Multi-View Attention Network for Visual Dialog

Sungjin Park 1,†, Taesun Whang 2,4,†, Yeochan Yoon 3 and Heuiseok Lim 1,*

1 Department of Computer Science and Engineering, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841, Korea; ec_park@korea.ac.kr
2 Wisenut Inc., 49, Daewangpangyo-ro 644beon-gil, Bundang-gu, Seongnam-si 13493, Gyeonggi-do, Korea; taesunwhang@wisenut.co.kr
3 Electronics and Technology Research Institute (ETRI), 161, Gajeong-dong, Yuseong-gu, Daejeon 34129, Korea; ycyoon@etri.re.kr
* Correspondence: limhseok@korea.ac.kr
† These authors contributed equally to this work.
‡ Work performed while at Korea University.

Abstract: Visual dialog is a challenging vision-language task in which a series of questions visually grounded by a given image are answered. To resolve the visual dialog task, a high-level understanding of various multimodal inputs (e.g., question, dialog history, and image) is required. Specifically, it is necessary for an agent to (1) determine the semantic intent of question and (2) align question-relevant textual and visual contents among heterogeneous modality inputs. In this paper, we propose Multi-View Attention Network (MVAN), which leverages multiple views about heterogeneous inputs based on attention mechanisms. MVAN effectively captures the question-relevant information from the dialog history with two complementary modules (i.e., Topic Aggregation and Context Matching), and builds multimodal representations through sequential alignment processes (i.e., Modality Alignment). Experimental results on VisDial v1.0 dataset show the effectiveness of our proposed model, which outperforms previous state-of-the-art methods under both single model and ensemble settings.

Keywords: visual dialog; attention mechanism; multimodal learning; vision-language

1. Introduction

As a part of the interdisciplinary research that combines natural language processing with computer vision, a wide variety of vision–language tasks (e.g., Visual Question Answering (VQA), image captioning, referring expressions, etc.) have been introduced in recent years. Considerable efforts in this field have advanced the capabilities of artificial intelligence agents a step further, but the agent’s comprehension of multimodal information is still far from human-level reasoning and cognitive ability [1,2].

The visual dialog task is similar to VQA in that it requires the agent to answer a question that is guided by an image, but this task differs in that the agent needs to answer a series of questions focusing on previous dialog as well as a given image. It is more challenging than other vision–language tasks because the agent is asked to selectively ground the visual contents related to the question topics, which change as the dialog proceeds. For example, Figure 1 illustrates how the question topics change during each dialog turn (e.g., “household goods”, “people”, and “food”). Furthermore, answering some questions that contain ambiguous expressions (e.g., question 5 (Q5): “how old do they look?”) can be more difficult because the agent should consider which entity “they” refers to (i.e., “kids” and “boys” in caption and Q4–A4 pair) and then ground the referent in the image. To address these issues, the dialog agent should be capable of determining the semantic intent of the question, clarifying referential ambiguities, and then leveraging grounded visual contents and identified semantic information.
Several recent researches have been performed to solve the visual dialog task from the perspective of visual co-reference resolution [3–6]. However, resolving the visual co-reference does not always lead to complete understanding of semantic intent and topics of the question. For example, when answering Q6 in Figure 1, the agent is required to not only resolve visual co-reference, but also explicitly understand the semantic intent of the question which is asking whether other “snacks” exist or not (i.e., the model should focus on “snacks” rather than “they”). From these observations, it is crucial to capture what the topic of the question is in order to accurately determine the semantic intent of the question.

To this end, this paper proposes Multi-View Attention Network (MVAN), which leverages question-guided contextual information and clues through the dialog history; and then effectively learns semantic alignments between visual and textual representations through the sequential alignment processes. MVAN consists of three main modules. First, the Context Matching module effectively represents contextual information of dialog history that is relevant to the question on a sentence level. This is because, in general, the semantic intent of a sentence tends to be determined by the context of the entire sequence as well as some words that are directly connected to the topic. Second, the Topic Aggregation module is capable of capturing topic-guided clues from the dialog history on a word level. This takes advantage of the fact that topic-related representation is well-constructed by directly using the original word embeddings of each word. Both modules adaptively propagate textual information that interacts with the semantic intent of the current question by attention mechanism and gate function. Lastly, the Modality Alignment module performs two sequential aligning processes to learn semantic alignments between each textual output of previous modules with visual features. Since the alignments between contextual representation and visual contents can be implicit and noisy, the Modality Alignment module first learns the soft mapping between heterogeneous inputs based on the topic-guided clues, then sequentially aligns them with the surrounding contextual information.

The main contributions of this paper are as follows. (1) We propose MVAN that consists of two complementary views and combine them with visual contents through multiple alignment steps. (2) Experimental results on VisDial v1.0 show that our proposed model outperforms the previous state-of-the-art methods. (3) Extensive experiments such as multi-task learning and fine-tuning on dense annotations demonstrate the superiority of our model. (4) We show that MVAN explicitly understands the semantic intent of the question through the visualization of reasoning process for each module, which leads to reasonable interpretations of employing heterogeneous multimodal inputs.
2. Related Work

Visual Dialog

Visual dialog is a task proposed by Das et al. [7] that requires the dialog agent to answer the current question by exploiting both the image and dialog history. Das et al. [7] also introduced encoder-decoder models such as late fusion, hierarchical recurrent network, and memory network as baseline methods. Most of the previous approaches are predominantly based on attention mechanisms to fuse representations of the given multimodal inputs [8–11]. Another direction of research that is inspired by graphical networks focuses on learning the inherent relations among image, dialog history, and question [12–14].

