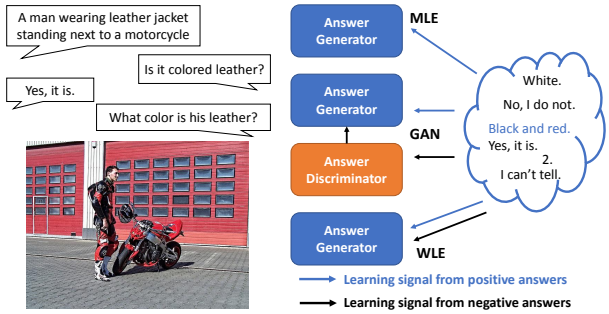
Generative Visual Dialogue System via Weighted Likelihood Estimation

Heming Zhang[1*], **Shalini Ghosh**[2], **Larry Heck**[2], **Stephen Walsh**[2],
Junting Zhang[1], **Jie Zhang**[3] and **C.-C. Jay Kuo**[1]
[1]University of Southern California
[2]Samsung Research America
[3]Arizona State University
{hemingzh, juntingz, cckuo}@usc.edu, {shalini.glosh, larry.h, s1.walsh}@samsung.com,
jiezhang.joena@asu.edu

Abstract

The key challenge of generative Visual Dialogue (VD) systems is to respond to human queries with informative answers in natural and contiguous conversation flow. Traditional Maximum Likelihood Estimation-based methods only learn from positive responses but ignore the negative responses, and consequently tend to yield safe or generic responses. To address this issue, we propose a novel training scheme in conjunction with weighted likelihood estimation method. Furthermore, an adaptive multi-modal reasoning module is designed, to accommodate various dialogue scenarios automatically and select relevant information accordingly. The experimental results on the VisDial benchmark demonstrate the superiority of our proposed algorithm over other state-of-the-art approaches, with an improvement of 5.81% on recall@10.



(a) (b)

Figure 1: (a) An example from the VisDial dataset, and (b) comparison between MLE, GAN and WLE, where positive responses are highlighted in blue. The MLE-based generator learn from data in positive answers only. The GAN-based generator learn from data in negative answers through discriminators indirectly. Our WLE-based generator learn from data in both positive and negative answers.

1 Introduction

Artificial Intelligence (AI) has witnessed rapid resurgence in recent years, due to many innovations in deep learning. Exciting results have been obtained in computer vision (*e.g.*, image classification [Simonyan and Zisserman, 2015; He *et al.*, 2016], detection [Ren *et al.*, 2015; Lin *et al.*, 2017; Zhang *et al.*, 2018a], etc.) as well as natural language processing (NLP) (*e.g.*, [Wen *et al.*, 2016; Li *et al.*, 2017; Zhang *et al.*, 2018b], etc.). Good progress has also been made by researchers in vision-grounded NLP tasks such as image captioning [You *et al.*, 2016; Krishna *et al.*, 2017] and visual question answering [Antol *et al.*, 2015; Malinowski *et al.*, 2015]. Proposed recently, the Visual Dialogue (VD) [Das *et al.*, 2017] task leads to a higher level of interaction between vision and language. In the VD task, a machine conducts natural language dialogues with humans by answering questions grounded in an image. It requires not only reasoning on vision and language, but also generating consistent and natural dialogues.

Existing VD systems can be summarized into two tracks [Das *et al.*, 2017]: generative models and discriminative models. The system adopting the generative model can generate responses while that using the disriminative model only chooses responses from a candidate set. Although discriminative models achieved better recall performance on the benchmark dataset [Das *et al.*, 2017], they are not as applicable as generative models in real world scenarios since candidate responses may not be available. In this work, we focus on the design of generative VD systems for broader usage.

One main weakness of existing generative models trained by the maximum likelihood estimation (MLE) method is that they tend to provide frequent and generic responses like 'Don't know'. This happens because the MLE training paradigm latches on to frequent generic responses [Lu *et al.*, 2017]. They may match well with some but poorly for others. There are many possible paths a dialogue may take in the future — penalizing generic poor responses can eliminate candidate dialogue paths and avoids abuse of frequent responses. This helps bridge the large performance gap between generative/discriminative VD systems.

