SeqDialN: Sequential Visual Dialog Networks in Joint Visual-Linguistic Representation Space

Liu Yang[1], Fanqi Meng[2], Xiao Liu[3], Ming-Kuang Daniel Wu[4], Vicent Ying[4]
Xianchao Xu[1]
[1]Intel China Research Center
[2]University of Science and Technology of China
[3]University of California, Davis
[4]Stanford University
{liu.y.yang, james.xu}@intel.com, farrell@mail.ustc.edu.cn
xioliu@ucdavis.edu, danielwu@alumni.stanford.edu
vhying@stanford.edu

Abstract

The key challenge of the visual dialog task is how to fuse features from multimodal sources and extract relevant information from dialog history to answer the current query. In this work, we formulate a visual dialog as an information flow in which each piece of information is encoded with the joint visual-linguistic representation of a single dialog round. Based on this formulation, we consider the visual dialog task as a sequence problem consisting of ordered visual-linguistic vectors. For featurization, we use a Dense Symmetric Co-Attention network (Nguyen and Okatani, 2018) as a lightweight vison-language joint representation generator to fuse multimodal features (i.e., image and text), yielding better computation and data efficiencies. For inference, we propose two Sequential Dialog Networks (SeqDialN): the first uses LSTM (Hochreiter and Schmidhuber, 1997) for information propagation (IP) and the second uses a modified Transformer (Vaswani et al., 2017) for multi-step reasoning (MR). Our architecture separates the complexity of multimodal feature fusion from that of inference, which allows simpler design of the inference engine. On VisDial v1.0 test-std dataset, our best single generative SeqDialN achieves 62.54% NDCG[1] and 48.63% MRR[2]; our ensemble generative SeqDialN achieves 63.78% NDCG and 49.98% MRR, which set a new state-of-the-art generative visual dialog model. We fine-tune discriminative SeqDialN with *dense annotations*[3] and boost the performance up to 72.41% NDCG and 55.11% MRR. In this work, we discuss the extensive experiments we have conducted to demonstrate the effectiveness of our model components. We also provide visualization for the reasoning process from the relevant conversation rounds and discuss our fine-tuning methods. The code is available at https://github.com/xiaoxiaoheimei/SeqDialN.

1 Introduction

Visual Dialog has attracted increasing research interest as an emerging field, bringing together aspects of computer vision, natural language processing, and dialog systems. In this task, an AI agent is required to hold a meaningful dialog with humans in natural, conversational language about visual content. Specifically, given an image, a dialog history, and a query about the image, the agent has to ground the query in image, infer context from history, and answer the query accurately (Das et al., 2017).

Our work is inspired by the use of visual-linguistic joint representation to erase the modality gap, where we embed the visual signals into the text snippets for each dialog round. In this way, we convert a visual dialog into an ordered vector sequence, where each vector is the joint visual-linguistic representation of a specific dialog round. Rather than using ViLBERT (Lu et al., 2019), we chose Dense Symmetric Co-Attention (Nguyen and Okatani, 2018) as a lightweight joint visual-linguistic representation generator. In contrast to VisDial-BERT (Murahari et al., 2019), which concatenates all rounds of the dialog history into a single textual input for ViLBERT(Lu et al., 2019), we keep each dialog round separate. Keeping this inherent sequential structure from the visual dialog allows us to reason across the dialog history to find the most query-relevant dialog rounds. By viewing visual dialog task as a vector sequence, We propose two sequential networks to tackle the problem.

Fig. 1 illustrates a conceptual overview

[1]Normalized Discounted Cumulative Gain
[2]Mean Reciprocal Rank
[3]Relevance scores for 100 answer options corresponding to each question on a subset of the training set, publicly available on visualdialog.org/data

Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 8–17

of the proposed method. The visual features and language embeddings are learned from two independent domains. They are fed into the Dense Symmetric Co-Attention Network (Nguyen and Okatani, 2018) to produce a **visual-linguistic vector sequence** in the joint visual-linguistic feature space. Our baseline model, the Information Propagation Network (SeqIPN), which uses a LSTM (Hochreiter and Schmidhuber, 1997) to summarize the visual-linguistic sequence, outperforms other well-known baselines (Das et al., 2017; Lu et al., 2017), on NDCG metric by a large margin > 0.5. Multi-step reasoning network (SeqMRN) is based on Transformer (Vaswani et al., 2017). We expect the multi-head attention mechanism of Transformer better captures the relationship within the visual linguistic sequence. We achieve multi-step reasoning by stacking several Transformers to refine attentions in high level semantic space. SeqMRN outperforms VisDial-BERT (Murahari et al., 2019) by $> 1.5\%$ on NDCG when trained with comparable amount of data, while using 30% less parameters. The pipeline in Fig.1 facilitates the combination of different word embeddings and SeqDialN models. In this work, we compare two kinds of pre-trained word representations: GloVe(Pennington et al., 2014) and DistilBert (Sanh et al., 2019). The ablation test shows that SeqMRN with DistilBert embedding yields the best performance. Further experiment reveals SeqDialN sets a new state-of-the-art **generative** visual dialog model.

VLDialog and NDCGFinetune(Murahari et al., 2019; Qi et al., 2019b) tune with *dense annotations*[3]. Training on the *dense annotation*[3] makes these models perform very well on the NDCG metric but poorly on the others because the *dense annotation*[3] dataset doesn't correlate well with the original ground-truth answer to the question (Murahari et al., 2019). In this work, we propose a reweighting method to mitigate the damage to non-NDCG metrics in fine-tuning process, which make our best model outperform (Murahari et al., 2019; Qi et al., 2019b,a) on MRR by a large margin at the cost of a little lower NDCG than them.

