The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Modality-Balanced Models for Visual Dialogue

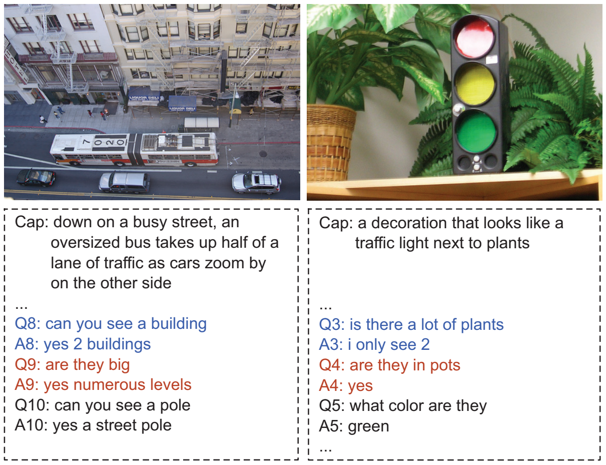
Hyounghun Kim, Hao Tan, Mohit Bansal
Department of Computer Science
University of North Carolina at Chapel Hill
{hyounghk, airsplay, mbansal}@cs.unc.edu

Abstract

The Visual Dialog task requires a model to exploit both image and conversational context information to generate the next response to the dialogue. However, via manual analysis, we find that a large number of conversational questions can be answered by only looking at the image without any access to the context history, while others still need the conversation context to predict the correct answers. We demonstrate that due to this reason, previous joint-modality (history and image) models over-rely on and are more prone to memorizing the dialogue history (e.g., by extracting certain keywords or patterns in the context information), whereas image-only models are more generalizable (because they cannot memorize or extract keywords from history) and perform substantially better at the primary normalized discounted cumulative gain (NDCG) task metric which allows multiple correct answers. Hence, this observation encourages us to explicitly maintain two models, i.e., an image-only model and an image-history joint model, and combine their complementary abilities for a more balanced multimodal model. We present multiple methods for this integration of the two models, via ensemble and consensus dropout fusion with shared parameters. Empirically, our models achieve strong results on the Visual Dialog challenge 2019 (rank 3 on NDCG and high balance across metrics), and substantially outperform the winner of the Visual Dialog challenge 2018 on most metrics.

1 Introduction

When we pursue conversations, context is important to keep the topic consistent or to answer questions which are asked by others, since most new utterances are made conditioned on related mentions or topic clues in the previous utterances in the conversation history. However, conversation history is not necessarily needed for all interactions, for instance, someone can change topics during a conversation and can ask a sudden new question which is not related to the context. This is similar to the setup in the Visual Dialog task (Das et al. 2017), in which one agent (say the 'asker') keeps asking questions and the other one (say the 'answerer') keeps answering the questions based on an image for multiple rounds. The asker can ask a question from the conversation context. Then the answerer should answer the question by considering the conversation history as well as the image information, e.g., if the asker asks a question, "Are they in pots?" (Q4 in Fig. 1), the answerer should find a clue in the past question-answer pairs "Is there a lot of plants?" - "I only see 2." (Q3-A3 in Fig. 1) and figure out what 'they' means first to answer the question correctly. On the other hand, some questions in this task are independent of the past conversation history, e.g., "Can you see a building?" (Q8 in Fig. 1), where the answerer does not need to look at conversation context and can answer the question only based on the image information.



Cap: down on a busy street, an oversized bus takes up half of a lane of traffic as cars zoom by on the other side
...
Q8: can you see a building
A8: yes 2 buildings
Q9: are they big
A9: yes numerous levels
Q10: can you see a pole
A10: yes a street pole

Cap: a decoration that looks like a traffic light next to plants
...
Q3: is there a lot of plants
A3: i only see 2
Q4: are they in pots
A4: yes
Q5: what color are they
A5: green
...

Figure 1: Examples of Visual Dialog Task. Some questions only need an image to be answered (Q8-A8 and Q3-A3 pairs in blue from each example, respectively), but others need conversation history (Q9-A9 and Q4-A4 pairs in orange from each example, respectively).

We first conduct a manual investigation on the Visual Dialog dataset (VisDial) to figure out how many questions can be answered only with images and how many of them need conversation history to be answered.[1] This investiga-

[1] We also conduct the same manual investigation to see how many questions can be answered by only looking at conversation

tion shows that around 80% of the questions can be answered only with images. Moreover, on the model side, we verify this observation by building a model that uses only images to answer questions. As expected, this image-only model works very well on the *primary* task metric of NDCG (evaluated on dense annotations which consider multiple similar answers as correct ones with similarity weights on them) without any help from the conversation history (see Table 2). However, we find that the image-only model does not get higher scores on other metrics such as mean reciprocal rank (MRR), recall@k, and mean rank (evaluated on single ground-truth answers). Because the image-only model does not use any conversation-history information, we hypothesize that this scoring behavior might be related to the amount of history information available, and hence we also conduct additional experiments by building an image-history joint model and train it with different lengths of history features. From these experiments, we see a tendency that a model with the less amount of history features gets a higher NDCG score (with lower values for other metrics), whereas a model with more history information has the opposite behavior. Previously, Massiceti et al. (2018) argued that the Visdial dataset has an answer bias such that a simple model without vision or dialogue history could achieve reasonable results. However, our motivation is different from theirs. The purpose of our paper is to find characteristics of existing multimodal models on the dataset (which are biased towards the language information in the dialogue history), analyze behaviors of these models on different metrics, as well as employ this analysis to build better, less biased models that achieve more balanced scores.

