Making History Matter: History-Advantage Sequence Training for Visual Dialog

Tianhao Yang[1]   Zheng-Jun Zha[1*]   Hanwang Zhang[2]

[1]University of Science and Technology of China, [2]Nanyang Technological University

jshmyth@mail.ustc.edu.cn, zhazj@ustc.edu.cn, hanwangzhang@ntu.edu.sg

Abstract

*We study the multi-round response generation in visual dialog, where a response is generated according to a visually grounded conversational history. Given a triplet: an image, Q&A history, and current question, all the prevailing methods follow a codec (i.e., encoder-decoder) fashion in a supervised learning paradigm: a multimodal encoder encodes the triplet into a feature vector, which is then fed into the decoder for the current answer generation, supervised by the ground-truth. However, this conventional supervised learning does NOT take into account the **impact of imperfect history**, violating the conversational nature of visual dialog and thus making the codec more inclined to learn history bias but not contextual reasoning. To this end, inspired by the actor-critic policy gradient in reinforcement learning, we propose a novel training paradigm called **History-Advantage Sequence Training** (HAST). Specifically, we intentionally impose wrong answers in the history, obtaining an adverse critic, and see how the historic error impacts the codec's future behavior by **History Advantage** — a quantity obtained by subtracting the adverse critic from the gold reward of ground-truth history. Moreover, to make the codec more sensitive to the history, we propose a novel attention network called **History-Aware Co-Attention Network** (HACAN) which can be effectively trained by using HAST. Experimental results on three benchmarks: VisDial v0.9&v1.0 and GuessWhat?!, show that the proposed HAST strategy consistently outperforms the state-of-the-art supervised counterparts.*

1. Introduction

Visual dialog is one of the most comprehensive task for benchmarking the AI's comprehension of natural language grounded by a visual scene [6]. A good visual dialog agent should accomplish a complex series of reasoning sub-tasks: contextual visual perception [36, 23, 5, 30, 19, 37, 31], lan-
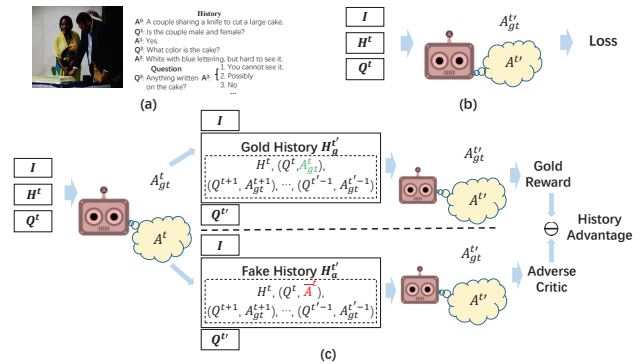
---

*Corresponding author.

Figure 1. (a) A typical visual dialog task. Particularly, the initial answer A^0 denotes the given image captioning. (b) The conventional training process at round t: given an image I, a history H^t, and question Q^t, the loss is supervised by the ground-truth answer A_{gt}^t. (c) The proposed History-Advantage Sequence Training (HAST) paradigm: the reward (*i.e.* $-$ loss) is a History-Advantage, which is more focused on the impact caused by a wrong answer $\bar{A}^t$ to the future round t', by comparing the difference between the Gold Reward from gold history $H_g^{t'}$ and the Adverse Critic from fake history $H_a^{t'}$.

guage modeling [14, 4], and co-reference resolution in dialog history [30, 19] (*e.g.*, identify what is "that"). Thanks to the end-to-end deep neural networks in their respective sub-tasks, state-of-the-art visual dialog systems can be built by assembling them into a codec framework [36, 6, 23]. The encoder encodes a triplet input — history question-answer sentences, an image, and a current question sentence — into a vector representation; then, the decoder fuses those vectors and decodes them into answer sentences (*e.g.*, by generative language models [6, 23, 36] or discriminative candidate ranking [15, 19]).

So far, one may identify that the key difference between Visual Dialog (VisDial) and the well-known Visual Question&Answering (VQA) [3] is the exploitation of the *history*. As shown in Figure 1(a) and Figure 1(b), VisDial can be cast into a multi-round VQA given additional lan-

guage context of history question-answering pairs. Essentially, at each round, the response generated by the agent is "thrown away" and the history is artificially "corrected" by the ground-truth answers [6] for the next round. Note that this ground-truth history setting is reasonable because it steers the conversation to be evaluable; otherwise, any other response may digress the conversation into a never-ending open-domain chitchat [27, 13]. However, we argue that *by only exploiting the ground-truth history is ineffective for the codec training*. For example, the ground-truth answer only tells the model that "*you can not see it*" is good, but neglect to show how badly other answers will impact the dialog. Therefore, the resultant model is easily over-grounded in the insufficient ground-truth data, but not learning visual reasoning [21].

