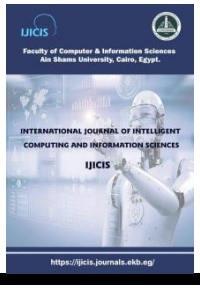
**IJICIS, Vol.24, No.2, 58-79**
**DOI: 10.21608/ijicis.2024.295310.1339**

**International Journal of Intelligent Computing and Information Sciences**

https://ijicis.journals.ekb.eg/



# A COMPARATIVE STUDY ON REINFORCEMENT LEARNING-BASED VISUAL DIALOG SYSTEMS

Ghada M. Elshamy
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
ghada.magdy@cis.asu.edu.eg

Marco Alfonse
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
marco_alfonse@cis.asu.edu.eg

Islam Hegazy
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
islheg@cis.asu.edu.eg

Mostafa M. Aref
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
mostafa.aref@cis.asu.edu.eg

Received 2024-06-04; Revised 2024-07-20; Accepted 2024-07-20

***Abstract:*** *Recently the conjunction between vision and language has created many intersecting tasks as visual question-answering systems, image captioning, etc. Specifically, dialog systems that depend on a visual scene play an important role in improving human-computer interaction technology. At the same time, reinforcement learning has emerged as a very successful paradigm for a variety of machine learning tasks, especially those tasks that aim to develop smart and humanoid machines. In this paper, we show how reinforcement learning is applied to conversational agents to build a powerful visual dialog agent. Visual Dialog task requires the agent to have a meaningful conversation about visual content in natural language. For a given image, its caption, dialog history (question/answer pairs), and a question about this scene, the agent should comprehend the question, extract the relevant context from the history, and ground this information on the image to correctly answer the current question. Two main visual dialog tasks have been introduced which are a free-form dialog task known as "Visual Dialog" and a goal-oriented dialog task formulated as a guessing game. Two datasets have been introduced to address these tasks which are VisDial dataset and GuessWhat?! datasets. For evaluation, some approaches use the accuracy metric while others use four metrics that have been proposed for the sake of this task. Several approaches are proposed for tackling this task based on supervised learning or reinforcement learning or even combining both techniques. This paper represents a comparative study of eleven important reinforcement learning approaches for visual dialog.*

***Keywords:*** *Visual Dialog, Guessing Game, Guess What?!, Guess Which, Attention Mechanism.*

***Corresponding Author**: Ghada M. Elshamy
Computer Science Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt
Email address: ghada.magdy@cis.asu.edu.eg

## 1. Introduction

Recently, the artificial intelligence (AI) field has witnessed great progress in the conjunction between vision and language tasks. This intersection began with the introduction of the image captioning systems [1], video description [2], storytelling [3], visual question answer (VQA) systems [4], and visual dialog task which is our focus here. The visual dialog (VD) task was first introduced by Das et a. [5] to build an AI agent that is capable of seeing, understanding, and talking in a natural language like humans.

The main goal of the visual dialog task is that for a given image *I*, dialog history (*H),* and a question (*Q)* about the image, the agent has to correctly answer this question by understanding the context from the history and relating this information to the image. Two paradigms have been introduced for this task: a free-form dialog task and a goal-oriented dialog task. In the former, the agent should answer the question of its peer about the given image, while in the latter, the dialog is held between two AI agents, i.e., an oracle (A-bot) and a questioner (Q-bot). The second one is usually formulated as a guessing game which was first introduced by De Vries et al. [6]. In this game, the A-bot produces answers while the Q-bot generates questions (q-gen) and guesses the target object (guesser) from the image or even an image from a pool of images. Different approaches have been proposed to address this task including reinforcement learning (RL) approaches which are the focus of this research.

Whatever the setting of the game, free-form, or goal-oriented, the agent could produce a question/answer via either a discriminative setting or a generative setting. In the discriminative setting, the agent learns to choose the correct sentence from a pool of candidate sentences. In the generative setting, the agent learns to generate a sentence as relevant and informative as possible according to the image and textual information. The paper is structured as follows: section 2 lists the different reinforcement learning approaches for visual dialog tasks. Section 3 describes the benchmark datasets for the VD task. Section 4 describes the evaluation metrics used by the proposed approaches. Section 5 represents a summary and discussion of the mentioned approaches. Finally, section 6 represents the conclusion and future work.

## 2. Reinforcement Learning-based Visual Dialog Approaches

Das et al. [7] introduced the first reinforcement-based paradigm for goal-driven visual dialog agents. They introduced a cooperative game between two AI agents which were the Q-bot and the A-bot. Given a pool of images, the two agents began to communicate in natural language together about an image that was revealed only to the A-bot while the Q-bot only had a 1-sentence caption. The Q-bot asked a series of questions about the target image while the A-bot answered the former questions till the Q-bot guessed the correct image. To apply the proposed approach, they first trained the agents to communicate with ungrounded vocabularies on a synthetic dataset via pure RL training. This dataset was composed of symbolic representations (e.g. X, Y, Z) that have no pre-specified meaning. During the training, they observed that the agents built mutual understanding as they mapped each symbol to defined attributes (e.g. X: color, Y: shape, Z: style, 1: red, etc.) and communicated in that way as shown in Figure 1.

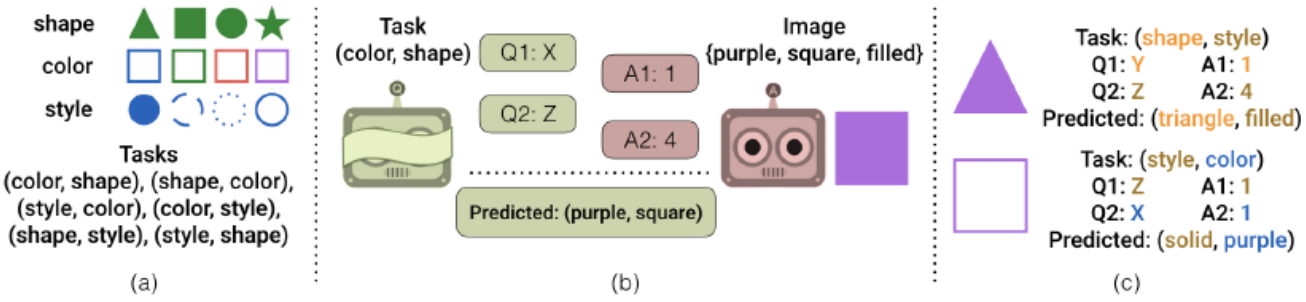



Figure. 1: The sanity check game where (a) represents dataset image attributes and the game tasks. (b) agents interaction during pure RL training. (c) represents how grounding emerges [7].

This step enabled the agents to build their communication way so that they have a strong understanding of each other. Secondly, they initiated the game using VisDial v0.5 [5] dataset via supervised learning (SL) and curriculum learning. The policy network proposed is illustrated in Figure 2. They evaluated their work through the discriminative setting on A-bot's ability to rank the correct answer from the given 100-candidate answers.

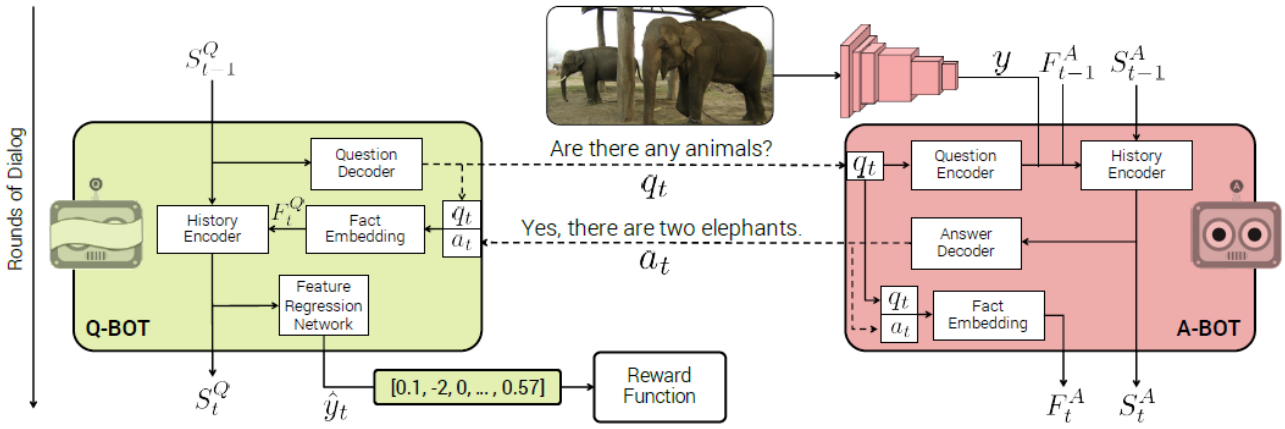



Figure. 2: Policy network for A-bot and Q-bot proposed by [7].

They observed that their proposed agent slightly outperforms the SL agent regarding the evaluation metrics, yet the generated answers were more informative. Also, the Q-bot guessing ability improved as it got the correct guess by 95%. Although RL has proven its efficiency, the model performance degraded gradually at a certain round due to repeated questions.