Since NDCG measures more of a model's generalization ability (because it allows multiple similar answers), while the other metrics measure a model's preciseness, we interpret the results of these above experiments to mean that a model with more history information tends to predict correct answers by memorizing keywords or patterns in the history while a model with less history information (i.e., the image-only model) is better at generalization by avoiding relying on such exact-match extracted information. We think that an ideal model should have more balanced behavior and scores over all the metrics rather than having higher scores only for a certain metric and such a model could be considered as the one with both preciseness and generalization. To this end, we propose two models, an image-only and an image-history-joint model. We analyze that the answers these two models produce are complementarily good, and better at different metrics. Hence, we integrate these two models (image-only and image-history-joint) in two ways: consensus-dropout-fusion and ensemble. Our final consensus-dropout-fusion ensemble model scores strongly on both NDCG and recall metrics for the VisDial v1.0 test dataset, and these scores outperform the state-of-the-art of the Visual Dialog challenge 2018 on most metrics. Also, our model shows competitive balanced results in the Visual Dialog challenge 2019 (test-std leaderboard rank 3 based on NDCG metric and high balance across metrics).

2 Related Work

Visual Question Answering (VQA) Visual question answering is a task in which a machine is asked to answer a question about an image. The recent success of deep neural networks and massive data collection (Antol et al. 2015) has made the field more active. One of the most challenging parts of the task is to ground the meaning of text on visual evidence. Co-attention (Lu et al. 2016) is proposed to integrate information from different modalities (i.e., image and language) and more advanced approaches have shown good performance (Yu et al. 2017a; Nam, Ha, and Kim 2017; Nguyen and Okatani 2018). A bilinear approach has also been proposed to replace simple addition or concatenation approaches for fusing the two modalities (Gao et al. 2016; Fukui et al. 2016; Kim et al. 2017; Ben-Younes et al. 2017). In our work, we employ multi-modal factorized bilinear pooling (MFB) (Yu et al. 2017b) to fuse a question and image-history features.

Visual Dialog The Visual Dialog task (Das et al. 2017) can be seen as an extended version of the VQA task, with multiple rounds of sequential question-answer pairs as dialog history, including an image caption, which should be referred to before answering a given question. This conversation history can help a model better predict correct answers by giving direct or indirect clues for the answers, or proper context for co-reference resolution. However, having conversation history also means that a model should extract relevant information from the history and introduces another challenge to the task. Many approaches have been proposed to handle this challenge. Niu et al. (2018) tries to extract the clues from history recursively while Wu et al. (2018) and Guo, Xu, and Tao (2019) employ co-attention to fuse visual, history, and question features. In our work, we employ Seo et al. (2017)'s approach to fuse visual and history features before they are attended by a question. Our joint model with fused features has much information from history and we find that it is in complementary relation with our image-only model. Thus, we combine the two models to take the most appropriate information from each model to answer questions.

3 Models

In the Visual Dialog task (Das et al. 2017), two agents interact via natural language with respect to an image. The asker keeps asking about the image given an image caption without seeing the image. The other agent (i.e., answerer) keeps answering the questions by viewing the image. They conduct multiple rounds of conversation accumulating question-answer pairs which are called 'history' (Figure 1). The full history HISTORY consists of question-answer pairs as well as an image caption which describes the given image, such that at a current time point t, the previous history is $\text{HISTORY}_t = \{C, (Q_1, A_1), (Q_2, A_2), ..., (Q_{t-1}, A_{t-1})\}$, where C is the image caption and Q_{t-1} and A_{t-1} are the

history. It turns out that only 1% of the questions (2 from 200 questions) can be answered. This motivates us to focus on an image-history joint model (instead of a history-only model) and merge this with an image-only model.

