature vector for the region of the object indexed by m.

Textual Memory In the ℓ-th dialogue turn, the dialog history H_ℓ consists of the caption C and $\ell - 1$ rounds of QA pairs (Q_j, A_j) $(j = 1, \ldots, \ell - 1)$. For each dialog-history snippet j (the caption is considered as the first one with $j = 0$), it is first represented as a matrix $\mathbf{M}_h^{(j)} = [\boldsymbol{h}_0^{(j)}, \ldots, \boldsymbol{h}_{K-1}^{(j)}] \in \mathbb{R}^{n_h \times K}$ via a bidirectional Long Short-Term Memory (BiLSTM) network (Hochreiter and Schmidhuber, 1997), where K is the maximum length of the dialog-history snippet. Then, a self-attention mechanism is applied to learn the attention weight of every word in the snippet, identifying the key words and ruling out irrelevant information. Specifically,

$$\begin{aligned} \boldsymbol{\omega}_j &= \text{softmax}(\boldsymbol{p}_\omega^T \cdot \tanh(\mathbf{W}_h \mathbf{M}_h^{(j)}))\,, \\ \boldsymbol{u}_j &= \boldsymbol{\omega}_j \cdot (\mathbf{M}_h^{(j)})^T\,, \end{aligned} \tag{3}$$

[1]We have also tried using an adaptive number of detected objects for an image. Results are very similar to the results with $M = 36$.

where $\boldsymbol{\omega}_j \in \mathbb{R}^{1\times K}, \boldsymbol{p}_\omega \in \mathbb{R}^{n_h\times 1}$, $\mathbf{W}_h \in \mathbb{R}^{n_h\times n_h}$, and $\boldsymbol{u}_j \in \mathbb{R}^{1\times n_h}$. After applying the same BiLSTM to each dialog-history snippet, the textual memory is then represented as $\mathbf{M}_d = [\boldsymbol{u}_0^T, \ldots, \boldsymbol{u}_{\ell-1}^T] \in \mathbb{R}^{n_h\times \ell}$.

3.2 Multi-step Reasoning Module

The multi-step reasoning framework is implemented via an RNN, where the hidden state $\boldsymbol{s}_t$ represents the current representation of the question, and acts as a query to retrieve visual and textual memories. The initial state $\boldsymbol{s}_0$ is a self-attended question vector $\boldsymbol{q}$. Let $\boldsymbol{v}_t$ and $\boldsymbol{d}_t$ denote the attended image representation and dialog-history representation in the t-th step, respectively. A one-step reasoning pathway can be illustrated as $\boldsymbol{s}_t \rightarrow \boldsymbol{v}_t \rightarrow \boldsymbol{d}_t \rightarrow \boldsymbol{s}_{t+1}$, which is performed T times. Details are described below.

Self-attended Question Similar to textual memory construction, a question Q (the subscript ℓ for Q_ℓ is omitted to reduce confusion) is first represented as a matrix $\mathbf{M}_q = [\boldsymbol{q}_0, \ldots, \boldsymbol{q}_{K'-1}] \in \mathbb{R}^{n_h\times K'}$ via a BiLSTM, where K' is the maximum length of the question. Then, self attention is applied,

$$\boldsymbol{\alpha} = \text{softmax}(\boldsymbol{p}_\alpha^T \cdot \tanh(\mathbf{W}_q \mathbf{M}_q)), \ \boldsymbol{q} = \boldsymbol{\alpha}\mathbf{M}_q^T,$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{1\times K'}, \boldsymbol{p}_\alpha \in \mathbb{R}^{n_h\times 1}$, and $\mathbf{W}_q \in \mathbb{R}^{n_h\times n_h}$. $\boldsymbol{q} \in \mathbb{R}^{1\times n_h}$ then serves as the initial hidden state of the RNN, *i.e.*, $\boldsymbol{s}_0 = \boldsymbol{q}$.