To reach this goal, we propose a novel weighted likelihood estimation (WLE) based training scheme. Specifically, instead of assigning equal weights to each training sample as done in the MLE, we assign a different weight to each training sample. The weight of a training sample is determined by its positive response as well as the negative ones. By in-

*Contact Author

corporating supervision from both positive and negative responses, we enhance answer diversity in the resulting generative model. The proposed training scheme is effective in boosting the VD performance and easy to implement.

Another challenge for VD systems is effective reasoning based on multi-modal inputs. Previous work pre-defined a set of reasoning paths based on multi-modal inputs. The path is specified by a certain sequential processing order, *e.g.*, human queries followed by the dialogue history and then followed by image analysis [Lu *et al.*, 2017]. Such a pre-defined order is not capable of handling different dialogue scenarios, e.g., answering a follow-up question of 'Is there anything else on the table?'. We believe that a good reasoning strategy should determine the processing order by itself. Here, we propose a new reasoning module, where an adaptive reasoning path accommodates different dialogue scenarios automatically.

There are three major contributions of this work. First, an effective training scheme for the generative VD system is proposed, which directly exploits both positive and negative responses using an unprecedented likelihood estimation method. Second, we design an adaptive reasoning scheme with unconstrained attention on multi-modal inputs to accommodate different dialogue scenarios automatically. Third, our results demonstrate the state-of-the-art performance on the VisDial dataset [Das *et al.*, 2017]. Specifically, our model outperforms the best previous generative-model-based method [Wu *et al.*, 2018] by 3.06%, 5.81% and 5.28 with respect to the recall@5, the recall@10 and the mean rank performance metrics, respectively.

2 Related Work

Visual dialogue. Different visual dialogue tasks have been examined recently. The VisDial dataset [Das *et al.*, 2017] is collected from free-form human dialogues with a goal to answer questions related to a given image. The GuessWhat task [De Vries *et al.*, 2017] is a guessing game with goal-driven dialogues so as to identify a certain object in a given image by asking yes/no questions. In this work, we focus on the VisDial task. Most previous research on the VisDial task follows the encoder-decoder framework in [Sutskever *et al.*, 2014]. Exploration on encoder models includes late fusion [Das *et al.*, 2017], hierarchical recurrent network [Das *et al.*, 2017], memory network [Das *et al.*, 2017], history-conditioned image attentive encoder (HCIAE) [Lu *et al.*, 2017], and sequential co-attention (CoAtt) [Wu *et al.*, 2018]. Decoder models can be classified into two types: (a) Discriminative decoders rank candidate responses using cross-entropy loss [Das *et al.*, 2017] or n-pair loss [Lu *et al.*, 2017]; (b) Generative decoders yield responses using MLE [Das *et al.*, 2017], which can be further combined with adversarial training [Lu *et al.*, 2017; Wu *et al.*, 2018]. The latter involves a discriminator trained on both positive and negative responses, and its discriminative power is then transferred to the generator via auxiliary adversarial training.

Weighted likelihood estimation. Being distinct from previous generative work that uses either MLE or adversarial training, we use WLE and develop a new training scheme for VD systems in this work. WLE has been utilized for different purposes. For example, it was introduced in [Warm, 1989] to remove the first-order bias in MLE. Smaller weights are assigned to outliers for training to reduce the effect of outliers [Ning *et al.*, 2015]. The binary indicator function and the similarity scores are compared for weighting the likelihood in visual question answering (VQA) in [Hu *et al.*, 2018]. We design a novel weighted likelihood remotely related to these concepts, to utilize both positive and negative responses.

Hard example mining. Hard example mining methods are frequently seen in object detection algorithms, where the amount of background samples is much more than the object samples. In [Rowley, 1999], the proposed face detector is trained until convergence on sub-datasets and applied to more data to mine the hard examples alternatively. Online hard example mining is favored by later work [Shrivastava *et al.*, 2016; Lin *et al.*, 2017], where the softmax-based cross entropy loss is used to determine the difficulty of samples. We adopt the concept of sample difficulty and propose a novel way to find hard examples without the preliminary of softmax-based cross entropy.