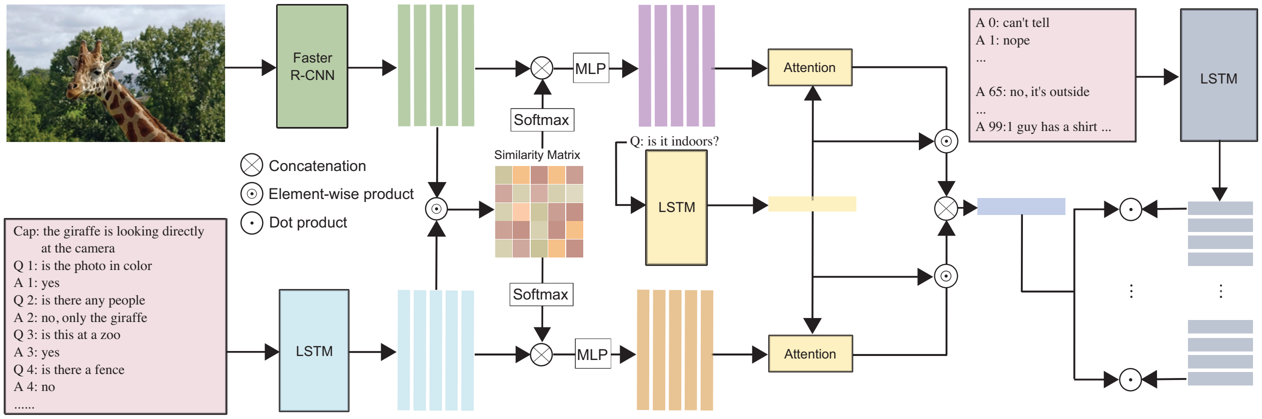


Figure 2: The architecture of the image-history joint model. The visual features are obtained from Faster R-CNN and the history features are encoded via LSTM. They are fused together via the similarity matrix calculated using cross-attention. The fused features are combined with a question feature and dot products are calculated between the combined feature and candidate answers to rank the answers.

question and answer at round $t-1$, respectively. Then, given a new current time-stamp question Q_t, the history HISTORY$_t$, and the image, the model has to rank 100 candidate answers from the answerer's perspective.

3.1 Features

Visual Features: For visual features, we use object features which are extracted from an image by using Faster R-CNN (Ren et al. 2015). The visual feature, $V_{rcnn} \in \mathbb{R}^{k \times d_v}$, is a matrix whose rows correspond to objects, where k is the number of objects (k=36 in our experiment), d_v is dimension size of visual feature (d_v = 2048 for ResNet backbone).
Question Features: The word sequence of a question at round r, $W_{q_r} = \{w_{q_r1}, w_{q_r2}, ..., w_{q_rT_{q_r}}\}$ is encoded via an LSTM-RNN (Hochreiter and Schmidhuber 1997),

$$h_t^{q_r} = \text{LSTM}_q(w_{q_rt}, h_{t-1}^{q_r}) \tag{1}$$

and, we take the last hidden state as a question representation: $q_r = h_{T_{q_r}}^{q_r}$, where T_{q_r} is the length of the question at round r.
History Features: History H_r is a history feature at round r encoded from concatenation of a question and a ground truth answer, such that

$$\begin{aligned} W_{h_r} &= \{w_{q_{r-1}1}, .., w_{q_{r-1}T_{q_{r-1}}}, w_{a_{r-1}1}, .., w_{a_{r-1}T_{a_{r-1}}}\} \\ &= \{w_{h_r1}, w_{h_r2}, ..., w_{h_rT_{h_r}}\} \end{aligned} \tag{2}$$

where $T_{a_{r-1}}$ is the length of the answer of round $r-1$, and the length of history at round r is $T_{h_r} = T_{q_{r-1}} + T_{a_{r-1}}$. The history H_r is also encoded with an LSTM,

$$h_t^{h_r} = \text{LSTM}_h(w_{h_rt}, h_{t-1}^{h_r}) \tag{3}$$

We also take the last hidden state as history representation at round r: $H_r = h_{T_{h_r}}^{h_r}$. Note that the first history feature H_1 comes from the image caption C.

3.2 Image-Only Model

We first build a model which only uses visual features to answer questions. We employ a state-of-the-art 'bottom-up and top-down' approach from Anderson et al. (2018), in which we apply the attention mechanism over detected object features. We also adopt the multi-modal factorized bilinear pooling (MFB) method (Yu et al. 2017b) to calculate attention weights over the visual features with respect to a question feature. From projected visual features and a question feature, we obtain $z \in \mathbb{R}^{k \times d_m}$ by applying MFB:

$$V = \text{Linear}_{d_v \times d}(V_{rcnn}) \tag{4}$$

where $\text{Linear}_{d_v \times d}$ is a linear projection which projects points from a d_v-dimension space to a d-dimension space.

$$z_r = \text{MFB}(V, q) = \sum_{i=1}^{m} ((M_i V^\top) \odot (N_i q_r \cdot \mathbb{1}_k^\top))^\top \tag{5}$$

where $M, N \in \mathbb{R}^{d_m \times d \times m}$ are trainable parameters, d is the dimension of projected visual features and a question feature, d_m is dimension of the fused feature, and m is the number of factors. $\mathbb{1}_k \in \mathbb{R}^k$ is a vector whose elements are all one. Following Yu et al. (2017b), we also apply the power normalization and ℓ_2 normalization to obtain $\hat{z}_r$. After applying linear projection, the softmax operation is applied to get a weight vector α: $\alpha_r = \text{softmax}(L\hat{z}_r^\top)$. We then get a visual representation vector, v_r by weighted summing the projected visual features: $v_r = \sum_{i=1}^{k} \alpha_{ri} V_i$, where $L \in \mathbb{R}^{1 \times d_m}$ is trainable parameter, and V_i is the i-th row vector of visual feature matrix V. The visual representation vector and a question feature vector are combined with element-wise product after linear projection. After one more linear projection, we get the final feature, $f_{v_r}^{q_r}$ which is further used to rank answers.

$$f_{v_r}^{q_r} = \text{fc}_f(\text{fc}_v(v_r) \odot \text{fc}_q(q_r)) \tag{6}$$

where fc$_*$ is an fully-connected layer.