The reasoning pathway $\boldsymbol{s}_t \rightarrow \boldsymbol{v}_t \rightarrow \boldsymbol{d}_t \rightarrow \boldsymbol{s}_{t+1}$ includes the following steps: (i) $(\boldsymbol{s}_t, \boldsymbol{d}_{t-1}) \rightarrow \boldsymbol{v}_t$; (ii) $(\boldsymbol{s}_t, \boldsymbol{v}_t) \rightarrow \boldsymbol{d}_t$; and (iii) $(\boldsymbol{v}_t, \boldsymbol{d}_t) \rightarrow \boldsymbol{s}_{t+1}$.

Query and History Attending to Image Given $\boldsymbol{s}_t$ and the previous attended dialog history representation $\boldsymbol{d}_{t-1} \in \mathbb{R}^{1\times n_h}$, we update $\boldsymbol{v}_t$ as follows:

$$\begin{aligned}\boldsymbol{\beta} &= \text{softmax}(\boldsymbol{p}_\beta^T \cdot \tanh(\mathbf{W}_v \mathbf{M}_v + \mathbf{W}_s \boldsymbol{s}_t^T + \mathbf{W}_d \boldsymbol{d}_{t-1}^T)), \\ \boldsymbol{v}_t &= \boldsymbol{\beta} \cdot \mathbf{M}_v^T, \end{aligned} \quad (4)$$

where $\boldsymbol{\beta} \in \mathbb{R}^{1\times M}, \boldsymbol{p}_\beta \in \mathbb{R}^{n_h\times 1}, \mathbf{W}_v \in \mathbb{R}^{n_h\times n_h}, \mathbf{W}_s \in \mathbb{R}^{n_h\times n_h}$ and $\mathbf{W}_d \in \mathbb{R}^{n_h\times n_h}$. The updated $\boldsymbol{v}_t$, together with $\boldsymbol{s}_t$, is used to attend to the dialog history.

Query and Image Attending to History Given $\boldsymbol{s}_t \in \mathbb{R}^{1\times n_h}$ and the attended image representation $\boldsymbol{v}_t \in \mathbb{R}^{1\times n_h}$, we update $\boldsymbol{d}_t$ as follows:

$$\begin{aligned}\boldsymbol{\gamma} &= \text{softmax}(\boldsymbol{p}_\gamma^T \cdot \tanh(\mathbf{W}_d' \mathbf{M}_d + \mathbf{W}_s' \boldsymbol{s}_t^T + \mathbf{W}_v' \boldsymbol{v}_t^T)), \\ \boldsymbol{d}_t &= \boldsymbol{\gamma} \cdot \mathbf{M}_d^T, \end{aligned} \quad (5)$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{1\times \ell}, \boldsymbol{p}_\gamma \in \mathbb{R}^{n_h\times 1}, \mathbf{W}_v' \in \mathbb{R}^{n_h\times n_h}, \mathbf{W}_s' \in \mathbb{R}^{n_h\times n_h}$ and $\mathbf{W}_d' \in \mathbb{R}^{n_h\times n_h}$. The updated $\boldsymbol{d}_t$ is fused with $\boldsymbol{v}_t$ and then used to update the RNN query state.

Multimodal Fusion Given the query vector $\boldsymbol{s}_t$, we have thus far obtained the updated image representation $\boldsymbol{v}_t$ and the dialog-history representation $\boldsymbol{d}_t$. Now, we use Multimodal Factorized Bilinear pooling (MFB) (Yu et al., 2017c) to fuse $\boldsymbol{v}_t$ and $\boldsymbol{d}_t$ together. Specifically,

$$\boldsymbol{z}_t = \text{SumPooling}(\mathbf{U}_v \boldsymbol{v}_t^T \circ \mathbf{U}_d \boldsymbol{d}_t^T, k), \quad (6)$$

$$\boldsymbol{z}_t = \text{sign}(\boldsymbol{z}_t)|\boldsymbol{z}_t|^{0.5}, \ \boldsymbol{z}_t = \boldsymbol{z}_t^T / ||\boldsymbol{z}_t||, \quad (7)$$

where $\mathbf{U}_v \in \mathbb{R}^{n_h k\times n_h}, \mathbf{U}_d \in \mathbb{R}^{n_h k\times n_h}$. The function $\text{SumPooling}(\boldsymbol{x}, k)$ in (6) means using a one-dimensional non-overlapped window with the size k to perform sum pooling over $\boldsymbol{x}$. (7) performs power normalization and ℓ_2 normalization. The whole process is denoted in short as:

$$\boldsymbol{z}_t = \text{MFB}(\boldsymbol{v}_t, \boldsymbol{d}_t) \in \mathbb{R}^{1\times n_h}. \quad (8)$$