For this reason, Murahari et al. [8] introduced an additional objective to the previously proposed model by [7] to reduce this repetition hence giving the dialog more lifetime. They formulated a penalty called the "**smooth-L1 penalty**" which aimed to minimize the similarity between generated dialog state embeddings. The Smooth-L1 penalty was applied to the Q-bot states $(s_{t-1}^{Q}, s_t^{Q})$ aiming to maximize this penalty on $\Delta_t$ in addition to the overall objective function. Eq. (1) illustrates the smooth-L1 penalty term:

$$\Delta_t = abs(\|s_{t-1}^{Q}\|_2 - \|s_t^{Q}\|_2)$$
$$f(\Delta_t) = \begin{cases} 0.5\Delta_t^2, & \Delta_t < 0.1 \\ 0.1 - (\Delta_t - 0.05), & otherwise \end{cases} \quad (1)$$

where $\sum_{t=2}^{N} f(\Delta_t)$ was the additional objective term and N is the number of dialog rounds. This penalty was maximized to force the Q-bot to ask diverse questions by keeping $\Delta_t$ as large as possible so that the A-bot can explore more state space during RL training. They managed to produce a more diverse dialog with slight improvement regarding the evaluation metrics.

Strub et al. [9] proposed a deep RL approach for the goal-oriented visual dialog task. The motive of their proposed model was to optimize the q-gen generated questions via an RL training environment. To achieve this goal, they formulated GuessWhat?! game as a Markov Decision Process (MDP) with the REINFORCE policy gradient method. Firstly, they trained the oracle, the q-gen, and the guesser separately with cross-entropy loss in the SL paradigm. This step aimed to train the oracle and the guesser to act as the environment that would reward the q-gen in the RL step. While the q-gen was trained in the SL step to get a reasonable initial policy value for the RL step. For the oracle, they used the same network proposed by [6] where the input embeddings were concatenated as a single vector representation and fed into a single multi-layer perceptron followed by a softmax layer for the final answer distribution. The q-gen and the guesser, each was implemented as a single LSTM layer to process the input embeddings followed by a softmax layer to get the distribution over the tokens in the case of the q-gen or to guess the object in the case of the guesser.

Switching to RL, the environment was the pretrained (in SL) oracle and guesser modules. The agent was the q-gen which they aimed to improve, the policy function was the REINFORCE method and finally, the reward function was a zero-one reward where it returned 1 if the guessing was correct and zero otherwise. To handle the policy gradient variance problem, they developed a baseline function of the current state ($b_{\phi_h}(x_t)$). This baseline was a single MLP layer, its input was the hidden states of the q-gen LSTM network and was trained to predict the expected reward value. The training objective was to minimize the loss between the expected reward and the discounted reward as shown in Eq.(2).

$$L(\phi_h) = \langle [b_{\phi_h}(x_t) - \sum_{t'=t}^{T} \gamma^{t'} r_{t'}]^2 \rangle_{\tau_h} \tag{2}$$

Their proposed model managed to achieve a 10% accuracy enhancement on GuessWhat?! dataset over previous works on the same task. As limitations, they observed that the REINFORCE policy method did not provide good exploration and exploitation abilities for the q-gen agent, whereas the agent at some point, kept repeating the same question thus, it was not able to stop the conversation. In addition, it sometimes stopped the conversation too early especially when there were many objects in the image of the same category.

Considering the policy optimization limitations stated by [9], Rui Zhao and Volker Tresp [10] proposed a novel policy gradient approach class called Tempered Policy Gradients (TPGs) to improve goal-oriented dialog agents. To handle the exploration and exploitation problem, they introduced three instances from the TPGs class: single, parallel, and, dynamic. In addition, to improve the guessing ability of the guesser, they integrated a memory attention mechanism into the guesser network as shown in Figure 3. Also, they used a modified seq2seq model for question-level training in the q-gen network to improve the quality of the generated questions as shown in Figure 4.

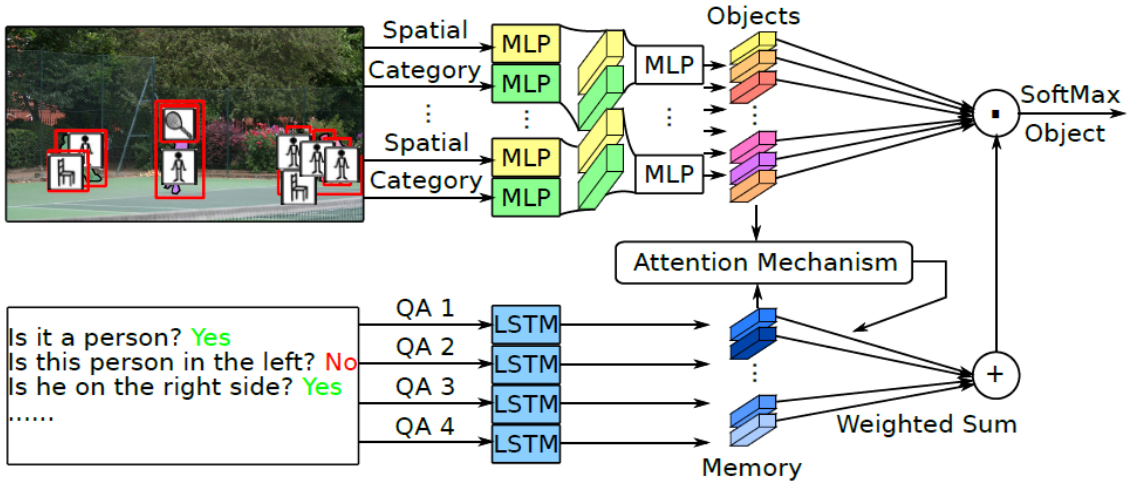



Figure. 3: The Guesser module proposed by [10]

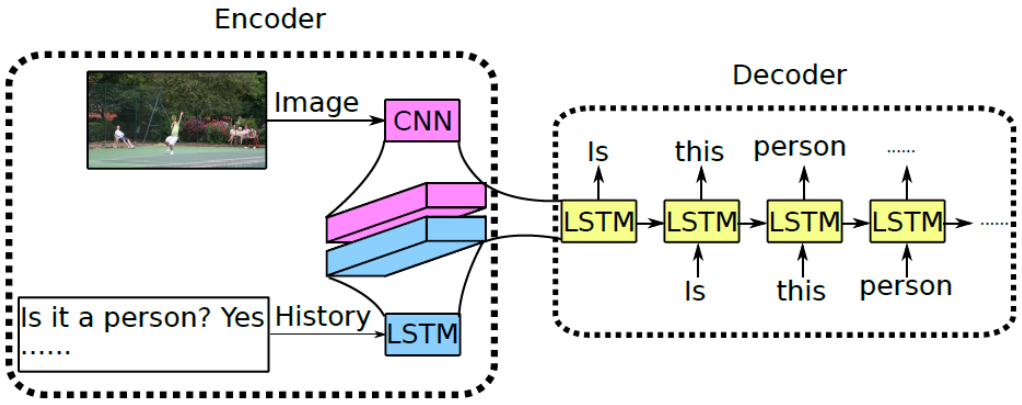



Figure. 4: The q-gen module proposed by [10]

For each TPG instance, a temperature (τ) parameter was introduced, it was a hyperparameter that affected the probability distribution of the sampled tokens for the generated question. Higher values of τ (i.e. > 1) gave more diverse and coherent text and reduced the repetition, while lower values (i.e. < 1) gave more deterministic text, however in their case, higher values were the choice. Single TPG had one temperature parameter where τ > 1 that was applied for all sampled tokens. The parallel TPG used multiple single TPGs concurrently each with a different τ parameter. Last but not least, the dynamic TPG used a heuristic function to assign the τ value for each token distribution. Parallel-TPG gave better results than single-TPG, yet it required more computational power. On the other hand, Dynamic-TPG gave better results than the other two and with less computational power than the parallel-TPG. The three modules, the oracle, the q-gen, and the guesser, were trained in a supervised manner as usual and then switched to the RL environment.

The difference between this work and [9] was that the RL here was turned on for both the q-gen and the guesser modules while the oracle was the RL environment. For each game episode, if the guesser had guessed the object correctly the reward function returned 1, otherwise, it returned zero and the guesser parameters were updated using the REINFORCE rule. For the q-gen, the reward was distributed uniformly among all generated question tokens and was updated using the TPGs instances. The

proposed model was evaluated using GuessWhat?! dataset and their powerful Dynamic-TPG, it outperformed all its previous models reaching the highest accuracy of 74.31%.

Zhang et al. [11] proposed a new approach for the visual dialog task with two main contributions which were a novel framework based on multimodal hierarchical RL (HRL) and a state adaptation technique for dialog state representation improvement. The HRL resembled Feudal RL where it used a hierarchical dialog policy that combined two RL modules within a control flow using variants of Deep Q-Network (DQN) and Deep Reinforcement Relevance Network (DRRN). Double DQN was used to learn the higher-level policy for either question selection or image guessing. DRRN was used to have better performance when dealing with discrete natural language action space. The proposed framework was a discriminative agent that should select the most relevant question from a given pool of questions. The game framework was composed of four main modules; visual dialog semantic embedding, visual dialog state tracking, hierarchical policy learning, and question selection as shown in Figure 5. The last module was the **Simulator** which simulated the game environment for the proposed framework. The game was constructed from 1000 sets where each set consisted of two similar images from VisDial dataset. The simulator also tracked the game state and provided the reward signals and answers related to the target image.