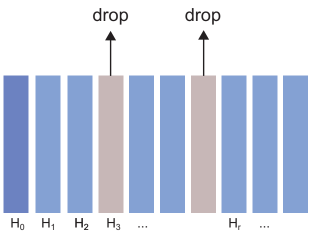


Figure 3: Round Dropout: history features are dropped randomly. H_0 is the image caption, H_r is the history feature at round r. Dropout is not applied to the image caption feature.

Answer Selection For each round, there are 100 candidate answers. The l-th answer at round r,

$$A_{rl} = \{w_{rl1}, w_{rl2}, ...w_{rlT_{a_{rl}}}\} \tag{7}$$

is encoded in the same way as question and history.

$$h_t^{a_{rl}} = \text{LSTM}_a(w_{rlt}, h_{t-1}^{a_{rl}}) \tag{8}$$

$$a_{rl} = h_{T_{a_{rl}}}^{a_{rl}} \tag{9}$$

where $T_{a_{rl}}$ is the length of the l-th candidate answer. Scores for each candidate answer are calculated by dot product between fused feature $f_{v_r}^{q_r}$ and each candidate answer representation, a_{rl}: $s_{rl} = f_{v_r}^{q_r} \cdot a_{rl}$.

3.3 Image-History Joint Model

We calculate the similarity matrix, $S_r \in \mathbb{R}^{k \times r}$ between visual and history features following Seo et al. (2017).

$$(S_r)_{ij} = w_s^\top [V_i; H_j; V_i \odot H_j] \tag{10}$$

where $w_s \in \mathbb{R}^{3d}$ is trainable parameter and H_j is the j-th row vector of the history feature $H_{1:r}$. From the similarity matrix, the new fused history representation is:

$$V_r^h = \text{softmax}(S_r^\top)V \tag{11}$$

$$H_{1:r}^f = [H_{1:r}; V_r^h; H_{1:r} \odot V_r^h] \tag{12}$$

Similarly, the new fused visual representation is:

$$H_r^v = \text{softmax}(S_r)H_{1:r} \tag{13}$$

$$V_r^f = [V; H_r^v; V \odot H_r^v] \tag{14}$$

These fused features are then fed to the MFB module and attended over w.r.t. a question feature, respectively, following the same process as a visual feature in the image-only model. The weighted-summed features are combined with a question feature through element-wise product and concatenated together to produce the integrated representation:

$$f_{v_r}^{q_r} = \text{fc}_v(v_r^f) \odot \text{fc}_q(q_r) \tag{15}$$

$$f_{h_r}^{q_r} = \text{fc}_h(h_r^f) \odot \text{fc}_q(q_r) \tag{16}$$

$$f_{v_r h_r}^{q_r} = \text{fc}_f([f_{v_r}^{q_r}; f_{h_r}^{q_r}]) \tag{17}$$

where v_r^f and h_r^f are weighted-sum of fused features with respect to a question feature. Figure 2 depicts the whole process of the joint model in this section.

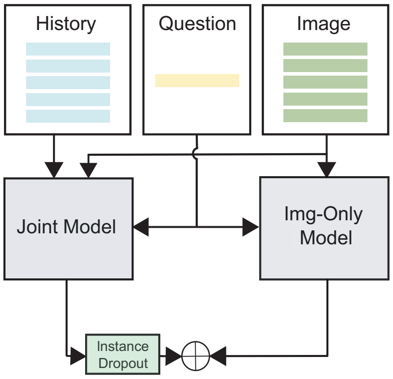


Figure 4: Consensus Dropout Fusion. Logits from both image-only model and joint model are added to produce combined one. Instance dropout is applied to the logit from joint model to prevent strong coupling. The two models share many portions of parameters and are trained together.

Round Dropout To prevent the model from over-relying on history information, we propose a novel dropout approach in which some rounds of history features are dropped out (Figure 3). To be specific, we randomly pick up to 3 rounds of history from entire history except image caption feature and throw them away.

$$N_D^r = \begin{cases} \max(0, N_h^r - 2) & \text{if } N_h^r \leq 5 \\ 3 & \text{otherwise} \end{cases} \tag{18}$$

where N_h^r is number of history features at round r and N_D^r is the number of history features to drop at round r.