There are also other methods for multimodal fusion, such as MCB (Fukui et al., 2016) and MLB (Kim et al., 2017). We use MFB in this paper due to its superior performance in VQA.

Image and History Updating RNN State The initial state $\boldsymbol{s}_0$ is set to $\boldsymbol{q}$, which represents the initial understanding of the question. The question representation is then updated based on the current dialogue history and the image, via an RNN with Gated Recurrent Unit (GRU) (Cho et al., 2014):

$$\boldsymbol{s}_{t+1} = \text{GRU}(\boldsymbol{s}_t, \boldsymbol{z}_t). \quad (9)$$

This process forms a cycle completing one reasoning step. After performing T steps of reasoning, multimodal fusion is then used to obtain the final context vector:

$$\boldsymbol{c} = [\text{MFB}(\boldsymbol{s}_T, \boldsymbol{v}_T), \text{MFB}(\boldsymbol{s}_T, \boldsymbol{d}_T), \text{MFB}(\boldsymbol{v}_T, \boldsymbol{d}_T)]. \quad (10)$$

3.3 Answer Decoding Module

Discriminative Decoder The context vector $\boldsymbol{c}$ is used to rank answers from a pool of candidates $\mathcal{A}$ (the subscript ℓ for $\mathcal{A}_\ell$ is omitted). Similar to how we obtain the self-attended question vector in Sec. 3.2, a BiLSTM, together with the self-attention mechanism, is used to obtain a vector representation for each candidate $A_j \in \mathcal{A}$, resulting in $\boldsymbol{a}_j \in \mathbb{R}^{1\times n_h}$, for $j = 1, \ldots, N$. Based

on this, a probability vector $\boldsymbol{p}$ is computed as $\boldsymbol{p} = \text{softmax}(\boldsymbol{s})$, where $\boldsymbol{s} \in \mathbb{R}^N$, and $\boldsymbol{s}[j] = \boldsymbol{c}\boldsymbol{a}_j^T$. During training, ReDAN is optimized by minimizing the cross-entropy loss[2] between the one-hot-encoded ground-truth label vector and the probability distribution $\boldsymbol{p}$. During evaluation, the answer candidates are simply ranked based on the probability vector $\boldsymbol{p}$.

Generative Decoder Besides the discriminative decoder, following Das et al. (2017a), we also consider a generative decoder, where another LSTM is used to decode the context vector into an answer. During training, we maximize the log-likelihood of the ground-truth answers. During evaluation, we use the log-likelihood scores to rank answer candidates.

Rank Aggregation Empirically, we found that combining the ranking results of discriminative and generative decoders boosts the performance a lot. Two different rank aggregation methods are explored here: (i) average over ranks; and (ii) average over reciprocal ranks. Specifically, in a dialog session, assuming $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_K$ represents the ranking results obtained from K trained models (either discriminative, or generative). In the first method, the average ranks $\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{r}_k$ are used to re-rank the candidates. In the second one, we use the average of the reciprocal ranks of each individual model $\frac{1}{K}\sum_{k=1}^{K}1/\boldsymbol{r}_k$ for re-ranking.

4 Experiments

In this section, we explain in details our experiments on the VisDial dataset. We compare our ReDAN model with state-of-the-art baselines, and conduct detailed analysis to validate the effectiveness of our proposed model.

