# History for Visual Dialog: Do we really need it?

**Shubham Agarwal[1]\*, Trung Bui[2], Joon-Young Lee[2], Ioannis Konstas[1] and Verena Rieser[1]**

[1]The Interaction Lab, Heriot-Watt University, Edinburgh, UK
[2]Adobe Research, San Jose, CA, US
{sa201, i.konstas, v.t.rieser}@hw.ac.uk
{bui, jolee}@adobe.com

## Abstract

Visual Dialog involves "understanding" the dialog history (what has been discussed previously) and the current question (what is asked), in addition to grounding information in the image, to generate the correct response. In this paper, we show that co-attention models which explicitly encode dialog history outperform models that don't, achieving state-of-the-art performance (72 % NDCG on val set). However, we also expose shortcomings of the crowd-sourcing dataset collection procedure by showing that history is indeed only required for a small amount of the data and that the current evaluation metric encourages generic replies. To that end, we propose a challenging subset (VisDialConv) of the VisDial val set and provide a benchmark of 63% NDCG.

## 1 Introduction

Recently, there has been an increased interest in visual dialog, i.e. dialog-based interaction grounded in visual information (Chattopadhyay et al., 2017; De Vries et al., 2017; Seo et al., 2017; Guo et al., 2018; Shekhar et al., 2018; Kottur et al., 2019; Haber et al., 2019). One of the most popular test beds is the Visual Dialog Challenge (*VisDial*) (Das et al., 2017), which involves an agent answering questions related to an image, by selecting the answer from a list of possible candidate options. According to the authors, nearly all interactions (98%) contain dialog phenomena, such as co-reference, that can only be resolved using dialog history, which makes this a distinct task from previous Visual Question Answering (VQA) challenges, e.g. (Antol et al., 2015). For example, in order to answer the question "About how many?" in Figure 1, we have to infer from what was previously said, that the conversation is about the skiers.

---

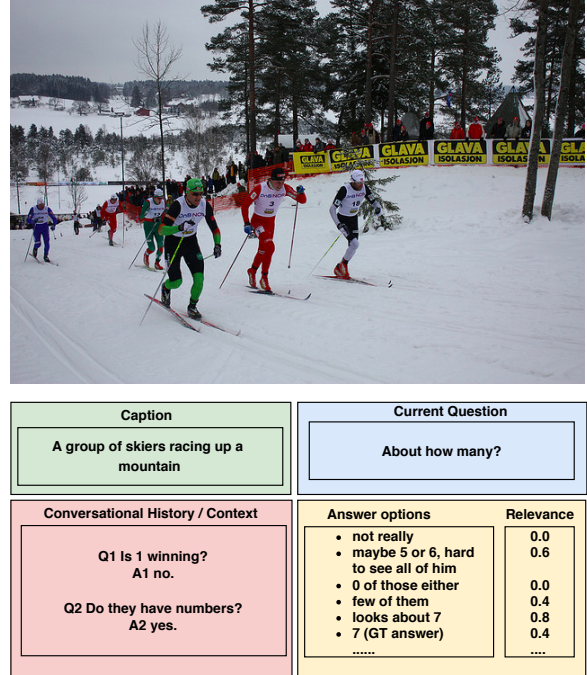\*This work was carried out during the internship at Adobe Research.



Figure 1: Visual Dialog task according to (Das et al., 2017) as a ranking problem, where for the current question (blue), the agent ranks list of 100 candidate answers (yellow). Relevance weights for each candidate were collected via crowd-sourcing. Previous dialog history (red) together with the caption (green) forms the contextual information for the current turn.

In the original paper, Das et al. (2017) find that models which structurally encode dialog history, such as Memory Networks (Bordes et al., 2016) or Hierarchical Recurrent Encoders (Serban et al., 2017) improve performance. However, "naive" history modelling (in this case an encoder with late fusion/concatenation of current question, image and history encodings) might actually hurt performance. Massiceti et al. (2018) take this even further, claiming that VisDial can be modeled without taking history or even visual information into account. Das et al. (2019) rebutted by showing that both features are still needed to achieve state-of-the-

art (SOTA) results and an appropriate evaluation procedure has to be used.

In this paper, we show that competitive results on VisDial can indeed be achieved by replicating the top performing model for VQA (Yu et al., 2019b) – and effectively treating visual dialog as multiple rounds of question-answering, without taking history into account. However, we also show that these results can be significantly improved by encoding dialog history, as well as by fine-tuning on a more meaningful retrieval metric. Finally, we show that more sophisticated dialog encodings outperform naive fusion on a subset of the data which contains "true" dialog phenomena according to crowd-workers. In contrast to previous work on the VisDial dataset, e.g. (Kottur et al., 2018; Agarwal and Goyal, 2018; Gan et al., 2019; Guo et al., 2019; Kang et al., 2019), we are the first to conduct a principled study of dialog history encodings. Our contributions can thus be summarized as follows:

- We present SOTA results on the VisDial dataset using transformer-based Modular Co-Attention (MCA) networks. We further show that models encoding dialog history outperform VQA models on this dataset.
- We show that curriculum fine-tuning (Bengio et al., 2009) on annotations of semantically equivalent answers further improves results.
- We experiment with different dialog history encodings and show that early fusion, i.e. dense interaction with visual information (either via *grounding* or *guided attention*) works better for cases where conversational historical context is required.
- We release a crowd-sourced subset containing verified dialog phenomena and provide benchmark results for future research.

## 2 Visual Dialog Models

In this section, we extend Modular Co-Attention Networks, which won the VQA challenge 2019 (Yu et al., 2019b) and adapt it to visual dialog. Different from previous co-attention networks (Kim et al., 2018; Nguyen and Okatani, 2018), MCA networks use guided attention to model dense relations between the question and image regions for better visual grounding. In the following, we explore MCA networks with different input encodings following a '[model]-[input]' convention to refer to our MCA model variants; see Figure 3 for an overview. Whenever unspecified, images

are represented as a bag of bottom-up features, i.e. object level representations (see Section 3).

### 2.1 Modular Co-Attention networks

The MCA module with multi-modal fusion as depicted in Figure 2, is common to all our architectures. Inspired by the transformers (Vaswani et al., 2017), the MCA network (Yu et al., 2019b) is a modular composition of two basic attention units: self-attention and guided attention. These are arranged in an encoder-decoder composition in the MCA module (Figure 2), which performed best for VQA (Yu et al., 2019b).

#### 2.1.1 Self-Attention and Guided-Attention

The Self-Attention (SA) unit in transformers (Vaswani et al., 2017) is composed of a multi-head attention layer followed by a feed-forward layer. When applied to vision, the SA unit can be viewed as selecting the most relevant object-level image features for the downstream task. Specifically, the scaled dot product attention takes as input key, query and value (usually same modality's embedded representations) and outputs a self-attended vector (Eq.1). Multi-head attention provides multiple representation spaces to capture different linguistic/grounding phenomena, which are otherwise lost by averaging using a single head.

$$Att(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_K}})V$$
$$MHAtt(Q, K, V) = Concat(head_1, \ldots head_n)W^O$$
$$head_i = Att(QW_i^Q, KW_k^K, VW_i^V)$$
(1)

The Guided-Attention (GA) unit conditions the attention on different sequences. The key and value come from one modality, while the query comes from a different modality similar to the decoder architecture in Transformers (Vaswani et al., 2017). Similar to Eq. 1, the GA unit outputs features $f_i = Att(X, Y, Y)$ where $X \in \mathcal{R}^{m \times d_x}$ comes from one modality and $Y \in \mathcal{R}^{n \times d_y}$ from the other. Residual connection (He et al., 2016) and layer normalization (Ba et al., 2016) are applied to the output of both the attention and feed-forward layers similar to (Vaswani et al., 2017; Yu et al., 2019b) in both the SA and GA units.