In this paper, we propose a novel training strategy that utilizes the history response in a more efficient way, that is, to make the codec model more sensitive to the history dialog such as co-reference resolution and context. As illustrated in Figure 1(c), we intentionally impose wrong answers in the "tampered" history (*e.g.* replacing "*White with blue lettering, but hard to see it.*" with " *Yellow*" in the history) and see how the model behaves as compared to the "gold" history. Specifically, suppose that we are going to train a model at the t-th round, to gain more insights about the wrong answers, we maintain two parallel lines of dialogs: one with the ground-truth answer A_{gt}^t, and the other one with a probable mistake $\overline{A}_t$. Then, we run the two lines to a future round t' (both of them are filled with ground-truth answers from $t+1$ to t'). Thus, we can collect two rewards at round t': 1) *Gold Reward* (GR): a conventional ground-truth history reward for answer A_{gt}^t, the larger the better , and 2) *Adverse Critic* (AC): a proposed critic for fake history reward of $\overline{A}_t$, small AC implies large future impact by the mistake answer. Interestingly, their difference HA = GR − AC, which we call *History-Advantage* (HA), will tell the model how to reward A_{gt}^t: if HA > 0, large HA implies large impact of fake history, *i.e.*, very small AC; thus, we need to strengthen the training signal of gold history with A_{gt}^t; if HA ≤ 0, it implies the ineffectiveness of the gold history, *i.e.*, the current model cannot accurately rank correct answer candidates; thus, the gradient for A_{gt}^t will be in the opposite direction. In this way, we can collect a sequence of history-advantage training losses from $t+1$ to T. Therefore, we call the proposed training paradigm: History-Advantage Sequence Training (HAST).

Although the application scenario of HAST is regardless of specific codec models, for more effective training, we develop a novel codec dubbed: History-Aware Co-Attention (HACAN) encoder to address the essential co-reference and visual context in history encoding. In a nutshell, HACAN is a sequential model that contains two novel co-attention modules and one history-aware gate. Equipped with the the

proposed History-Advantage Sequence Training (HAST), we achieve a new state-of-the-art single-model on the real-world VisDial benchmarks: 0.6792 MRR on VisDial v0.9, 0.6422 MRR on VisDial v1.0, and 66.8% accuracy on GuessWhat?! [7]. We also achieve a top performing 0.5717 NDCG score on the official VisDial online challenge server. More ablative studies, qualitative examples, and detailed results are discussed in Section 5.

2. Related Work

Visual Dialog. Visual dialog is recently proposed in [6] and [7], which is a more challenging vision-language problem. Most vision-and-language problems are based on a single-round language interaction (*e.g.* image caption [9, 10, 17, 2] and visual question answering [2, 1, 3, 5, 24]). On the contrary, visual dialog task involves a multi-round dialog which is more complex. Das *et al.* [6] proposed a large-scale free-form visual dialog dataset, which consists of sequential open-ended questions and answers about arbitrary objects in the image. Another visual dialog task GuessWhat?! proposed by [7] focuses on a different aspect, which aims at object discovery with a set of yes/no questions. We apply the first setting in this paper. The proposed approaches for Visual Dialog are based on encoder-decoder structure, and can be categorized into three groups based on the design of encoder: (1) Fusion-based Models (LF [6], HRE [6], Sync [11]), the methods fuse image, question, and history features at different stages. (2) Attention-based Models (MN [6], HCIAE [23], CoAtt [24]), the methods establish attention mechanisms over image, question and history. (3) Approaches that deal with visual reference resolution (AMEM [30], CorefNMN [19], RvA [16], DAN [16]), the methods focus on explicit visual co-reference resolution in visual dialog.

Reinforcement Learning with Baseline. Reinforcement learning with baseline is widely applied in language generation, *e.g.*, image caption [29] and visual dialog [23, 36]. However, the historical response rounds in dialog are not considered as a sequence, which is in contrast our work focuses on. In particular, we follow the spirit of sequence training and design a history-advantage as a baseline to reward or penalize the future dialog rounds based on the current gold and fake history.

3. Our Codec Model

In this section, we formally introduce the visual dialog task and describe the details of our proposed model. We follow the definition introduced by Das *et al.* [6]. Given input as: 1) an image I, 2) a dialog history with a caption A^0 of the image and $t-1$ rounds of the dialog $\{A^0, (Q^1, A^1), ..., (Q^{t-1}, A^{t-1})\}$, where (Q^i, A^i) is the i-th round of the "ground-truth" question and answer pair,
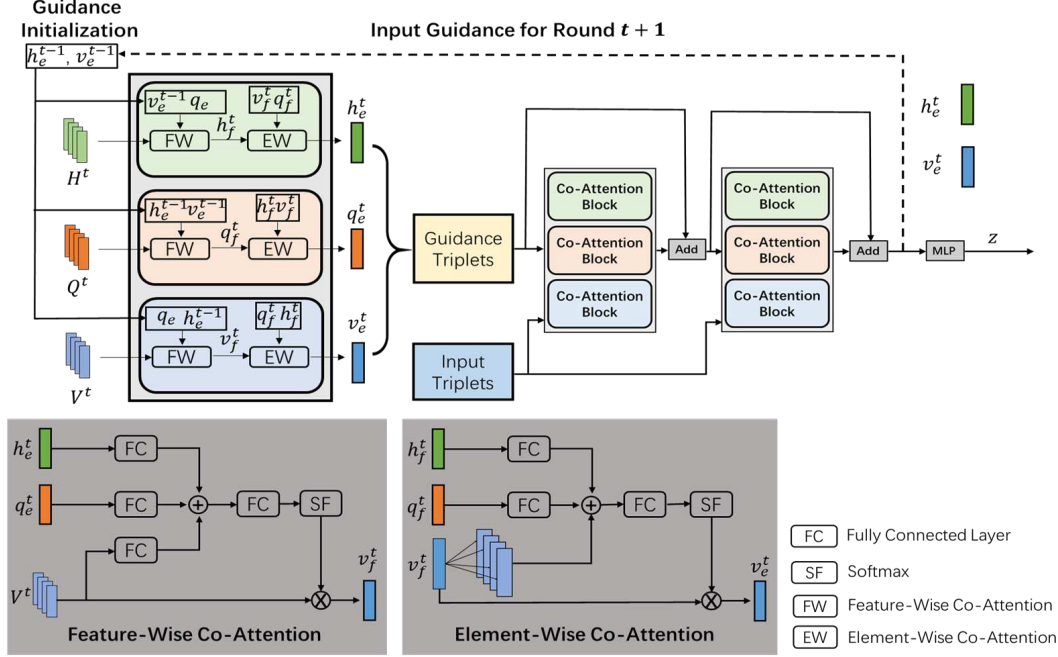
Figure 2. The framework of our proposed codec model. $\boldsymbol{H}^t$, $\boldsymbol{Q}^t$ and $\boldsymbol{V}^t$ are the input triplets (Section 3.1) extracted by CNNs and LSTMs. t denotes the current round of the dialog. Feature-Wise Co-Attention and Element-Wise Co-Attention are applied as blocks in parallel to encode the input triplets and generate the guidances of the follow-up attention layer. The history-awareness, $\boldsymbol{h}_e^{t-1}$ and $\boldsymbol{v}_e^{t-1}$, are used to initialize the guidances in the first block (Section 3.2). The final outputs of the encoder update the history-awareness and feed to the decoder as well. The decoder finally generates the response and ranks candidate answer options (Section 3.3).