The main contributions of this paper is three fold. (1) We formulate the visual dialog task as reasoning from a sequence in the joint visual-linguistic representation space. (2) We propose two sequential networks to tackle the visual dialog task in the joint visual-linguistic representation space. (3) We set a new state-of-the-art **generative** visual dialog model.

2 Related Work

2.1 VQA

VQA focuses on providing a natural language answer given an image and a free-form, open-ended question. Attention mechanisms have been deeply explored in VQA related work. In deep networks, the attention mechanism helps refine semantic meanings at different levels. SANs (Yang et al., 2016) create stacked attention networks, producing multiple attention maps in a sequential manner to imitate multi-step reasoning. (Lu et al., 2016) introduces co-attention between image regions and words in the question. (Yu et al., 2017) utilizes image-guided attention to extract the language concept of an image and then combines this with a novel multi-modal feature fusion of image and question.

Recently, Dense Co-Attention Network (DCN) (Nguyen and Okatani, 2018) proposes a symmetric co-attention layer to address VQA tasks. DCN is "dense symmetric" because it makes each visual region aware of the existence of each question word and vice versa. This fine-granularity co-attention enables DCN to discriminate subtle differences or similarities between vision and language features. In this work, we use DCN as the generator of joint visual-linguistic representation.

2.2 Visual Dialog

Previous research has tackled the visual dialog task from various theoretical perspectives. Early baselines include Late Fusion, Hierarchical Recurrent Encoder, and Memory Networks (Das et al., 2017). (Guo et al., 2019) proposes a two-stage method which filters out the obviously irrelevant answers in primary stage, then re-ranks the rest answers in synergistic stage. (Guo et al., 2019) won the visual dialog challenge[4] in 2018. Several models try to leverage the dialog structure to conduct explicit reasoning. GNN (Zheng et al., 2019) abstracts visual dialog as a fully connected graph where each node represents a single dialog round and each edge represents semantic dependency of the two connected nodes. Recursive Visual Attention(RvA) (Niu et al., 2019) designs sub-networks to infer the stopping condition when

[4] visdial/challenge2020

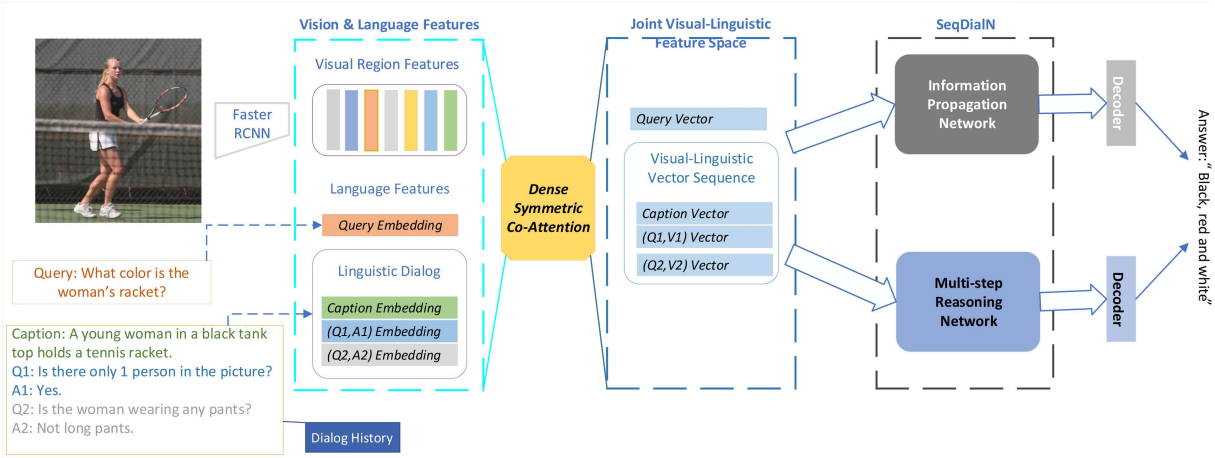

Figure 1: Conceptual architecture of sequential visual dialog network (SeqDialN).

recursively traversing the dialog stack to resolve visual co-reference relationships. RvA won the visual dialog challenge[4] in 2019 by fine-tuning with *dense annotations*[3]. ReDAN (Gan et al., 2019) develops a recurrent dual attention network to progressively update the semantic representations of query, vision, and history, making them co-aware through multiple steps to achieve multi-step reasoning. ReDAN (Gan et al., 2019) achieves 64.47% NDCG on the VisDial v1.0 test-std set, is still the highest score among all published work trained **without** *dense annotations*[3].

Based on ViLBERT (Lu et al., 2019), recent VisDial-BERT (Murahari et al., 2019) leverages the joint visual-linguistic representation to tackle visual dialog task. By fine-tuning with *dense annotations*, VisDial-BERT (Murahari et al., 2019) achieves state-of-the-art NDCG (74.47%) using a discriminative model. However, its non-NDCG performance is significantly lower. Futhermore, it's not easy to deploy a discriminative model in real applications. Similar performance degradation occurs to P1P2 (Qi et al., 2019a), which also trained with *dense annotations*[3].

3 Approach

The visual dialog task (Das et al., 2017) is formulated as follows: at time t, given a query Q_t grounded in image I, and dialog history (including the image caption C) $H_t = \{C, (Q_1, A_1), \cdots, (Q_{t-1}, A_{t-1})\}$ as additional context. For discriminative task, the goal is to rank 100 candidate answers $A_t = \{A_t^1, A_t^2, \cdots, A_t^{100}\}$. For generative task, the goal is to generate an answer in natural language. The task requires the agent to predict the ground truth answer and rank other feasible answers as high as possible.

