Submitted: 2024-11-24 | Revised: 2025-01-12 | Accepted: 2025-01-20

Keywords: visual dialog, transformers, ViLBERT, GPT, answer generation

Ghada M. Elshamy [1*], *Marco Alfonse* [1], *Islam Hegazy* [1], *Mostafa M. Aref* [1]
[1] Ain Shams University, Egypt, ghada.magdy@cis.asu.edu.eg, marco_alfonse@cis.asu.edu.eg, islheg@cis.asu.edu.eg, mostafa.aref@cis.asu.edu.eg
[*] Corresponding author: ghada.magdy@cis.asu.edu.eg

A multi-modal transformer-based model for generative visual dialog system

Abstract

Recent advancements in generative artificial intelligence have boosted significant interest in conversational agents. The visual dialog task, a synthesis of visual question-answering and dialog systems, requires agents capable of both seeing and chatting in natural language interactions. These agents must effectively understand cross-modal contextual information and generate coherent, human-like responses to a sequence of questions about a given visual scene. Despite progress, previous approaches often required complex architectures and substantial resources. This paper introduces a generative dialog agent that effectively addresses these challenges while maintaining a relatively simple architecture, dataset, and resource requirements. The proposed model employs an encoder-decoder architecture, incorporating ViLBERT for cross-modal information grounding and GPT-2 for autoregressive answer generation. This is the first visual dialog agent solely reliant on an autoregressive decoder for text generation. Evaluated on the VisDial dataset, the model achieves promising results, with scores of 64.05, 62.67, 70.17, and 15.37 on normalized discounted cumulative gain (NDCG), rank@5, rank@10, and the mean, respectively. These outcomes underscore the effectiveness of this approach, particularly considering its efficiency in terms of dataset size, architecture complexity, and generation process. The code and dataset are available at https://github.com/GhadaElshamy/MS-GPT-visdial.git , complete with usage instructions to facilitate replication of these experiments.

1. INTRODUCTION

Recent advancements in artificial intelligence have significantly sparked the development of vision-language tasks, with visual dialog systems emerging as a particularly promising area of research. These systems, designed to facilitate natural language interactions based on visual content, offer a compelling avenue for human-machine communication. The pursuit of creating AI agents capable of both understanding and chatting has driven substantial progress in visual dialog research as evidenced by the growing body of literature on the visual dialog domain (Das et al., 2016, 2017; de Vries et al., 2016; Fan et al., 2020; Jiang et al., 2020b; Kottur et al., 2018; Liu et al., 2022; Lu et al., 2017; Schwartz et al., 2019; Seo et al., 2017; Wu et al., 2018; Yang et al., 2019; J. Zhang et al., 2018; Zhao et al., 2021). The work of Das et al. (2016) was the first introduction of the visual dialog task and the VisDial dataset. The VisDial dataset is mostly used as the benchmark dataset for evaluating and advancing the capabilities of visual dialog systems.

Visual dialog (VD) task is a conversational task with a cross-modality nature, where it integrates natural language understanding with visual grounding. The VD agent is required to answer the questions posed by a human or another agent regarding a visual scene within the context of a free-form conversation. Given a dialog history consisting of question-answer pairs, a current question, and a visual scene, the agent is required to infer the context from the history and ground it on the visual scene to answer the question correctly. The ultimate objective of these systems is to develop AI agents that can effectively assist humans by leveraging their conversational and visual capabilities. Visual dialog systems have the potential to address a multitude of real-world applications, including providing support to visually impaired individuals in navigating their surroundings, aiding in the analysis of extensive surveillance data, and contributing to robotic missions for space exploration. Furthermore, this technology supports applications in biomedical engineering, such as analyzing medical images, and aiding in surgical planning and treatment decisions based on patients' medical histories. Additionally, it can be utilized in the field of economics for financial market analysis and risk assessment.