3) a follow-up question Q^t, and 4) a list of 100 candidate answer options $\{A_1^t, ..., A_{100}^t\}$ which contains one correct answer A_{gt}^t. The visual dialog model needs to sort the answer options and choose the right one when given the inputs. To perform response generation given the above task, as illustrated in Figure 2, our codec model includes three modules: 1) feature representation (Section 3.1), 2) the proposed History-Aware Co-Attention Network (HACAN) for the encoder (Section 3.2), and 3) a discriminative decoder for response generation by ranking (Section 3.3).

3.1. Feature Representation

Image Feature. We follow bottom-up attention mechanism to extract region-based image features as in [2]. We train Faster-RCNN [28] based on the ResNet-101 backbone [12] on Visual Genome [20] dataset. We choose top-K regions from each image and encode the regions as the visual feature $\boldsymbol{V}^t$, where t is the current round in the dialog. For fair comparison with some methods without region proposal network [28], we replace bottom-up attention mechanism with VGG model in the ablation study. More details are discussed in Section 5.

Language Feature. 1) Question Feature: We first embed the words of the follow-up question. LSTM is applied to generate a sequence of hidden states. Words play differ-

ent roles in the question. The operative words can tell the model the kind of the question and which instance attribute to consider. Especially in visual dialog task, history and the current question may have relationships and make word contribution mean more. As illustrated in Figure 1(a), the word "*written*" in the current question has relationship with "*blue lettering*" and helps the model focus on the last round in the history. Meanwhile, the last round reminds the model to pay more attention to the word "*written*". From this, we use the whole hidden state sequence generated by LSTM instead of last hidden state as some prior works do, which is denoted as $\boldsymbol{Q}^t$. 2) Answer Feature: We apply another LSTM to the word embeddings of the candidates, and use the the whole hidden state sequence as the answer feature. 3) History Feature: Each round of the question and the answer in the history is concatenated into a long "sentence". Another LSTM is applied to each round of the history after word embedding. We use the last hidden state of each round as the question-answer pair feature. The history feature is denoted as $\boldsymbol{H}^t$.

3.2. History-Aware Co-Attention Network

We propose a novel attention-based model called History-Aware Co-Attention Network (HACAN) to encode the input features described above with the co-attention

mechanism [36, 23, 24].

As illustrated in Figure 2, HACAN is composed by a sequence of attention blocks, each of them contains two attention-based modules: Feature-Wise Co-Attention module (FCA) and Element-Wise Co-Attention module (ECA). Given the input triplet $\{\boldsymbol{V}^t, \boldsymbol{Q}^t, \boldsymbol{H}^t\}$, FCA aims to attend to the relevant features in one set of inputs with the guidances from the other two inputs. ECA takes the outputs of FCA as the inputs. It aims to activate the relevant elements and restrain the irrelevant ones with the guidances from the other two inputs before the final fusion. We now describe the two modules in details.

Feature-Wise Co-Attention (FCA). We use additive attention function to compute the feature-wise attention, returning the attended feature as output. Without loss of generality, we take the FCA for V as an example, which is denoted as $Attend_f(\boldsymbol{V}^t, \boldsymbol{g}^Q, \boldsymbol{g}^H)$. Take the visual feature of image regions $\boldsymbol{V}^t = \{\boldsymbol{v}_1, ... \boldsymbol{v}_K\}$ as input, the visual attention is formulated as:

$$\boldsymbol{v}_i' = \tanh(W_1^f \boldsymbol{v}_i^t + W_2^f \boldsymbol{g}^Q + W_3^f \boldsymbol{g}^H), \qquad (1)$$

$$\alpha_i = \mathrm{softmax}(W^T \boldsymbol{v}_i'), \qquad (2)$$

$$\boldsymbol{v}_f^t = \sum_{i=1}^{K} \alpha_i \boldsymbol{v}_i \quad i = 1, ..., K, \qquad (3)$$

where $\boldsymbol{g}^Q, \boldsymbol{g}^H \in \mathbb{R}^d$ are the guidances from $\boldsymbol{Q}^t$ and $\boldsymbol{H}^t$. $W_1^f, W_2^f, W_3^f \in \mathbb{R}^{d \times d}$, $W \in \mathbb{R}^{d \times 1}$ and d is the feature dimension. As the formats of $\boldsymbol{Q}^t$ and $\boldsymbol{H}^t$ are consistent with $\boldsymbol{V}^t$, we can simply apply Eq. (1)-(3) to $Attend_f(\boldsymbol{Q}^t, \boldsymbol{g}^H, \boldsymbol{g}^V)$ and $Attend_f(\boldsymbol{H}^t, \boldsymbol{g}^V, \boldsymbol{g}^Q)$, and compute three modules in parallel.