As illustrated in Fig. 1, we rely on Faster-RCNN (Ren et al., 2015) to extract features corresponding to salient image regions (Anderson et al., 2018). The vision feature of image I is represented as $F_I \in R^{n_v \times d_v}$, where $n_v = 36$ being the number of object-like region proposals in the image and $d_v = 2048$ being the dimension of the feature vector. Q_t and each item in H is padded or truncated to the same length d_l. Thus, each sentence S is represented as $F_S \in R^{d_l \times d_e}$, where d_e being the dimension of the word embedding. To facilitate further discussion, we denote d_h as the dimension of the hidden state throughout this section.

3.1 Visual Dialog as Visual-Linguistic Vector Sequence

Dense Co-Attention Network (DCN) (Nguyen and Okatani, 2018) proposes using contents in sub-grids of a convolutional neuron network as visual region features. However, we turn to use Faster R-CNN proposals (Ren et al., 2015; Anderson et al., 2018) because people usually talk about objects in their conversations, so Faster R-CNN proposals better suit for the purpose of object identification. Given an image I with vision feature $F_I \in R^{n_v \times d_v}$ and a sentence S with embedding $F_S \in R^{d_l \times d_e}$, we define $DCN(I, S) \in R^{d_h}$ the Dense Co-attention (Nguyen and Okatani, 2018) representation of I and S. We define an instance of t round visual dialog by a tuple $D = (I, H_t, Q_t)$. Using DCN, we convert dialog history H_t into the visual-linguistic vector sequence $\widehat{H}_t$ as:

$$\begin{aligned} \widehat{C} &= DCN(I, C) \\ \widehat{L}_i &= DCN(I, (Q_i, A_i)), i = 1, \cdots, t-1 \\ \widehat{H}_t &= \{\widehat{C}, \widehat{L}_1, \cdots, \widehat{L}_{t-1}\} \end{aligned} \quad (1)$$

Let $\widehat{Q}_t = DCN(I, Q_t)$, the original visual dialog then turns into a new tuple $\widehat{D} = (\widehat{H}_t, \widehat{Q}_t)$ in the joint visual-linguistic representation space. Note that the sequential structure of $\widehat{H}_t$ is exactly the same as that of H_t and image I no longer exists in $\widehat{D}$ as an explicit domain.

To facilitate discussion in section 3.2, we define the question history $\mathcal{Q}_t$ by:

$$\begin{aligned} \widehat{Q}_i &= DCN(I, Q_i), 1 \le i \le t \\ \mathcal{Q}_t &= \{\widehat{Q}_1, \cdots, \widehat{Q}_{t-1}, \widehat{Q}_t\} \end{aligned} \tag{2}$$

Note, $\mathcal{Q}_t$ includes the visual-linguistic vector of the query Q_t.

3.2 SeqIPN: Information Propagation Network

As illustrated in Fig. 2, Information Propagation Network is a 2-layer LSTM. After converting the visual dialog into a tuple $\widehat{D} = (\widehat{H}_t, \widehat{Q}_t)$ in the joint visual-linguistic representation space, we apply a LSTM to the visual-linguistic vector sequence $\widehat{H}_t$ and use the hidden state at time t as the summary of visual-linguistic history. Specifically:

$$R_L = LSTM(\widehat{H}_t)[t],\ R_L \in R^{d_h} \tag{3}$$

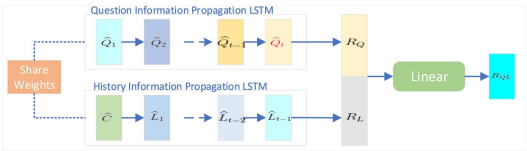

Figure 2: Architecture of Information Propagation Network (SeqIPN)

We apply the same LSTM to question history $\mathcal{Q}_t$ and use $\widehat{Q}_t$'s hidden state R_Q as the context aware query. Experiment shows introducing R_Q can slightly drop the MRR ($< 1\%$) but increase NDCG a lot ($> 1.5\%$). The observation can be explained as R_Q is the query distorted by LSTM, which fools the discriminator and results in the MRR drop. However, the impact is controllable because LSTM's forget gate makes the impact of previous questions gradually fade away along the propagation. On the other hand, R_Q collects more semantic information to broaden the scope of candidate answers, which results in the NDCG increase.

$[R_L, R_Q] \in R^{2d_h}$ is linearly projected to $R_{QL} \in R^{d_h}$ as the final representation of $\widehat{D}$. R_{QL} is fed into the decoder to predict answer.

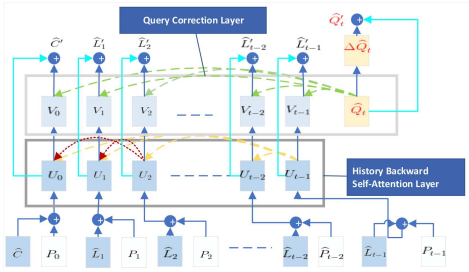

Figure 3: Conceptual architecture of Multistep Reasoning Network (SeqMRN).