On the other hand, several approaches that explicitly resolve ambiguous references are based on a visual co-reference resolution. Kottur et al. [4] used neural module networks [15] to effectively link references and ground relevant visual contents at word level. Kang et al. [5] adapted a self-attention mechanism [16] based on sentence-level representation to resolve referential ambiguity. Niu et al. [6] proposed a recursive attention mechanism to capture question-relevant dialog history and ground-related visual contents to the image. Most existing works [4,5] that use only word or sentence representations have limitations in identifying the semantic intent of the question. Unlike this, our proposed model considers both topic-related clues and contextual information to effectively capture the semantic intent of the question. In addition, MVAN adaptively integrates dialog history and visual contents by performing sequential alignment steps rather than exploiting only dialog history and visual contents that meet specific recursion conditions [6]. More recently, Murahari et al. [17] and Wang et al. [18] introduced a fine-tuning method using a pre-trained model, as it is observed that pre-trained language model architectures (e.g., BERT [19]) also effectively perform on vision-language tasks [20]. Also, Qi et al. [21] proposed causal intervention algorithms that can be applicable to other visual dialog models and Agarwal et al. [22] proposed curriculum fine-tuning inspired by the work of Bengio et al. [23].

The curriculum fine-tuning approach is adopted in several existing works [17,18,22,24] to boost the performance of a specific metric, Normalized Discounted Cumulative Gain (NDCG), but this significantly deteriorates the rest of the metrics, such as Mean Reciprocal Rank (MRR), recall, and mean rank. Many researches have struggled to improve performance only in NDCG by fine-tuning the model on dense annotations since the previous visual dialog challenge (https://visualdialog.org/challenge/2018 (accessed on 26 March 2021)) have selected the challenge winners by the ranking of NDCG performance. However, most recent challenge (https://visualdialog.org/challenge/2020 (accessed on 26 March 2021)) picked winners based on both the NDCG and MRR evaluation metrics so that we focus on building a model that boosts the performance with respect to all evaluation metrics in this work.

3. Model

In the visual dialog task [7], a dialog agent is given a set of multimodal inputs for each dialog turn $t$. This input set consists of an image $I$, a current question $Q_t$, the dialog history set $H_t = \{C, (Q_1, A_1^t), \ldots, (Q_{t-1}, A_{t-1}^t)\}$, which contains image caption $C$ and $t-1$ consecutive question-answer pairs, and a set of answer candidates $A_t = \{A_1^t, A_2^t, \ldots, A_{100}^t\}$. The agent then is required to answer the question by either discriminating or generating a correct answer.

3.1. Multimodal Representation

3.1.1. Visual Features

We employ a bottom-up attention mechanism [25] to represent the objects appearing in an image. Visual features of object regions $\{v_k\}_{k=1}^{n_o} \in \mathbb{R}^{d_v \times n_o}$, where $n_o$ is the number of detected objects ranging from 10 to 100, are adaptively extracted from a Faster-RCNN [26] that is pre-trained with Visual Genome [27].
3.1.2. Language Features

We first embed three different text inputs: The current question, dialog history, and answer candidates. The word embedding layer is initialized with pre-trained GloVe embeddings [28]. We then feed the word embeddings into a Bi-Directional Long Short-Term Memory (BiLSTM) to encode a sequential representation of each embedding. Specifically, each word in question $Q$ is embedded as $\{w_{q}^{i}\}_{i=1}^{n_{q}} \in \mathbb{R}^{d_{w} \times n_{q}}$, where $n_{q}$ is the number of words in the question. Each word embedding is fed into the BiLSTM layer as follows:

$$\overrightarrow{u_{q}^{i}} = \text{LSTM}(w_{q}^{i}, \overrightarrow{u_{q}^{i-1}})$$  \hspace{1cm} (1a)

$$\overleftarrow{u_{q}^{i}} = \text{LSTM}(w_{q}^{i}, \overleftarrow{u_{q}^{i+1}})$$  \hspace{1cm} (1b)

The sequential representation of each token is constructed by concatenating the hidden states of the forward and backward LSTMs, denoted as $u_{q}^{i} = [\overrightarrow{u_{q}^{i}}, \overleftarrow{u_{q}^{i}}]$. Meanwhile, the sequential representation for each dialog history $u_{hr}^{i} = \{u_{hr}^{i}j\}_{j=1}^{n_{hr}} (0 \leq r \leq t - 1)$ is constructed using the question construction process with different BiLSTM layers. The maximum sequence lengths of the question and dialog history are set to $n_{q} = 20$ and $n_{hr} = 40$, respectively. For the answer candidates, we use a different uni-directional LSTM to represent them because their sequence lengths are shorter than those of the questions.

3.2. Multi-View Attention Network

We propose MVAN, which considers the semantic intent and topic of the question simultaneously and effectively aligns the textual and visual information through multiple alignment processes. Figure 2 describes the MVAN architecture, which consists of three components: (1) Context Matching, (2) Topic Aggregation, and (3) Modality Alignment.

3.2.1. Context Matching Module

Generally, the semantic intent of a sentence not only relies on particular words that implicitly point to the topic of the sentence but also tends to be determined by the context of the entire sequence. Therefore, we build the Context Matching module that adaptively integrates the question and its relevant history at sentence level. Contextual representation is constructed by concatenating the last hidden states of the forward and backward LSTMs.
for the question and dialog history, denoted as $s^q = \{\tilde{w}_1^q, \tilde{w}_n^q\}$ and $s^h = \{\tilde{w}_1^h, \tilde{w}_{rn}^h\}$, respectively. We then apply an attention mechanism to focus on question-relevant history. The Context Matching module takes contextual representation of the question $s^q \in \mathbb{R}^{d_q \times 1}$ and dialog history $s^h = \{s_0^h, s_1^h, ..., s_{t-1}^h\} \in \mathbb{R}^{d_h \times t}$ and outputs question-relevant history features as follows:

$$
\begin{align}
zh^S &= W^T (f^S_q(s^q) \circ f^S_h(s^h)) + b \\
a_r^S &= \text{softmax}(zh^S) \\
s^h &= \sum_{t=0}^{t-1} a_r^S s^h_t,
\end{align}
$$

where $\circ$ is element-wise multiplication, $W \in \mathbb{R}^{d_q \times 1}$ is a projection matrix, and $f^S_q(\cdot)$ and $f^S_h(\cdot)$ denote non-linear transformation functions which convert $d_q$ to $d_f$ dimensions. We apply a gate function to automatically filter out dialog history that is irrelevant to current question as follows:

$$
\begin{align}
gate^c &= \sigma(W^c \cdot [s^q, s^h]) + b^c \\
e^c &= \gate^c \circ [s^q, s^h],
\end{align}
$$

where $\sigma(\cdot)$ is a sigmoid function and $W^c, b^c$ are trainable parameters. Note that $e^c \in \mathbb{R}^{d_f \times 1}$ is a context-matching representation that selectively combines the contextual information of the question and question-relevant dialog history.

### 3.2.2. Topic Aggregation Module

The topic of the question is generally expressed in a single word or phrase and likely to be connected with clues (i.e., the topic of previous questions in the dialog history). We design the Topic Aggregation module to combine the clues associated with the question topic by exploiting the initial word embeddings (i.e., GloVe) to represent their original meaning. Specifically, this module leverages word-level sequential representation of the question and dialog history, $\{u^q_i\}_{i=1}^{n_q} \in \mathbb{R}^{d_u \times n_q}$ and $\{u^h_{r,j}\}_{j=1}^{m_h} \in \mathbb{R}^{d_u \times m_h}$, respectively. The dot product attention mechanism is employed to selectively focus words that are relevant to the question topic from the dialog history as follows:

$$
\begin{align}
zh^W &= f^W_q(u^q_i) \top f^W_h(u^h_{r,j}) \\
a_{r,j}^W &= \exp(z_{r,j}^W) / \sum_{j=1}^{m_h} \exp(z_{r,j}^W) \\
\tilde{a}_r^W &= \sum_{j=1}^{m_h} a_{r,j}^W u^h_{r,j},
\end{align}
$$

where $f^W_q(\cdot)$ and $f^W_h(\cdot)$ are non-linear transformation functions which convert $d_u$ to $d_f$ dimensions. The question-guided history feature for each round $\tilde{a}_r^W$ is computed by a weighted sum of their word embeddings, which represent the original meanings of words. The attended representation $\{\tilde{u}_{r,j}^h\}_{i=1}^{n_h} \in \mathbb{R}^{d_u \times n_h}$ is computed by aggregating overall history $\{\tilde{a}_r^W\}_{r=0}^{t-1}$, weighted by the attention scores of the Context Matching module $a_r^S$ as follows:

$$
\tilde{u}_r^h = \sum_{r=0}^{t-1} a_r^S \tilde{a}_r^W.
$$

Similar to the Context Matching module, the gate operation adaptively filters out irrelevant topic-guided clues at word level.

$$
\begin{align}
gate^T &= \sigma(W^T_{gate}[\tilde{u}_r^h, \tilde{a}_r^W]) + b^T_{gate} \\
e^T &= \gate^c \circ [\tilde{u}_r^h, \tilde{a}_r^W],
\end{align}
$$
where $W_{gate}^T \in \mathbb{R}^{d_w \times 2d_w}$ and $b_{gate}^T \in \mathbb{R}^{d_w \times 1}$ are trainable parameters. Note that $\{e_i^T \}_{i=1}^{n_q} \in \mathbb{R}^{d_w \times n_q}$ is topic-aggregation representations that encode adaptive information of the question topic and various history clues associated with it.

### 3.2.3. Modality Alignment Module

Given the output representations of the Context Matching module $e^C \in \mathbb{R}^{d_f \times 1}$ and Topic Aggregation module $\{e_i^T \}_{i=1}^{n_q} \in \mathbb{R}^{d_w \times n_q}$, the Modality Alignment module aligns them with visual features $\{v_k \}_{k=1}^{n_v} \in \mathbb{R}^{d_v \times n_v}$ via two-step alignments. First, topic-view alignment performs soft alignment between heterogeneous modalities at topic level before mapping high-level contextual representation with the visual features. We utilize dot-product attention to represent the visual relevant topic-aggregation embeddings as follows:

$$z_{ik}^T = f_{l}^{VW}(e_i^T) f_{l}^{V}(v_k)$$  \hspace{1cm} (7a)$$

$$a_{ik}^T = \exp(z_{ik}^T) / \sum_{i=1}^{n_q} \exp(z_{ik}^T)$$  \hspace{1cm} (7b)$$

$$\hat{e}_k^T = \sum_{i=1}^{n_q} a_{ik}^T e_i^T,$$  \hspace{1cm} (7c)

where $f_{l}^{VW}()$ and $f_{l}^{V}()$ are non-linear transformation functions to embed two different modality representations into the same $d_f$ dimensions. Then, we obtain the fused feature vectors by concatenating the visual and attended topic-aggregation features and using a Multi-Layer Perceptron (MLP):

$$m_k^T = \text{MLP}([v_k, \hat{e}_k^T]).$$  \hspace{1cm} (8)