Element-Wise Co-Attention (ECA). The outputs of FCAs are the sum of attended features. However, the features itself are not attended yet. We introduce an element-wise attention mechanism to attend the sum of features. It is worth noting that, each element in the feature (*e.g.*, $\boldsymbol{v}_i \in \mathbb{R}^d$ in $\boldsymbol{V}^t$) is a response activation of the neural networks, and reflects some attributes of the instance in a sense. We apply the attention mechanism in an element-wise manner with the guidance from the other two inputs. It can be regarded as selecting the relevant semantic attributes and discarding the irrelevant ones based on the guidances from other domains. We take $Attend_e(\boldsymbol{v}_f^t, \boldsymbol{q}_f^t, \boldsymbol{h}_f^t)$ as an example, and ECA is formulated as:

$$\boldsymbol{V}^e = \tanh(W_1^e \otimes \boldsymbol{v}_f^t + (W_2^e \boldsymbol{q}_f^t)\boldsymbol{E}^T + (W_3^e \boldsymbol{h}_f^t)\boldsymbol{E}^T), \quad (4)$$

$$\boldsymbol{v}_e^t = \sigma(W'^T \boldsymbol{V}^e) \odot \boldsymbol{v}_f^t, \qquad (5)$$

where $W_1^e \in \mathbb{R}^m$ which is learnable and all elements are initialized with value 1, $W_2^e, W_3^e \in \mathbb{R}^{d \times d}$ and $W' \in \mathbb{R}^m$.

$\boldsymbol{E} \in \mathbb{R}^m$ is a vector with all elements set to 1. $\otimes$ represents the outer product of vectors and $\odot$ represents the element-wise product. We use W_1^e to broadcast $\boldsymbol{v}^t$ to m times, and perform the attention function in parallel. It can be viewed as the additional-attention of multi-head attention in [35]. We compute $Attend_e(\boldsymbol{q}_f^t, \boldsymbol{h}_f^t, \boldsymbol{v}_f^t)$ and $Attend_e(\boldsymbol{h}_f^t, \boldsymbol{v}_f^t, \boldsymbol{q}_f^t)$ in the same way.

Gated History-Awareness. We observe that an ambiguous question often has relationship with its latest round in the history. From this, the encoding feature of the latest history v_e^{t-1}, h_e^{t-1} is a good choice to initialize the guidances in FCAs, which can be viewed as history-awareness. However, sometimes the current question has nothing to do with its latest history. A simple solution is to apply a gate function to control the history-awareness, which is formulated as:

$$\boldsymbol{q}_s^t = Attend_f(\boldsymbol{Q}^t, \boldsymbol{0}, \boldsymbol{0}), \qquad (6)$$

$$\boldsymbol{o} = MLP([f_g^q(\boldsymbol{q}_s^t), f_g^h(\boldsymbol{h}_e^{t-1})]), \qquad (7)$$

$$\boldsymbol{h}_e^{t-1} = \sigma(\boldsymbol{o})\boldsymbol{h}_e^{t-1}, \quad \boldsymbol{v}_e^{t-1} = \sigma(\boldsymbol{o})\boldsymbol{v}_e^{t-1}, \qquad (8)$$

where $\boldsymbol{q}_s^t$ is self-attention of the current question, $[\cdot]$ is a concatenation operation and $\sigma(\boldsymbol{o})$ is the gate value. With the gated history-awareness, each first FCA is guided by its latest history feature and update the attended features efficiently. For the follow-up modules, we take the outputs of ECAs as the guidance inputs in Eq. (1), so we can connect two attention modules recurrently and update the encoding feature $\boldsymbol{v}_e^t, \boldsymbol{h}_e^t, \boldsymbol{q}_e^t$, as illustrated in Figure 2. The shortcut connections in the layers help the model consider different levels in multi-hop visual reasoning. The outputs of final round of attention blocks, $\boldsymbol{v}_e^t, \boldsymbol{h}_e^t, \boldsymbol{q}_e^t$, are used to generate the answer response.

3.3. Response Generation

Now we introduce how to generate the answers for the visual dialog task. We concatenate the three features $\boldsymbol{v}_e^t, \boldsymbol{h}_e^t, \boldsymbol{q}_e^t$ together and use a linear transform followed by a tangent activation:

$$\boldsymbol{z} = \tanh(W_e[\boldsymbol{v}_e^t, \boldsymbol{h}_e^t, \boldsymbol{q}_e^t]), \qquad (9)$$

where $[,]$ is a concatenation operation. We encode candidate answer features using a self-attention mechanism. The self-attention mechanism is formulated like Eq. (6). We dot product the candidate answer features and $\boldsymbol{z}$ to calculate the similarities. We sort the answer candidates by the similarities and choose the top one with highest similarity as the prediction.

In the task of GuessWhat?!, the information of answer candidates are the localizations and categories of the objects. We concatenate the localizations and categories and embed them with a linear transform to obtain the answer

option features. We also calculate the similarities of answer features and the final encoding features by dot product.