#### 2.1.2 Modular Co-Attention Module

The following description of the MCA module is based on the question and the image, but can be

Figure 2: Modular Co-Attention (MCA) module with MCA-I (Section 2.1) as an example.

extended analogously to model the interaction between the question and history. First, the input (i.e. the question) is passed through multiple multi-head self-attention layers $L$, in order to get self-aware representations before acting as conditional signal to different modalities (visual or contextual history) similar to the auto-encoding procedure of Transformers. Then the final representation $X^L$ is used as the input for GA units to model cross-modal dependencies and learn the final conditioned representation $Y^L$.

### 2.1.3 Multi-modal fusion

The learned representations $X^L \in \mathcal{R}^{m \times d}$ and $Y^L \in \mathcal{R}^{n \times d}$ contain the contextualized and conditioned representations over the word and image regions, respectively. We apply attention reduction (Yu et al., 2019b) with a multi-layer perceptron (MLP) for $X^L$ (analogously for $Y^L$). We obtain the final multi-modal fused representation $z$:

$$\alpha^x = softmax(MLP^x(X^L))$$
$$\tilde{x} = \sum_m^{i=1} \alpha_i^x x_i^L \qquad (2)$$
$$z = LayerNorm(W_x^T \tilde{x} + W_y^T \tilde{y})$$

where $\alpha^x = [\alpha_1^x \ldots \alpha_m^x] \in \mathcal{R}^m$ are learned attention weights (same process for $\alpha^y$ and $\tilde{y}$) and $W_x \in \mathcal{R}^{d \times d_z}$, $W_y \in \mathcal{R}^{d \times d_z}$ are linear projection matrices (dimensions are the same for simplicity).

We call this model **MCA with Image component only; (MCA-I)**, since it only encodes the question and image features and therefore treats each question in Visual Dialog as an independent instance of VQA, without conditioning on the historical context of the interaction.

### 2.2 Variants with Dialog History

In the following, we extend the above framework to model dialog history. We experiment with late/shallow fusion of history and image (MCA-I-H), as well as modelling dense interaction between conversational history and the image representation (i.e. MCA-I-VGH, MCA-I-HGuidedQ).

**History guided Question (MCA-I-HGuidedQ):** The network in Figure 3a is designed to model co-reference resolution, which can be considered as the primary task in VisDial (Kottur et al., 2018). We first enrich the question embedding by conditioning on historical context using guided attention in the MCA module. We then use this enriched (co-reference resolved) question to model the visual interaction as described in Section 2.1.

**Visually grounded history with image representation (MCA-I-VGH):** Instead of considering conversational history and the visual context as two different modalities, we now ground the history with the image first, see Figure 3b. This is similar in spirit to maintaining a pool of visual attention maps (Seo et al., 2017), where we argue that different questions in the conversation attend to different parts of the image. Specifically, we pass the history to attend to object-level image features using the MCA module to get visually grounded contextual history. We then embed the question to pool the relevant grounded history using another MCA module.
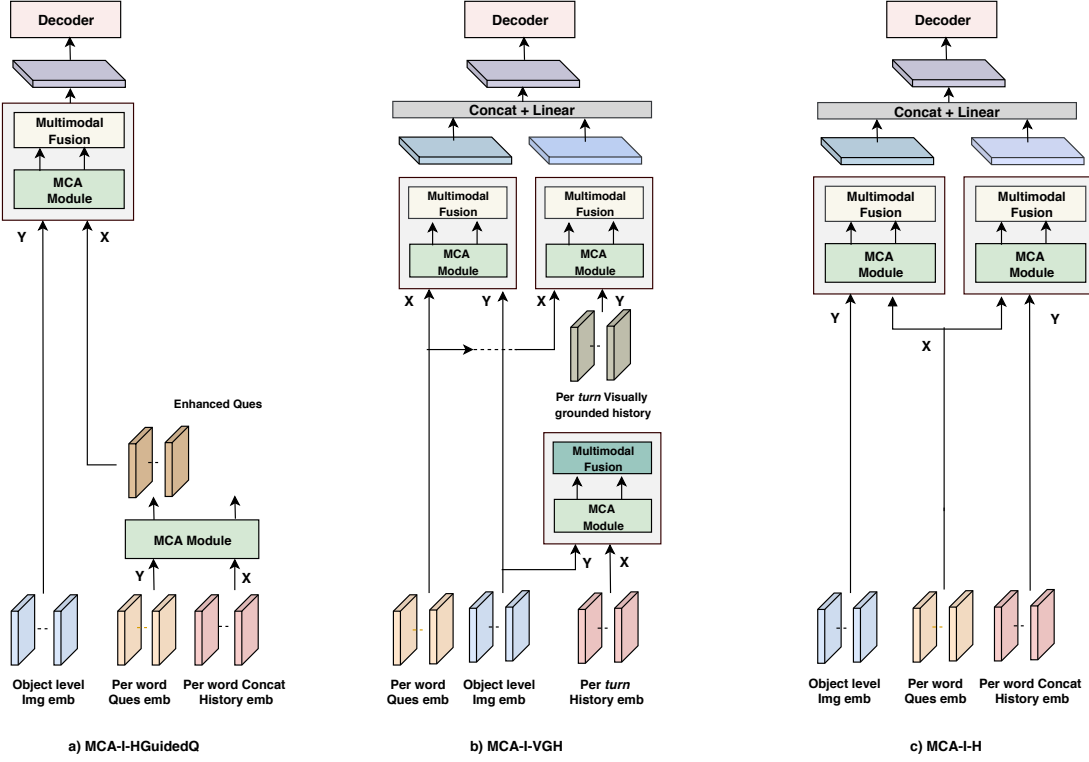
Figure 3: All models incorporating dialog history described in Section 2.2

In parallel, the question embedding is also used to ground the current visual context. At the final step, the respective current image and historical components are fused together and passed through a linear layer before decoding. Note, this model is generic enough to potentially handle multiple images in a conversation and thus could be extended for tasks e.g. conversational image editing, which is one of the target applications of visual dialog (Kim et al., 2017; Manuvinakurike et al., 2018a,b; Lin et al., 2018; El-Nouby et al., 2018).

**Two-stream Image and History component (MCA-I-H):** Figure 3c shows the model which maintains two streams of modular co-attention networks – one for the visual modality and the other for conversational history. We follow a similar architecture for the visual component as MCA-I and duplicate the structure for handling conversational history. At the final step, we concatenate both the embeddings and pass them through a linear layer.

## 2.3 Decoder and loss function

For all the models described above, we use a discriminative decoder which computes the similarity between the fused encoding and RNN-encoded answer representations which is passed through a softmax layer to get the probability distribution

over the candidate answers. We train using cross entropy over the ground truth answer:

$$L(\theta) = \frac{1}{N} \sum_{n=1}^{N=100} y_n log P(x_n, \theta) \qquad (3)$$

$N$ denotes the number of candidate answers which is set to 100 for this task, $y_n$ is the (ground truth) label which is 0 or 1 during the training procedure, or a relevance score of the options during fine-tuning (casting it as multi-label classification).

## 3 Implementation

We use PyTorch[1] (Paszke et al., 2017) for our experiments[2]. Following Anderson et al. (2018), we use bottom-up features of 36 proposals from images using a Faster-RCNN (Ren et al., 2015) pre-trained on Visual Genome (Krishna et al., 2017) to get a bag of object-level 2048-d image representations. Input question and candidate options are tokenized to a maximum length of 20 while the conversational history to 200. Token embeddings in text are initialized with 300-d GloVe vectors (Pennington et al., 2014) and shared among all text-based encoders. The RNN encodings are implemented using LSTMs (Hochreiter and Schmidhuber, 1997).

---

[1] https://pytorch.org/
[2] Code available at https://github.com/shubhamagarwal92/visdial_conv

We use the Adam optimizer (Kingma and Ba, 2015) both for training and fine-tuning. More training details can be found in Appendix A.