Note that $\{m_k^T \}_{k=1}^{n_v} \in \mathbb{R}^{d_v \times n_v}$ is a topic-view aligned representation that is aligned information across the visual contents with the salient question topics. This is passed to the context-view aligning step as follows:

$$\hat{m}_k^C = [m_k^T, e^C]$$  \hspace{1cm} (9a)$$

$$z_k^C = W^T \text{L2Norm}(f_m^C(\hat{m}_k^C) \circ f_l^C(e^C))) + b$$  \hspace{1cm} (9b)$$

$$a_k^C = \text{softmax}(z_k^C)$$  \hspace{1cm} (9c)$$

$$\hat{m}_k^C = \sum_{k=1}^{n_v} a_k^C m_k^C,$$  \hspace{1cm} (9d)

where $f_m^C(\cdot)$ and $f_l^C(\cdot)$ are non-linear transformation functions which convert $\hat{m}_k^C$ and $e^C$ into the same $d_f$ dimensions and $W \in \mathbb{R}^{d_f \times 1}$ is a projection matrix. Note that $\hat{m}_k^C \in \mathbb{R}^{(d_v+2d_i) \times 1}$ is a context-view aligned representation, which is realigned using context-matching representation. These multiple alignment processes allow the model to understand the semantic intent of the question with complementary views, and effectively align the corresponding heterogeneous multimodal inputs. Finally, this enhanced feature is fed into a single-layer Feed-Forward Neural Network (FFNN) with a ReLU activation function.

$$m^{enc} = \max(0, W^T \hat{m}_k^C + b),$$  \hspace{1cm} (10)

where $W \in \mathbb{R}^{(d_v+2d_i) \times d_{enc}}$ and $b \in \mathbb{R}^{d_{enc} \times 1}$ are trainable parameters. Note that $m^{enc} \in \mathbb{R}^{d_{enc} \times 1}$ is the multi-view aligned representation, which is fed into either a discriminative or generative decoder.

### 3.3. Answer Decoder

While existing studies mainly utilize only a discriminative decoder to rank answer candidates, we adopt not only a discriminative decoder but also a generative decoder. In the case of a discriminative decoder, it ranks all candidate answers by calculating the similarity score (dot-product) between the output representation and the sentence representation of each candidate. However, the generative decoder only takes the ground truth answer, and then predicts the next token for each time step (i.e., it does not use negative responses...
during training). The discriminative decoder is first described in Section 3.3.1 and the generative decoder is described in Section 3.3.2.

### 3.3.1. Discriminative Decoder

We use the last hidden states of the forward LSTM to encode sentence representations of answer candidates, denoted as \( s^a = \{ s^a_i \}_{i=1}^{100} \in \mathbb{R}^{d_a \times 100} \). We rank them according to the dot products of the candidates \( s^a \) and multi-view aligned representation \( m^{enc} \) (i.e., output representation of the encoder), then apply the softmax function to obtain the probability distribution of the candidates, denoted as \( p = \text{softmax}(s^a m^{enc}) \). Note that dimension of each answer candidate representation is the same as that of the encoder output. We use cross entropy loss as the discriminative objective function, formulated as:

\[
L_{\text{Disc}} = -\sum_{i=1}^{100} y_i \log p_i, \quad (11)
\]

where \( y_i \) is a one-hot encoded vector of the ground truth answer.

### 3.3.2. Generative Decoder

Unlike most previous approaches, which take only a discriminative approach, we also train our model in a generative manner [7]. During the training phase, we use a two-layer LSTM to predict the next token given the previous tokens in the ground truth answer. Specifically, the hidden state \( u^a_{gt,k} \) at \( k \)-th time step is computed given current word embedding \( w^a_k \) and previous hidden state \( u^a_{gt,k-1} \) as follows:

\[
u^a_{gt,k} = \text{LSTM}(w^a_k, u^a_{gt,k-1}). \quad (12)
\]

The initial hidden state of the LSTM \( u^a_{gt,0} \) is initialized with \( m^{enc} \) (i.e., output representation of the encoder). When generating each token, we compute the likelihood of the ground truth of each token \( \{p_k\}_{k=1}^{n_a} \) as follows:

\[
p_k = \text{softmax}_j(W^a_k u^a_{gt,k} + b_k), \quad (13)
\]

where \( W^a_k \in \mathbb{R}^{d_a \times |V|} \) is a projection matrix and \( |V| \) is the size of vocabulary. Note that \( j \) is the index of \( k \)-th word in the vocabulary and \( p_k \in \mathbb{R}^{|V|} \) is the probability score of \( k \)-th ground truth token. In the training phase, the model is optimized by minimizing the summation of the negative log-likelihood as follows:

\[
L_{\text{Gen}} = -\sum_{k=1}^{n_a} \log p_k. \quad (14)
\]

### 3.3.3. Multi-Task Learning

We perform multi-task learning by combining both discriminative and generative decoders to classify answers, denoted as \( L_{\text{Multi}} = L_{\text{Disc}} + L_{\text{Gen}} \). For the evaluation, we simply average the probability distributions of each decoder. Multi-task learning substantially improves performance with respect to the NDCG metric.

### 3.3.4. Fine-Tuning on Dense Annotations

In the discriminative setting, MVAN mainly focuses on finding the one-hot ground truth answer. However, many candidates may be semantically relevant to the ground truth so that they also have to be predicted with a high ranking. For example, more than one answer to Q6 in Figure 1 exists in the candidate pool (“no”, “no,”, “nope”, “i don’t think so”, and “not that i can see”). To evaluate how well the model is capable of generalizing, visual dialog challenge (https://visualdialog.org/challenge/2018 (accessed on 26 March 2021)) organizers released relevance scores (i.e., dense annotations) for each candidate answer, denoted as \( \{d_i\}_{i=1}^{100} \) where \( d_i \in [0, 1] \). Once MVAN is trained based on a discriminative decoder, we fine-tune the model on dense annotations (we utilize both training and validation set for
fine-tuning the model since only 2k cases are provided for each dataset) to predict more semantically relevant answers with higher rankings. Most existing works [17,18,22,24] have shown fine-tuning on dense annotations significantly boosts performance in NDCG but simultaneously deteriorates performance in non-NDCG metrics. To alleviate the trade-off between NDCG and non-NDCG metrics during the fine-tuning period, we design joint loss function by combining relevance loss and discriminative loss as follows:

\[ L_{Rel} = -\sum_{i=1}^{100} d_i \log p_i, \quad (15a) \]

\[ L_{Dense} = \alpha L_{Disc} + L_{Rel}. \quad (15b) \]

In order to examine the effect of the discriminative loss, we conduct various experiments by increasing the \( \alpha \) from 0 to 2 in 0.2 increments (see Figure 3). We decide the final \( \alpha \) as 1 to obtain the most balanced performance.