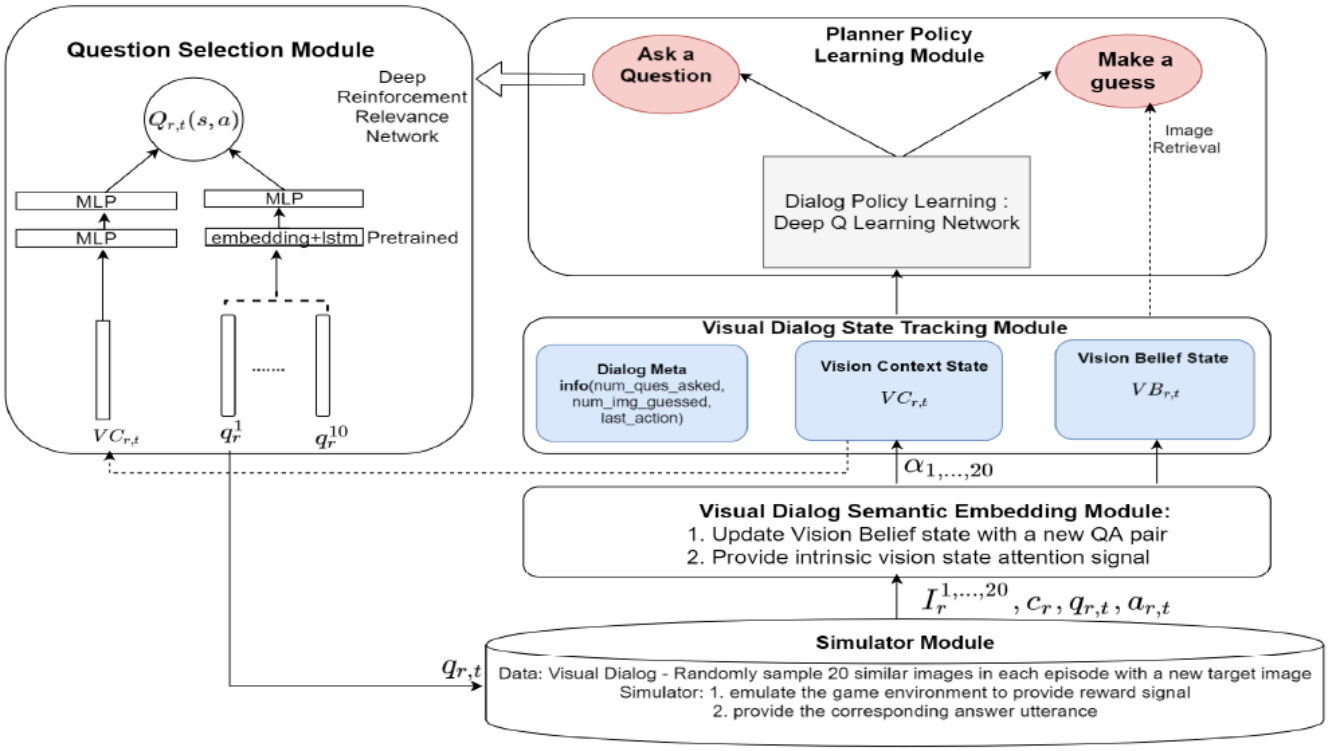



Figure. 5: The architecture of the proposed model by [11]

This work was evaluated by two novel metrics which were Win Rate and Average Number of Dialog Turns. Win Rate was the average of the summation of the prediction accuracy across 1000 sets of the game. They achieved a win rate of 76.3% on 7.22 Avg. turns. They had observed that their proposed model had the efficient ability for decision-making when compared to the oracle baselines proposed before which kept asking till the rounds terminated and then performed the gussing. This means that their agent had strong context awareness and environment adaptation ability.

Wu et al. [12] proposed a novel approach to improve the quality of the generated responses to be more human-like. They used Generative Adversarial Networks (GANs) in an RL framework. Regarding the GAN architecture, this work trained two sub-networks in an adversarial manner; generator and discriminator. The generator network was responsible for generating human-like responses based on the given image and dialog history. While the discriminator network used the previous outputs of the generator to differentiate between real and generated responses. In the reinforcement part, the generator aimed to fool the discriminator into believing that the former output was real. On the other hand, the discriminator output was used as a reward to the generator network to improve its generation ability. As technical contributions, they defined a sequential co-attention mechanism for the generative model and a memory attention mechanism for the discriminator model that depends on image, question, and dialog history features. The sequential co-attention encoder is illustrated in Figure 6.

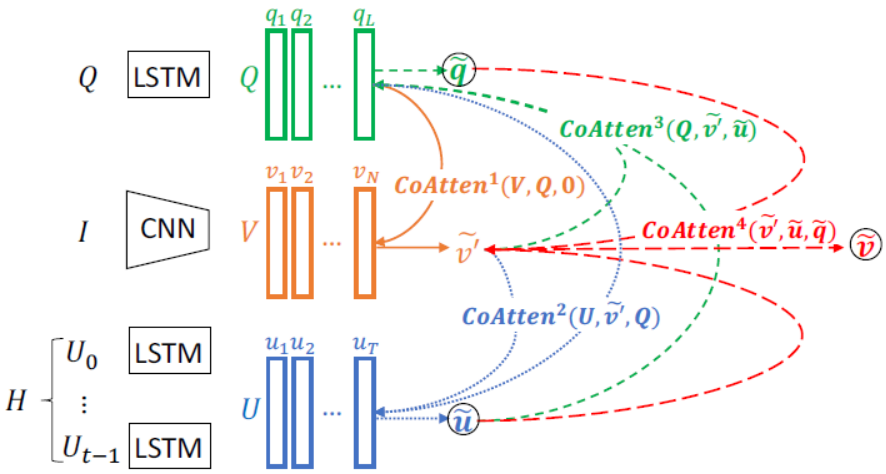



Figure. 6: The sequential co-attention encoder proposed by [12]

As we can see, each input attribute was co-attended on the other two attributes (e.g. image features were co-attended on both question and history features, etc.) to jointly reason over them. The co-attended features were shared with the discriminator network to enable it to focus on the region of interest in the image and the informative parts from the textual information.

Regarding the RL framework, the output of the discriminator was used as the reward for the generator to improve its generation quality. To achieve better results, instead of using one reward for the whole generated sequence, the Mont-Carlo (MC) search was applied with a REINFORCE roll-out policy π. This technique enabled generating probability for each generated token so that a different reward was calculated for each token. The objective of using such a technique was to give a high reward to the relevant token and a low one to the irrelevant token. The total reward is illustrated in Eq. (3) as:

$$r_{ak} = \frac{1}{N}\sum_{n=1}^{N} r\left(\{\tilde{v}, \tilde{u}, Q, \hat{A}^{n}_{1:K}\}\right) \tag{3}$$

where $N$ was the number of the generated tokens. In Addition, to build a generator as stable and efficient as possible, the "Teacher Forcing" strategy was applied in which the generator was fed with a

combination of the ground truth human responses and its own generated ones. The full model architecture is shown in Figure 7.

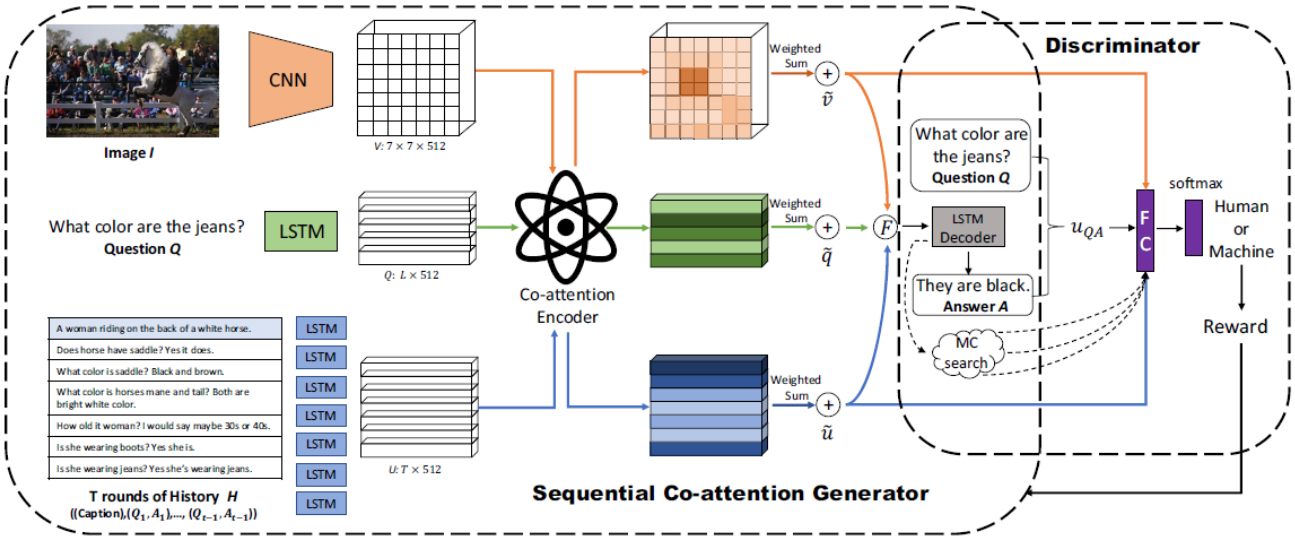



Figure. 7: The adversarial learning framework of the proposed model by [12]

The model was evaluated on VisDial dataset v0.9 with the same predefined evaluation metrics; MRR, Recall@k, and Mean Rank. For their generative model, they compared their work with state-of-the-art models, and it outperformed all previous work by 3.81% on R@1. In addition, their discriminative model outperformed the state-of-the-art ones by 1.81 on R@1 and 1.96 on R@5.