4. History-Advantage Sequence Training

HACAN described in Section 3.2 encodes the "ground truth" triplet and generates the responds. However, by only using the conventional supervised learning, HACAN does not take into account the contribution of "gold" history and the impact of imperfect history, which makes the model history-aware and explore semantic information in the history. To disentangle the individual contribution of a specific round in the history and make the codec model more sensitive to the history, an intuitive solution is to replace the default answer of the specific round in the history with imperfect answers. To this end, besides utilizing "gold" history, we intentionally impose wrong answers chosen by the model to generate the "tampered" history, and see how much better the model performs with "gold" history than the one with "tampered" history. We first describe the policy gradient for visual dialog in Section 4.1, and then we describe the history-advantage in the policy gradient in Section 4.2. Finally, we briefly introduce the training process in Section 4.3.

4.1. Policy Gradient for Visual Dialog

We transform visual dialog task to a simple two-step decision-making game: in step 1, given an image I, t-round history with caption $\{A^0, (Q^1, A^1), ..., (Q^{t-1}, A^{t-1})\}$ and the follow-up question Q^t, the agent needs to choose an answer from the candidates. In step 2, besides the inputs above and the choice of the agent, a judge is given the remaining rounds from $t+1$ to t', and needs to mark by answering the current question $Q^{t'}$ (t' is anywhere from $t+1$ to 10) correctly. The goal of the game is to choose the right answer of $Q^{t'}$, which is the same as visual dialog task described in Section 3. The difference is the choice of Q^t may be imperfect and impact the score of the game. We formulate the game as a decision-making problem. The action space is the answer candidates denoted as $\{A_1^t, ..., A_{100}^t\}$ and the state space is $\{\boldsymbol{V}^{t'}, \boldsymbol{Q}^{t'}, \boldsymbol{H}^{t'}\}$.

4.2. History-Advantage

We use the metric of the visual dialog task to compute the reward (*e.g.* MRR). We denote the gold reward (GR) as $R(\boldsymbol{V}^{t'}, \boldsymbol{Q}^{t'}, \boldsymbol{H}_g^{t'})$ and use $R(\boldsymbol{V}^{t'}, \boldsymbol{Q}^{t'}, \boldsymbol{H}_{a,i}^{t'})$ to represent the adverse critic (AC) when the agent choose the i-th negative answer in round t. As illustrated in Figure 1(c), their difference GR $-$ AC can reflect the influence of the "gold" round t when answering $Q^{t'}$. We definite history-advantage of the "gold" round t as:

$$GR - AC = A(t') = A(\boldsymbol{V}^{t'}, \boldsymbol{Q}^{t'}, \boldsymbol{H}^{t'}) =$$
$$R(\boldsymbol{V}^{t'}, \boldsymbol{Q}^{t'}, \boldsymbol{H}_g^{t'}) - \sum p(\bar{\boldsymbol{A}}_i^t)R(\boldsymbol{V}^{t'}, \boldsymbol{Q}^{t'}, \boldsymbol{H}_{a,i}^{t'}), \quad (10)$$

where $A(t')$ can estimate the contribution of the "gold" answer of round t on answering $Q^{t'}$. It plays a similar role as the "*advantage*" in actor-critic methods, and AC is a baseline in the history-advantage. When the total adverse critic is lower than gold reward, $A(t')$ is positive. It means the mistake does impact negatively on the future and the influence of the "gold" round is positive. The baseline can reduce the variance of gradient estimation in the training part as well and help the model take the contribution of the history into account. As t' is anywhere from $t + 1$ to 10, we compute $A(t')$ with different value of t', which can reduce the variance of gradient estimation as well. The sum of $A(t')$ can also be viewed as the contribution of the "gold" round in the whole dialog.

4.3. Training

Inspired by the policy gradient theorem [34], the history-advantage gradient can be simply denoted as:

$$\nabla_\theta J_g = \nabla_\theta \log \boldsymbol{p}(A_{gt}^t; \theta) \cdot \left(\frac{1}{10 - t} \sum_{t'=t+1}^{10} A(t')\right), \quad (11)$$

Following previous policy gradient works that use a supervised pre-training step as model initialization, we train our model with two-stage training. In the first training stage, we use a metric-learning multi-class N-pair loss $\mathcal{L}_{np}$ [33, 23].

In HAST, we denote the final gradient as:

$$\nabla_\theta J = \nabla_\theta J_g - \alpha \nabla_\theta \mathcal{L}_{np}, \quad (12)$$

where we incorporate N-pair loss (weighted by a trade-off α) for an end-to-end training. The whole training process is reviewed in Algorithm. 1.

5. Experiments

In the following we evaluate our proposed approach on three visual dialog datasets, VisDial v0.9 [6], VisDial v1.0 [6] and GuessWhat?! [7]. We first present the details about the datasets, evaluation metrics and the implementation details. Then we provide qualitative results and compare our methods with the state-of-the-art models.