Multi-modal reasoning. Multi-modal reasoning involves extracting and combining useful information from multi-modal inputs. It is widely used in the intersection of vision and language, such as image captioning [Xu *et al.*, 2015] and VQA [Xu and Saenko, 2016]. For the VD task, reasoning can be applied to images (I), questions (Q) and history dialogues (H). In [Lu *et al.*, 2017], the reasoning path adopts the order "Q $\rightarrow$ H $\rightarrow$ I". This order is further refined to "Q $\rightarrow$ I $\rightarrow$ H $\rightarrow$ Q" in [Wu *et al.*, 2018]. In the recent arxiv paper [Gan *et al.*, 2019], the reasoning sequence of "Q $\rightarrow$ I $\rightarrow$ H " is recurrently occurring to solve complicated problems. Unlike previous work that defines the reasoning path order a priori, we propose an adaptive reasoning scheme with no pre-defined reasoning order.

3 Proposed Generative VD System

In this section, we describe our approach to construct and train the proposed generative visual dialogue system. Following the problem formulation in [Das *et al.*, 2017], the input consists of an image I, a 'ground-truth' dialogue history $H_{t-1} = (\underbrace{C}_{h_0}, \underbrace{(Q_1, A_1)}_{h_1}, \cdots, \underbrace{(Q_{t-1}, A_{t-1})}_{h_{t-1}})$ with image caption C and a follow-up question Q_t at round t. N candidate responses $\mathcal{A}_t = \{A_t^1, A_t^2, \cdots, A_t^N\}$ are provided for both training and testing. Figure 1(a) shows an example from VisDial [Das *et al.*, 2017].

We adopt the encoder-decoder framework [Sutskever *et al.*, 2014]. Our proposed encoder, which involves an adaptive multi-modal reasoning module without pre-defined order, will be described in details in Sec. 3.1. The generative decoder receives the embedding of the input triplet $\{I, H_{t-1}, Q_t\}$ from the encoder and outputs a response sequence $\hat{A}_t$. Our VD system is trained using a novel training scheme with weighted likelihood estimation, which will be described in Sec. 3.2 with details.

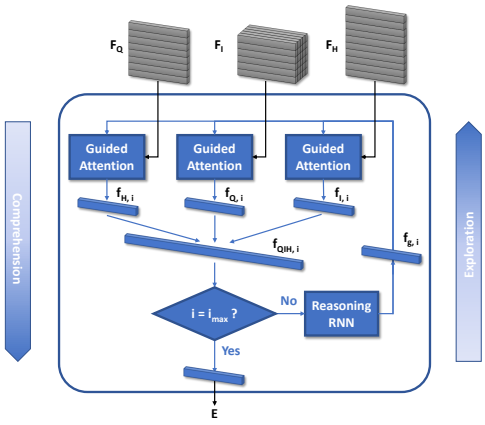


Figure 2: The adaptive multi-modal reasoning.

3.1 Adaptive Multi-modal Reasoning (AMR)

To conduct reasoning on multi-modal inputs, we first extract image feature $F_I \in \mathbb{R}^{N\times H\times W}$ by a convolutional neural network, where N is the length of the feature, and H and W are the height and width of the output feature map. The question feature $F_Q \in \mathbb{R}^{N\times l_Q}$ and history feature $F_H \in \mathbb{R}^{N\times l_H}$ are obtained by recurrent neural networks, where l_Q and l_H are the length of the question and the history, respectively.

Our reasoning path consists of two main steps, namely the comprehension step and the exploration step, in a recurrent manner. In the comprehension step, useful information from each input modality is extracted. It is apparent that not all the input information is equally important in the conversation. Attention mechanism is thus useful to extract relevant information. In the exploration step, the relevant information is processed and the following attention direction is determined accordingly. Along the reasoning path, these two steps are performed alternatively.

In [Lu *et al.*, 2017; Wu *et al.*, 2018], the comprehension and exploration steps are merged together. The reasoning scheme focuses on one single input modality at each time and follows a pre-defined reasoning sequence through each input modality. However, this pre-defined order cannot accommodate various dialogue scenarios in real world. For example, a question of "How many people are there in the image?" should yield a short reasoning sequence like

$$\underset{\text{the word 'people'}}{\texttt{question}} \rightarrow \underset{\text{regions of people}}{\texttt{image}},$$

whereas a question of "Is there anything else on the table?" should result in a long reasoning sequence such as

$$\underset{\text{the word 'table'}}{\texttt{question}} \rightarrow \underset{\text{regions of table}}{\texttt{image}} \rightarrow \underset{\text{the word 'else'}}{\texttt{question}} \rightarrow \underset{\text{context for 'else'}}{\texttt{history}}.$$

To overcome the drawback of pre-defined reasoning sequence, we propose an adaptive multi-modal reasoning module as illustrated in Figure 2.