Y. Chang and W. Peng [13] had a different opinion about the REINFORCE policy method from [9], [10] regarding its effectiveness on the questioner (asking and guessing). From their point of view, pretraining the questioner using SL was not effective in achieving the optimal policy which would guarantee a satisfying performance. Relying on this hypothesis, they pretrained their proposed model using imitation learning (IL) and then finetuned it in the RL step using REINFORCE policy method. As imitation learning requires experts to guide the agent during the learning process, they developed two experts i.e., information gain and target posterior experts, based on the information theoretic approach known as “Answerer in Questioner Mind (AQM)” [14]. The AQM role was to help the q-gen to ask questions that would reduce the uncertainty for guessing the target object. This was done by adding an approximate oracle, a probabilistic model that simulates the true oracle behaviour, inside the Q-bot network. The information gain expert (IGE) aimed to maximize the amount of the information gained by the q-gen to generate more relevant and coheret questions. The target posterior expert (TPE) guided the q-gen to generate questions that would lead to the correct gusseing. Both experts used the answer generated from the approximate oracle to achieve better policy during training the q-gen and the guesser. Figure 8 shows the difference between the old and the proposed guesser model with the approximate oracle.

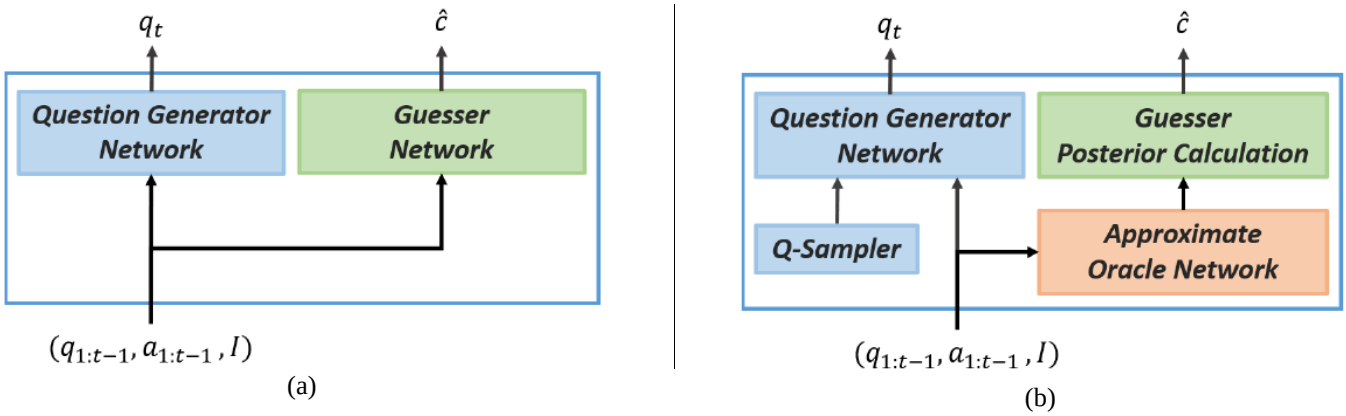



Figure. 8: (a) Q-bot proposed previously by [9]. (b) the proposed Q-bot by [13].

After the imitation learning step, the RL step was turned on to improve the Q-bot to successfully predict the correct object. The Q-bot model took the RL agent role while the true oracle took the environment role that rewarded the former for its behavior. The REINFORCE policy method was used with zero-one reaward. Experiments were conducted on GuessWhat?! dataset, the proposed questioner model was evaluated with three different oracles; indA, depA, and trueA. indA was trained from the training data, and depA oracle was trained on images and questions from the training data with the answers from the true oracle. trueA was the same as the true oracle. The questioner was pretrained with the different oracle settings for 5 IL question/answer rounds (4 with IGE and 1 with TPE) and then optimized by RL from question/answer rounds ranging from T =5 to T=10. trueA oracle model outperformed indA and depA oracles, with a prediction accuracy of 82.45%.

Fan, Hehe, et al. [15] introduced a novel approach for the visual dialog task. They proposed a recurrent attention network with an attention processor for memorizing temporal textual information and spatial information respectively. They followed the codec architecture (encoder-decoder) model. The encoder was composed of their proposed modules which were a recurrent attention network and attention processor. The decoder was composed of generative (*G*) and discriminative (*D*) networks. *G* was trained on sentence-level training with the guidance of D in a reinforcement manner where D's output was the reward for *G*. This methodology aimed to overcome the problems with word-by-word level training found in previous work. The model was evaluated on VisDial v0.9 dataset.

Firstly, the encoder was responsible for encoding input information from the image and the dialog history (*I* and *H*) and attending to the most relevant and important parts of the current question ($q_t$). The dialog network and the attention processor were the main components of the encoder module. The dialog network modeled the dialog history (question-answer pairs) rounds as a temporal context via an LSTM network. The Image features and the caption features were concatenated and used as initial input for the dialog network. The attention processor attended to the spatial information highlighted by the output signal from the dialog network to generate a state vector for the decoder. Secondly, the decoder module, either the generative or the discriminative one, used the encoder's output state vector to get an answer for ($q_t$). Regarding reinforcing $G$ by $D$, $G$ and $D$ were trained in two different ways, joint training, and reinforced training. In joint training, each of them was updated separately as shown in Figure 9 (a). In RL, *D* transferred its knowledge to improve *G* by learning the optimal policy that would maximize the reward (*r)* provided by *D* as illustrated in Figure 9 (b).

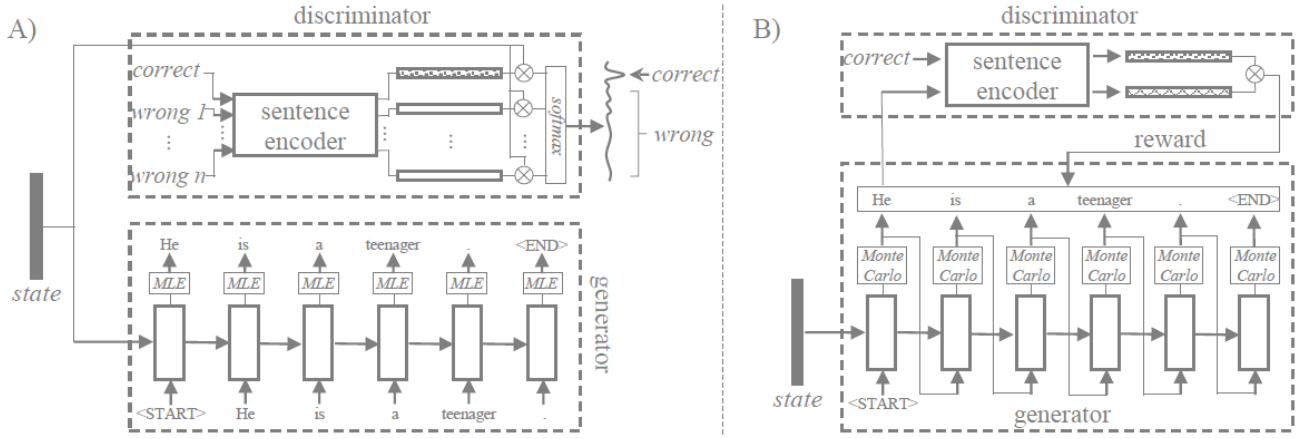



Figure. 9: (a) The joint training of both D and G. (b) Reinforcement learning between *G* and *D*. [15]

For $G$'s training, sentence-level training was applied as a word sequence generation process. The objective of this RL training was to learn the appropriate policy for maximizing $r$. At training time, the hybrid loss shown in Eq. (4) was minimized.

$$L = \alpha L_D + \beta L_G - \gamma J \tag{4}$$

where $L$ was the total loss function, $L_D$ *and* $L_G$ were the *D's* and the *G's* losses respectively. $J$ was the RL policy objective function. And $\alpha, \beta, \gamma$ were three positive factors. The model was evaluated on VisDial dataset v0.9 with the same predefined metrics by [5]. Their proposed model (attention + joint + RL) achieved promising results but not as good as [12].

Zhao et al. [16] proposed a novel approach for goal-oriented visual dialog task. It was based on an image guessing game "GuessWhich", which was first introduced by [7]. It was composed of two AI agents Q-bot and A-bot and followed the same training process proposed in [7]. An Attentive Memory Network (AMN) was proposed to alleviate the generation of irrelevant and repeated questions. Their proposed model outperformed the state-of-the-art methods on VisDial v1.0 dataset. The proposed methodology was a round of interaction between Q-bot and A-bot as shown in Figure 10. The main contribution was to inject an AMN into A-bot and Q-bot networks. The $AMN^Q$ of the Q-bot was composed of a memory module ($M_t^Q$) and a fusion module. $M_t^Q$ was queried by the embedded fact ($f_t^Q$) to produce an attention weight vector ($W^Q$) to compute the attention history vector ($h_{t-1}^Q$). The fusion module took the advantage of the caption $C$ across all rounds along with ($h_{t-1}^Q$) and ($f_t^Q$). The $AMN^A$ of the A-bot was composed of a memory module and a fusion module like $AMN^Q$. The A-bot memory module ($M_t^A$) was queried by the embedded fact ($q_t$) to produce an attention weight vector ($W^A$) to compute the attention history vector ($h_{t-1}^A$). The fusion module fused ($q_t, h_{t-1}^A$) and the features of the input image.