4.1 Experimental Setup

Dataset We evaluate our proposed approach on the recently released VisDial v1.0 dataset[3]. Specifically, the training and validation splits from v0.9 are combined together to form the new training data in v1.0, which contains dialogs on $123,287$ images from COCO dataset (Lin et al., 2014). Each dialog is equipped with 10 turns, resulting in a total of 1.2M question-answer pairs. An additional $10,064$ COCO-like images are further collected from Flickr, of which $2,064$ images are used as the validation set (val v1.0), and the rest 8K are used as the test set (test-std v1.0), hosted on an evaluation server[4] (the ground-truth answers for this split are not publicly available). Each image in the val v1.0 split is associated with a 10-turn dialog, while a dialog with a flexible number of turns is provided for each image in test-std v1.0. Each question-answer pair in the VisDial dataset is accompanied by a list of 100 answer candidates, and the goal is to find the correct answer among all the candidates.

Preprocessing We truncate captions/questions/answers that are longer than 40/20/20 words, respectively. And we build a vocabulary of words that occur at least 5 times in train v1.0, resulting in $11,319$ words in the vocabulary. For word embeddings, we use pre-trained GloVe vectors (Pennington et al., 2014) for all the captions, questions and answers, concatenated with the learned word embedding from the BiLSTM encoders to further boost the performance. For image representation, we use bottom-up-attention features (Anderson et al., 2018) extracted from Faster R-CNN (Ren et al., 2015) pre-trained on Visual Genome (Krishna et al., 2017). A set of 36 features is created for each image. Each feature is a 2048-dimentional vector.

Evaluation Following Das et al. (2017a), we use a set of ranking metrics (Recall@k for $k = \{1, 5, 10\}$, mean rank, and mean reciprocal rank (MRR)), to measure the performance of retrieving the ground-truth answer from a pool of 100 candidates. Normalized Discounted Cumulative Gain (NDCG) score is also used for evaluation in the visual dialog challenge 2018 and 2019, based on which challenge winners are picked. Since this requires dense human annotations, the calculation of NDCG is only available on val v1.0, test-std v1.0, and a small subset of 2000 images from train v1.0.

Training details All three BiLSTMs used in the model are single-layer with 512 hidden units. The number of factors used in MFB is set to 5, and we use mini-batches of size 100. The maximum number of epochs is set to 20. No dataset-specific tuning or regularization is conducted except dropout (Srivastava et al., 2014) and early

[2] We have also tried the N-pair ranking loss used in Lu et al. (2017). Results are very similar to each other.

[3] As suggested in https://visualdialog.org/data, results should be reported on v1.0, instead of v0.9.

[4] https://evalai.cloudcv.org/web/challenges/challenge-page/161/overview

Model	NDCG	MRR	R@1	R@5	R@10	Mean
MN-D (Das et al., 2017a)	55.13	60.42	46.09	78.14	88.05	4.63
HCIAE-D (Lu et al., 2017)	57.65	62.96	48.94	80.50	89.66	4.24
CoAtt-D (Wu et al., 2018)	57.72	62.91	48.86	80.41	89.83	4.21
ReDAN-D (T=1)	58.49	63.35	49.47	80.72	90.05	4.19
ReDAN-D (T=2)	59.26	63.46	49.61	80.75	89.96	4.15
ReDAN-D (T=3)	59.32	64.21	50.60	81.39	90.26	4.05
Ensemble of 4	**60.53**	**65.30**	**51.67**	**82.40**	**91.09**	**3.82**

Table 1: Comparison of ReDAN with a discriminative decoder to state-of-the-art methods on VisDial v1.0 validation set. Higher score is better for NDCG, MRR and Recall@k, while lower score is better for mean rank. All these baselines are re-implemented with bottom-up features and incorporated with GloVe vectors for fair comparison.