3.3 SeqMRN: Multi-step Reasoning Network

Transformer (Vaswani et al., 2017) was originally developed for sequence to sequence task using an encoder-decoder architecture.In this work, we modify Transformer's encoder by replacing its self-attention with the decoder's masked self-attention, while keeping other modules unchanged. We focus on the modifications to enable multi-step reasoning via Transformer. For simplicity, we define three functions $Query(), Key()$, and $Value()$. Given a vector $\boldsymbol{v} \in R^{d_h}$, $Query(\boldsymbol{v}), Key(\boldsymbol{v})$, and $Value(\boldsymbol{v})$ are vectors in R^{d_h} and represent $\boldsymbol{v}$'s query, key, and value described in (Vaswani et al., 2017) respectively.

Fig. 3 is a conceptual architecture of the proposed Multi-step Reasoning Network(SeqMRN). $\{P_0, \cdots, P_{t-1}\}$ are position features defined in (Vaswani et al., 2017). Given dialog tuple $\widehat{D} = (\widehat{H}_t, \widehat{Q}_t)$, the position aware visual-linguistic sequence $\mathcal{U}_t$ is defined by:

$$\begin{aligned} \mathcal{U}_t &= \{U_0, U_1, \cdots, U_{t-1}\} \\ U_0 &= \widehat{C} + P_0 \\ U_i &= \widehat{L}_i + P_i,\ 1 \le i \le t-1 \end{aligned} \tag{4}$$

3.3.1 History Backward Self-Attention Layer

As illustrated in Fig. 3, this layer applies masked self-attention within the position aware sequence $\mathcal{U}_t$. This layer allows a single dialog round to gather relevant information from **previous** conversations and embed the information into its own representation.

Specifically, for $U_i, 0 \le i \le t-1$, its attention logits with respect to all the other rounds of dialog is defined by:

$$\boldsymbol{\tau^i} : \tau^i_j = \begin{cases} Key(U_j)^T Query(U_i) & j \le i \\ -\infty & i < j \end{cases} \tag{5}$$

where $\boldsymbol{\tau}^i \in R^t$. Then, the context aware visual-

linguistic sequence $\mathcal{V}_t$ is defined by:

$$\begin{aligned} \boldsymbol{w}^i &= softmax(\boldsymbol{\tau}^i/\sqrt{d_h}),\ \boldsymbol{w}^i \in R^t \\ \mathcal{V}_t &= \{V_0, \cdots, V_{t-1}\} : V_i = \sum_{j=0}^{t-1} \boldsymbol{w}^i[j] \cdot U_j \end{aligned} \tag{6}$$

3.3.2 Query Correction Layer

In this layer, the query $\widehat{Q}_t$ renews its knowledge about the context based on $\mathcal{V}_t$. The attention weights reflect how $\widehat{Q}_t$ distributes its focus over $\mathcal{V}_t$, which enables reasoning across the dialog history.

Specifically, the query's attention logits with respect to $\mathcal{V}_t$ is defined by:

$$\begin{aligned} \boldsymbol{u} : u_j = Key(V_j)^T Query(\widehat{Q}_t)/\sqrt{d_h} \\ 0 \le j \le t-1 \end{aligned} \tag{7}$$

However, we don't want history information in $\mathcal{V}_t$ to overpower the query's own semantic meaning, thus we augment $\widehat{Q}_t$ by self-attention weight u_q:

$$u_q = Key(\widehat{Q}_t)^T Query(\widehat{Q}_t)/\sqrt{d_h} \tag{8}$$

Then, the query's correction $\triangle\widehat{Q}_t$ is defined as:

$$\begin{aligned} \boldsymbol{w} &= softmax([\boldsymbol{u}; u_q]), \boldsymbol{w} \in R^{t+1} \\ \triangle\widehat{Q}_t &= \sum_{i=0}^{t-1} w_i V_i + w_t \widehat{Q}_t \end{aligned} \tag{9}$$

Note that Question Correction Layer keeps $\mathcal{V}_t$ unchanged. Contrary to SeqIPN, we don't use question history $\mathcal{Q}_t$ in SeqMRN because attention mechanism can make $\widehat{Q}_t$ indistinguishable from other questions in $\mathcal{Q}_t$.

3.3.3 Multi-step Reasoning

History Backward Self-Attention Layer and Question Correction Layer form the building blocks of our proposed Multi-step Reasoning Network. As illustrated in Fig. 3, residual connection is used.

$$\begin{aligned} \widehat{Q}'_t &= \widehat{Q}_t + \triangle\widehat{Q}_t \\ \widehat{C}' &= V_0 + U_0 \\ \widehat{L}'_i &= V_i + U_i, 1 \le i \le t-1 \end{aligned} \tag{10}$$

where the results $\widehat{Q}'_t$, $\widehat{C}'$ and $\widehat{L}'_t$ are vectors in R^{d_h}.

We have refined the dialog tuple $\widehat{D} = (\widehat{H}_t, \widehat{Q}_t)$ to be a new tuple $\widehat{D}' = (\widehat{H}'_t, \widehat{Q}'_t)$, where $\widehat{H}'_t = \{\widehat{C}', \widehat{L}'_1, \cdots, \widehat{L}'_{t-1}\}$. Members in $\widehat{D}'$ are more environment aware than their corresponding members in $\widehat{D}$. We achieve multistep reasoning by stacking several such building blocks to progressively refine $\widehat{D}$. We consider $\widehat{L}'_{t-1}$ of the last block as the summary of dialog history and consider $\widehat{Q}'_t$ of the last block as the context aware query. We project $[\widehat{Q}'_t; \widehat{L}'_{t-1}]$ to $R_{QL} \in R^{d_h}$ as the final representation of $\widehat{D}$.