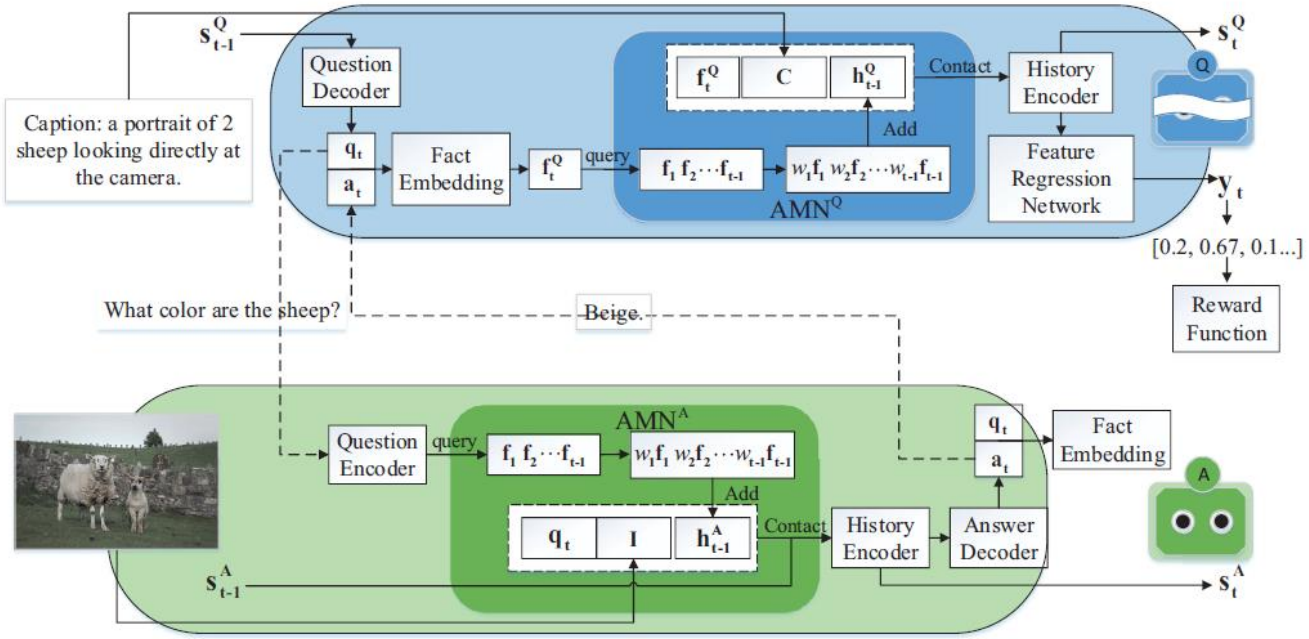



Figure. 10: The framework proposed by [16]

This newly injected AMN helped the Q-bot to generate more diverse and relevant questions and enabled the A-bot to generate more precise and informative answers. The proposed framework was evaluated on VisDial v1.0 dataset using the same evaluation metrics introduced by [5] This work outperformed the state-of-the-art on dialog generation across all metrics. Their model archives 10% and 30% enhancement on R@10 and MRR respectively.

Zhao et al. [17] proposed a structured knowledge-aware network (SKANet) to improve reasoning ability by incorporating the commonsense knowledge derived from ConceptNet [18]. ConceptNet is an external knowledge base that enables machines to understand the meanings of words and entities and their relations in natural language. This SKANet consisted of an image knowledge-aware module and a caption knowledge-aware module. They applied their proposed approach to the two VD sub-tasks: free-form visual dialog in the SL paradigm and the guessing game in the RL paradigm. For the free-form VD sub-task, their proposed architecture consisted of three sub-modules which were the two sub-modules of the SKANet and a multi-modality fusion module as shown in Figure 11. The objects' features, question's features and the dialog's features were processed through the multi-modality fusion attention mechanisms that grounded the question relevant information from the dialog history on the visual features to generate the multi-fused features $F^{vh}$. At the same time the visual features were fed into the image knowledge-aware module to build the image sub-graph to produce the visual knowledge features $F_v^g$. Also the caption features were firstly processed via a concept grounding to perform tokens matching from the ConceptNet graph, then the resulted tokens were fed to the caption knowledge-aware module to generate caption knowledge features $F_c^g$. The two generated features $F_v^g$ *and* $F_c^g$ were then queried by the question to focus on the question relevant features.

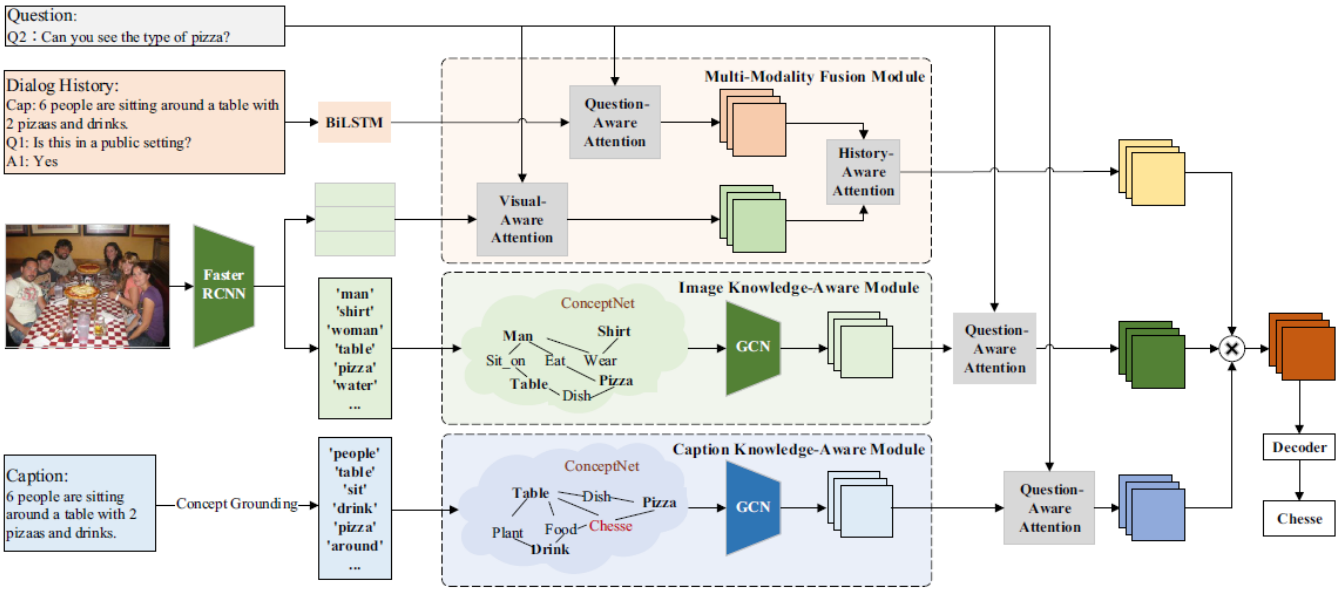



Figure. 11: SKANet architecture proposed by [17]

There were two steps to build the knowledge-aware graphs which were graph construction and GCN encoding. The graph construction was a collection of nodes created by matching the visual objects and entities of the image and caption from the ConceptNet. Then the constructed graph was encoded via GCN [19] that extracted the relational context among the graphs' entities. Finally, the three generated features $F^{vh}, F_v^g$ *and* $F_c^g$ were fused through element-wise multiplication and the resulted features were fed into the decoder. For the guessing game, the SKANet sub-modules were applied to the A-bot and Q-bot separately. The image knowledge-aware module was applied to the A-bot network to emphasize its ability to resolve the visual scene entities. The caption knowledge-aware module was applied to the Q-bot to generate more relevant and diverse questions and also to make the correct image guess. They trained the A-bot and Q-bot with SL followed by RL in a curriculum learning paradigm same as [7] did.

They evaluated their proposed approaches on VisDial dataset v0.9 and v1.0 test sets for the free-form VD sub-task using the same predefined metrics by Das et al. [5]. They observed that their proposed model could deal with complex scenarios that their prior work had failed with. For the guessing game, they evaluated their model on VisDial v1.0 validation set on its ability for answer generation task.

Xu et al. [20] saw that to improve the generated questions quality, the objects' categories information should be considered. They proposed a question generation model called object category visual dialog (OCVD) to improve the question generation diversity and coherency for the GuessWhat?! game. They developed their proposed methodology through four steps as shown in Figure 12. First, they employed an object information extraction module to extract not only the object features and their bounding boxes as prior works did but also their categories information. This step aimed to emphasize the comprehension of the input objects. Second, they used the extracted information in the category selection process where the categories were sorted based on their appearance frequencies in the image to identify which category was more likely to be the target object's category. Then, they recursively applied the selection mechanism on the candidate categories list for each round.