Model	NDCG	MRR	R@1	R@5	R@10	Mean
MN-G (Das et al., 2017a)	56.99	47.83	38.01	57.49	64.08	18.76
HCIAE-G (Lu et al., 2017)	59.70	49.07	39.72	58.23	64.73	18.43
CoAtt-G (Wu et al., 2018)	59.24	49.64	40.09	59.37	65.92	17.86
ReDAN-G (T=1)	59.41	49.60	39.95	59.32	65.97	17.79
ReDAN-G (T=2)	60.11	49.96	40.36	59.72	66.57	17.53
ReDAN-G (T=3)	60.47	50.02	40.27	59.93	66.78	17.40
Ensemble of 4	**61.43**	**50.41**	**40.85**	**60.08**	**67.17**	**17.38**

Table 2: Comparison of ReDAN with a generative decoder to state-of-the-art generative methods on VisDial val v1.0. All the baseline models are re-implemented with bottom-up features and incorporated with GloVe vectors for fair comparison.

stopping on validation sets. The dropout ratio is 0.2. The Adam algorithm (Kingma and Ba, 2014) with learning rate 4×10^{-4} is used for optimization. The learning rate is halved every 10 epochs.

4.2 Quantitative Results

Baselines We compare our proposed approach with state-of-the-art models, including Memory Network (MN) (Das et al., 2017a), History-Conditioned Image Attentive Encoder (HCIAE) (Lu et al., 2017) and Sequential Co-Attention model (CoAtt) (Wu et al., 2018). In their original papers, all these models used VGG-Net (Simonyan and Zisserman, 2014) for image feature extraction, and reported results on VisDial v0.9. Since bottom-up-attention features have proven to achieve consistently better performance than VGG-Net in other tasks, we re-implemented all these models with bottom-up-attention features, and used the same cross-entropy loss for training. Further, unidirectional LSTMs are used in these previous baselines, which are replaced by bidirectional LSTMs with self-attention mechanisms for fair comparison. All the baselines are also further incorporated with pre-trained GloVe vectors. We choose the best three models on VisDial v0.9 as the baselines:

- **MN** (Das et al., 2017a): (i) mean pooling is performed over the bottom-up-attention features for image representation; (ii) image and question attend to the dialog history.
- **HCIAE** (Lu et al., 2017): (i) question attends to dialog history; (ii) then, question and the attended history attend to the image.
- **CoAtt** (Wu et al., 2018): (i) question attends to the image; (ii) question and image attend to the history; (iii) image and history attend to the question; (iv) question and history attend to the image again.

Results on VisDial val v1.0 Experimental results on val v1.0 are shown in Table 1. "-D" denotes that a discriminative decoder is used. With only one reasoning step, our ReDAN model already achieves better performance than CoAtt, which is the previous best-performing model. Using two or three reasoning steps further increases the performance. Further increasing the number of reasoning steps does not help, thus results are not shown. We also report results on an ensemble of 4 ReDAN-D models. Significant improvement was observed, boosting NDCG from 59.32 to 60.53, and MRR from 64.21 to 65.30.

In addition to discriminative decoders, we also evaluate our model with a generative decoder. Results are summarized in Table 2. Similar to Table 1, ReDAN-G with T=3 also achieves the best performance. It is intuitive to observe that ReDAN-D achieves much better results than ReDAN-G on MRR, R@k and Mean Rank, since ReDAN-D is a discriminative model, and utilizes much more information than ReDAN-G. For ex-