Let λ denote any multi-modal feature type (image, question or history), and $F_\lambda \in \mathbb{R}^{N\times M}$ denote the features to be attended, where M is the number of features. The guided attention operation that paying attention according to the given guide is denoted as $\mathbf{f}_\lambda = \text{GuidedAtt}(F_\lambda, \mathbf{f}_g)$, where $\mathbf{f}_g \in \mathbb{R}^{N\times 1}$ is the attention guiding feature. The guided attention can be expressed as:

$$E_\lambda = \tanh(W_\lambda F_\lambda + W_g \mathbf{f}_g \mathbf{1}^T) \quad (1)$$

$$\mathbf{a}_\lambda = \text{softmax}(E_\lambda^T \mathbf{w}_{att}) \quad (2)$$

$$\mathbf{f}_\lambda = F_\lambda \mathbf{a}_\lambda, \quad (3)$$

where W_λ, W_g and $\mathbf{w}_{att}$ are learnable weights, $\mathbf{1}$ is a vector with all elements set to 1.

In time step i, the image features F_I, the question features F_Q and the history features F_H are attended separately by their own guided attention blocks. During the comprehension step, the outputs of the guided attention blocks $\mathbf{f}_{I,i}$, $\mathbf{f}_{Q,i}$ and $\mathbf{f}_{H,i}$, *i.e.* the extracted information from each modality, are merged into $\mathbf{f}_{QIH,i}$. During the exploration step, the merged vector is processed in the reasoning RNN block, which generates the new attention guiding feature $\mathbf{f}_{g,i}$ to guide the attention in time step $i+1$. The final embedding feature $\mathbf{E}$ is

$$\mathbf{E} = \tanh(W \mathbf{f}_{QIH,i_{max}}), \quad (4)$$

where W is learnable weights, i_{max} is the maximum number of recurrent steps.

Through this mechanism, the reasoning RNN block maintains a global view of the multi-modal features and reasons what information should be extracted in the next time step. The information extraction order and subject are therefore determined adaptively along the reasoning path.

3.2 WLE-based Training Scheme

As the discriminative VD models are trained to differentiate positive and negative responses, they perform better on the standard discriminative benchmark. In contrast, the generative visual dialogue models are trained to only maximize the likelihood of positive responses. The MLE loss function is expressed as:

$$L_{MLE} = \textstyle\sum_m -\log(p_m^{pos}), \quad (5)$$

where p_m^{pos} denotes the estimated likelihood of the positive response of sample m. There is only one positive response per sample provided for training in the VisDial task. However, there are many possible paths a dialogue may take in the future, the MLE approach therefore favors the frequent and generic responses when the training data is limited [Lu *et al.*, 2017]. In the VisDial task, negative responses are selected from positive responses to other questions, including frequent and generic responses. Incorporating the negative responses to maximize the learning from all available information is thus essential to improve the generative models.

We propose a WLE based training scheme to utilize the negative responses and remedy the bias of MLE. Rather than treating each sample with equal importance, we assign a weight α_m to each estimated log-likelihood as:

$$L_{WLE} = \textstyle\sum_m -\alpha_m \log(p_m^{pos}). \quad (6)$$

We can interpret the weighted likelihood as a hard sample mining process. We are inspired by OHEM [Shrivastava *et al.*, 2016] and focal loss [Lin *et al.*, 2017] designed for object detection, where hard samples are mined using their loss

Model	MRR	R@1	R@5	R@10	Mean
LF [Das *et al.*, 2017]	0.5199	41.83	61.78	67.59	17.07
HREA [Das *et al.*, 2017]	0.5242	42.28	62.33	68.17	16.79
MN [Das *et al.*, 2017]	0.5259	42.29	62.85	68.88	17.06
HCIAE [Lu *et al.*, 2017]	0.5467	44.35	65.28	71.55	14.23
FlipDial [Massiceti *et al.*, 2018]	0.4549	34.08	56.18	61.11	20.38
CoAtt [Wu *et al.*, 2018]	0.5578	**46.10**	65.69	71.74	14.43
Coref [Kottur *et al.*, 2018]	0.5350	43.66	63.54	69.93	15.69
Ours	**0.5614**	44.49	**68.75**	**77.55**	**9.15**

Table 1: Performance of generative models on VisDial 0.9. 'Mean' denotes mean rank, for which lower is better. All the models use VGG as backbone except for Coref which uses ResNet.

values and receive extra attention. Rather than using the preliminary softmax cross entropy loss for discriminative learning, we propose to use likelihood estimation to mine the hard samples. If the current model cannot predict the likelihood for a sample well, it indicates that this sample is hard for the model. Then we should increase the weight for this hard sample and vice versa.