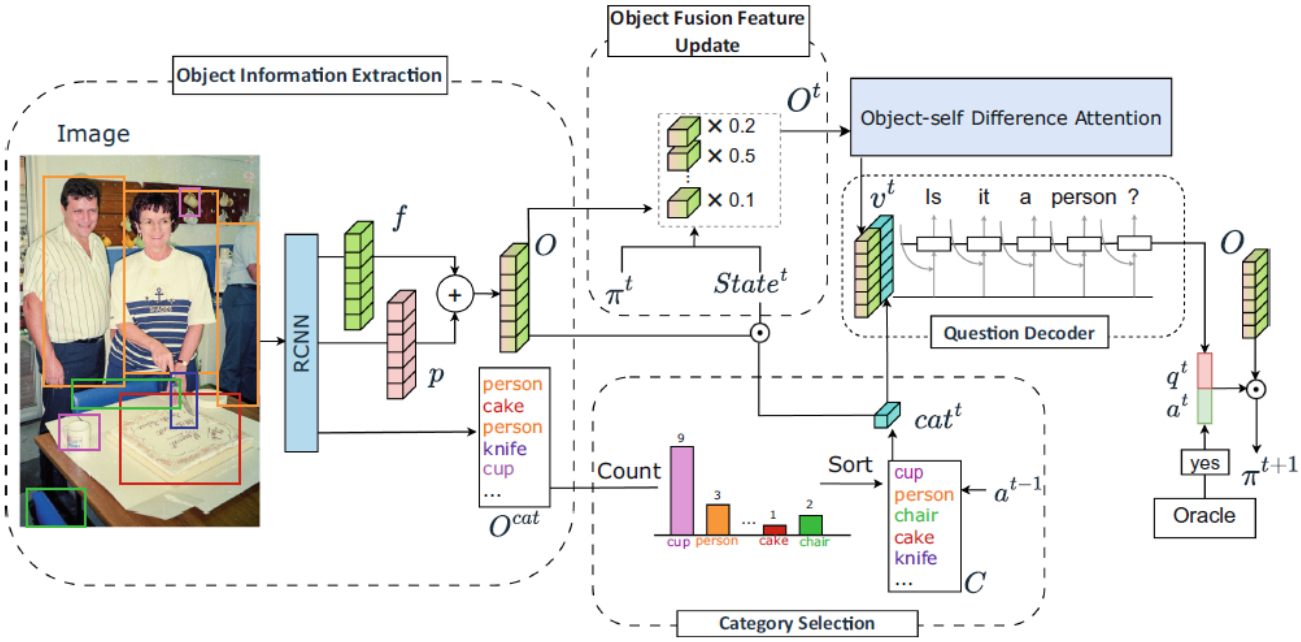



Figure. 12: The OCVD approach proposed by [20]

At the same time, they encouraged the agent to focus on the object's features that belonged to the selected category. This was done by calculating the similarity score between the objects' features ($O$) and the selected category ($cat^t$) via a softmax layer to get the category-level attention distribution ($Score^t$). In addition, they calculated the object-level attention distribution ($\pi^t$) based on the question-answer pair and the object's features of the previous round. They used both $Score^t$ *and* $\pi^t$ to calculate the update function for the object features representation ($O^t$). Third, they obtained the final visual features representation via object-self difference attention mechanism to capture the visual differences between the objects' features. Fourth, the obtained visual features were concatenated with the selected category features for the current round and fed into the question decoder to generate a more improved and coherent question sentence.

They evaluated their model on GuessWhat?! game under different settings. They managed to achieve a 72.1% success rate on new objects and a 67.9% success rate on new games, which indicated the effectiveness of the OCVD approach. New objects were images from the training set but with different target objects. New games were images with target objects from the test set.

## 3. Dataset Description

Two datasets have been used as the benchmark for the VD tasks which are VisDial and GuessWhat?! datasets. Both datasets were collected using Amazon Mechanical Turk (AMT). The VisDial dataset was first introduced by [5] and it has three versions: v0.5, v0.9, and v1.0 (used in recent works). It contains one dialog/image (10 question-answer pairs) on images from MS-COCO [21] dataset. VisDial dataset serves the idea of training an AI agent to be able to produce an answer for the given question during the conversation. However, some of the mentioned works have manipulated VisDial dataset to be used in a guessing game rather than training an agent to answer a given question during the conversation. A sample from VisDial dataset is represented in Figure 13.


**Caption**: A statue depicting a bear breaking into a car.
**Person A (1)**: how big is statue
**Person B (1)**: about size of real full grown bear
**Person A (2)**: so is car full size then as well
**Person B (2)**: yes replica of car
**Person A (3)**: is statue all 1 color
**Person B (3)**: no brown and black
**Person A (4)**: what color is car
**Person B (4)**: dark red
**Person A (5)**: where is this, do you think
**Person B (5)**: in wooded area someplace
**Person A (6)**: do you see any people in image
**Person B (6)**: yes 1 man
**Person A (7)**: how old is man
**Person B (7)**: 35-40

Figure. 13: Example from VisDial dataset [5]

The GuessWhat?! dataset was first introduced by [6], it consists of ~155K dialogs with a total of ~800K visual question-answer pairs on ~66K images. GuessWhat?! dataset was introduced to serve as a cooperative guessing game between two AI agents, an oracle (A-bot) and a questioner (Q-bot). The game was about guessing an object from a given image through conversational question answering. In the beginning, both agents could see the same image, however, the target object was only visible to the A-bot while the Q-bot should guess what was this object. The Q-bot began to questionnaire its peer during the conversation to be able to deduce the target object and the A-bot answered the former's questions with either "yes" or "no" as illustrated in Figure 14. When the Q-bot was satisfied with the gained information, it stopped asking and started to guess the object. If it was a correct guess, it won otherwise it lost with a penalty in return to force the Q-bot to ask more relevant questions.

The challenge here is mainly the role of the Q-bot, as it performed two subtasks, i.e., generating questions (q-gen) and guessing the target object (guesser). The first subtask required it to ask relevant questions as much as possible without being affected by previously irrelevant generated questions (i.e., visual contextual comprehension challenge). The second one required the agent to be aware of the important parts in the context to be able to guess the correct object (i.e., multi-modal reasoning challenge).

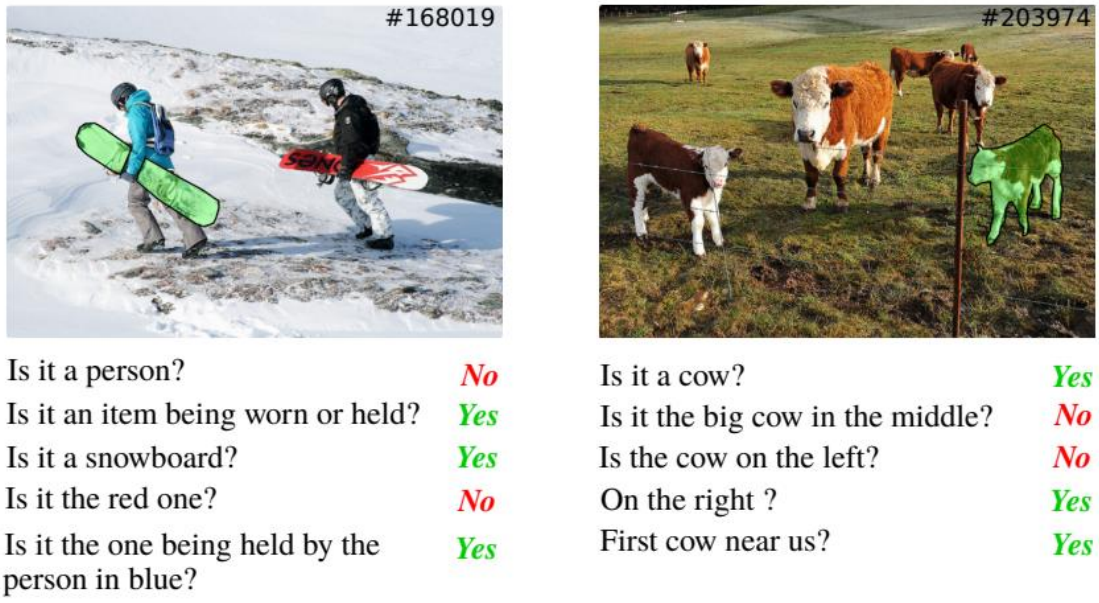



Figure. 14: Example of GuessWhat?! dataset [6]

The goal of this game was to build agents that have cognition and understanding abilities as much as humans. Many experiments on this game were mainly conducted on GuessWhat?! dataset. Table 1 presents a full description of both datasets.