3.4 Decoder Module

3.4.1 Discriminative Decoder

For each candidate anwer $A_t^j \in A_t$, a LSTM is applied to A_t^j to obtain its representation $R_j \in R^{d_h}$. The score of A_t^j is defined by $s_j = R_j^T R_{QL}$. Like (Guo et al., 2019), we optimize the N-pair loss (Sohn, 2016):

$$\mathcal{L}_D = log(\sum_{j=1}^{100} exp\frac{s_j - s_{gt}}{\tau}) \tag{11}$$

where s_{gt} is the score of the ground truth answer, and we set $\tau = 0.25$.

3.4.2 Generative Decoder

Inspired by attention based NMT (Luong et al., 2015), we develop an attention based decoder. The decoder is a LSTM initialized by R_{QL}. At time t, we compute similarity weights between current hidden state and the hidden states of previous timestamps instead of directly using the hidden state to generate the distribution over vocabulary. Then, the distribution is generated based on the weighted sum of hidden states.

3.5 Reweighting Method in Fine-tuning with Dense Annotations

VisDial v1.0 training dataset provides a subset named *dense annotations*[3] which contains 2K dialog instances. For each instance in *dense annotations*, two human annotators assign each of its candidate answer with a relevance score based on the ground-truth answer. (Qi et al., 2019b) finetunes with *dense annotations* using a generalized cross entropy loss:

$$\mathcal{L}_G = -\sum_{j=1}^{100} y_j log(softmax(\boldsymbol{s})[j]) \tag{12}$$

where $\boldsymbol{s}$ is the score vector of candidate answers, y_j is the relevance score label of the j^{th} candidate answer. However, blindly optimizing this objective will significantly hurt non-NDGC metrics.

To mitigate this issue, we propose a reweighting method to make the fine-tuning process aware of the importance of the ground truth answer. Specifically, we update the relevance label $\boldsymbol{y}$ by:

$$y'_i = \begin{cases} \frac{y_i+2}{3}, & i = index_{gt} \\ \frac{y_i}{3}, & otherwise \end{cases} \quad (13)$$

where $index_{gt}$ is the index of the ground truth answer.

4 Experiments

Using the VisDial v1.0 dataset, we experiment with 4 types of SeqDiaN: SeqIPN with GloVe Embedding (Pennington et al., 2014) (SeqIPN-GE), SeqIPN with DistilBert Embedding (Sanh et al., 2019) (SeqIPN-DE), SeqMRN with GloVe Embedding (SeqMRN-GE) and SeqMRN with DistilBert Embedding (SeqMRN-DE). For each type, we consider both discriminative and generative models. We trained Dense Symmetric Co-Attention Network (Nguyen and Okatani, 2018) from scratch. We use NDCG[1], MRR[2], recall (R@1, 5, 10), and mean rank to evaluate the models' performance.

In discriminative task, the model ranks the 100 candidate answers based on discriminative score, which is defined as the dot product similarity between the representation of dialogue and that of candidate answer.

In training and evaluation phases, to simplify the framework, the generative task is to rank the 100 candidate answers too. Given a candidate answer A, its generative score is defined as $\frac{lld_A}{\sqrt{|A|}}$, where lld_A is the answer's log-likelihood and $|A|$ is the answer's length. Based on generative score, the rank of 100 candidate answers is well defined, as well as the sparse metric MRR and Recall. However, in inference phase, we obtain the answer via distribution over vocabulary and beam search at every step as usual.

4.1 Quantitative Results

4.1.1 Model Comparison

We compare the performance between SeqDialN models of different configurations. We use Memory Network (MN) (Das et al., 2017), History-Conditioned Image Attentive Encoder (HCIAE)(Lu et al., 2017), Sequential Co-Attention Model (CoAtt)(Wu et al., 2018) and ReDAN (Gan et al., 2019) as baselines in this study because published work (Gan et al., 2019) reports the performance of these models with both discriminative and generative decoders.

Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
MN-D(Das et al., 2017)	55.13	60.42	46.09	78.14	88.05	4.63
HCIAE-D(Lu et al., 2017)	57.65	62.96	48.94	80.50	89.66	4.24
CoAtt-D(Wu et al., 2018)	57.72	62.91	48.86	80.41	89.83	4.21
ReDAN-D(T=1)(Gan et al., 2019)	58.49	63.35	49.47	80.72	90.05	4.19
ReDAN-D(T=2)(Gan et al., 2019)	59.26	63.46	49.61	80.75	89.96	4.15
ReDAN-D(T=3)(Gan et al., 2019)	59.32	**64.21**	50.60	81.39	90.26	4.05
SeqIPN-GE-D	58.44	58.74	44.87	75.49	85.30	5.56
SeqIPN-DE-D	58.18	59.49	45.58	76.08	86.40	5.15
SeqMRN-GE-D	59.73	61.32	47.59	78.03	87.04	5.08
SeqMRN-DE-D	**60.17**	57.98	44.46	74.16	84.50	5.86
Model	**NDCG↑**	**MRR↑**	**R@1↑**	**R@5↑**	**R@10↑**	**Mean↓**
MN-G(Das et al., 2017)	56.99	47.83	38.01	57.49	64.08	18.76
HCIAE-G(Lu et al., 2017)	59.70	49.07	39.72	58.23	64.73	18.43
CoAtt-G(Wu et al., 2018)	59.24	49.64	40.09	59.37	65.92	17.86
ReDAN-G(T=1)(Gan et al., 2019)	59.41	49.60	39.95	59.32	65.97	17.79
ReDAN-G(T=2)(Gan et al., 2019)	60.11	49.96	40.36	59.72	66.57	17.53
ReDAN-G(T=3)(Gan et al., 2019)	60.47	**50.02**	40.27	59.93	66.78	17.40
SeqIPN-GE-G	63.30	48.77	38.36	59.29	68.24	13.36
SeqIPN-DE-G	60.72	47.86	38.16	57.08	64.89	15.27
SeqMRN-GE-G	63.01	49.22	38.75	59.62	68.47	13.00
SeqMRN-DE-G	**64.15**	49.72	39.33	60.17	69.73	12.37

Table 1: Performance of SeqDialN models on VisDial v1.0 validation set. Left: discriminative SeqDialN. Right: generative SeqDialN. ↑ indicates higher is better. ↓ indicates lower is better.