Given both positive and negative responses for training, we propose to assign weights as:

$$\beta_{m,n} = 1 - \frac{\log(p^{neg}_{m,n})}{\log(p^{pos}_{m})}, \tag{7}$$

$$\tilde{\beta}_m = \exp\left(\tau \max_n \left(\beta_{m,n}\right)\right), \tag{8}$$

$$\alpha_m = \max\left(\tilde{\beta}_m,\ \gamma\right), \tag{9}$$

where $p^{neg}_{m,n}$ denotes the n-th negative response of sample m, τ and γ are hyper-parameters to shape the weights.

We can also view the proposed loss function as a ranking loss. We assign a weight to a sample by comparing the estimated likelihood of its positive and negative responses. $\beta_{m,n}$ measures the relative distance of likelihood between the positive response and the n-th negative response of sample m. If the likelihood of a positive response is low comparing to the negative responses, we should penalize more by increasing the weight for this sample. If the estimated likelihood of a positive sample is already very high, we should lower its weight to reduce the penalization.

4 Experiments

4.1 Dataset

We evaluate our proposed model on the VisDial dataset [Das *et al.*, 2017]. In VisDial v0.9, on which most previous work has benchmarked, there are in total 83k and 40k dialogues on COCO-train and COCO-val images, respectively. We follow the methodology in [Lu *et al.*, 2017] and split the data into 82k for `train`, 1k for `val` and 40k for `test`. In the new version VisDial v1.0, which was used for the Visual Dialog Challenge 2018, `train` consists of the previous 123k images and corresponding dialogues. 2k and 8k images with dialogues are collected for `val` and `test`, respectively.

Question	Is there any people do you see ?	What type of bag it is?	What color are the kites?	Is it sunny in the pic?
Human Response	0	A leather bag	All different colors	Ys
Rank 1	No	A leather bag	All different colors	Yes
Rank 2	No people	UNK	White and Red	Overcast
Rank 3	No people in picture	PC	Yellow and green	Partly
Rank 4	0	Tennis	Blue and white	Kind of
Rank 5	No, there aren't any people around	Not sure	Yellow and green	It is sunny
Rank 6	No 1 else	A lab	Red, black and blue	No, cloudy
Rank 7	'no	Folding wooden	They are green	No, very overcast
Rank 8	Nope	Banquet maybe	Green and white	No kind of overcast
Rank 9	No, I can't see any	An orange chair	Both are black	Ys
Rank 10	Nope, just a bear	Restaurant	They are both black	Looks like a overcast

Figure 3: Examples of top-10 responses ranked by our model. When there are multiple correct responses to the question, our model may choose other candidates that are semantically similar to the human response. The human responses are highlighted in blue.

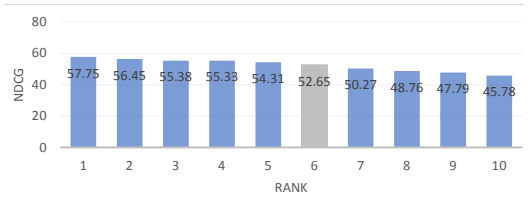


Figure 4: Results of the top-10 teams in the first visual dialog challenge. As the only team in top-10 uses generative visual dialogue system, we are ranked as the 6th place (highlighted with gray color). Our NDCG score is comparable with other discriminative systems.

Each question is supplemented with 100 candidate responses, among which only one is the human response for this question. Following the evaluation protocol in [Das *et al.*, 2017], we rank the 100 candidate responses by their estimated likelihood and evaluate the models using standard retrieval metrics: (1) mean rank of the human response, (2) recall rate of the human response in top-k ranked responses for $k = 1, 5, 10$, (3) mean reciprocal rank (MRR) of the human response, (4) normalized discounted cumulative gain (NDCG) of all correct responses (only available for v1.0).

4.2 Implementation Details

We follow the procedures in [Lu *et al.*, 2017] to pre-process the data. The captions, questions and answers are truncated at 24, 16 and 8 words for VisDial v0.9, and 40, 20 and 20 words for VisDial v1.0. Vocabularies are built afterwards from the words that occur at least five times in `train`. We use 512D word embeddings, which are trained from scratch and shared by question, dialogue history and decoder LSTMs.