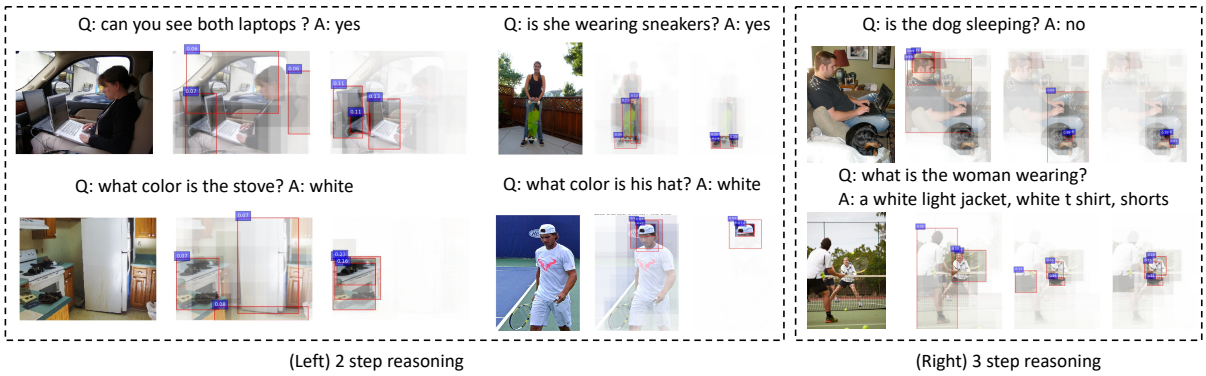


Figure 3: Visualization of learned attention maps in multiple reasoning steps.

Model	Ens. Method	NDCG	MRR	R@1	R@5	R@10	Mean
4 Dis.	Average	60.53	65.30	51.67	82.40	91.09	3.82
4 Gen.	Average	61.43	50.41	40.85	60.08	67.17	17.38
1 Dis. + 1 Gen.	Average	63.85	53.53	42.16	65.43	74.36	9.00
1 Dis. + 1 Gen.	Reciprocal	63.18	59.03	42.33	78.71	88.13	4.88
4 Dis. + 4 Gen.	Average	65.13	54.19	42.92	66.25	74.88	8.74
4 Dis. + 4 Gen.	Reciprocal	64.75	61.33	45.52	80.67	89.55	4.41
ReDAN+ (Diverse Ens.)	Average	**67.12**	56.77	44.65	69.47	79.90	5.96

Table 3: Results of different rank aggregation methods. Dis. and Gen. is short for discriminative and generative model, respectively.

ample, ReDAN-D uses both positive and negative answer candidates for ranking/classification, while ReDAN-G only uses positive answer candidates for generation. However, interestingly, ReDAN-G achieves better NDCG scores than ReDAN-D (61.43 vs 60.53). We provide some detailed analysis in the question-type analysis section below.

4.3 Qualitative Analysis

In addition to the examples illustrated in Figure 1b, Figure 3 provide six more examples to visualize the learned attention maps. The associated dialog histories are omitted for simplicity. Typically, the attention maps become sharper and more focused throughout the reasoning process. During multiple steps, the model gradually learns to narrow down to the image regions of key objects relevant to the questions ("*laptops*", "*stove*", "*sneakers*", "*hat*", "*dog's eyes*" and "*woman's clothes*"). For instance, in the top-right example, the model focuses on the wrong region ("*man*") in the 1st step, but gradually shifts its focus to the correct regions ("*dog's eyes*") in the later steps.

4.4 Visual Dialog Challenge 2019

Now, we discuss how we further boost the performance of ReDAN for participating Visual Dialog Challenge 2019[5].

Rank Aggregation As shown in Table 1 and 2, ensemble of discriminative or generative models increase the NDCG score to some extent. Empirically, we found that aggregating the ranking results of both discriminative and generative models readily boost the performance. Results are summarized in Table 3. Combining one discriminative and one generative model already shows much better NDCG results than ensemble of 4 discriminative models. The ensemble of 4 discriminative and 4 generative models further boosts the performance. It is interesting to note that using average of the ranks results in better NDCG than using reciprocal of the ranks, though the reciprocal method achieves better results on the other metrics. Since NDCG is the metric we mostly care about, the method of averaging ranking results from different models is adopted.

Finally, we have tried using different image feature inputs, and incorporating relation-aware encoders (Li et al., 2019) into ReDAN to further boost the performance. By this diverse set of ensembles (called ReDAN+), we achieve an NDCG score of 67.12% on the val v1.0 set.