In Table 1, "-D" stands for discriminative model and "-G" for generative model. SeqMRN-DE-D and SeqMRN-DE-G outperform all baselines and other SeqDialN models on NDCG[1] for both discriminative and generative cases. Especially for the generative case, SeqMRN-DE-G outperforms the second place ReDAN-G(T=3) by $> 3.6\%$ NDCG. Meanwhile, the MRR difference between ReDAN-G(T=3) and SeqMRN-DE-G is merely 0.3, SeqMRN-DE-G still outperforms ReDAN-G(T=3) on average performance. We arrive at the conclusion that SeqMRN-DE-G is a new state-of-the-art **generative** visual dialog model.

SeqIPN with GloVe Embedding is the simplest SeqDialN. However, SeqIPN-GE-D achieves better NDCG than well-known discriminative baselines such as MN-D, HCIAE-D and CoAtt-D. In addition, SeqIPN-GE-G even outperforms all generative baselines on NDCG. The model simplicity and performance gain together validate the merit of considering visual dialog as a visual-linguistic vector sequence.

4.1.2 Ensemble SeqDialN Analysis

In this section, we add VisDialBERT(Murahari et al., 2019) as a baseline. At this stage, the comparison is conducted between models trained **without** *dense annotation*[3].

As discriminative SeqDialN and generative SeqDialN rank the 100 candidate answers via discriminative score and generative score respectively, the uniform task definition facilitates the ensemble process. Given a set of SeqDialN models, we sim-

Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
ReDAN: 4 Dis. + 4 Gen.(Gan et al., 2019)	65.13	54.19	42.92	66.25	74.88	8.74
ReDAN+ (Diverse Ens.)(Gan et al., 2019)	67.12	56.77	44.65	69.47	79.90	5.96
VisDial-BERT: w/L-only(Murahari et al., 2019)	62.64	67.86	54.54	84.34	92.36	3.44
VisDial-BERT: w/CC+VQA(Murahari et al., 2019)	64.94	**69.10**	55.88	85.50	93.29	3.25
SeqDialN: 4 Dis.	64.66	64.67	51.74	80.49	89.10	4.34
SeqDialN: 4 Gen.	65.55	50.69	40.61	60.50	69.35	12.94
SeqMRN-DE-D + SeqIPN-GE-G	67.26	56.41	44.44	69.67	79.51	7.44
SeqDialN: 4 Dis + 4 Gen	**68.61**	58.11	45.94	71.66	81.22	6.73

Table 2: Comparison of SeqDialN to state-of-the-art visual dialog models on VisDial v1.0 validation set.

ply average scores of all models to obtain the new score to rank the 100 candidate answers and evaluate the metrics based on the new rank.

In Table 2, "SeqDialN: 4 Dis." is an ensemble of the 4 types of discriminative SeqDialN models while "SeqDialN: 4 Gen." an ensemble of the 4 types of generative SeqDialN models. Our best model outperforms ReDAN and ReDAN+ by significant margin on both NDCG ($> 1.5\%$) and MRR ($> 1\%$). Our model also outperforms VisDial-BERT(Murahari et al., 2019) by $> 3.5\%$ NDCG despite the latter being pretrained on several large-scale datasets.

VisDial-BERT(Murahari et al., 2019) has roughly 250M parameters, the configuration "w/L-only" is trained only on VisDial v1.0-train set, which is more suitable to compare with SeqDialN. SeqIPN-GE-G has less than 69M parameters but it can outperform "w/L-only" on NDCG ($> 0.5\%$). The ensemble configuration (SeqMRN-DE-D + SeqIPN-GE-G) has roughly the same parameters as "w/L-only" and it further outperforms "w/L-only" by $> 4\%$ NDCG. Actually, it even outperforms "w/CC+VQA" by $> 2\%$ NDCG. The advantage of VisDial-BERT (Murahari et al., 2019) is the high MRR score it achieves.

We also evaluate SeqDialN on VisDial v1.0 test-std set. Table 3 shows the comparison between our model and state-of-the-art visual dialog models trained **without** *dense annotations*[3]. SeqDialN achieves state-of-the-art performance on NDCG, even a single generative SeqDialN can outperform most previous work on that metric. At present, SeqDialN doesn't perform well on MRR, which is partly because it is hard for generative models to produce exactly the same answer as the ground truth, even when conditioned on the same semantic scenarios.