3.4 Combining Image-Only & Image-History Joint Models

Since each of our models has different abilities, we exploit their complementary abilities together by combining them in two ways. The first is our novel consensus dropout fusion which integrates the two models in training time. The other way is to build an ensemble model from the two models at test time.

Consensus Dropout Fusion In order to integrate the image-only model and the image-history joint model into one model, we propose a novel integration method called consensus dropout fusion. Our consensus dropout fusion is the combination of a consensus method and an instance dropout method (Figure 4).

Consensus We employ a consensus method in which logits from each model are added to produce the final logit following Wang et al. (2016)'s approach.

$$L_{IJ} = L_I + L_J \tag{19}$$

where L_I and L_J are the logit from image-only model and image-hitory joint model, respectively, and L_{IJ} is the new logit obtained by adding the two logits.

Instance Dropout To allow the image-only model to have a stronger effect producing more balanced results over all metrics, we apply dropout to instances of the logit of the joint model. To be specific, when we add two logits, we multiply L_J by I_{drop},

$$L_J^{drop} = I_{drop} \odot L_J \tag{20}$$

$$I_{drop} = (\mathbb{1}_{(N \times R)} \odot \xi) \cdot \mathbb{1}_d^\top \tag{21}$$

$$\xi_i \sim \frac{1}{1-p}\text{Bernoulli}(1-p) \tag{22}$$

where $\mathbb{1}_{(N \times R)} \in \mathbb{R}^{(N \times R)}$ and $\mathbb{1}_d \in \mathbb{R}^d$ are all-ones vectors of $(N \times R)$ and d dimension, respectively. N is the training batch size and R is the length of rounds of the conversation history. The dropout mask, ξ, is calculated following Srivastava et al. (2014)'s work.

Ensemble We also integrate our 2 models via an ensemble. We train each model separately and combine them at test time. To be specific, we take logits from the pre-trained models and select the answer with the highest sum of logits.

4 Experimental Setup

4.1 Dataset

We use the VisDial v1.0 (Das et al. 2017) dataset to train our models, where one example has an image with its caption, 9 question-answer pairs, and follow-up questions and candidate answers for each round. At round r, the caption and the previous question-answer pairs become conversational context. The whole dataset is split into 123,287/2,000/8,000 images for train/validation/test, respectively. Unlike the images in the train and validation sets, the images in the test set have only one follow-up question and candidate answers and their corresponding conversational context.

4.2 Metrics

For evaluation, the Visual Dialog task employs four metrics. NDCG is the primary metric of the Visual Dialog Challenge which considers multiple similar answers as correct ones. The other three are MRR, recall@k, and mean rank where they only consider the rank of a single answer. Our experiments show the scores of NDCG and non-NDCG metrics from our image-only and joint models have a trade-off relationship due to their different ability (as shown in Sec.5.2) in completing Visual Dialog tasks: the image-only model has a high NDCG and low non-NDCG values while the joint model has a low NDCG and high non-NDCG values.

4.3 Training Details

In our models, the size of word vectors is 300, the dimension of visual feature is 2048, and hidden size of LSTM units which are used for encoders of questions, context history, and candidate answers is 512. We employ Adam (Kingma and Ba 2015) as the optimizer. We set the initial learning rate to 0.001 and decrease it by 0.0001 per epoch until 8th epoch and decay by 0.5 from 9th epoch on. For round dropout, we set the maximum number of history features to be dropped to 3 and we tune the p value to 0.25 for our instance dropout in the consensus dropout fusion module. Cross-entropy is used to calculate the loss.

	Only Img.	Need Hist.
% of Questions	81.0 %	19.0 %

Table 1: Human evaluation on questions of VisDial v1.0 val set. Percentage of questions which can be answered only from image or need help from conversation history is calculated by the manual investigation.

5 Analysis and Results

In this section, we first discuss how many questions are answered only from image and how many of them need image and history jointly to be answered by conducting a manual investigation. We find that a large portion of questions in the VisDial dataset can be answered by only using images. Next, to verify the observation from the manual investigation, we perform a follow-up experiment and find a trade-off relation between the amount of history features and the metric scoring trend of models. We then analyze the answers from two models (image-only and image-history joint model) and show they are in complementary relation. Lastly, we show each model can make up for the other by being combined in consensus dropout fusion or in an ensemble model.

5.1 Human Evaluation: Is Image Alone Enough?

We conduct a human evaluation on image, history, and question. To be specific, we randomly select 100 images (which leads to 1000 questions) from the validation set for the evaluation and count the number of questions which can be answered only with images and the number of questions which need conversation context to be answered (ground-truth answers are provided to check if the answers can be inferred given corresponding questions and images instead of providing all the 100 candidate answers). Two annotators conduct the experiment independently and questions on which both annotators mark as being able to be answered only with images are classified as only-image questions otherwise as need-history questions. The inter-annotation agreement (kappa) is 0.74.[2] As shown in Table 1, around 80%[3] of the questions can be answered only from images. Conversely, this also implies that a model needs conversation context to better perform the task. However, as discussed in Sec.1, using only history is not enough either (only 1% of the questions can be answered) and thus history should be used jointly with images. Note that we consider a question with a pronoun as answerable only with an image if the pronoun can be inferred (co-reference) from the corresponding image (e.g., a question mentions 'he' and the image has only one person who is a boy).