![Figure 3](image_url). Performance of Multi-View Attention Network (MVAN) according to the ratio of relevance loss and discriminative loss on the VisDial v1.0 validation set.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets

We used the VisDial v1.0 dataset (https://visualdialog.org/data (accessed on 26 March 2021)) [7] to evaluate our proposed model. VisDial v1.0 training splits consist of 123k MS-COCO [29] images and an additional 10k images from the Flickr dataset were utilized to construct the validation and test splits, which contain 2k and 8k images, respectively. Each image has 10 consecutive question-answer pairs. Each question is paired with a group of 100 correct answers, one of which is ground true. Additionally, VisDial v1.0 provides dense annotations for 100 answer candidates corresponding to each question on subsets of the training and validation splits.

4.1.2. Evaluation Metrics

Following the work of Das et al. [7], we evaluated our proposed model using several retrieval metrics: (1) Mean rank of the ground truth response (Mean), (2) recall at k \((k = \{1, 5, 10\})\), which is denoted as R@k and evaluates where the ground truth is positioned in the sorted list, and (3) MRR [30]. NDCG was also introduced as a primary metric in the VisDial v1.0 dataset, and decreases when the model gives a low ranking to candidate answers with high relevance scores. MRR evaluates the precision of the model by ranking where a ground truth answer is positioned, whereas NDCG evaluates relative relevance of the predicted answers. In this work, we compare our model with the baselines mainly based on both NDCG and MRR metrics following the most recent visual dialog challenge (the VisDial challenge (2018~2019) selects the winners based solely on NDCG metric, but the winners are picked based on the average rank of NDCG and MRR from the 2020 challenge). Specifically, we rank all the baselines including ours on each metric individually (i.e., NDCG and MRR), then average ranks across two metrics (denoted as AVG in the following results).
4.1.3. Training Details

Our model is implemented using the PyTorch framework [31] based on open source code (https://github.com/batra-mlp-lab/visdial-challenge-starter-pytorch (accessed on 26 March 2021)) from the work of Das et al. [7]. The question and dialog history are represented using different BiLSTMs with 512 hidden states. The maximum sequence lengths of the question and dialog history are set to 20 and 40, respectively. We set the batch size to 32 and apply the Adam optimizer [32] with an initial learning rate of 0.00001, which is gradually increased to 0.001 until epoch 2, then decay at epochs 6 and 7 with the decay rate 0.1. Our code is publicly available (https://github.com/taesunwhang/MVAN-VisDial (accessed on 26 March 2021)).

4.2. Quantitative Results

4.2.1. Discriminative Setting

We compare the results of our proposed model with previously published results on the VisDial v1.0 dataset for the following methods: Late Fusion (LF) [7], Hierarchical Recurrent Encoder (HRE) [7], Memory Network (MN) [7], Graph Neural Network (GNN) [12], Co-reference Neural Module Network (CorefNMN) [4], Recursive Visual Attention (RVA) [6], Synergistic Network [9], Dual Encoding Visual Dialogue (DualVD) [33], Context-Aware Graph (CAG) [14], History-Aware Co-Attention (HACAN) [34], Consensus Dropout Fusion (CDF) [11], Dual Attention Network (DAN) [5], and Factor Graph Attention (FGA) [13].

Table 1 reports the quantitative results on the VisDial v1.0 under the discriminative decoder setting. For VisDial v1.0, our MVAN model outperforms the previous state-of-the-art methods with respect to NDCG and AVG and shows competitive results in non-NDCG metrics. Specifically, MVAN achieves significant improvements in NDCG from 57.59 to 59.37 and is ranked first as a result of the average rank of NDCG and MRR, compared to the state-of-the-art baseline. We also report the results for an ensemble of 10 independent models that were trained with random initial seeds. This leads to further performance improvements in NDCG from 59.37 to 60.92 and MRR from 64.84 to 66.38. These results indicate that MVAN not only has an accurate prediction ability, as indicated by the non-NDCG metric results (i.e., MRR, R@k, and Mean), but it has a powerful generalization capability given the result of NDCG score because this metric considers several relevant answers to be correct.

Table 1. Results on VisDial v1.0 (test-std). † denotes ensembles. Existing works [17,18] based on pre-trained models are not reported for the fair comparison. The numbers in the brackets indicate rankings of the models in Normalized Discounted Cumulative Gain (NDCG) and Mean Reciprocal Rank (MRR) metrics.