For a fair comparison with previous work, we adopt the simple LSTM decoder with a softmax output which models the likelihood of the next word given the embedding feature and previous generated sequence. We also set all LSTMs to have single layer with 512D hidden state for consistency with other works. We extract image features from pre-trained CNN models (VGG [Simonyan and Zisserman, 2015] for VisDial v0.9, ResNet [He *et al.*, 2016] or bottom-up features [Anderson *et al.*, 2018] for VisDial v1.0), and train the rest of our model from scratch. We use the Adam optimizer with the base learning rate of 4×10^{-4}.

Model	MRR	R@1	R@5	R@10	Mean
MN [Das *et al.*, 2017]	0.4799	38.18	57.54	64.32	18.60
HCIAE [Lu *et al.*, 2017]	0.4910	39.35	58.49	64.70	18.46
CoAtt [Wu *et al.*, 2018]	0.4925	39.66	58.83	65.38	18.15
ReDAN [Gan *et al.*, 2019]	0.4969	**40.19**	59.35	66.06	17.92
Ours	**0.5015**	38.26	**62.54**	**72.79**	**10.71**

Table 2: Performance of generative models on VisDial v1.0 val. Results of previous work are reported by ReDAN.

Model	MRR	R@1	R@5	R@10	Mean
HCIAE-MLE	0.5386	44.06	63.55	69.24	16.01
HCIAE-GAN	0.5467	44.35	65.28	71.55	14.23
HCIAE-WLE	0.5494	43.43	66.88	75.59	9.93
AMR-MLE	0.5403	44.17	63.86	69.67	15.49
AMR-WLE	**0.5614**	**44.49**	**68.75**	**77.55**	**9.15**
Model	ΔMRR	ΔR@1	ΔR@5	ΔR@10	ΔMean
HCIAE-MLE	—	—	—	—	—
HCIAE-GAN	+0.0081	+0.29	+1.73	+2.31	-1.78
HCIAE-WLE	+0.0108	-0.92	+3.33	+6.35	-6.08
AMR-MLE	—	—	—	—	—
AMR-WLE	**+0.0211**	**+0.32**	**+4.89**	**+7.88**	**-6.34**

Table 3: Ablation study on VisDial 0.9. Top: absolute values. Bottom: improvement from MLE models.

4.3 Experiments Results and Analysis

Baselines

We compare our proposed model to several baselines and the state-of-the-art generative models. In [Das *et al.*, 2017], three types of encoders are introduced. Late Fusion **(LF)** extracts features from each input separately and fuses them in the later stage. Hierarchical Recurrent Encoder **(HRE)** uses hierarchical recurrent encoder for dialogue history and **HREA** adds attention to the dialogue history on top of the hierarchical recurrent encoder. Memory Network **(MN)** uses memery bank to store the dialogue history and find corresponding memory to answer the question. History-Conditioned Image Attentive Encoder **(HCIAE)** is proposed in [Lu *et al.*, 2017] to attend on image and dialogue history and trained with generative adversarial training (GAN). Another concurrent work with GAN [Wu *et al.*, 2018] proposes a co-attention model (**CoAtt**) that attends to question, image and dialogue history. **FlipDial** [Massiceti *et al.*, 2018] uses VAE for sequence generation. We also compare to a neural module network approach **Coref** [Kottur *et al.*, 2018] in which only the performance with ResNet [He *et al.*, 2016] backbone is reported. **ReDAN** [Gan *et al.*, 2019] is recently proposed method which involves a multi-step reasoning path with pre-defined order.

Results on VisDial v0.9

Table 1 compares ours results to other reported generative baselines. Our model performs the best on most of the evaluation metrics. Comparing to HCIAE [Lu *et al.*, 2017], our model shows comparable performance on R@1, and outperforms on MRR, R@5, R@10 and mean rank by 1.47%, 3.47%, 6%, 5.08, respectively. Our model also outperforms CoAtt [Wu *et al.*, 2018], which achieved the previous best results for generative models. Our result surpass it with large margins on R@5, R@10 and mean rank by 3.06%, 5.81% and 5.28, respectively.