Table 1: Dataset Description.

| Dataset | | No. of Images | Dialog Pairs | Answer Distribution | Dataset Distribution | | |
|---|---|---|---|---|---|---|---|
| | | | | | Training | Validation | Testing |
| **VisDial [5]** | v0.5 | ~ 68k images from COCO-train-val v0.5. | 10 rounds(q-a pairs) per image | - | ~ 50,729 (~75%) | ~ 7,663 (~11%) | ~ 9,628 (~14%) |
| | v0.9 | ~ 123k COCO-train-val v0.9. | 10 rounds (q-a pairs) per image | | ~ 80,000 (~65%) | ~ 3000 (~2%) | ~40000 (~33%) |
| | v1.0 | ~ 134,000 images from MS-COCO dataset. | ~ 1.4M question/answer pairs (10 rounds (Q\A-pairs)/image. | | ~ 92% | ~ 2% | ~ 6% |
| **GuessWhat?! [6]** | | 66,537 images and 134,073 objects from MS-COCO dataset | 821,889 question/answer pairs (~ 12 rounds/image) | No: 52.2% Yes: 45.6% NA: 2,2% | 70% | 15% | 15% |

## 4. Metrics

Many metrics are used to evaluate visual dialog tasks including model accuracy, mean rank (Mean), mean reciprocal rank (MRR) and recall @k (R@K). Das et al. [5] introduced the last three metrics as retrieval metrics. These retrieval metrics rely on the concept of ranking that tries to rank a list

of items (e.g. a list of 100 candidate answers) based on their relevance in the given task (answer generation or prediction). The mean rank measures the overall individual ranks. It lies on the interval $MR \in [1, \infty)$ where a lower value is better and is computed as shown in Eq.(5).

$$MR = \frac{1}{|I|}\sum_{r \in I} r \tag{5}$$

where *I* denoted the set of all ranks and *r* is the individual rank. The mean reciprocal rank measures the arithmetic mean of the MR, which can be calculated as the inverse of the mean of ranks as shown in Eq. (6). It lies in interval MRR∈(0,1] where a higher value is better.

$$MRR = \frac{1}{|I|}\sum_{r \in I} r^{-1} = \left(\frac{|I|}{\sum_{r \in I} r^{-1}}\right)^{-1} \tag{6}$$

The recall @K measures the existence of human response in the top-k-ranked answers as shown in Eq. (7). The lower value is better.

$$R@K = \frac{1}{|I|}\sum_{r \in I} \mathbb{I}[r \leq K] \tag{7}$$

## 5. Discussion

Visual dialogue is a complex vision language task that processes the visual and textual features simultaneously for sentence generation. To attain this task, the agent must acquire the abilities of contextual comprehension, multimodal reasoning, visual coreference resolution, and relationship mining. Specifically, the VD task requires the model to understand the intent of the question in terms of the relevant regions in the image and be guided by the dialog history. In addition, the agent must be able to resolve the pronouns to its previously mentioned entities and relate them to the objects in the image, which is known as the visual coreference resolution. Therefore, we can agree that generalization visual coreference resolution on multimodal features along with visual coreference resolution, represents considerable challenges for the VD task. Moreover, VD may suffer from a lack of generalization ability and robustness due to the dataset bias. Whereas the visual dialog model may overly rely on the relation between the question and answer and remember the pattern between them, therefore ignoring the exploration of the image content. Solving the cross-modality bias and enhancing the generalization ability leads to a better visual dialog system.

Here we focus on the reinforcement learning-based approaches for visual dialog task. After studying these approaches we can categorize them into two categories: goal-driven conversation and free-form conversation. Also, we discussed the benchmark datasets for VD which are VisDial dataset and GuessWhat?! dataset. Let's agree on that, to use RL in this task the agents should be pre-trained before applying RL techniques to provide a kick-start policy for the agent due to the wide action space for this task. All mentioned work above has pre-trained their model with different techniques and then finetuned it with RL. Table 2 and Table 3 represent performance measures for the mentioned RL techniques in this survey.

In our opinion, we see that the reinforcement learning approach is much more suitable to address this task than the SL approach. RL addresses the decision-making process for an agent without human guidance. The agent should decide what to do to perform the given task to maximize its reward. For the VD task, an AI agent is required to engage in a conversation in the form of a question-answer in a human-like manner. In this task, the AI agent tries to learn the optimal behavior when interacting in the environment to obtain the maximum reward and get the task done. The optimal behavior varies according to the setting of the VD task here either asking or answering questions. However, in the end, the agent should behave like a human as much as it could. For this reason, we find that RL is more appropriate for the VD task.

Table 2: Comparative results of the mentioned RL work using accuracy metric.

| Reference | Dataset | Methodology | Accuracy |
|---|---|---|---|
| **Stub et al. [9]** | GuessWhat?! (val \ test - sets) | Deep RL | 53.3% -52.3% |
| **Rui Zhao and Volker Tresp [10]** | GuessWhat?! | Deep Neural Network + Deep RL + TPGs + Memory-Attention (tow-hop attention) for the Guesser model. | 74.31% |
| **Zhang et al.** [11] | VisDial | Multimodal hierarchical RL (double DQN + DRRN) and a state adaptation technique for dialog state representation improvement. | 76.3% |
| **Y. Chang and W. Peng [13]** | GuessWhat?! | Imitation Learning followed by RL training. | 82.45% |
| **Zhao et al. [17]** | VisDial v1.0 (val set) | Structured knowledge-aware network (SKANet) + RL (image guessing performance of Q-bot) | 95% |
| **Xu et al. [20]** | GuessWhat?! | OCVD for question generation + RL | 72.1% -67.9% |

Table 3: Comparative results of the mentioned RL work using proposed metrics by Das et al. **[5]**

| Reference | Dataset | Methodology | MRR | R@1 | R@5 | R@10 | Mean Rank |
|---|---|---|---|---|---|---|---|
| **Das et al. [7]** | VisDial v0.5 (test-set) | Sanity check step (pure RL) followed by curriculum learning | 43.7 | - | 53.67 | 60.48 | 21.13 |
| **Murahari et al. [8]** | VisDial v1.0 (val / test - sets) | Adding Smooth-L1 penalty to the main objective function | val. Set | | | | |
| | | | 46.46 | 36.31 | 56.26 | 62.53 | 19.35 |
| | | | test set | | | | |
| | | | 45.64 | 34.85 | 56.55 | 63.43 | 18.96 |
| **Wu et al. [12]** | VisDial v0.9 | GANs + attention mechanisms for the generative and the discriminator model | Generative | | | | |
| | | | 55.78 | 46.10 | 65.69 | 71.74 | 14.43 |
| | | | Discriminative | | | | |
| | | | 63.98 | 50.29 | 80.71 | 88.81 | 4.47 |
| **Fan, Hehe, et al. [15]** | VisDial | Recurrent attention network with an attention processor. | Generative | | | | |
| | | | 51.54 | 43.42 | 59.31 | 62.04 | 22.24 |
| | | | Discriminative | | | | |
| | | | 60.40 | 46.52 | 77.10 | 86.34 | 5.10 |
| **Zhao, Lei, et al. [16]** | VisDial v0.9 (val-set) | The same model proposed by Das et al. [7] + attentive memory network (AMN) | 61.3 | 42.17 | 63.24 | 68.65 | 15.82 |
| **Zhao et al. [17]** | VisDial v1.0 (val set) | Structured knowledge-aware network (SKANet) + RL (generated answer evaluation) | 47.06 | 37.19 | 56.52 | 63.31 | 18.88 |

From the previously stated approaches in section 4, we can find that [7], and [9] have proposed a simple RL approach combined with SL pretraining to improve the quality of the generated text whether it was generated answers as in [7] or generated questions as in [9]. We can observe that they outperformed their previous generative SL work but the agent at some point stuck into repeating the same text regardless of the current context. This might be because of the agent's weak reward policy and weak reasoning ability. The repetition problem indicated that the model had a problem in correlating the image content with the asked questions. [8] added a loss term to the overall loss function introduced by [7] which was L1-penalty for the Q-Gen network, they managed to generate less repetitive text, yet it did not drastically forbid the repetition. [10] focused on the quality of the generated text by introducing novel policy class "TPG" which led to nearly 5% improvement in RL training. In addition, they defined an attention memory netowrk in the guesser module to improve its reasoning and coreference resolution abilities, therefore the guesser could efficiently know when to stop asking and to start guessing the correct object. This attention network improved the model accuracy by nearly 7% with an overall accuracy of 74.31%.

[11] tried to make use of the efficient policy gradient techniques along with attention mechanism for context awareness as [10] did. For their proposed game, they used double DQN & Q-value policy methods to increase the Q-bot agent exploration and exploitation ability and stats adaption technique which implements the memory attention mechanism to improve the agent ability of environment adaption and context-awareness given variant of visual inputs (e.g. different images). [12] also tried to improve the visual dialog agent by enhancing the policy optimization using GANs and considering the discriminator as the critic function for the generator. They used REINFORCE with intermediate policy update and not just the vanilla REINFORCE as it had been proved to be weak in this task. Also for more improvement, they applied attention in a more focused way than the previous work did. They co-attended each input feature on the other input feature to ensure a more relevant generated sequence to the image and the given textual information.