5.1. Datasets

VisDial v0.9 [6] contains about 123k image-caption-dialog tuples. The images are all from MS COCO [22] with multiple objects. The dialog of each image has 10 question-answer pairs, which were collected by pairing two

Algorithm 1: Discriminative Model with History-Advantage Sequence Training

---
Require: Supervised Pre-trained Model HACAN
 1: **for** Round $t = 1, \cdots, T - 1$ **do**
 2:   Compute
       $\{p(\bar{A}_1^t), \cdots, p(\bar{A}_{N-1}^t)\} \sim \text{HACAN}(\boldsymbol{V}^t, \boldsymbol{Q}^t, \boldsymbol{H}^t)$
       where N is the number of answer candidates.
 3:   **for** $t' = t + 1, \cdots, T$ **do**
 4:     **for** $i = 1, \cdots, N - 1$ **do**
 5:       Fake the history $H^{t'}$ with $\bar{A}_i^t$
 6:       Compute $R(\boldsymbol{V}^{t'}, \boldsymbol{Q}^{t'}, \boldsymbol{H}_g^{t'})$ and
              $R(\boldsymbol{V}^{t'}, \boldsymbol{Q}^{t'}, \boldsymbol{H}_{a,i}^{t'})$ using HACAN
 7:     **end for**
 8:   **end for**
 9:   Compute the gradient $\nabla_\theta J$ with Eq. (12)
10:   $\theta \leftarrow \theta + \delta \nabla_\theta J$
11: **end for**

---

people on Amazon Mechanical Turk to chat with each other about the image. Specifically, the "questioner" is required to "imagine the scene better" by sequentially asking questions about the hidden image. The "answerer" then observes the picture and answers questions.

VisDial v1.0 [6] is an extension of VisDial v0.9 [6]. Images for the training set are all from COCO train2014 and val2014. The dialogs in validation and test sets were collected on about 10k COCO-like images from Flickr. The test set is split into two parts, 4k images for test-std and 4k images for test-challenge. Answers are already provided for the train and val set, but not in the test set. For the test-std and test-challenge phases, the results must be submitted to the evaluation server.

We also evaluated our proposed model on GuessWhat?! dataset [7]. The dataset contains 67k images collected from MS COCO [22] and 155k dialogs including about 820k question-answer pairs. The guesser game in GuessWhat?! is to predict the correct object in object options through a multi-round dialog.

5.2. Evaluation Metrics

For VisDial v0.9, we followed the evaluation protocol established in [6] and used the retrieval setting to evaluate the responses at each round in the dialog. Specifically, for each question we sorted the answer options and used Recall@k, mean reciprocal rank (MRR) and mean rank of the ground truth answer to evaluate the model. Recall@k is the percentage of questions for which the correct answer option is ranked in the top k predictions of a model. Mean rank is the average rank of the ground truth answer option. Mean reciprocal rank is the average of 1/rank of the ground truth answer option. For the test set of VisDial v1.0, we also

evaluated our model using the newly introduced normalized discounted cumulative gain (NDCG). NDCG is invariant to the order of options with identical relevance and to the order of options outside of the top K, where K is the number of answers marked as correct by at least one annotator. For GuessWhat?! dataset, we used classification accuracy to evaluate our model.

5.3. Implementation Details

Language Processing. We first tokenized the questions and answers using the Python NLTK toolkit and constructed a vocabulary of words that appear at least 5 times in the training split. All the words were embedded to a 300-dimension vector initialized by pre-trained GloVe [26] embeddings. The LSTMs of the question and history were two-layered, while they were one-layered for the answers. The hidden states in all LSTM were 512-d.

Training Details. We pretrained our codec using the supervised training for 15 epochs before starting HAST. We used Adam optimizer [18] and started the supervised training with the base learning rate of 1×10^{-3} and decreasing to 5×10^{-4} after 10 epochs. In HAST, the base learning rate was 1×10^{-4}, and decayed every 5 epochs with an exponential rate of 0.5. We set the hyper-parameter m to 16, and $W_1^e \in \mathbb{R}^{16}$ in Eq. (4). We set α to 1 in Eq. (12).

In Eq. (10), we needed to sum over all the negative answer candidates to calculate the advantage for HAST. It cost quite a lot of time to train (about 99 evaluations for each incorrect answers). However, we noticed that only a few negative answers have non-ignorable probabilities. For reducing the time cost, we made a trade-off between the accuracy and the speed and summed over the top-5 negative answers chosen by our model. The experiment showed that we saved 95% training time with a slight performance drop.

5.4. Ablative Study

Components in HACAN. We present a few variants of our model to verify the contribution of each component: **(1)** HACAN w/o FCA is a baseline without FCA and ECA. Image features are guided by the question, the attention weight is computed by Eq. (1) with only question guidance. **(2)** HACAN w/o ECA-1 is a model with one FCA and Gated History-Awareness. The model has no ECA. **(3)** HACAN w/o ECA-16 is a model with one FCA, followed by one ECA with $m = 1$ in Eq. (5). The ECA is not multihead. **(4)** HACAN w/o RS is a model with one FCA, followed by one ECA with $m = 16$. The numbers of multihead is 16. **(5)** HACAN is a whole model that we stack two attention modules three times in a residual manner.

The first and second row in Table 3 show that, HACAN w/o FCA uses question-guided image feature (like the style in VQA task), but is not history-aware. Benefiting from FCA and Gated History-Awareness, the model takes