# 4 Task Description

## 4.1 Dataset

We use VisDial v1.0 for our experiments and evaluation.[3] The dataset contains 123K/2K/8K dialogs for train/val/test set respectively. Each dialog is crowd-sourced on a different image, consisting of 10 rounds of dialog turns, totalling approx. 1.3M turns. Each question has also been paired with a list of 100 automatically generated candidate answers which the model has to rank. To account for the fact that there can be more than one semantically correct answer (e.g. "Nope", "No", "None", "Cannot be seen"), "dense annotations" for 2k/2k turns of train/val of the data have been provided, i.e. a crowd-sourced relevance score between 0 and 1 (1 being totally relevant) for all 100 options.

## 4.2 Evaluation protocol

As the Visual Dialog task has been posed as a ranking problem, standard information retrieval (IR) metrics are used for evaluation, such as Recall@{1,5,10} to measure performance in the top N results (higher better), mean reciprocal rank (MRR) of the Ground-Truth (GT) answer (higher better), and Mean rank of the GT answer (lower better). Normalized Discounted Cumulative Gain (NDCG) is another measure of ranking quality, which is commonly used when there is more than one correct answer (provided with their relevance).

## 4.3 Training details

**Sparse Annotation Phase:** We first train on sparse annotations, i.e. only 1 provided ground-truth answer, which is available for the whole training set. Here the model learns to select only one relevant answer.

**Curriculum Fine-tuning Phase:** Dense annotations, i.e. crowd-sourced relevance weights, are provided for 0.16% of training set, which we use to fine-tune the model to select multiple semantically equivalent answers. This acts like a curriculum learning setup (Elman, 1993; Bengio et al., 2009),

where selecting one answer using sparse annotation is an easier task and fine-tuning more difficult.[4]

## 4.4 Baselines

**MCA-I-HConcQ and MCA-H:** MCA-I-HConcQ is a naive approach of concatenating raw dialog history to the question while keeping the rest of the architecture the same as MCA-I. MCA-H on the other hand considers this task as only conversational (not visual) dialog with MCA module on history instead of image.

**RvA:** We reproduce the results of Niu et al. (2019)'s Recursive Visual Attention model (RvA), which won the 2019 VisDial challenge. Their model browses the dialog history and updates the visual attention recursively until the model has sufficient confidence to perform visual co-reference resolution. We use their single model's open-source implementation and apply our fine-tuning procedure on the val set in Table 1. When reporting on the test set results in Table 2, we use the leaderboard scores published online which contains further unpublished enhancements based on ensembling (MReaL-BDAI).

# 5 Results

In the following, we report results on the VisDial v1.0 val set, (Table 1), as well as the test-std set[5], (Table 2). For measuring significance (reported on $p \leq 0.05$), we use Kruskal-Wallis (Kruskal and Wallis, 1952) and Wilcoxon signed rank test (Wilcoxon, 1992) with Bonferroni correction (Bonferroni, 1936). We report results in terms of NDCG, which is the main metric of the challenge.

MCA-I-H is our best performing model. It achieves state-of-the-art performance: It outperforms the RvA baseline by almost 5 NDCG points on the val set and by over 7 points on the test set. On the official challenge test set, MCA-I-H ranks $2^{nd}$: it improves over 7 NDCG over the best single model but loses by 2 points against a 6-strong RvA ensemble model (2019 winning entry).

---

[3]Following the guidelines on the dataset page we report results only on v1.0, instead of v0.9. VisDial v1.0 has been consistently used for Visual Dialog Challenge 2018 and 2019.

[4]While 'instance-level' curriculum learning is defined in terms of 'harder dialogs', in our work, we used 'dataset/task-level' curriculum finetuning. Our suggested method is a combination of curriculum learning and fine tuning (pre-training and adjusting to a specific downstream task). As such, we use the term 'curriculum fine-tuning' i.e. adaptation by NDCG aware curriculum during fine-tuning.

[5]We only report results for our best preforming models as the number of allowed submissions to the challenge is limited.

| Model | Sparse annotation Phase | | | | | | Curriculum Fine-tuning | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NDCG ↑ | MRR ↑ | R@1 ↑ | R@5 ↑ | R@10 ↑ | Mean ↓ | NDCG ↑ | MRR ↑ | R@1 ↑ | R@5 ↑ | R@10 ↑ | Mean ↓ |
| RvA (Challenge winners; single model) | 55.86 | 64.42 | 50.71 | 81.50 | 90.15 | 4.06 | 67.90 | 51.92 | 36.57 | 70.69 | 83.61 | 5.85 |
| MCA-H | 51.67 | 59.65 | 45.21 | 77.01 | 86.79 | 4.92 | 64.06 | 38.16 | 22.86 | 54.99 | 71.24 | 9.19 |
| MCA-I | 59.94 | 59.67 | 45.95 | 76.15 | 86.24 | 5.24 | 70.82 | 37.34 | 21.22 | 56.13 | 72.74 | 9.23 |
| MCA-I-HConcQ | 60.65 | 64.08 | 50.83 | 80.74 | 89.62 | 4.22 | 70.81 | 40.75 | 24.53 | 60 | 75.11 | 8.13 |
| MCA-I-HGuidedQ | 60.17 | 64.36 | 50.99 | 80.95 | 89.93 | 4.17 | 71.32 | 44.1 | 28.44 | 61.74 | 76.53 | 7.83 |
| MCA-I-VGH | **62.44** | 61.25 | 47.5 | 78.16 | 87.8 | 4.74 | 72.0 | 40.22 | 24.38 | 58.8 | 73.77 | 8.44 |
| **MCA-I-H** | 60.27 | **64.33** | **51.12** | 80.91 | 89.65 | 4.24 | **72.22** | 42.38 | 26.94 | 60.17 | 75.2 | 8.2 |
| MCA-I-H-GT | 60.27 | 64.33 | 51.12 | 80.91 | 89.65 | 4.24 | 72.18 | 46.92 | 32.09 | 63.85 | 78.06 | 7.37 |

Table 1: Results on VisDial v1.0 val set. Here 'I' denotes image modality while 'H' refers to the use of dialog history. Our baseline models are defined in Section 2.1 and 4.4. MCA variants with dialog history follow the same order as Section 2.2. MCA-I-H-GT refers to the model with corrected dense annotations (see Section 6.2)

| Model | NDCG ↑ | MRR ↑ | R@1 ↑ | R@5 ↑ | R@10 ↑ | Mean ↓ |
|---|---|---|---|---|---|---|
| RvA | 55.59 | 63.03 | 49.03 | 80.40 | 89.83 | 4.18 |
| MS-D365-AI (Ensemble-2nd) | 64.78 | 54.23 | 42.88 | 65.38 | 76.12 | 6.50 |
| MReaL-BDAI (Ensemble-1st) | 74.57 | 53.37 | 40.96 | 66.45 | 79.70 | 6.60 |
| MCA-I | 70.97 | 35.65 | 19.32 | 54.57 | 71.39 | 9.51 |
| MCA-I-VGH | 71.33 | 38.92 | 22.35 | 58.42 | 74.5 | 8.69 |
| **MCA-I-H** | **72.47** | 37.68 | 20.67 | 56.67 | 72.12 | 8.89 |

Table 2: Evaluation on test-std set with results from the online leaderboard. Winners are picked on NDCG. MReaL-BDAI (2019 winning entry) is an ensemble of 6 RvA models. Runner-up MS-D365AI (unpublished) also used ensembling. Note all our submitted MCA models use curriculum fine-tuning and no ensembling.

Compared to MCA-I, which treats the task as multiple rounds of VQA, encoding history improves results, but only significantly for MCA-I-VGH in the sparse annotation phase. After fine-tuning, MCA-I-VGH and MCA-I-H perform equally. MCA-I-H implements a late/shallow fusion of history and image. Architectures which model dense interaction between the conversational history and the image representations (i.e. MCA-I-VGH, MCA-I-HGuidedQ) perform comparably; only MCA-HConcQ performs significantly worse. Note that MCA-I also outperforms the baselines and current SOTA by a substantial margin (both in the sparse annotation phase and curriculum fine-tuning phase), while, counter-intuitively, there is not a significant boost by adding conversational history. This is surprising, considering that according to Das et al. (2017), 38% of questions contain a pronoun, which would suggest that these questions would require dialog history in order to be "understood/grounded" by the model.