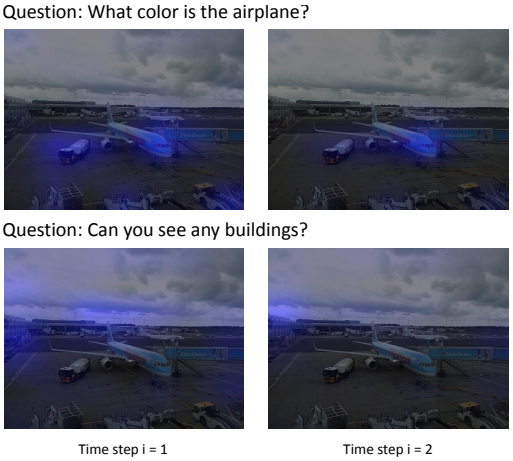


Figure 5: Visualization of image attention heatmaps for different questions and reasoning steps. Regions of attention are highlighted in blue.

While our model demonstrates remarkable improvement on R@5, R@10 and mean rank, MRR shows moderate gain while R@1 is slightly behind. We attribute this to the fact that there could be more than one correct response among the candidates while only one is provided as *the* correct answer. As demonstrated by the examples of top-10 responses in Figure 3, our model is capable of ranking multiple correct answers to higher places. However, the single human answer is not necessarily ranked the 1st, thus greatly affecting R@1.

Results on VisDial v1.0

In the Visual Dialog Challenge 2018, all correct responses in `test` are annotated by humans and considered in the evaluation. Figure 4 represents the top-10 results. Our model, as the only generative model in the top-10, ranked as the 6th among those discriminative models. It also verifies our claim that our low R@1 score on v0.9 is because the evaluation only considers the human response but ignore all other correct responses. We used ResNet features for the challenge. Since ReDAN only reports its generative performance on VisDial v1.0 `val` with bottom-up features, we also present our results using the same setting in Table 2. We list the results of previous work in Table 2 as reported in [Gan *et al.*, 2019]. Similar to the results on VisDial v0.9, our proposed method outperforms previous methods on MRR, R@5, R@10 and Mean.

Ablation Study

Our model contains two main novel components, namely the adaptive multi-modal reasoning module and the WLE based training scheme. To verify the contribution of each component, we compare the following models : (a) **HCIAE-MLE** is the HCIAE model trained via MLE; (b) **HCIAE-GAN** is the HCIAE model trained via MLE and GAN; (c) **HCIAE-WLE** is the HCIAE model trained via WLE; (d) **AMR-MLE** is our AMR model trained via MLE; (e) **AMR-WLE** is our

Images	Questions	Human responses	MLE responses	WLE responses
	What color ducks ?	They are geese they have black heads and	Black and white	Black and white
	Are they large or small ?	Large i guess , or normal sized	They are large	They are large
	Any grass ?	Yes	Yes , in the background	Yes , a lot of grass
	Any people ?	No	Yes , there are a lot of people	No
	What color grass ?	It is green with brownish dead spots	Green	Green and brown
	Any vehicles on the lot ?	Yes	Yes , quite a few	Yes , there are a lot of cars
	How many ?	I can see 6	I see 2	I see 2
4 ducks are in a grassy island of a parking lot with their heads down	What kind ?	Some suvs and some sedans	I ca n't tell	I ca n't tell what it is ,
	Do they look new or old ?	They look new	They look new	They look new
	What color is the lot?	Normal black asphalt	It is gray	It 's gray
	Are zebras in zoo ?	No zebras not in zoo	No	No , they are in wild
	Are they adults or babies ?	I would say they are adults	They look like adults	They look like adults
	Are they standing next to each other ?	1 lead in front , while other 1	Yes	They are side by side
	Is there water source in image ?	No	No water	No water is visible
	Is it sunny out ?	It does look like sunny	Yes	It appears to be
	Any trees ?	There are few trees but is not in	There are trees in background	There are trees in background
	Are zebras eating ?	No , they are walking	No	No they are not
2 zebras standing in tall savannah grass near forest brush	Any other animals in picture ?	No	No , just zebras	No , just zebras
	Is there any type of enclosure ?	No	There is no fence	There is no fence
	Is grass green ?	No	No , it is brown	No , it 's brown
	Are there umbrellas ?	Yes , 1 umbrella	Yes	Yes
	How many people are there ?	I see 5	More than 10	A lot of people
	What color is the umbrella ?	I ca n't tell , the image is	The photo is black and white	The picture is in black and white
	Are they male or female ?	Looks like 4 females and 1 male	I ca n't tell	They are too far away to tell
	Is it an old time picture ?	Modern picture , just in black and white	I do n't think so	No , it 's old
	Are they on a sidewalk ?	They 're in a parking lot	Yes	No , they are on a sidewalk
	Is there any buildings ?	Yes , across the street from	Yes , quite a few	Yes , in the background
	Is it a city street ?	Yes	Yes	Yes , it 's a city street
People walk near a street in the rain	Do they appear to know each other ?	Looks like 4 of them do , and	No	No , they look like they are walking
	Are the buildings businesses ?	Yes	I do n't think so	I ca n't tell , the buildings are