	VisDial v1.0(test-std)						VisDial v0.9(val)				
	NDCG	MRR	R@1	R@5	R@10	Mean	MRR	R@1	R@5	R@10	Mean
LF w/o RPN [6]	0.4531	0.5542	40.95	72.45	82.83	5.95	0.5807	43.82	74.68	84.07	5.78
HRE [6]	0.4546	0.5416	39.93	70.45	81.50	6.41	0.5846	44.67	74.50	4.22	5.72
MN [6]	0.4750	0.5549	40.98	72.30	83.30	5.92	0.5965	45.55	76.22	85.37	5.46
HCIAE [23]	-	-	-	-	-	-	0.6222	48.48	78.75	87.59	4.81
AMEM [30]	-	-	-	-	-	-	0.6227	48.53	78.66	87.43	4.86
CoAtt [24]	-	-	-	-	-	-	0.6398	50.29	80.71	88.81	4.47
CorefNMN [19]	0.5470	0.6150	47.55	78.10	88.80	4.40	0.6410	50.92	80.18	88.81	4.45
RvA w/o RPN [25]	0.5176	0.6060	46.25	77.88	87.83	4.65	0.6436	50.40	81.36	89.59	4.22
Ours w/o RPN	0.5281	0.6174	47.91	78.59	87.81	4.63	0.6451	50.72	81.18	89.23	4.32
LF [6]	0.5163	0.6041	46.18	77.80	87.30	4.75	-	-	-	-	-
RvA [25]	0.5559	0.6303	49.03	80.40	89.83	4.18	0.6634	52.71	82.97	**90.73**	**3.93**
Sync[11]	0.5732	0.6220	47.90	80.43	**89.95**	**4.17**	-	-	-	-	-
DAN [16]	**0.5759**	0.6320	49.63	79.75	89.35	4.30	0.6638	53.33	82.42	90.38	4.04
Ours	0.5717	**0.6422**	**50.88**	**80.63**	89.45	4.20	**0.6792**	**54.76**	**83.03**	90.68	3.97

Table 1. Retrieval performance of discriminative models on the test-standard split of VisDial v1.0 and the validation set of VisDial v0.9. RPN indicates the usage of region proposal network.
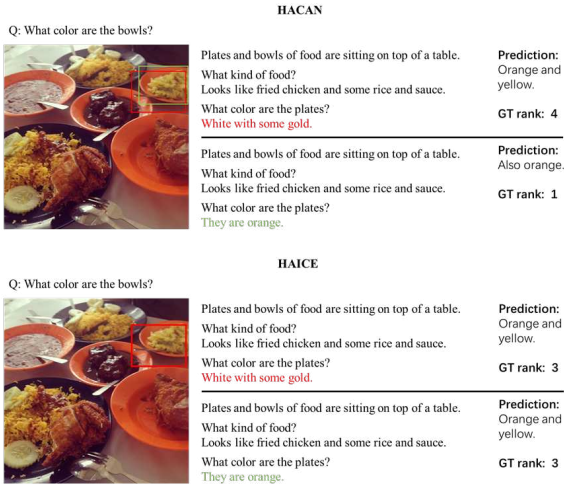

Figure 3. Qualitative results on VisDial dataset. We visualize the behaviors of our model and HCIAE with historical changes. Incorrect history and the region chosen by model with tampered history are marked with red. GT rank denotes the rank of ground-truth answer in the sorted list.
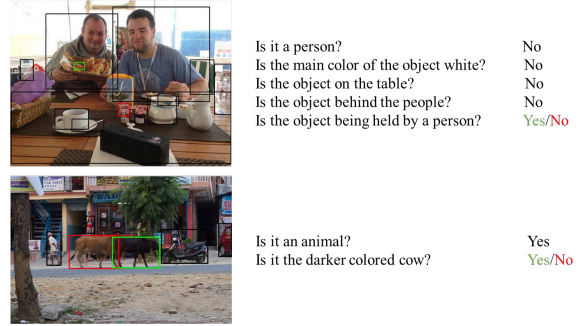

Figure 4. Qualitative results of our model on GuessWhat?!. The green bounding boxes highlight the right predictions with gold histories. The red bounding boxes highlight the wrong predictions with tamperred histories.

history-aware features and achieves approximately 3.5% improvements on MRR, more in line with the conversational nature of visual dialog. In HACAN w/o ECA-16, two types of attention modules guide the model not only which features to focus (*e.g.* feature-level attended histories), but also the attended attributes of features, advancing the attention mechanism to a more fine-grained level. The hyper-parameter m can be viewed as the number of heads in multi-head attention [35], and improves the performance by approximately 0.5 point on MRR. We stacked two attention modules up to three layers in a residual manner to obtain multi-level abstraction of history-aware and question-aware features, this achieved the best performance in our ablative experiments. We believe the ablative experiments demon-

strate that: (1) different from the VQA task, the model for visual dialog task relies on history-aware representations. (2) FCA and ECA are efficient to compute both reliable question-aware and history-aware visual attention.

Effectiveness of HAST. We did ablative experiments to demonstrate the effectiveness of HAST. Results are showed in Table 4 and Table 5. Our model achieves approximately another 1 point improvement on R@1 and 0.5% on MRR on VisDial dataset. Beyond that, we applied HAST to some additional ablative models with official codes[1]. We denote HCIAE-D-MLE [23] and HCIAE-D-NP-ATT [23] as HCIAE-M and HCIAE. The encoder models are attention-based. Interestingly, we found that both models achieve improvements with our proposed HAST. The results validate that: taking the impact of imperfect history into account, the history-aware models explore the contributions of "gold" history and deal with the relationship between visual and linguistic representations better.

---

[1] https://github.com/jiasenlu/visDial.pytorch

The qualitative results shown in Figure 3 and Figure 4 demonstrate the following advantages of our HACAN model with HAST:

History Sensitive. Our model is sensitive to the history. In Figure 3, the upper half part shows that, with historical changes, our model HACAN generates different responses. In more details, with the contribution of "gold" history, the "ground truth" answer achieves a higher rank order. Thanks to FCA and ECA, the visual attention influenced by historical changes is more precise when given the "gold" history. Rather, with historical changes, HCIAE behaves the same and does not benefit from the contribution of "gold" history.