4.1.3 Fine-tuning with Dense Annotations

We fine-tune discriminative SeqDialN with *dense annotations*[3]. Table 4 shows the proposed reweighting method greatly mitigates performance drop in our fine-tuning experiment. We list the

Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
GNN(Zheng et al., 2019)	52.82	61.37	47.33	77.98	87.83	4.57
CorefNMN(Kottur et al., 2018)	54.70	61.50	47.55	78.10	88.80	4.40
RvA(Niu et al., 2019)	55.59	63.03	49.03	80.40	89.83	4.18
DualVD(Jiang et al., 2020)	56.32	63.23	49.25	80.23	89.70	4.11
HACAN(Yang et al., 2019)	57.17	64.22	50.88	80.63	89.45	4.20
SN(Guo et al., 2019)	57.32	62.20	47.90	80.43	89.95	4.17
SN†(Guo et al., 2019)	57.88	63.42	49.30	80.77	90.68	3.97
NMN(Kottur et al., 2018)	58.10	58.80	44.15	76.88	86.88	4.81
DAN(Kang et al., 2019)	57.59	63.20	49.63	79.75	89.35	4.30
DAN†(Kang et al., 2019)	59.36	64.92	51.28	81.60	90.88	3.92
ReDAN†(Gan et al., 2019)	61.86	53.13	41.38	66.07	74.50	8.91
VisDial-BERT: w/CC+VQA(Murahari et al., 2019)	63.87	**67.50**	53.85	84.68	93.25	3.32
ReDAN+ †(Gan et al., 2019)	64.47	53.74	42.45	64.68	75.68	6.64
SeqMRN-DE-G (single)	62.54	48.63	37.90	59.95	69.03	12.47
SeqDialN: 4 Gen.	63.78	49.98	39.50	60.48	69.27	12.97
SeqMRN-DE-D + SeqIPN-GE-G	65.56	55.66	43.23	69.15	79.93	7.44
SeqDialN: 4 Dis. + 4 Gen.	**66.91**	56.84	44.30	70.85	80.93	6.87

Table 3: Comparison of SeqDialN to state-of-the-art visual dialog models on VisDial v1.0 test-std set. ↑ indicates higher is better. ↓ indicates lower is better. † denotes ensembles. All models have been trained **without** *dense annotations*[3]

Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
SeqMRN-DE-D	70.23	38.33	23.04	55.17	71.51	9.29
SeqMRN-DE-D*	**70.72**	**53.59**	**42.35**	**65.05**	**77.73**	**7.27**
SeqIPN-DE-D	69.12	37.93	23.10	53.83	69.84	9.70
SeqIPN-DE-D*	**69.68**	**52.2**	**41.13**	**62.94**	**75.54**	**7.78**

Table 4: Using reweighting method to lessen performance drop on VisDial v1.0 validate set. * denotes fine-tuning with reweighting method.

fine-tuning statistics for one SeqIPN and one SeqMRN as representatives.

Table 5 compares SeqDialN with state-of-the-art models trained **with** *dense annotations*. On VisDial v1.0 test-std set, our model achieves comparable NDCG as others while outperforming them on MRR. It is interesting to note that VisDial-BERT (Murahari et al., 2019) outperforms our model on MMR by $> 5\%$ before fine-tuning. After fine-tuning however, our model outperforms it on MRR by nearly 5%. This observation validates the effectiveness of the reweighting method in preserving a model's overall performance when trained with *dense annotations*[3]. In addition, we find fine-tuning generative models don't improve NDCG as much as discriminative case.

4.2 Ablation Study

We note SeqMRN yeilds the best performance in the single model comparison, we conduct further experiments to analyze contribution of its components. For simplicity, We train discriminative SeqMRN in different configurations to 13 epochs without fine-tuning.

Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
MReal-BDAI†(Qi et al., 2019b)	74.02	52.62	40.03	68.85	79.15	6.76
P1P2†(Qi et al., 2019a)	**74.91**	49.13	36.68	62.96	78.55	7.03
VisDial-BERT: w/CC+VQA(Murahari et al., 2019)	74.47	50.74	37.95	64.13	80.00	6.28
SeqDialN: 4 Dis.	72.41	**55.11**	43.23	67.65	79.77	6.55

Table 5: Comparison of SeqDialN to state-of-the-art visual dialog models on VisDial v1.0 test-std set. All models have been trained **with** *dense annotations*[3]

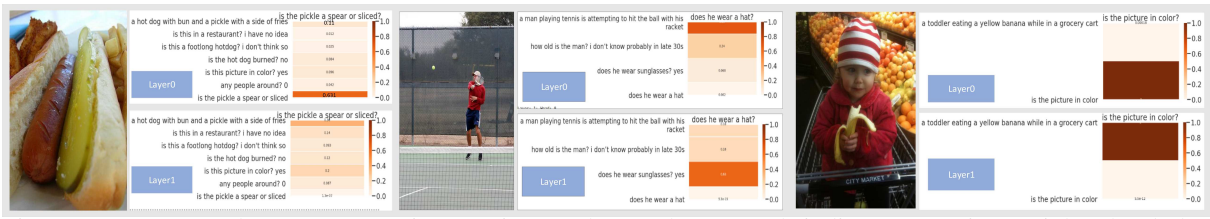

Figure 4: SeqMRN: learn to reason in attention stacks. Color strength indicates attention weight, the darker highlighting the higher attention paid.

4.2.1 Effectiveness of visual-linguistic joint representation

We close the modules in DCN (Nguyen and Okatani, 2018) which apply cross modality attention between vision and language features. Thus the two modalities are fused in a simple summation way in DCN.

In this configuration, the two modalities won't be aware of the existence of each other until the masked self-attention step in Transformer. Item named SeqMRN-DE-D-LateFusion in Table 6 shows its performance, which drops on all metrics. Especially on NDCG, it drops 3.14%.

This experiment demonstrates the positive impact of our early fusion, as we say, the visual-linguistic joint representation. Further analysis reveals early fusion helps enhance the model's capability to filter out irrelevant answers. We find that each image in *dense annotation*[3] of VisDial v1.0 has on average 12.68 answers with non-zero relevant-score. On average, We find SeqMRN-DE-D-LateFusion ranks 5.58 (44.00%) zero relevant-score answers into the top 12.68 predictions, while this number of SeqMRN-DE-D is 5.36 (42.27%).

Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
SeqMRN-DE-D	**59.49**	**61.53**	**47.68**	**78.67**	**87.88**	**4.79**
SeqMRN-DE-D-NoQC	59.08	61.25	47.34	78.58	87.72	4.86
SeqMRN-DE-D-LateFusion	56.35	61.14	47.11	78.29	87.48	4.83

Table 6: Ablation Study on VisDial v1.0 validation set.

4.2.2 Effectiveness of Query Correction Layer

In Table 6, the item SeqMRN-DE-D-NoQC shows the performance of the configuration by closing the Query Correction Layer illustrated in section 3.3.2. We see that performance drops on all metrics as well.

We find Query Correction Layer enhances the model's capability to integrate history information based on the given query, thus it helps answer the query which requires dialog history. (Agarwal et al., 2020) points out that not all questions in VisDial v1.0 dataset need dialogue history to answer. They have proposed a dataset named VisDialConv(Agarwal et al., 2020), which is actually a subset of VisDial v1.0 validation dataset including 97 instances which answer needs the reference to dialog history.

We run both SeqMRN-DE-D and SeqMRN-DE-D-NoQC on VisDialConv dataset. SeqMRN-DE-D gets 51.11% NDCG and SeqMRN-DE-D-NoQC gets 50.22%, the former has 1.77% relative improvement. As illustrated in Figure 5, the score distribution of the two models are similar, which concentrates in range [0.2, 0.9]. However, SeqMRN-DE-D scores significantly more instances in range [0.6, 0.7] than the other. SeqMRN-DE-D also scores less instances in the low range [0.0, 0.2] but scores more instances in the high range [0.8, 1]. These observations support the conclusion that Query Correction Layer helps answer history related questions.

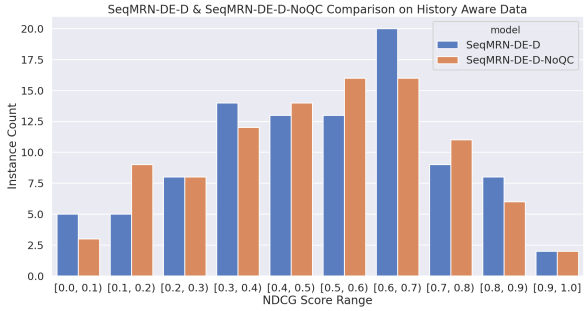

Figure 5: NDCG Distribution Comparison on VisDialConv

4.3 Qualitative Analysis

We use the 3 examples in Fig. 4 to illustrate SeqMRN's reasoning capability. On the left, the question asks: *"Is the pickle a spear or sliced?"*. In SeqMRN's first reasoning block (layer0), the question focus on preserving its own information (its self attention weight being 0.671). However, in the second reasoning block (layer1), the question pays more attention to the first round which has *"pickle"* related information. This example demonstrates the attention gets the right "correction" in Query Correction Layer.

In the middle, the question asks: *"Does he wear a hat?"* Due to the word *"he"*, in SeqMRN's first

reasoning block (layer0), the attention is on the caption (0.69), which has words *"man"* and *"his"*. However, in the second reasoning block (layer1), the attention turns to the round *"does he wear sunglasses? yes"*. Note the semantic similarity between *"wear sunglasses"* and *"wear hat"* (they are both wearables on the head). This example shows the attention making decisions based upon refined knowledge about the context in a deeper stack.

On the right, the question asks: *"Is the picture in color?"* In SeqMRN's first reasoning block, the attention focuses on itself. However, in the second reasoning block, the attention switches to the caption. Most likely in deeper stack, it make the inference like: *only a color image makes a banana look "yellow"*.

5 Conclusion

We presented Sequential Visual Dialog Network (SeqDialN) based on a novel idea that treats dialog rounds as a visual-linguistic vector sequence. We explore both discriminative and generative models and set up a new state-of-the-art **generative** visual dialog model. Even though our model is trained only on VisDial v1.0 dataset, it achieves competitive performance against other models trained on much larger vision-language datasets, which facilitates its deployment in industrial environment.

References

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? *arXiv preprint arXiv:2005.07493*.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Zhe Gan, Yu Cheng, Ahmed Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6463–6474.

Dalu Guo, Chang Xu, and Dacheng Tao. 2019. Image-question-answer synergistic network for visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10434–10443.

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

X. Jiang, J. Yu, Z. Qin, Y. Zhuang, X. Zhang, Y. Hu, and Q. Wu. 2020. Dualvd: An adaptive dual encoding model for deep visual understanding in visual dialogue. *AAAI*.

Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. 2019. Dual attention networks for visual reference resolution in visual dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2024–2033.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

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

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2019. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. *arXiv preprint arXiv:1912.02379*.

Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6087–6096.

Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. 2019. Recursive visual attention in visual dialog. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6679–6688.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2019a. Two causal principles for improving visual dialog.

Jiaxin Qi, Yulei Niu, Hanwang Zhang, Jianqiang Huang, Xian-Sheng Hua, and Ji-Rong Wen. 2019b. Learning to answer: Fine-tuning with generalized cross entropy for visual dialog challenge 2019. [Online; accessed November 12, 2019].

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pages 1857–1865.

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 *Advances in neural information processing systems*, pages 5998–6008.

Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. 2018. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6106–6115.

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

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830.

Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. 2019. Reasoning visual dialogs with structural and partial observations. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6669–6678.