Furthermore, curriculum fine-tuning significantly improves performance with an average improvement of 11.7 NDCG points, but worsens performance in terms of the other metrics, which only consider a single ground truth (GT) answer.

## 6 Error Analysis

In the following, we perform a detailed error analysis, investigating the benefits of dialog history encoding and the observed discrepancy between the NDCG results and the other retrieval based metrics.

### 6.1 Dialog History

We performed an ablation study whereby we did not include the caption as part of historical context and compare with the results in Table 1. The performance dropped from (NDCG 72.2, MRR 42.3) to (NDCG 71.6, MRR 40.7) using our best performing MCA-I-H model after finetuning. Since the crowd-sourced conversation was based on the caption, the reduced performance was expected.

In order to further verify the role of dialog history, we conduct a crowd-sourcing study to understand which questions require dialog history, in order to be understood by humans. We first test our history-encoding models on a subset (76 dialogs) of the recently released VisPro dataset (Yu et al., 2019a) which focuses on the task of Visual Pronoun Resolution.[6] Note that VisPro also contains non-referential pleonastic pronouns, i.e. pronouns used as "dummy subjects" when e.g. talking about the weather ("Is it sunny?").

We thus create a new crowd-sourced dataset[7], which we call *VisDialConv*. This is a subset of the VisDial val-set consisting of 97 dialogs, where the crowd-workers identified single turns (with dense annotations) requiring historical information. In particular, we asked crowd-workers whether they could provide an answer to a question given an image, without showing them the dialog history, and select one of the categories in Table 4 (see further details in Appendix B).

In order to get reliable results, we recruited 3 crowd-workers per image-question pair and only kept instances where at least 2 people agreed. Note that we only had to discharge 14.5% of the origi-

---

[6] We use the intersection of dialogs in VisDial val set and VisPro to create this subset.

[7] Data collection code available at https://github.com/shubhamagarwal92/visdialconv-amt

| Model | Sparse annotation Phase | | | | | | Curriculum Fine-tuning | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NDCG ↑ | MRR ↑ | R@1 ↑ | R@5 ↑ | R@10 ↑ | Mean ↓ | NDCG ↑ | MRR ↑ | R@1 ↑ | R@5 ↑ | R@10 ↑ | Mean ↓ |
| **VisPro subset dataset** | | | | | | | | | | | | |
| MCA-I | 59.80 | 57.88 | 45.39 | 72.24 | 82.76 | 5.84 | 69.82 | 36.2 | 20 | 54.08 | 70.92 | 10.02 |
| MCA-I-HConcQ | 61.08 | **61.79** | **48.95** | **77.5** | **86.58** | **4.72** | 68.44 | 38 | 22.24 | 55.79 | 71.71 | 9.17 |
| MCA-I-HGuidedQ | 61.35 | 60.13 | 47.11 | 75.26 | 86.18 | 5.23 | 68.29 | 36.59 | 21.05 | 53.29 | 70.13 | 9.76 |
| MCA-I-VGH | 61.68 | 59.33 | 46.18 | 75.53 | 86.71 | 5.07 | 68.97 | 39.21 | 23.68 | **57.11** | 70.53 | **8.83** |
| **MCA-I-H** | **61.72** | 59.62 | 45.92 | 77.11 | 86.45 | 4.85 | **70.87** | **39.8** | **25.39** | 55.13 | 70.39 | 9.42 |
| **VisDialConv (Crowd-sourced subset) dataset** | | | | | | | | | | | | |
| MCA-I | 52.07 | 55.55 | 41.65 | 72.47 | 83.81 | 5.92 | 58.65 | 36.2 | 20.52 | 53.3 | 68.25 | 10.32 |
| MCA-I-HConcQ | 54.84 | 62.06 | 47.42 | **80.1** | **88.87** | **4.37** | 61.42 | 37.92 | 21.86 | 55.67 | **73.3** | **9.01** |
| MCA-I-HGuidedQ | 53.81 | **62.29** | **48.35** | **80.1** | 88.76 | 4.42 | **62.92** | 38.07 | 22.58 | 54.74 | 70.82 | 9.5 |
| MCA-I-VGH | **55.48** | 58.45 | 44.54 | 74.95 | 86.19 | 5.18 | 60.63 | 38.1 | 22.89 | 53.71 | 70.31 | 9.49 |
| **MCA-I-H** | 53.01 | 61.24 | 47.63 | 79.07 | 87.94 | 4.77 | 59.89 | **39.73** | **25.15** | **56.49** | 71.86 | 9.53 |

Table 3: Automatic evaluation on the subsets of VisPro and VisDialConv dataset. We found history based MCA models to outperform significantly compared to the MCA-I model. On VisDialConv, MCA-I-VGH still outperform all other models in spare annotation phase while MCA-I-HGuidedQ performs the best after fine-tuning.

| Annotation | Count | Percentage |
|---|---|---|
| VQA turns | 594 | 67.12% |
| **History required** | **97** | **10.96%** |
| Common Sense | 94 | 10.62% |
| Guess | 59 | 6.67% |
| Cant tell | 34 | 3.84% |
| Not relevant | 7 | 0.79% |

Table 4: Results of crowd-sourcing study to understand whether humans require dialog history to answer the question. 'VQA turns' indicate that humans could potentially answer correctly without having access to the previous conversation while 'History required' are the cases identified requiring dialog context. We also identified the cases requiring world knowledge/ common sense, guessing and questions not relevant to the image.

nal 1035 image-question pairs, leaving us with 885 examples. The results in Table 4 show that only 11% required actual dialog historical context according to the crowd-workers. Most of the time (67% cases), crowd-workers said they can answer the question correctly without requiring history.

The results in Table 3 are on the subset of 97 questions which the crowd-workers identified as requiring history.[8] They show that history encoding models (MCA-I-HGuidedQ / MCA-I-HConcQ / MCA-I-H / MCA-I-VGH) significantly outperform MCA-I, suggesting that this data cannot be modelled as multiple rounds of VQA. It can also be seen that all the models with dense (early) interaction of the historical context outperform the one with late interaction (MCA-I-H) in terms of NDCG. Models with dense interactions appear to be more reliable in choosing other correct relevant answers because of the dialog context.

---

[8]We took care to only include examples from Visdial val set in both Vispro and VisDialConv subsets. Also note, there are only 8 overlapping instances between Vispro and Visdial-Conv subsets.

Our best performing model on VisDialConv is MCA-I-HGuidedQ and achieves a NDCG value of 62.9 after curriculum fine-tuning. However, on the VisPro subset, we observe that MCA-I-H still outperforms the other models. Interestingly, on this set, MCA-I also outperforms other history encoding models (except for MCA-I-H).

In sum, our analysis shows that only a small subset of the VisDial dataset contains questions which require dialog history, and for those, models which encode history lead to better results. We posit that this is due to the fact that questions with pleonastic pronouns such as "Is it sunny/daytime/day..." are the most frequent according to our detailed analysis in Appendix C about the dialog phenomena.

| Relevance of GT | Train | | Val | |
|---|---|---|---|---|
| | Count | Percent | Count | Percent |
| 1 | 1057 | 52.85% | 643 | 31.15% |
| 0.8 | - | - | 397 | 19.23% |
| 0.6 | - | - | 330 | 15.99% |
| 0.5 | 526 | 26.30% | - | - |
| 0.4 | - | - | 281 | 13.61% |
| 0.2 | - | - | 227 | 11.00% |
| 0 | 417 | **20.85%** | 186 | **9.01%** |
| Total | 2000 | 100% | 2064 | 100% |

Table 5: Relevance score (dense annotation) provided for 2k/2k train/val QA turns. We find that 20% of the ground truth answers were marked as irrelevant (0 score) and partially relevant (0.5 score) by the human annotators for train set. This can be attributed to human errors made while collecting the original data as well as when crowd-sourcing the dense annotations.