Looking for better policy value to start with, [13] saw that SL pertaining didn't kick-start the agent with enough and strong policy value. They proposed IL for kick-starting their RL agent using the recent information theoretic approach (AQM). They aimed to enhance the reasoning and understanding between the environment agents which enabled them to invent their own communication way. They achieved promising enhancement with an overall accuracy of 82.45% compared to [11] and with a less complex network. [15] tried to improve the agent with a less complex network architecture, they proposed a recurrent attention network with an attention processor for memorizing temporal textual information and spatial information respectively. They trained their model on both discrimination and generation settings. In the generation setting, they guided the generator network by the discriminator one but not in an adversarial manner. Although the proposed model outperformed some of the prior work, it underperformed others like [12] in the discrimination task. [16] followed the same training methodology proposed in [7] with an additional attentive memory network for both Q-bot and A-bot networks, and used the vanilla REINFORCE as the RL policy algorithm. Surprisingly this work achieved comparative results with all previously mentioned works although, it used just an attentive memory with REINFORCE policy in a simple network architecture. [17] proposed a knowledge-aware model that boosted the guessing game percentile for the Q-bot achieving 95%, and achieved promising results on the answer generation task, however, it was observed that the model struggled in some complex scenarios in the answer generation task. [20] had different points of view to deal with the question generation, where they focused on deep comprehension of the objects' category information. Although they shifted the focus to a new point (i.e. object category information) they did not

outperform the prior work with worth percentiles. As observed from these previous works, the attention memory network plays an important role in developing context-aware AI agents for conversational task.

The attention memory in deep learning networks simulates the cognitive attention in human brains which enables the machine to mimic human interactions and understating within the surrounding environment. Recently transformer network [22] has been proposed to solve seq2seq tasks while handling long-range dependencies effectively. It introduces a memory attention mechanism to allow the agent to focus on the most relevant parts of the input sequences. Different from basic seq2seq models, it implements the encoder-decoder network in some different way as the encoder passes more information into the decoder guaranteeing more understanding and information grounding. Many transformer-based models have been introduced addressing different tasks related to natural language processing as BERT [23], RoBERTa [24], ViLBERT [25], and GPT [26]. These transformer-based models could be a replacement for the basic seq2seq models defined for the VD task to improve the cognitive ability of the agent.

## 6. Conclusion and Future Work

In this study, we represent eleven of the most recent reinforcement learning-based approaches for visual dialog task. Also, we have discussed the two benchmark datasets which are VisDial and GuessWhat?! datasets that are used by the mentioned approaches. For evaluation, some approaches use the accuracy metric while others use four metrics that have been proposed for the sake of this task: R@k, NDCG, MRR, and Mean. From what we have discussed above we can conclude that the VD task is a complex multi-modal task that requires efficient methodologies to be developed. Using an efficient policy gradient, such as the proximal policy gradient, could enhance the exploration and exploitation of the RL agent to enable it to think and act like a human during the conversation. Moreover, attention mechanisms are an important factor that fosters the agent's awareness of the environment resulting in more efficient reasoning, understanding, and yet image-relevant text generation. From our point of view, building a successful VD agent is not dependent on how complex your architecture is, it is about what is the specific methodology that leads to major enhancement as we saw in Zhao, Lei, et al., using simple methodology and vanilla REINFORCE policy and it achieves very promising results compared to others who developed more complex model like using GANs.

For future work, as we believe that the attention memory network plays a main role in this task, the transformer network will affect the efficiency of the VD agent positively. The transformer model has proved its ability in many seq2seq tasks like machine translation and its architecture same as the VD task requires encoder-decoder architecture. Many transformer-based models have been introduced as BERT, RoBERTa, ViLBERT, and GPT. Recently GPT has made a big influence in natural language modeling that cannot be neglected! We suggest that training a transformer-based model in an RL environment could improve the performance of the VD task more.

## References

[1] J. Zhong, Y. Cao, Y. Zhu, J. Gong, and Q. Chen, ‘Multi-channel weighted fusion for image captioning’, *The Visual Computer*, vol. 39, no. 12, pp. 6115–6132, Dec. 2023, doi: https://doi.org/10.1007/s00371-022-02716-7.

[2] L. Zhu and Y. Yang, 'ActBERT: Learning Global-Local Video-Text Representations', in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 8743–8752. doi: 10.1109/CVPR42600.2020.00877.

[3] Y. Song, C. Yang, W. Gai, Y. Bian, and J. Liu, 'A new storytelling genre: combining handicraft elements and storytelling via mixed reality technology', *The Visual Computer*, vol. 36, no. 10, pp. 2079–2090, Oct. 2020, doi: 10.1007/s00371-020-01924-3.

[4] R. Cadène, C. Dancette, H. Ben-younes, M. Cord, and D. Parikh, 'RUBi: Reducing Unimodal Biases in Visual Question Answering', in *Neural Information Processing Systems*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:195584122

[5] A. Das *et al.*, 'Visual Dialog', in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 1080–1089. doi: 10.1109/CVPR.2017.121.

[6] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville, 'GuessWhat?! Visual Object Discovery through Multi-modal Dialogue', in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 4466–4475. doi: 10.1109/CVPR.2017.475.

[7] A. Das, S. Kottur, J. M. F. Moura, S. Lee, and D. Batra, 'Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning', in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 2970–2979. doi: 10.1109/ICCV.2017.321.

[8] V. Murahari, P. Chattopadhyay, D. Batra, D. Parikh, and A. Das, 'Improving Generative Visual Dialog by Answering Diverse Questions', 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)*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 1449–1454. doi: 10.18653/v1/D19-1152.

[9] F. Strub, H. De Vries, J. Mary, B. Piot, A. Courville, and O. Pietquin, 'End-to-end optimization of goal-driven and visually grounded dialogue systems', in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Melbourne, Australia: International Joint Conferences on Artificial Intelligence Organization, Aug. 2017, pp. 2765–2771. doi: 10.24963/ijcai.2017/385.

[10] R. Zhao and V. Tresp, 'Learning Goal-Oriented Visual Dialog via Tempered Policy Gradient', in *2018 IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece: IEEE, Dec. 2018, pp. 868–875. doi: 10.1109/SLT.2018.8639546.

[11] J. Zhang, T. Zhao, and Z. Yu, 'Multimodal Hierarchical Reinforcement Learning Policy for Task-Oriented Visual Dialog', in *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 140–150. doi: 10.18653/v1/W18-5015.

[12] Q. Wu, P. Wang, C. Shen, I. Reid, and A. V. D. Hengel, 'Are You Talking to Me? Reasoned Visual Dialog Generation Through Adversarial Learning', in *2018 IEEE/CVF Conference on Computer*

*Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 6106–6115. doi: 10.1109/CVPR.2018.00639.

[13] Y.-W. Chang and W.-H. Peng, 'Learning Goal-Oriented Visual Dialog Agents: Imitating and Surpassing Analytic Experts', in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, Shanghai, China: IEEE, Jul. 2019, pp. 520–525. doi: 10.1109/ICME.2019.00096.

[14] S.-W. Lee, Y.-J. Heo, and B.-T. Zhang, 'Answerer in Questioner's Mind: Information Theoretic Approach to Goal-Oriented Visual Dialog', in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/0829424ffa0d3a2547b6c9622c77de03-Paper.pdf

[15] H. Fan, L. Zhu, Y. Yang, and F. Wu, 'Recurrent Attention Network with Reinforced Generator for Visual Dialog', *ACM Transactions on Multimedia Computing Communications and Applications*, vol. 16, no. 3, pp. 1–16, Aug. 2020, doi: 10.1145/3390891.

[16] L. Zhao, X. Lyu, J. Song, and L. Gao, 'GuessWhich? Visual dialog with attentive memory network', *Pattern Recognition*, vol. 114, p. 107823, Jun. 2021, doi: https://doi.org/10.1016/j.patcog.2021.107823.

[17] L. Zhao, L. Gao, Y. Guo, J. Song, and H. Shen, 'SKANet: Structured Knowledge-Aware Network for Visual Dialog', in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, Shenzhen, China: IEEE, Jul. 2021, pp. 1–6. doi: 10.1109/ICME51207.2021.9428279.

[18] R. Speer, J. Chin, and C. Havasi, 'ConceptNet 5.5: An Open Multilingual Graph of General Knowledge', *AAAI*, vol. 31, no. 1, Feb. 2017, doi: 10.1609/aaai.v31i1.11164.

[19] T. N. Kipf and M. Welling, 'Semi-supervised classification with graph convolutional networks', presented at the 5th International Conference on Learning Representations, Toulon, France, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[20] F. Xu, Y. Zhou, Z. Zhong, and G. Li, 'Object Category-Based Visual Dialog for Effective Question Generation', in *Computational Visual Media*, vol. 14593, F.-L. Zhang and A. Sharf, Eds., in Lecture Notes in Computer Science, vol. 14593. , Singapore: Springer Nature Singapore, 2024, pp. 316–331. doi: 10.1007/978-981-97-2092-7_16.

[21] T.-Y. Lin *et al.*, 'Microsoft COCO: Common Objects in Context', in *Computer Vision - ECCV 2014*, Springer International Publishing, 2014, pp. 740–755. doi: https://doi.org/10.1007/978-3-319-10602-1_48.

[22] A. Vaswani *et al.*, 'Attention is All you Need', in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', in *Proceedings of the 2019 Conference of NAACL-HLT*,

Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: https://doi.org/10.18653/v1/N19-1423.

[24] Y. Liu *et al.*, 'RoBERTa: A Robustly Optimized BERT Pretraining Approach', 2019, *arXiv*. doi: 10.48550/ARXIV.1907.11692.

[25] J. Lu, D. Batra, D. Parikh, and S. Lee, 'ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks', in *Advances in Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, BC, Canada: Curran Associates Inc., 2019. [Online]. Available: http://papers.neurips.cc/paper/by-source-2019-16

[26] A. Radford and K. Narasimhan, 'Improving Language Understanding by Generative Pre-Training', 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245