[5] https://visualdialog.org/challenge/2019

Model	NDCG	MRR	R@1	R@5	R@10	Mean
ReDAN+ (Diverse Ens.)	**64.47**	53.73	42.45	64.68	75.68	6.63
ReDAN (1 Dis. + 1 Gen.)	61.86	53.13	41.38	66.07	74.50	8.91
DAN (Kang et al., 2019)	59.36	64.92	51.28	81.60	90.88	3.92
NMN (Kottur et al., 2018)	58.10	58.80	44.15	76.88	86.88	4.81
Sync (Guo et al., 2019)	57.88	63.42	49.30	80.77	90.68	3.97
HACAN (Yang et al., 2019)	57.17	64.22	50.88	80.63	89.45	4.20
FGA†	57.13	69.25	55.65	86.73	94.05	3.14
USTC-YTH‡	56.47	61.44	47.65	78.13	87.88	4.65
RvA (Niu et al., 2018)	55.59	63.03	49.03	80.40	89.83	4.18
MS ConvAI‡	55.35	63.27	49.53	80.40	89.60	4.15
CorefNMN (Kottur et al., 2018)	54.70	61.50	47.55	78.10	88.80	4.40
FGA (Schwartz et al., 2019)	54.46	67.25	53.40	85.28	92.70	3.54
GNN (Zheng et al., 2019)	52.82	61.37	47.33	77.98	87.83	4.57
LF-Att w/ bottom-up†	51.63	60.41	46.18	77.80	87.30	4.75
LF-Att‡	49.76	57.07	42.08	74.83	85.05	5.41
MN-Att‡	49.58	56.90	42.43	74.00	84.35	5.59
MN‡	47.50	55.49	40.98	72.30	83.30	5.92
HRE‡	45.46	54.16	39.93	70.45	81.50	6.41
LF‡	45.31	55.42	40.95	72.45	82.83	5.95

Table 4: Comparison of ReDAN to state-of-the-art visual dialog models on the blind test-std v1.0 set, as reported by the test server. (†) taken from https://evalai.cloudcv.org/web/challenges/challenge-page/161/leaderboard/483. (‡) taken from https://evalai.cloudcv.org/web/challenges/challenge-page/103/leaderboard/298.

Question Type	All	Yes/no	Number	Color	Others
Percentage	100%	75%	3%	11%	11%
Dis.	59.32	60.89	44.47	58.13	52.68
Gen.	60.42	63.49	41.09	52.16	51.45
4 Dis. + 4 Gen.	65.13	68.04	46.61	57.49	57.50
ReDAN+	67.12	69.49	50.10	62.70	58.50

Table 5: Question-type analysis of the NDCG score achieved by different models on the val v1.0 set.

Results on VisDial test-std v1.0 We also evaluate the proposed ReDAN on the blind test-std v1.0 set, by submitting results to the online evaluation server. Table 4 shows the comparison between our model and state-of-the-art visual dialog models. By using a diverse set of ensembles, ReDAN+ outperforms the state of the art method, DAN (Kottur et al., 2018), by a significant margin, lifting NDCG from 59.36% to 64.47%.

Question-Type Analysis We further perform a question-type analysis of the NDCG scores achieved by different models. We classify questions into 4 categories: *Yes/no*, *Number*, *Color*, and *Others*. As illustrated in Table 5, in terms of the NDCG score, generative models performed better on Yes/no questions, while discriminative models performed better on all the other types of questions. We hypothesize that this is due to that generative models tend to ranking short answers higher, thus is beneficial for Yes/no questions. Since Yes/no questions take a majority of all the questions (75%), the better performance of generative models on the Yes/no questions translated into an overall better performance of generative models. Aggregating the ranking results of both discriminative and generative models results in the mutual enhancement of each other, and therefore boosting the final NDCG score by a large margin. Also, we observe that the Number questions are most difficult to answer, since training a model to count is a challenging research problem.

5 Conclusion

We have presented Recurrent Dual Attention Network (ReDAN), a new multimodal framework for visual dialog, by incorporating image and dialog history context via a recurrently-updated query vector for multi-step reasoning. This iterative reasoning process enables model to achieve a fine-grained understanding of multimodal context, thus boosting question answering performance over state-of-the-art methods. Experiments on the VisDial dataset validate the effectiveness of the proposed approach.

Acknowledgements We thank Yuwei Fang, Huazheng Wang and Junjie Hu for helpful discussions. We thank anonymous reviewers for their constructive feedbacks.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *CVPR*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. Evaluating visual conversational agents via cooperative human-ai games. In *HCOMP*.

Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. 2018. Language-based image editing with recurrent attentive models. In *CVPR*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *EMNLP*.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-attention neural networks for reading comprehension. In *ACL*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual dialog. In *CVPR*.

Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *ACL*.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *CVPR*.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. *arXiv preprint arXiv:1809.08267*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. 2015. Draw: A recurrent neural network for image generation. In *ICML*.

Dalu Guo, Chang Xu, and Dacheng Tao. 2019. Image-question-answer synergistic network for visual dialog. *arXiv preprint arXiv:1902.09774*.

Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, and Rogerio Schmidt Feris. 2018. Dialog-based interactive image retrieval. In *NIPS*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *ICLR*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*.

Drew A Hudson and Christopher D Manning. 2018. Compositional attention networks for machine reasoning. In *ICLR*.

Unnat Jain, Svetlana Lazebnik, and Alexander G Schwing. 2018. Two can play this game: visual dialog with discriminative question generation and answering. In *CVPR*.

Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. 2019. Dual attention networks for visual reference resolution in visual dialog. *arXiv preprint arXiv:1902.09368*.

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Hadamard product for low-rank bilinear pooling. In *ICLR*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *ICLR*.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *ECCV*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.

Sang-Woo Lee, Yu-Jung Heo, and Byoung-Tak Zhang. 2018. Answerer in questioner's mind for goal-oriented visual dialogue. In *NIPS*.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *EMNLP*.

Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Relation-aware graph attention network for visual question answering. *arXiv preprint arXiv:1903.12314*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2018. Stochastic answer networks for machine reading comprehension. In *ACL*.

Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *NIPS*.

Daniela Massiceti, N Siddharth, Puneet K Dokania, and Philip HS Torr. 2018. Flipdial: A generative model for two-way visual dialogue. In *CVPR*.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. In *NIPS*.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *CVPR*.

Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. 2018. Recursive visual attention in visual dialog. *arXiv preprint arXiv:1812.02664*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2018. Coqa: A conversational question answering challenge. In *EMNLP*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander Schwing. 2019. Factor graph attention. *arXiv preprint arXiv:1904.05880*.

Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. 2017. Visual reference resolution using attention memory for visual dialog. In *NIPS*.

Ravi Shekhar, Tim Baumgartner, Aashish Venkatesh, Elia Bruni, Raffaella Bernardi, and Raquel Fernandez. 2018. Ask no more: Deciding when to guess in referential visual dialogue. In *COLING*.

Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. Reasonet: Learning to stop reading in machine comprehension. In *KDD*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *NIPS*.

Alessandro Sordoni, Philip Bachman, Adam Trischler, and Yoshua Bengio. 2016. Iterative alternating neural attention for machine reading. *arXiv preprint arXiv:1606.02245*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*.

Florian Strub, Harm De Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *IJCAI*.

Florian Strub, Mathieu Seurin, Ethan Perez, Harm de Vries, Philippe Preux, Aaron Courville, Olivier Pietquin, et al. 2018. Visual reasoning with multi-hop feature modulation. In *ECCV*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *CVPR*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. 2018. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *CVPR*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. 2019. Making history matter: Gold-critic sequence training for visual dialog. *arXiv preprint arXiv:1902.09326*.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *CVPR*.

Adams Wei Yu, Hongrae Lee, and Quoc V Le. 2017a. Learning to skim text. *arXiv preprint arXiv:1704.06877*.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017b. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017c. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*.

Heming Zhang, Shalini Ghosh, Larry Heck, Stephen Walsh, Junting Zhang, Jie Zhang, and C-C Jay Kuo. 2019. Generative visual dialogue system via adaptive reasoning and weighted likelihood estimation. *arXiv preprint arXiv:1902.09818*.

Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, Jianfeng Lu, and Anton Van Den Hengel. 2018. Goal-oriented visual question generation via intermediate rewards. In *ECCV*.

Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. 2019. Reasoning visual dialogs with structural and partial observations. *arXiv preprint arXiv:1904.05548*.