Reliable Contextual Reasoning. Our model addresses contextual reasoning reliably using HAST. Focusing more on the influences of the different histories in the dialog, HACAN learns more contextual reasoning. In Figure 4, with different facts, HACAN chooses the corresponding regions in the image correctly.

5.5. Comparison with the State-of-the-art

We compared our model with the state-of-the-art methods on VisDial v0.9 and v1.0. Early methods use grid-based CNN(*e.g.* VGG-16 [32]) features. For fair comparison, we replaced the bottom-up attention features [2] with ImageNet pre-trained VGG-16 features. The upper half of Table 1 reports the results of the methods with the VGG-16 features, and the bottom half reports the results with bottom-up attention features.

Compared Methods. The state-of-the-art methods can be categorized into three groups based on the design of encoder: (1) Fusion-based Models (LF [6], HRE [6], Sync [11]). (2) Attention-based Models (MN [6], HCIAE [23], CoAtt [24]). (3) Approaches that deal with visual reference resolution in VisDial (AMEM [30], CorefNMN [19], RvA [16], DAN [16])

Our method outperforms the state-of-the-art methods across most of the metrics. Specifically, our method achieves more than 1 point improvement on R@1, and 1% increase on MRR. We also achieve a new state-of-the-art single-model on the official VisDial online challenge server[2]. Furthermore, we conducted supplementary experiments on the guesser task of GuessWhat?!. Table 2 shows that our method is comparable to the state-of-the-art methods.

6. Conclusion

In this paper we develop a codec model equipped with History-Aware Co-Attention Network (HACAN) for the visual dialog task. HACAN contains Feature-Wise Co-Attention module and Element-Wise Co-Attention module to address the co-reference and visual context in question

---

Model	Train err	Val err	Test err
LSTM [7]	27.9%	37.9%	38.7%
HRED [7]	32.6%	38.2%	39.0%
LSTM+VGG [7]	**26.1%**	38.5%	39.2%
HRED+VGG [7]	27.4%	38.4%	39.6%
ATT [8]	26.7%	33.7%	34.2%
Ours	**26.1%**	**32.3%**	**33.2%**

Table 2. Results on the guesser game of GuessWhat?!.

Model	MRR	R@1	R@5	R@10	Mean
HACAN w/o FCA	0.5837	44.52	74.77	84.84	5.56
HACAN w/o ECA-1	0.6181	48.29	78.23	87.76	4.77
HACAN w/o ECA-16	0.6285	49.26	79.41	88.72	4.53
HACAN w/o RS	0.6323	49.61	79.96	89.05	4.40
HACAN	**0.6391**	**50.44**	**80.67**	**89.71**	**4.32**

Table 3. Performance of ablative models on the validation set of VisDial v1.0.

Model	MRR	R@1	R@5	R@10	Mean
HCIAE-M w/o HAST [23]	0.6156	47.67	78.50	87.54	4.68
HCIAE-M [23]	0.6177	47.95	78.70	87.97	4.61
HCIAE w/o HAST [23]	0.6227	48.58	79.19	88.16	4.58
HCIAE [23]	0.6243	48.71	79.14	88.72	4.53
HACAN w/o HAST	0.6391	50.44	80.67	89.71	4.32
HACAN	**0.6445**	**51.20**	**80.76**	**89.92**	**4.25**

Table 4. Performance of ablative models on the validation set of VisDial v1.0. HAST indicates the usage of History-Advantage Sequence Training.

Model	Train err	Val err	Test err
HRED w/o HAST [7]	32.6%	38.2%	39.0%
HRED [7]	31.8%	37.7%	38.4%
HRED+VGG w/o HAST [7]	27.4%	38.4%	39.6%
HRED+VGG [7]	26.8%	37.7%	38.9%
HACAN w/o HAST	26.9%	33.6%	34.1%
HACAN	**26.1%**	**32.3%**	**33.2%**

Table 5. Performance of ablative models on the guesser game of GuessWhat?!. HAST indicates the usage of History-Advantage Sequence Training.

and history encoding. We propose a novel training strategy dubbed History-Advantage Sequence Training (HAST) that utilizes the history response to make the codec model more sensitive to the history dialog. Extensive experiments on the real-world datasets, VisDial and GuessWhat?!, achieve a new state-of-the-art single-model on the benchmarks.

Acknowledgments

References

[1] Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. Sort story: Sorting jumbled images and captions into stories. In *EMNLP*, 2016.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and vqa. In *CVPR*, 2018.

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

[4] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *NIPS*, 2014.

[5] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.

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

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

[8] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *CVPR*, 2018.

[9] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[10] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *CVPR*, 2015.

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

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

[13] Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. Towards an open-domain conversational system fully based on natural language processing. In *COLING*, 2014.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.

[15] Unnat Jain, Svetlana Lazebnik, and Alexander G Schwing. Two can play this game: visual dialog with discriminative question generation and answering. In *CVPR*, 2018.

[16] Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. Dual attention networks for visual reference resolution in visual dialog. *arXiv preprint arXiv:1902.09368*, 2019.

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[19] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *ECCV*, 2018.

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

[21] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *BEHAV BRAIN SCI*, 2017.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[23] Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *NIPS*, 2017.

[24] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016.

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

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

[27] Silvia Quarteroni and Suresh Manandhar. Designing an interactive open-domain question answering system. *NAT LANG ENG*, 2009.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[29] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.

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

[31] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, 2016.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[33] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016.

[34] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[36] Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *CVPR*, 2018.

[37] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.