## 6.2 Dense Annotations for NDCG

Here, we investigate the discrepancy between the NDCG results and the other retrieval-based methods. First, we find that the annotation scales differs: while there is a 3-way annotation on the train set, the val set defines 6 possible relevance classes, see Table 5. This affects the evaluation results of our

| Image | Dialog | MCA-I-H | MCA-I-VGH |
|---|---|---|---|
| A bag of chips and a apple and orange. $N_{Rel}$: 15 | Q What kind of chips are they? A Chili cheese corn chips. Q Is the bag open or still sealed? A Sealed. Q Is it next to the apple and orange? A Yes. Q Are they all on a table? GT: Yes. Rel: 1.0 | ♣ $R_{GT}$:1 ; NDCG: 65.56 (1.0) Yes. (1.0) Yes they are on a table. (0.0) Maybe , it's a close up. (0.0) Can't see a table. (0.2) I think so, it is a close up. ♦ $R_{GT}$:2 ; NDCG: 69.94 (0.8) I think so. (1.0) Yes. (0.2) It appears to be. (0.4) I would think so. (0.2) I think so, it is a close up. | ♣ $R_{GT}$:1 ; NDCG: 83.93 (1.0) Yes. (1.0) Yes they are on a table. (0.0) Yes they are. (0.0) Can't see a table. (0.2) I think so, it is a close up. ♦ $R_{GT}$:4 ; NDCG: 84.15 (0.8) I think so. (0.8) They appear to be. (0.4) Probably. (1.0) Yes. (1.0) Yes they are. |
| A remote controller is hidden in a console inside of an arm rest. $N_{Rel}$: 8 | Q Can you see the remote? A Yes i can. Q What color is it? A It is black. Q Can you tell what it is for? A It appears to be a phone. Q What kind of furniture is it in? GT: Looks like a car console. Rel: 0.4 | ♣ $R_{GT}$:1 ; NDCG: 63.19 (0.4) Looks like a car console. (0.4) It looks like a chair on a train or a bus. (0.0) There are tables. (0.0) Looks like an outdoor space. (0.2) It's a cubicle with shelves. ♦ $R_{GT}$:3 ; NDCG: 79.2 (0.4) I cannot tell. (0.4) I can't tell. (0.4) Looks like a car console. (0.2) Not sure. (0.4) Can't tell. | ♣ $R_{GT}$:2; NDCG: 58.99 (0.0) A cell phone, i can't see it close up. (0.4) Looks like a car console. (0.4) It looks like a chair on a train or a bus. (0.2) It's a cubicle with shelves. (0.0) The picture does not show 1. ♦ $R_{GT}$:7 ; NDCG: 82.22 (0.4) I cannot tell. (0.4) Can't tell. (0.4) I can't tell. (0.2) Not sure. (0.0) The picture does not show 1. |

Figure 4: Top-5 ranked predictions (relevance in parentheses) of MCA-I-H and MCA-I-VGH after both sparse annotation and curriculum fine-tuning phase. $R_{GT}$ defines the rank of Ground Truth (GT) predicted by the model. We also calculate NDCG of rankings for current question turn. $N_{Rel}$ denotes number of candidate answer options (out of 100) with non-zero relevance (dense annotations). Here ♣ and ♦ represents predictions after sparse annotation and curriculum fine-tuning respectively.

model, for which we can't do much.

Next, a manual inspection reveals that the relevance weight annotations contain substantial noise: We find that ground truth answers were marked as irrelevant for about 20% of train and 10% of val set. Thus, our models seem to get "confused" by fine-tuning on this data. We, therefore, manually corrected the relevance of only these GT answers (in dense annotations of train set only, but not in val set). Please see Appendix D for further details. The results in Table 1 (for MCA-I-H-GT) show that the model fine-tuned on the corrected data still achieves a comparable NDCG result, but substantially improves stricter (single answer) metrics, which confirms our hypothesis.

Finally, due to the noisy signal they receive during fine-tuning, our models learn to select "safe" answers[9], such as "I can't tell" (see examples in

Figure 4), which rank high according to (the more forgiving) NDCG, but perform poorly for stricter metrics like MRR and Recall.

## 7 Discussion and Related Work

Our results suggest that the VisDial dataset only contains very limited examples which require dialog history. Other visual dialog tasks, such as GuessWhich? (Chattopadhyay et al., 2017) and GuessWhat?! (De Vries et al., 2017) take place in a goal-oriented setting, which according to Schlangen (2019), will lead to data containing more natural dialog phenomena. However, there is very limited evidence that dialog history indeed matters for these tasks (Yang et al., 2019). As such, we see data collection to capture visual dialog phenomena as an open problem.

Nevertheless, our results also show that encoding dialog history still leads to improved results. This is in contrast with early findings that a) "naive" encoding will harm performance (Das et al. (2017);

---

[9] We show the statistics of top-ranked predictions by our MCA-I-H model on our VisdialConv subset (i.e. 97 dialogs of the Visdial val set). Read as: (Response, count, %) (Yes, 14, 14%) (No, 11, 11.34%) (I cannot tell, 9, 9.27%) (Nope, 3, 3%) (Not that I see, 2, 2.06%) (Red and white, 2, 2.06%) (Not sure, 2, 2.06%) (I can't tell, 2, 2.06%). This shows that

at least 13.3% of answers are non-commital (I cannot tell, Not sure, I can't tell).

see MCA-I-HConcQ in Table 1), or that b) history is not necessary (Massiceti et al., 2018).

Furthermore, we find that our model learns to provide generic answers by taking advantage of the NDCG evaluation metric. Learning generic answers is a well-known problem for open-domain dialog systems, e.g. (Li et al., 2016). While the dialog community approaches these phenomena by e.g. learning better models of coherence (Xu et al., 2018), we believe that evaluation metrics also need to be improved for this task, as widely discussed for other generation tasks, e.g. (Liu et al., 2016; Novikova et al., 2017; Reiter, 2018). As a first step, BERT score (Zhang et al., 2019) could be explored to measure ground-truth similarity replacing the noisy NDCG annotations of semantic equivalence.

## 8 Conclusion and Future Work

In sum, this paper shows that we can get SOTA performance on the VisDial task by using transformer-based models with Guided-Attention (Yu et al., 2019b), and by encoding dialog history and fine-tuning we can improve results even more.

Of course, we expect pre-trained visual BERT models to show even more improvements on this task, e.g. Vilbert (Lu et al., 2019), LXMert (Tan and Bansal, 2019), UNITER (Chen et al., 2019) etc. However, we also show the limitations of this shared task in terms of dialog phenomena and evaluation metrics. We, thus, argue that progress needs to be carefully measured by posing the right task in terms of dataset and evaluation procedure.

## Acknowledgments

## References

Shubham Agarwal and Raghav Goyal. 2018. Ensemble based discriminative models for visual dialog challenge 2018. In *Visual Dialog Challenge*.

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*.

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

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. In *CoRR abs/1607.06450*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ACM*.

Carlo Bonferroni. 1936. Statistical class theory and probability calculation. *Publications of the Higher Institute of Economic and Commercial Sciences of Florence*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. In *CoRR abs/1605.07683*.

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 *CoRR abs/1708.05122*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. UNITER: Learning universal image-text representations. In *CoRR abs/1909.11740*.

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

Abhishek Das, Devi Parikh, and Dhruv Batra. 2019. Response to "visual dialogue without vision or dialogue"(massiceti et al., 2018). *CoRR abs/1901.05531*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W. Taylor. 2018. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *ICCV*.

Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*.

Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. In *ACL*.

Dalu Guo, Chang Xu, and Dacheng Tao. 2019. Image-question-answer synergistic network for visual dialog. In *CVPR*.