<table>
<thead>
<tr>
<th>Model</th>
<th>AVG ↓</th>
<th>NDCG ↑</th>
<th>MRR ↑</th>
<th>R@1 ↑</th>
<th>R@5 ↑</th>
<th>R@10 ↑</th>
<th>Mean ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>LF [7]</td>
<td>12</td>
<td>45.31</td>
<td>55.42</td>
<td>40.95</td>
<td>72.45</td>
<td>82.83</td>
<td>5.95</td>
</tr>
<tr>
<td>HRE [7]</td>
<td>12</td>
<td>45.46</td>
<td>54.16</td>
<td>39.93</td>
<td>70.45</td>
<td>81.50</td>
<td>6.41</td>
</tr>
<tr>
<td>MN [7]</td>
<td>11</td>
<td>47.50</td>
<td>55.49</td>
<td>40.98</td>
<td>72.30</td>
<td>83.30</td>
<td>5.92</td>
</tr>
<tr>
<td>GNN [12]</td>
<td>10</td>
<td>52.82</td>
<td>61.30</td>
<td>47.33</td>
<td>77.98</td>
<td>87.83</td>
<td>4.57</td>
</tr>
<tr>
<td>CorefNMN [4]</td>
<td>9</td>
<td>54.70</td>
<td>61.50</td>
<td>47.55</td>
<td>78.10</td>
<td>88.80</td>
<td>4.40</td>
</tr>
<tr>
<td>RVA [6]</td>
<td>8</td>
<td>55.59</td>
<td>63.03</td>
<td>49.03</td>
<td>80.40</td>
<td>89.83</td>
<td>4.18</td>
</tr>
<tr>
<td>DualVD [33]</td>
<td>7</td>
<td>56.32</td>
<td>63.23</td>
<td>49.25</td>
<td>80.23</td>
<td>89.70</td>
<td>4.11</td>
</tr>
<tr>
<td>Synergistic [9]</td>
<td>6</td>
<td>57.32</td>
<td>62.20</td>
<td>47.90</td>
<td>80.43</td>
<td>89.95</td>
<td>4.17</td>
</tr>
<tr>
<td>CAG [14]</td>
<td>5</td>
<td>56.64</td>
<td>63.49</td>
<td>49.85</td>
<td>80.63</td>
<td>90.15</td>
<td>4.11</td>
</tr>
<tr>
<td>DAN [5]</td>
<td>4</td>
<td>57.59</td>
<td>63.20</td>
<td>49.63</td>
<td>79.75</td>
<td>89.35</td>
<td>4.30</td>
</tr>
<tr>
<td>HACAN [34]</td>
<td>3</td>
<td>57.17</td>
<td>64.22</td>
<td>50.88</td>
<td>80.63</td>
<td>89.45</td>
<td>4.20</td>
</tr>
<tr>
<td>FGA [13]</td>
<td>2</td>
<td>56.90</td>
<td>66.20</td>
<td>52.75</td>
<td>82.92</td>
<td>91.07</td>
<td>3.80</td>
</tr>
<tr>
<td>MVAN (ours)</td>
<td>1</td>
<td>59.37</td>
<td>64.84</td>
<td>51.45</td>
<td>81.12</td>
<td>90.65</td>
<td>3.97</td>
</tr>
<tr>
<td>Synergistic †[9]</td>
<td>5</td>
<td>57.88</td>
<td>63.42</td>
<td>49.30</td>
<td>80.77</td>
<td>90.68</td>
<td>3.97</td>
</tr>
<tr>
<td>CDF †[11]</td>
<td>2</td>
<td>59.49</td>
<td>64.40</td>
<td>50.90</td>
<td>81.18</td>
<td>90.40</td>
<td>3.99</td>
</tr>
<tr>
<td>DAN †[5]</td>
<td>2</td>
<td>59.36</td>
<td>64.92</td>
<td>51.28</td>
<td>81.60</td>
<td>90.88</td>
<td>3.92</td>
</tr>
<tr>
<td>FGA †[13]</td>
<td>2</td>
<td>57.20</td>
<td>69.30</td>
<td>55.65</td>
<td>86.73</td>
<td>94.05</td>
<td>3.14</td>
</tr>
<tr>
<td>MVAN †(ours)</td>
<td>1</td>
<td>60.92</td>
<td>66.38</td>
<td>53.20</td>
<td>82.45</td>
<td>91.85</td>
<td>3.68</td>
</tr>
</tbody>
</table>
4.2.2. Multi-Task Learning

As shown in Table 2, we report the results of our MVAN model, which is trained using multi-task learning. Our proposed approach performs better with respect to all metrics than Recurrent Dual Attention Network (ReDAN) [10], which averages the ranking results of the discriminative and generative model, and Light-weight Transformer for Many Inputs (LTMI) [24], which employs multi-task learning but uses only discriminative decoder outputs for evaluation.

<table>
<thead>
<tr>
<th>Model</th>
<th>AVG</th>
<th>NDCG</th>
<th>MRR</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReDAN [10]</td>
<td>2</td>
<td>61.86 (2)</td>
<td>53.13 (3)</td>
<td>41.38</td>
<td>66.07</td>
<td>74.50</td>
<td>8.91</td>
</tr>
<tr>
<td>LTMI * [24]</td>
<td>2</td>
<td>60.92 (3)</td>
<td>60.65 (2)</td>
<td>47.00</td>
<td>77.03</td>
<td>87.75</td>
<td>4.90</td>
</tr>
<tr>
<td>MVAN * (ours)</td>
<td>1</td>
<td><strong>63.15</strong> (1)</td>
<td><strong>63.02</strong> (1)</td>
<td><strong>49.43</strong></td>
<td><strong>79.48</strong></td>
<td><strong>89.40</strong></td>
<td><strong>4.38</strong></td>
</tr>
<tr>
<td>ReDAN+ † [10]</td>
<td>3</td>
<td>64.47 (2)</td>
<td>53.74 (3)</td>
<td>42.45</td>
<td>64.68</td>
<td>75.68</td>
<td>6.64</td>
</tr>
<tr>
<td>LTMI †,† [24]</td>
<td>1</td>
<td><strong>66.53</strong> (1)</td>
<td>63.19 (2)</td>
<td>49.18</td>
<td>80.45</td>
<td>89.75</td>
<td>4.14</td>
</tr>
<tr>
<td>MVAN †,† (ours)</td>
<td>2</td>
<td>63.22 (3)</td>
<td><strong>66.28</strong> (1)</td>
<td><strong>53.87</strong></td>
<td><strong>82.08</strong></td>
<td><strong>89.65</strong></td>
<td><strong>4.61</strong></td>
</tr>
</tbody>
</table>