Figure 6: Qualitative results on `test`. The questions and answers are truncated at 16 and 8, respectively, same as our data pre-processing.

final model with both key components.

The results are listed in Table 3. The effectiveness of the proposed reasoning scheme is demonstrated in the HCIAE-MLE vs. AMR-MLE and HCIAE-WLE vs. AMR-WLE comparisons where our model outperforms HCIAE on all metrics. The importance of our proposed WLE is highlighted in the comparison between HCIAE-WLE and HCIAE-GAN. HCIAE-WLE performs better on all metrics. Specifically, the improvement on the HCIAE model by WLE is more than twice of that by GAN on R@10 (6.35 vs. 2.31) and mean rank (6.08 vs. 1.78). Our proposed training scheme is therefore also compatible and effective with other encoders.

Qualitative Results

Examples of image attention heatmaps are visualized in Figure 5, which demonstrate the adaptive reasoning focuses for different questions and reasoning time steps. For example, for the second question, the attention on image was first at a large area of background, then moved to more focused region to answer the question 'any buildings'.

Figure 6 shows some qualitative results on `test`. Our generative model is able to generate more non-generic answers. As evidently shown in the comparison between MLE and WLE, the WLE results are more specific and human-like.

5 Conclusion

In this work, we have presented a novel generative visual dialogue system. It involves an adaptive reasoning module for multi-modal inputs. The proposed reasoning module does not have any pre-defined sequential reasoning order and can accommodate various dialogue scenarios. The generative visual dialogue system is trained using weighted likelihood estimation, for which we design a new training scheme for generative visual dialogue systems.

References

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[Das *et al.*, 2017] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2017.

[De Vries *et al.*, 2017] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C

Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, volume 1, page 3, 2017.

[Gan *et al.*, 2019] Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. *arXiv preprint arXiv:1902.00579*, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hu *et al.*, 2018] Hexiang Hu, Wei-Lun Chao, and Fei Sha. Learning answer embeddings for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5428–5436, 2018.

[Kottur *et al.*, 2018] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–169, 2018.

[Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[Li *et al.*, 2017] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*, 2017.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *International conference on computer vision*, 2017.

[Lu *et al.*, 2017] Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems*, pages 314–324, 2017.

[Malinowski *et al.*, 2015] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2015.

[Massiceti *et al.*, 2018] Daniela Massiceti, N Siddharth, Puneet K Dokania, and Philip HS Torr. Flipdial: A generative model for two-way visual dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[Ning *et al.*, 2015] Kefeng Ning, Min Liu, and Mingyu Dong. A new robust elm method based on a bayesian framework with heavy-tailed distribution and weighted likelihood function. *Neurocomputing*, 149:891–903, 2015.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[Rowley, 1999] Henry A Rowley. Neural network-based face detection. Technical report, Carnegie-Mellon Univ Pittsburgh PA Dept of Computer Science, 1999.

[Shrivastava *et al.*, 2016] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[Warm, 1989] Thomas A Warm. Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3):427–450, 1989.

[Wen *et al.*, 2016] Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*, 2016.

[Wu *et al.*, 2018] Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[Xu and Saenko, 2016] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[You *et al.*, 2016] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2016.

[Zhang *et al.*, 2018a] Heming Zhang, Xiaolong Wang, Jingwen Zhu, and C-C Jay Kuo. Accelerating proposal generation network for fast face detection on mobile devices. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 326–330. IEEE, 2018.

[Zhang *et al.*, 2018b] Jie Zhang, Xiaolong Wang, Dawei Li, and Yalin Wang. Dynamically hierarchy revolution: dirnet for compressing recurrent neural network on mobile devices. *IJCAI*, 2018.