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

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The photobook dataset: Building common ground through visually-grounded dialogue. In *ACL*.

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

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

Vidur Joshi, Matthew Peters, and Mark Hopkins. 2018. Extending a parser to distant domains using a few dozen partially annotated examples. In *ACL*.

Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. 2019. Dual attention networks for visual reference resolution in visual dialog. In *EMNLP*.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *NeurIPS*.

Jin-Hwa Kim, Devi Parikh, Dhruv Batra, Byoung-Tak Zhang, and Yuandong Tian. 2017. Codraw: Visual dialog for collaborative drawing. In *CoRR abs/1712.05558*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. 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*.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In *NAACL*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*.

William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL*.

Tzu-Hsiang Lin, Trung Bui, Doo Soon Kim, and Jean Oh. 2018. A multimodal dialogue system for conversational image editing. In *NeurIPS Conv AI workshop*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.

Sharid Loáiciga, Liane Guillou, and Christian Hardmeier. 2017. What is it? disambiguating the different readings of the pronoun 'it'. In *EMNLP*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *CoRR abs/1908.02265*.

Ramesh Manuvinakurike, Jacqueline Brixey, Trung Bui, Walter Chang, Doo Soon Kim, Ron Artstein, and Kallirroi Georgila. 2018a. Edit me: A corpus and a framework for understanding natural language image editing. In *LREC*.

Ramesh Manuvinakurike, Trung Bui, Walter Chang, and Kallirroi Georgila. 2018b. Conversational image editing: Incremental intent identification in a new dialogue task. In *SIGDial*.

Daniela Massiceti, Puneet K Dokania, N Siddharth, and Philip HS Torr. 2018. Visual dialogue without vision or dialogue. *CoRR abs/1812.06417*.

Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *CVPR*.

Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. 2019. Recursive visual attention in visual dialog. In *CVPR*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *EMNLP*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NeurIPS-W*.

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

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*.

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

David Schlangen. 2019. Grounded agreement games: Emphasizing conversational grounding in visual dialogue settings. In *CoRR abs/1908.11279*.

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

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*.

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

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. *Breakthroughs in statistics*.

Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity". In *EMNLP*.

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

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019a. What you see is what you get: Visual pronoun coreference resolution in dialogues. In *EMNLP*.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019b. Deep modular co-attention networks for visual question answering. In *CVPR*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with bert. In *CoRR abs/1904.09675*.

## A   More implementation details

We built our implementation upon starter code in PyTorch which the VisDial organisers kindly provided.[10] We follow the guidelines of Teney et al. (2018) and used static 36 as the number of object proposals in our experiments (though our model can handle dynamic number of proposals).

We experimentally determined the learning rates of 0.0005 for training MCA models and 0.0001 for fine-tuning and reducing it by 1/10 after every 7 and 10 epochs out of a total of 12 epochs for training and 1/5 after 2 epochs for fine-tuning.

We use pytorch's LambdaLR scheduler while training and ReduceLROnPlateau for the fine-tuning procedure. Dropout of 0.2 is used for regularization and we perform early stopping and saved the best model by tracking the NDCG value on val set. Layer normalisation (Ba et al., 2016) is used for stable training following (Vaswani et al., 2017; Yu et al., 2019b). Attention reduction consisted of 2 layer MLP (fc(d)-ReLU-Dropout(0.2)-fc(1)).

We also experimented with different contextual representations, including BERT (Devlin et al., 2019); However we didn't observe any improvement, similar to the observation by (Tan and Bansal, 2019).

For the results on the validation set, only the training split is used. To report results on test-std set, both the training and val set are used for training. For curriculum fine-tuning we use multi-class cross entropy loss where weighted by the relevance score. All our MCA modules have 6 layers and 8 heads, which we determined via a hyper parameter search. Table 7 shows more details.

| Annotation | Text |
| --- | --- |
| VQA turns | I can confidently tell the correct answer just seeing the image. |
| **History required** | I want to know what was discussed before to answer confidently. Cannot answer with just the question and image. Need more information (context) from previous conversation. |
| Common Sense | I can answer it but by inferring using common sense. |
| Guess | I can only guess the answer. |
| Cant tell | I can't tell the answer. |
| Not relevant | Not relevant question for this image. |

Table 6: Mapping of human annotation with the actual text shown to the user.

---

[10] https://github.com/batra-mlp-lab/ visdial-challenge-starter-pytorch.

| Model | Training | | | | | | Curriculum Fine-tuning | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NDCG | MRR | R@1 | R@5 | R@10 | Mean | NDCG | MRR | R@1 | R@5 | R@10 | Mean |
| MCA-I-H (L6 H8) | 60.27 | 64.33 | 51.12 | 80.91 | 89.65 | 4.24 | 72.22 | 42.38 | 26.94 | 60.17 | 75.2 | 8.2 |
| MCA-I-H (L2 H4) | 58.99 | 64.46 | 51.14 | 81.03 | 89.91 | 4.19 | 70.57 | 42.48 | 26.3 | 61.3 | 76.05 | 8.06 |
| MCA-I-H (L6 H2) | 60.13 | 60.63 | 46.7 | 77.55 | 87.47 | 4.8 | 70.42 | 39.17 | 23.3 | 57.64 | 73.48 | 8.69 |

Table 7: Hyper-parameter tuning for number of layers and number of heads. The results in the main paper are reported with 6 Layers(L6) and 8 Heads (H8) for all MCA models.

## B   AMT Interface

Here, we provide more details on the crowd-sourcing study described in Section 6.1. Figure 6 shows the instructions shown to the turkers. We also setup a qualification test consisting of 2 test images (in Figure 7) to assess whether turkers understood the task properly. This allowed us to have an automated quality check for the annotations. Each HIT consisted of 15 images. For the actual task (e.g. Fig. 8), users were shown just the image and the current question – without any previous historical context – and asked to choose one of the answers as shown in Table 6. Our AMT interface[11] used AWS boto3 library in python.

## C   Diversity and dialog phenomena in VisDial dataset

We also did an analysis of the top-20 questions (Figure 9) and answers (Figure 10) in the training set. 'Yes'/'No' binary answers form the major chunk (19.15% and 21.2% respectively) of ground truth answers. Color related answers (such as White, Brown in the top-20 answers) form 4% of all the ground truth answers. Numbered answers (such as 0, 1, 2 ,3) form 1.3% while 'Can't tell' form another 1.2%.

As evident in the top-20 questions, weather related questions (such as 'Is it sunny/daytime/day/night?'), color related ('What color is it/his hair/the table?') and basic conversational-starters ('Can you see any people?') form the major portion.

We also tried to analyze the top-20 answers (Figure 11) which had non-zero relevance in the dense annotations. Specifically, we took all 2k example turns of training set with dense annotations for each of 100 options. We find that generic answers such as 'Can't tell', binary answers 'Yes/No' and their semantically equivalent answers 'Not that i can see' are mostly given non-zero relevance by crowd-workers.

We tried to calculate the statistics of the pronouns and ellipsis which we consider essential (but not complete) phenomena in a dialog dataset. Figure 12 shows the number of pronouns in a dialog. We find that major chunk consisted of 2-6 pronouns in all the 10 questions across the dialog. We tried to distinguish between the usage of 'it' as pleonastic and non-pleonastic pronouns (discussed in (Loáiciga et al., 2017)). For e.g. in the sentence: 'It is raining'. Here, though, 'it' would be identified as a pronoun, but it doesn't refer to anything. Notice the drift in distribution of the number of pronouns (All pronouns vs Non-pleonastic). We also tried to identify the cases of ellipsis (methodology explained further) and found that majority questions (82%) doesn't contain any case of ellipsis in the dialog. We define simple heuristics to identify dialog phenomena. Specifically, our heuristics can be listed as:

- We use constituency parser (Joshi et al., 2018) to parse each question[12]. If the parsed tree doesn't contain 'Sentence' as the root ('S', 'SQ', 'SBARQ', 'SINV'), we consider it a case of ellipsis.
- We use spaCy[13] to extract the pronouns in all the questions of a dialog.
- To distinguish between different usage of 'it', we mark all the co-occurrences of manually defined weather identifiers ('rainy', 'sunny', 'daytime', 'day', 'night') as pleonastic.
- Though 'other' is a pronoun, it is not tagged by standard taggers. We explicitly deal with these cases to tag 'other' as a case of pronoun. For e.g. 'What about the other?'