[2] Kappa of 0.74 is considered 'substantial' agreement: https://en.wikipedia.org/wiki/Cohens_kappa

[3] We compute statistical significance via bootstrap test (Efron and Tibshirani 1994) and find that in 99,975 of 100K trials (i.e., p < 0.0005), the percentage of only-image questions is over 75%.

Models	NDCG	MRR	R@1	R@5	R@10	Mean
FULL	57.81	**64.47**	**50.87**	**81.38**	90.03	**4.10**
H-5	58.24	64.29	50.61	81.35	**90.22**	4.10
H-1	59.29	62.86	49.07	79.76	89.08	4.35
Img-only	**61.04**	61.25	47.18	78.43	88.17	4.61

Table 2: Performance of models with the different amount of history on validation dataset of VisDial v1.0 (Round dropout is not applied to the joint model in these experiments. FULL: full image-history joint model, H-k: image-history joint model with k history, Img-only: image-only model. For H-k models we include image caption feature for a fair comparison with the full joint model).

	Img-Only Model	Joint Model	Intersection	Union
R@1	47.18	50.87	41.57	56.48
NDCG	61.04	58.97	55.65	64.36

Table 3: Intersection and Union of the answers from image-only model and joint model which contribute to scoring for R@1 and NDCG metrics.

5.2 Reduced Question-Answer Rounds

We next run our joint model with various lengths of history. To be specific, we make our joint model use only k previous history features to answer a question. As shown in Table 2, there is a trade-off between the values of metrics and the number of history features. As the number of history features the joint model uses is increased, the score of NDCG is decreased while other metrics are increased. On the other hand, as the number of history features the joint model uses is decreased the score of NDCG is increased while other metrics are decreased. If we see the Visual Dialog primary task metric of NDCG as a barometer of the model's ability to generalize and the other metrics can be seen as an indicator of preciseness, this means that decreased size of history gives a model the ability of generalization at the cost of preciseness. From this tendency, the image-only model has the highest NDCG score.

5.3 Complementary Relation

If the image-only model is good at NDCG, can we exploit its ability by combining it with the joint model? To figure out this possibility, we compare each answer from the image-only model and the joint model. To be specific, for R@1, we list up the correct answers from each model and count answers which are in both sets, i.e., the intersection. From the intersection, we obtain the union of the two sets. For NDCG, there is not one single correct answer. So we roughly calculate the intersection by taking minimum values between the two models' scores and averaging them. As we can see in Table 3, the intersections do not take the entire score of either model for both metrics. This could mean image-only and joint models have room to be improved by combining them together.

5.4 Model Combination Results

Considering the complementary relation between image-only model and joint model, combining the two models would be a good approach to take the best from the both. So, we integrate these two models via two methods: consensus dropout fusion and ensemble (see Sec.3.4).

Models	NDCG	MRR	R@1	R@5	R@10	Mean
Img-Only	61.04	61.25	47.18	78.43	88.17	4.61
Joint	58.97	64.57	50.87	81.58	90.30	4.05
CDF	59.93	64.52	50.92	81.31	90.00	4.10
Ensemble	61.20	64.67	51.00	81.60	90.37	4.03

Table 4: Performance of the consensus dropout fusion model and the ensemble model between our image-only model and joint model on the validation dataset of VisDial v1.0 (Img-Only: image-only model, Joint: image-history joint model, CDF: consensus dropout fusion model).

Consensus Dropout Fusion Results As shown in Table 4, consensus dropout fusion improves the score of NDCG by around 1.0 from the score of the joint model while still yielding comparable scores for other metrics. Unlike ensemble way, consensus dropout fusion does not require much increase in the number of model parameters.

Ensemble Model Results As also shown in Table 4, the ensemble model seems to take the best results from each model. Specifically, the NDCG score of the ensemble model is comparable to that of the image-only model and the scores of other metrics are comparable to those of the image-history joint model. From this experiment, we can confirm that the two models are in complementary relation.

5.5 Final Visual Dialog Test Results

For the evaluation on the test-standard dataset of VisDial v1.0, we try 6 image-only model ensemble and 6 consensus dropout fusion model ensemble. As shown in Table 5, our two models show competitive results compared to the state-of-the-art on the Visual Dialog challenge 2018 (DL-61 was the winner of the Visual Dialog challenge 2018). Specifically, our image-only model shows much higher NDCG score (60.16). On the other hand, our consensus dropout fusion model shows more balanced results over all metrics while still outperforming on most evaluation metrics (NDCG, MRR, R@1, and R@5). Compared to results of the Visual Dialog challenge 2019, our models also show strong results. Although ReDAN+ (Gan et al. 2019) and MReaL–BDAI show higher NDCG scores, our consensus dropout fusion model shows more balanced results over metrics while still having a competitive NDCG score compared to DAN (Kang, Lim, and Zhang 2019), with rank 3 based on NDCG metric and high balance rank based on metric average.[4]