4.2.3. Fine-Tuning on Dense Annotations

Table 3 reports the results of fine-tuning on dense annotations. This fine-tuning on relevance scores for each candidate answer significantly boosts the performance of NDCG. However, existing works [17,18,21,22,24] show that the performance of non-NDCG metrics deteriorates in contrast to the case of NDCG. To alleviate the inconsistency between NDCG and non-NDCG metrics, we fine-tune the model with the joint loss (see Equation (15)). Figure 3 shows the performance change of MRR and NDCG by varying $\alpha$, which is the weight of the discriminative loss, from 0 to 2. As $\alpha$ increases from 0 to 1, MRR is rapidly improved while NDCG remained high. On the contrary, NDCG sharply decreases with small improvement in MRR when $\alpha$ increases from 1 to 2. Based on this observation, we set $\alpha$ to 1 and obtain significant improvements across two metrics. In the baseline comparisons, our MVAN model is ranked first in MRR regardless of whether it is a single or ensemble model (single: 56.06, ensemble: 65.16). In addition, MVAN is ranked first and second in AVG for single and ensemble model comparison, respectively, despite a little lower NDCG.

<table>
<thead>
<tr>
<th>Model</th>
<th>AVG</th>
<th>NDCG</th>
<th>MRR</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCA [22]</td>
<td>4</td>
<td>72.47 (4)</td>
<td>37.68 (4)</td>
<td>20.67</td>
<td>56.67</td>
<td>72.12</td>
<td>8.89</td>
</tr>
<tr>
<td>Visdial-BERT [17]</td>
<td>1</td>
<td>74.47 (2)</td>
<td>50.74 (2)</td>
<td>37.95</td>
<td>64.13</td>
<td>80.00</td>
<td>6.28</td>
</tr>
<tr>
<td>VD-BERT [18]</td>
<td>1</td>
<td><strong>74.54</strong> (1)</td>
<td>46.72 (3)</td>
<td>33.15</td>
<td>61.58</td>
<td>77.15</td>
<td>7.18</td>
</tr>
<tr>
<td>MVAN (ours)</td>
<td>1</td>
<td>73.07 (3)</td>
<td><strong>56.06</strong> (1)</td>
<td><strong>44.38</strong></td>
<td><strong>68.50</strong></td>
<td><strong>81.18</strong></td>
<td><strong>5.98</strong></td>
</tr>
<tr>
<td>P1_P2 † [21]</td>
<td>4</td>
<td>74.91 (2)</td>
<td>49.13 (5)</td>
<td>36.68</td>
<td>62.96</td>
<td>78.55</td>
<td>7.03</td>
</tr>
<tr>
<td>LTMI † [24]</td>
<td>4</td>
<td>74.88 (3)</td>
<td>52.14 (4)</td>
<td>38.93</td>
<td>66.60</td>
<td>80.65</td>
<td>6.63</td>
</tr>
<tr>
<td>MReal-BDAI † [21]</td>
<td>2</td>
<td>74.02 (4)</td>
<td>52.62 (2)</td>
<td>40.03</td>
<td>68.85</td>
<td>79.15</td>
<td>6.76</td>
</tr>
<tr>
<td>VD-BERT † [18]</td>
<td>1</td>
<td><strong>75.35</strong> (1)</td>
<td>51.17 (4)</td>
<td>38.90</td>
<td>62.82</td>
<td>77.98</td>
<td>6.69</td>
</tr>
<tr>
<td>MVAN † (ours)</td>
<td>2</td>
<td>71.40 (5)</td>
<td><strong>65.16</strong> (1)</td>
<td><strong>52.88</strong></td>
<td><strong>79.50</strong></td>
<td><strong>88.55</strong></td>
<td><strong>4.27</strong></td>
</tr>
</tbody>
</table>

4.2.4. Number of Dialog History

We experimented with the amount of dialog history to evaluate the impact of dialog history on the model performance of the two major metrics (i.e., MRR and NDCG). The results in Figure 4 are the average of the MRR and NDCG performance for each dialog turn in the validation set and show that as the amount of dialog history information increases, the MRR tends to gradually improve, but the NDCG score deteriorates. This quantity...
analysis shows that history information decreases the NDCG score but substantially boosts the other metrics.

![Figure 4. Performance of MVAN with different amounts of dialog history on the VisDial v1.0 validation set.](image)

4.3. Ablation Study

We conducted ablation studies on the VisDial v1.0 validation splits to evaluate the influence of each component in our model. Modality Alignment module is not ablated because this module handles the visual features. We use the same discriminative decoder model for all ablations to exclude the impact of multitask learning.

In Table 4, the first rows of each block indicate the impact of each module in our model. Since the two modules (i.e., Context Matching and Topic Aggregation) are interdependent, we employ simple visual features instead of topic-aggregation representation for MVAN w/o Topic Aggregation, whereas we simply remove context-matching representation for MVAN w/o Context Matching. Both models obtain slightly lower performance with respect to all evaluation metrics than MVAN. We can hence infer that the two modules are complementary with respect to each other and our model integrates these complementary characteristics well for the task.