## D   Corrected dense annotations

We maintain the whole relevance list, however we change the relevance of only the ground truth (GT) to 1 instead of 0/0.5 in the train annotations (only 943 values). This was done to avoid extra gradient

---

[11]We built upon the repo: https://github.com/jcjohnson/simple-amt.

[12]https://github.com/allenai/allennlp/blob/master/allennlp/pretrained.py

[13]https://spacy.io/usage/linguistic-features

| Image | Dialog | MCA-I-H | MCA-I-VGH |
|---|---|---|---|
|  A surfer crouches as they ride a cresting wave. $N_{Rel}$: 15 | Q Is the photo in color? <br> A Yes. <br> Q Any other people? <br> GT: No. <br> Rel: 0.8 | ♣ $R_{GT}$:1 ; NDCG 83.32 <br> (0.8) No. <br> (0.2) O. <br> (1.0) Nope. <br> (0.8) No there's not. <br> (0.4) Just the 1. <br><br> ♦ $R_{GT}$:2; NDCG 91.2 <br> (1.0) Nope. <br> (0.8) No. <br> (0.8) Not that i can see. <br> (0.8) Not that i see. <br> (0.8) No there's not. | ♣ $R_{GT}$:1 ; NDCG 74.98 <br> (0.8) No. <br> (0.2) O. <br> (1.0) Nope. <br> (0.2) 0 at all. <br> (0.8) Not that i can see. <br><br> ♦ $R_{GT}$:1 ; NDCG 85.24 <br> (0.8) No. <br> (0.8) Not that i can see. <br> (0.8) Not that i see. <br> (0.8) No there's not. <br> (1.0) Nope. |
|  An apple and orange are sitting in a white box with size measurements. $N_{Rel}$: 4 | Q What color is the apple? <br> A It is red and yellow. <br> Q What color is the orange? <br> A It is dark orange. <br> Q What is the size of the box? <br> A Can't tell. <br> Q Where is the box? <br> GT: Can't tell. <br> Rel: 1.0 | ♣ $R_{GT}$:9 ; NDCG 0.0 <br> (0.0) On the table. <br> (0.0) In a container by the window. <br> (0.0) On table. <br> (0.0) Use no paper here. <br> (1.0) I cannot tell. <br> ♦ $R_{GT}$:5 ; NDCG 73.59 <br> (1.0) I cannot tell. <br> (0.0) On the table. <br> (1.0) I can't tell. <br> (0.6) Not sure. <br> (1.0) Can't tell. | ♣ $R_{GT}$:9 ; NDCG 0.0 <br> (0.0) On table. <br> (0.0) On the table. <br> (0.0) In a container by the window. <br> (0.0) In a kitchen. <br> (1.0) I cannot tell. <br> ♦ $R_{GT}$:4 ; NDCG 98.84 <br> (1.0) I cannot tell. <br> (1.0) I can't tell. <br> (0.6) Not sure. <br> (1.0) Can't tell. <br> (0.0) In a container by the window. |
|  A clock with a rose on its corner sits on the wall. $N_{Rel}$: 3 | Q Is it a real rose? <br> A I don't think so. <br> Q Is it an old fashion clock? <br> A No, not really. <br> Q Is it a digital clock? <br> A No. <br> Q Is it hanging on the wall or leaning? <br> GT: It's hanging. <br> Rel: 1.0 | ♣ $R_{GT}$:1 ; NDCG: 81.55 <br> (1.0) It's hanging. <br><br> (0.0) Yes, it's attached to the side of the building. <br> (0.0) Yes. <br><br> (0.0) It is cut out, but it is definitely sitting on something. <br> (0.0) It looks like. <br> ♦ $R_{GT}$:2 ; NDCG 51.45 <br> (0.0) It looks like. <br><br> (1.0) It's hanging. <br> (0.0) Can't tell. <br> (0.0) Unclear. <br> (0.0) I think so. | ♣ $R_{GT}$:2 ; NDCG 51.45 <br> (0.0) No it is not mounted on the wall. <br> (1.0) It's hanging. <br><br> (0.0) It is cut out, but it is definitely sitting on something. <br> (0.0) Yes, it's attached to the side of the building. <br> (0.0) On the rail. <br> ♦ $R_{GT}$:3 ; NDCG 40.78 <br> (0.0) No it is not mounted on the wall. <br> (0.0) Not sure. <br> (1.0) It's hanging. <br> (0.0) Can't tell. <br> (0.0) I can't tell. |

Figure 5: Top-5 ranked predictions (relevance in parentheses) of MCA-I-H and MCA-I-VGH after both sparse annotation and curriculum fine-tuning phase. $R_{GT}$ defines the rank of Ground Truth (GT) predicted by the model and NDCG of rankings for current question turn. $N_{Rel}$ denotes number of candidate answer options (out of 100) with non-zero relevance (dense annotations). Here ♣ and ♦ represents predictions after sparse annotation and curriculum fine-tuning respectively.

information that the model will receive because of noise in the dataset, since these examples were already seen during the spare annotation phase. Val annotations remains unaffected for fair comparison. As expected, this simple correction increase the ground truth related metrics such as R{1,5,10} drastically.

## Instructions

- This task is fairly simple. In this task, you will be shown a picture and a question. Your task is to specify whether the question can be answered independently just seeing the image without any other context or information about the previous conversation.
- You have to select either of the following responses:
  - I can confidently tell the correct answer just seeing the image
  - I want to know what was discussed before to answer confidently. Cannot answer with just the question and image. Need more information (context) from previous conversation.
  - I can answer it but by inferring using common sense
  - I can only guess the answer
  - I can't tell the answer
  - Not relevant question for this image

  Good luck and have fun! Do not forget to press submit at the end! Otherwise all your progress will be lost.

### Some examples:

Question: "Are there any dogs in the image?"
You can confidently tell the correct answer just by looking at the image as at least one the dog is clearly visible present.
Question: "What is its color?"
Want to know what was discussed before to answer confidently. Do not assume anything apart from the image or question. Here you need more information on what the word 'its' is referring to, i.e. You need more information if the previous conversation was about the dog or boat.
Question: "What about the other?"
Want to know what was discussed before to answer confidently. Similarly, in this case of ambiguity, make no assumption here. We dont know if we are talking about mountain, boat or dog.
Question: "Is it a lake or sea?" / "How old is the dog?"
These are the type of questions where you can answer by common sense or just guessing. You actually dont need the previous context of the conversation.

Figure 6: Instructions for the AMT task.



is the plane above any bodies of water?

do you see other appliances?