Ensemble on More Models We also run an ensemble model from our image-only, joint, and consensus dropout fusion models (6 of each and total 18 models) and evaluate it on the test-standard dataset of the VisDial v1.0. This model's scores (NDCG: 59.90, MRR: 64.05, R@1: 50.28, R@5: 80.95, R@10: 90.60, Mean: 4.00) are in between our

[4] We are model name 'square' on https://evalai.cloudcv.org/web/challenges/challenge-page/161/leaderboard/483

	Models	NDCG	MRR	R@1	R@5	R@10	Mean
	LF (Das et al. 2017)	45.31	55.42	40.95	72.45	82.83	5.95
	HRE (Das et al. 2017)	45.46	54.16	39.93	70.45	81.50	6.41
	MN (Das et al. 2017)	47.50	55.49	40.98	72.30	83.30	5.92
	MN-att (Das et al. 2017)	49.58	56.90	42.43	74.00	84.35	5.59
	LF-att (Das et al. 2017)	49.76	57.07	42.08	74.83	85.05	5.41
	CorefNMN (Kottur et al. 2018)	54.7	61.5	47.55	78.10	88.80	4.40
Visual Dialog challenge 2018	RvA (Niu et al. 2018)	55.59	63.03	49.03	80.40	89.83	4.18
	USTC-YTH (Yang, Zha, and Zhang 2019)	57.17	64.22	50.88	80.63	89.45	4.20
	DL-61 (single) (Guo, Xu, and Tao 2019)	57.32	62.20	47.90	80.43	89.95	4.17
	DL-61 (ensemble) (Guo, Xu, and Tao 2019)	57.88	63.42	49.30	80.77	90.68	3.97
Visual Dialog challenge 2019	DAN (single) (Kang, Lim, and Zhang 2019)	57.59	63.20	49.63	79.75	89.35	4.30
	DAN (ensemble) (Kang, Lim, and Zhang 2019)	59.36	64.92	51.28	81.60	90.88	3.92
	ReDAN+ (ensemble) (Gan et al. 2019)	64.47	53.73	42.45	64.68	75.68	6.63
	MReaL–BDAI (not published)	74.02	52.62	40.03	65.85	79.15	6.76
	Our Image-Only (ensemble)	60.16	61.26	47.15	78.73	88.48	4.46
	Our Consensus Dropout Fusion (ensemble)	59.49	64.40	50.90	81.18	90.40	3.99

Table 5: Performance comparison between our models and other models on the test-standard dataset of VisDial v1.0. We run two ensemble models each from 6 image-only models and 6 consensus dropout fusion models.

Models	NDCG	MRR	R@1	R@5	R@10	Mean
CA	57.81	64.47	50.87	81.38	90.03	4.10
CA + RD	58.97	64.57	50.87	81.58	90.30	4.05

Table 6: The effect of round dropout: applying round dropout improves model's performance on NDCG by around 1.2 while also improving other metrics. (CA: cross-attention model (base model), RD: round dropout).

Models	NDCG	MRR	R@1	R@5	R@10	Mean
CDF (p=0.00)	59.40	64.61	51.01	81.73	90.30	4.06
CDF (p=0.15)	59.49	64.64	50.94	81.63	90.07	4.07
CDF (p=0.25)	59.93	64.52	50.92	81.31	90.00	4.10
CDF (p=0.35)	60.11	64.21	50.56	81.20	89.84	4.15

Table 7: Consensus dropout fusion and different dropout rates. With different dropout rates, consensus dropout fusion model yields different scores of all metrics. (CDF: consensus dropout fusion model).

Models	NDCG	MRR	R@1	R@5	R@10	Mean
Img+Img	61.97	62.24	48.20	79.49	88.83	4.41
Joint+Joint	59.84	65.60	52.06	82.46	90.87	3.88
Img+Joint	61.50	65.04	51.38	81.93	90.45	3.96

Table 8: Performance of ensemble models with different combinations. Img+Img model (3 Img models) has highest value of NDCG while Joint+Joint (3 Joint models) model highest values for other metrics. Img+Joint model (3 Img + 3 Joint models) has more balanced results (Img: image-only model, Joint: image-history joint model).

image-only ensemble model and our consensus dropout fusion ensemble model, i.e., this ensemble model has a higher NDCG than the consensus dropout fusion ensemble model and higher non-NDCG scores than the image-only ensemble model. This result shows that our image-only, joint, and consensus dropout fusion models make up for each other by being combined in an ensemble model as we expected.

6 Ablation Study

Round Dropout: As shown in Table 6, our round dropout (see Sec.3.3) improves the NDCG score by 1.2. A possible interpretation is that round dropout could help the model avoid from over-fitting to some patterns in the history features by intentionally dropping some of the features in the training session.