<table>
<thead>
<tr>
<th>Model</th>
<th>Context-Level History</th>
<th>Topic-Level History</th>
<th>NDCG ↑</th>
<th>MRR ↑</th>
<th>R@1 ↑</th>
<th>R@5 ↑</th>
<th>R@10 ↑</th>
<th>Mean ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVAN</td>
<td>✓</td>
<td>✓</td>
<td>60.17</td>
<td>65.33</td>
<td>51.86</td>
<td>82.40</td>
<td>90.90</td>
<td>3.88</td>
</tr>
<tr>
<td>w/o Topic Aggregation</td>
<td>✓</td>
<td>N/A</td>
<td>58.50</td>
<td>64.63</td>
<td>50.84</td>
<td>81.64</td>
<td>90.50</td>
<td>3.97</td>
</tr>
<tr>
<td>w/o Context Matching</td>
<td>N/A</td>
<td>✓</td>
<td>57.06</td>
<td>64.15</td>
<td>50.51</td>
<td>81.15</td>
<td>89.83</td>
<td>4.12</td>
</tr>
</tbody>
</table>

Recent approaches [11,17,24] reported that they observed a trade-off relationship between two primary metrics (i.e., NDCG and MRR) in the visual dialog task. We also found the trade-off relationship through ablative experiments with and without dialog history features (see Table 4). Specifically, adding dialog history features improves the MRR score by 3.54% on average, whereas NDCG score is decreased by 1.92% on average. We observe that the model has a tendency to predict the answers more precisely (i.e., it has a better MRR score) when the dialog history features are added. This may imply that question-related clues in the dialog history are important factors in reasoning the ground truth, but they hinder the model’s generalization ability (i.e., they lower the NDCG score).

4.4. Qualitative Analysis

To qualitatively demonstrate the advantages of our model, we visualize the attention scores of each module through examples from the VisDial v1.0 validation set in Figure 5.
The attention scores of the Context Matching module, highlighted in blue, show that our model selectively focuses on contextual information as the semantic intent of the question changes. The tendency for the caption (i.e., H0) to receive the highest attention score implies that the caption contains global information describing the image. In addition, the top three visual contents with high attention scores in each image lead to the potential interpretation that our model is capable of explicitly aligning the semantic intent (i.e., highlighted in yellow) of the question and visual contents through Modality Alignment module. In more detail, the attention scores of the dialog history, highlighted in red, indicate how our model captures topic-relevant clues through previous dialog history.

As shown in Figure 5a, comparing two examples, we see that the model no longer focuses on “6” and “people” in H0 because those words are not related to the topic of the current question (i.e., “drinking”). In the example in Figure 5b, when answering Q4 (the left dialog), the model pays more attention to the question-relevant clue such as “women” in H0, while no longer focusing on it when answering Q6 (the question topic changes from “tennis outfits” to “background”). These qualitative results show that our model successfully pays attention to visual and textual information connected to the semantic intent of the question.

4.5. Error Analysis

We analyzed examples in the VisDial v1.0 validation set for which our model obtained a score of 0 for the R@10 metric. The errors can be categorized into three groups: (1) Subjective judgment: Our model tends to make wrong predictions for the questions about age, weather, and appearance that could involve subjective judgment, but might be acceptable (Figure 6a,b). (2) Ambiguous questions: Our model may focus on the wrong visual contents, for instance the left and right side walls rather than the rear wall when faced with an ambiguous question (Figure 6c). (3) Wrong multimodal alignment: When the dialog history includes multiple entities (e.g., “boys”, “pizzas”, and “toppings”) that can be referenced by a single pronoun (i.e., “them”), MVAN may be confused as to which entity the pronoun refers (Figure 6d).
Figure 6. Error analysis on the VisDial v1.0 validation set. We analyze the examples for which the model scored 0 with respect to the R@10 metric. The errors are categorized into (1) subjective judgment, (2) ambiguous questions, and (3) wrong multimodal alignment. (a,b) belong to the first, (c) to the second, and (d) to the third, respectively.

5. Conclusions

In this paper, we introduced MVAN for the visual dialog task. MVAN could effectively determine the semantic intent of the current question and capture question-relevant information through complementary modules and sequential alignment processes. We used VisDial v1.0 to empirically evaluate our model, and as a result, our model outperformed existing state-of-the-art models. Moreover, we not only suggest plausible factors affecting a trade-off relationship of the evaluation metrics, but we enhanced the interpretability of multi-level attention through detailed visualization. As the main focus of our paper is on encoders, it limits in an in-depth exploration of decoders. In future, we plan to adopt the latest generative model for building decoder and incorporate vision-language pre-trained models [18,20] if the performance of MVAN can be further improved. In addition, we will study an efficient ranking method to improve performance both in NDCG and non-NDCG metrics.

Author Contributions: Conceptualization, S.P. and T.W.; data curation, S.P. and T.W.; formal analysis, T.W. and Y.Y.; funding acquisition/project administration/supervision, H.L.; investigation, S.P.; methodology/software, S.P. and T.W.; visualization, S.P.; review and editing, Y.Y.; original draft, S.P. and T.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-2018-0-01405) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), Institute for Information & communications Technology Planning & Evaluation (IITP), grant funded by the Korean government (MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques) and MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience program (IITP-2021-2020-0-01819) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The data can be found here: VisDial v1.0: https://visualdialog.org/data (accessed on 26 March 2021).

Conflicts of Interest: The authors declare no conflict of interest.
Appendix A. More Qualitative Results

Q4: are they wearing a **snowsuit**?
- **GT**: yes a tan snowsuit (1)
- **Prediction**: yes a tan snowsuit

Q6: is it **snowing** out?
- **GT**: no (1)
- **Prediction**: no

**H1**: is the **snowboarder** along? yes
**H2**: is it **daytime**? yes
**H3**: is the **border** wearing a **helmet**? no

**H0**: a **snowboarder** is traveling down a **snowy** hill

Q4: are they wearing a **snowsuit**?
- **GT**: yes a tan snowsuit (1)
- **Prediction**: yes a tan snowsuit

Q6: is it **snowing** out?
- **GT**: no (1)
- **Prediction**: no

**H1**: is the **snowboarder** along? yes
**H2**: is it **daytime**? yes
**H3**: is the **border** wearing a **helmet**? no

**H0**: a **snowboarder** is traveling down a **snowy** hill

**Fig. A1.** Visualization of the reasoning process for each module in MVAN.
Figure A2. Visualization of the reasoning process for each module in MVAN.

References