- ○ I can confidently tell the correct answer just seeing the image
- ○ Want to know what was discussed before in the conversation to answer
- ○ I can answer it but by inferring using common sense
- ○ I can only guess the answer
- ○ I cannot tell the answer
- ○ Not relevant question for this image

- ○ I can confidently tell the correct answer just seeing the image
- ○ Want to know what was discussed before in the conversation to answer
- ○ I can answer it but by inferring using common sense
- ○ I can only guess the answer
- ○ I cannot tell the answer
- ○ Not relevant question for this image

Figure 7: Qualification test consisting of 2 test images to allow the turkers to actually attempt the task
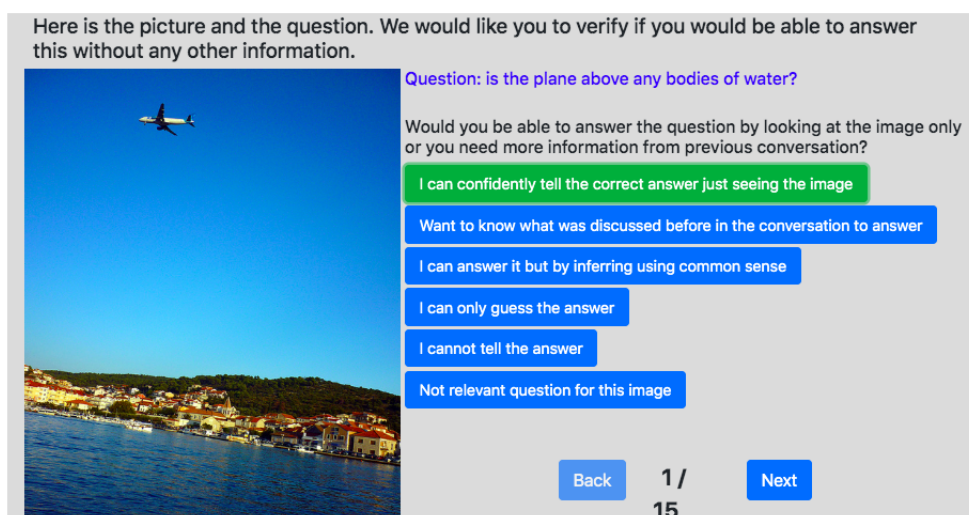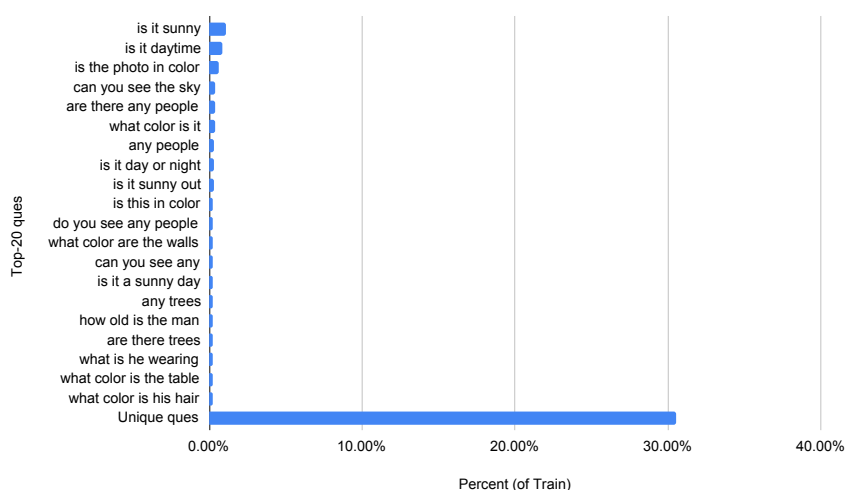
Figure 8: Sample task.



Figure 9: Top-20 questions in the training set. Of all the questions in the training set, only 30% questions are unique while weather related questions (like sunny, daytime, rainy) top the charts.
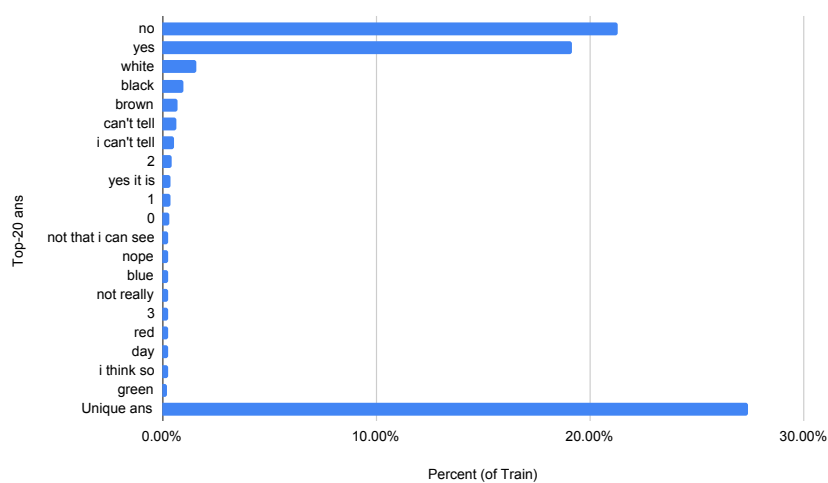


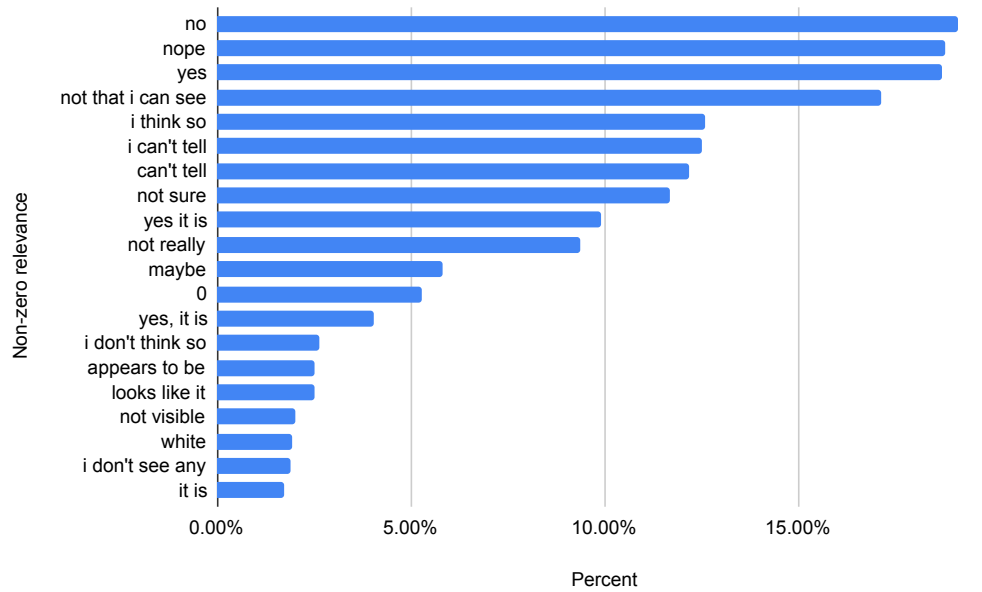Figure 10: Top-20 answers in the training set. Yes/No forms a major chunk in top 20 answers.

Figure 11: Top-20 answers with non-zero relevance in the dense annotations of training set. Generic and yes/no semantically equivalent answers mostly constitute the list. Percentage is calculated out of total 3652 unique answers which have non-zero relevance in train dense annotations set.
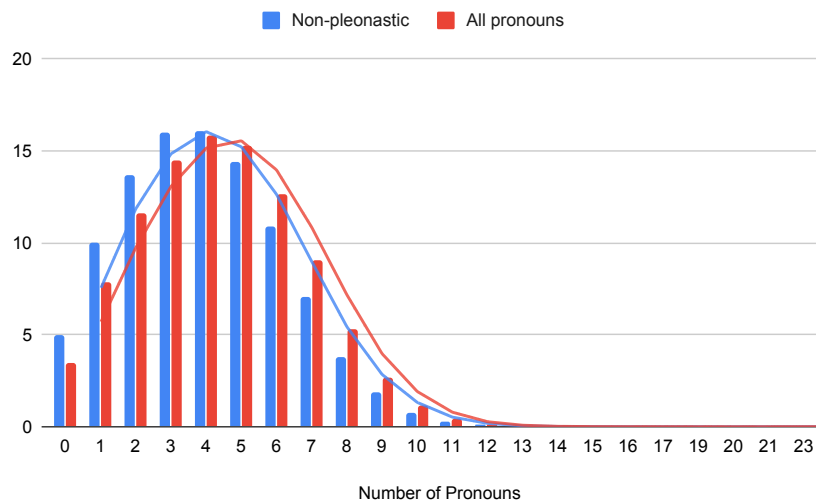


Figure 12: Number of pronouns in 10 questions of a dialog.