Consensus Dropout Fusion and Dropout Rate: We run our consensus dropout fusion model (see Sec.3.4) with different instance dropout rates to figure out how the dropout rates affect the performance of the model. As shown in Table.7, as the dropout rate increases the NDCG score is also increased while scores of non-NDCG metrics are decreased. By changing the dropout rate, we can modulate the influence of each model (image-only and joint models) over the combined model. We choose a value of 0.25 for the dropout rate since it yields more balanced scores over all metrics.

Ensemble Combination: We try different combinations from image-only and joint models to build ensemble models. The total number of models amounts to 3, i.e., image-only + image-only (I+I), joint + joint (J+J), and image-only + joint (I+J) ensemble models. As shown in Table 8, scores of the I+J ensemble model are comparable to same-kind ensemble models (I+I and J+J). To be specific, for the NDCG metric, the I+J model outperforms the J+J model, while, for other metrics (MRR, recall@k, and mean rank), the I+J model outperforms the I+I model. This might imply that the balanced scores (i.e., high scores over all metrics) of the I+J model are from the complementary relation between image-only and image-history joint model.

Output Examples: Due to space constraints and no supplementary allowed in AAAI rules, we provide detailed examples in the arxiv supplementary version, showing the coreference and memorization phenomena of the joint image-history model as well as image-only model's example outputs on image-only questions.

7 Conclusion

We first showed that current multimodal models on the Visual Dialog task over-rely on the dialogue history, and relatedly, image-only and image-history joint models achieve complementary performance gains. Hence, to balance the best abilities from each model, we proposed two ways of combining them: consensus dropout fusion and ensemble. Our consensus dropout fusion and ensemble model achieve strong ranks on multiple leaderboards. Specifically, the models show higher scores than the state-of-the-art results of the Visual Dialog challenge 2018 and more balanced scores than highest ranked results of the Visual Dialog challenge 2019. Given the characteristics of the dataset and current model behaviors, a potential future direction is to combine the power of the two models dynamically, e.g., learn to select a proper model based on the question type.

Acknowledgments

We thank the reviewers for their helpful comments. This work was supported by NSF Award #1840131, ARO-YIP Award #W911NF-18-1-0336, and faculty awards from Google, Facebook, Bloomberg, and Salesforce. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the funding agency.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV*.

Ben-Younes, H.; Cadene, R.; Cord, M.; and Thome, N. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*, 2612–2620.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual Dialog. In *CVPR*.

Efron, B., and Tibshirani, R. J. 1994. *An introduction to the bootstrap*. CRC press.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Gan, Z.; Cheng, Y.; Kholy, A. E.; Li, L.; Liu, J.; and Gao, J. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. *arXiv preprint arXiv:1902.00579*.

Gao, Y.; Beijbom, O.; Zhang, N.; and Darrell, T. 2016. Compact bilinear pooling. In *CVPR*, 317–326.

Guo, D.; Xu, C.; and Tao, D. 2019. Image-question-answer synergistic network for visual dialog. *arXiv preprint arXiv:1902.09774*.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Kang, G.-C.; Lim, J.; and Zhang, B.-T. 2019. Dual attention networks for visual reference resolution in visual dialog. *arXiv preprint arXiv:1902.09368*.

Kim, J.; On, K. W.; Lim, W.; Kim, J.; Ha, J.; and Zhang, B. 2017. Hadamard product for low-rank bilinear pooling. In *ICLR*.

Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Kottur, S.; Moura, J. M.; Parikh, D.; Batra, D.; and Rohrbach, M. 2018. Visual coreference resolution in visual dialog using neural module networks. In *ECCV*, 153–169.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, 289–297.

Massiceti, D.; Dokania, P. K.; Siddharth, N.; and Torr, P. H. 2018. Visual dialogue without vision or dialogue. *arXiv preprint arXiv:1812.06417*.

Nam, H.; Ha, J.-W.; and Kim, J. 2017. Dual attention networks for multimodal reasoning and matching. In *CVPR*.

Nguyen, D.-K., and Okatani, T. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *CVPR*.

Niu, Y.; Zhang, H.; Zhang, M.; Zhang, J.; Lu, Z.; and Wen, J.-R. 2018. Recursive visual attention in visual dialog. *arXiv preprint arXiv:1812.02664*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.

Seo, M. J.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2017. Bidirectional attention flow for machine comprehension. In *ICLR*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR* 15(1):1929–1958.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*.

Wu, Q.; Wang, P.; Shen, C.; Reid, I.; and van den Hengel, A. 2018. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *CVPR*.

Yang, T.; Zha, Z.-J.; and Zhang, H. 2019. Making history matter: Gold-critic sequence training for visual dialog. *arXiv preprint arXiv:1902.09326*.

Yu, D.; Fu, J.; Mei, T.; and Rui, Y. 2017a. Multi-level attention networks for visual question answering. In *CVPR*.

Yu, Z.; Yu, J.; Fan, J.; and Tao, D. 2017b. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*.