The VD task is a complex vision language task as it requires processing the inputs from vision and language modalities simultaneously. This processing includes contextual visual comprehension, language understanding, and visual coreference resolution. Contextual visual comprehension emphasizes single-modality feature extraction and multimodal reasoning. Reasoning aims to highlight relevant information between the input modalities to guarantee better understanding. Specifically, the VD task requires the agent not only to understand the textual intent but also requires grounding it on the visual information. Thus, multimodal reasoning can be a considerable challenge for the VD task. In addition, as a conversational task, there might be many pronouns resolution (e.g. it, they, her, etc.) referring to objects or people that are previously mentioned during the conversation which is difficult for the agent to understand and relate. This problem is known as visual coreference resolution which is also considered as a challenge for this task. Last but not least, the VD task encounters a dataset bias problem where the agent may rely excessively on the dataset pattern, i.e. relation between question and answer, which limits its exploration ability for the image content. Therefore, the agent generalization ability and robustness will be considerably restricted.

Subsequent to the introduction of the visual dialog task, extensive research efforts have been undertaken to address the challenges associated with this domain. Most of the proposed visual dialog models were trained from scratch on the VisDial dataset to produce the answer either via supervised learning such as Das et al. (2016), de Vries et al. (2016), Fan et al. (2020), Jiang et al. (2020b), Kottur et al. (2018), Liu et al. (2022), Lu et al. (2017), Schwartz et al. (2019), Yang et al. (2019) or deep reinforcement learning such as Das et al., 2017, Wu et al., 2018, J. Zhang et al., 2018, Zhao et al. (2021). Das et al. (2017) were the first to introduce a reinforcement learning-based model for free-form VD task. Different attention mechanisms-based approaches were proposed by Das et al. (2016), Fan et al. (2020), Kottur et al. (2018), Seo et al. (2017), Yang et al. (2019) to improve the multimodal reasoning and focus on the important relevant parts of the multimodal inputs. In addition to the attention mechanism, other approaches such as Kottur et al. (2018), Seo et al. (2017), Yang et al. (2019) addressed the visual coreference resolution problem in their proposed models. Furthermore, other approaches like (Lu et al., 2017; Wu et al., 2018) employed generative adversarial networks (GANs) to enhance attention mechanisms and generate more comprehensive responses. Zhang et al. (2018) utilized hierarchical reinforcement learning policy networks to augment the agent's capacity for exploration and exploitation, thereby optimizing reward acquisition and enhancing the overall quality of the dialogue. Moreover, some approaches, such as those proposed by Jiang et al. (2020b), Liu et al. (2022), Schwartz et al. (2019) proposed graph-based attention mechanisms to capture more details by extracting deeper relationships between multimodal features. While these studies incorporated attention mechanisms as an essential component and implemented various model enhancements, they remained constrained by limitations such as the generation of repetitive sentences and suboptimal model robustness.

Regarding these limitations, recent approaches such as C. Chen et al. (2022), F. Chen et al. (2021), Kang et al. (2023), Murahari et al. (2020), Y. Wang et al. (2020) decided to use transformer-based pre-trained models like BERT (Devlin et al., 2019) and ViLBERT (Lu et al., 2019) and finetuned these models on the visual dialog task. Transformer-based models (Vaswani et al., 2017) have demonstrated exceptional performance in natural language processing (NLP) tasks that demand strong reasoning abilities, a crucial requirement in this context. Furthermore, the efficacy of transformer-based models in computer vision tasks, including (Li et al., 2023; X. Lin et al., 2023), emphasizes their suitability for vision-language tasks. In addition, these models are pretrained on huge amount of data which helps in ameliorating the dataset bias problem. Employing BERT as the decoder for answer generation, as in the previous transformer-based approaches, may not be recommended due to its autoencoding nature. Although BERT is excellent at understanding and representing the context of a given text, it lacks the ability to generate new text sequences in a sequential manner as mentioned in (Wang & Cho, 2019). While it is applicable to adapt BERT for encoder-decoder architectures for text generation, its inherent limitations in producing diverse and lengthy responses constrain its conversational capabilities. Now comes our research question: What are the advantages and limitations of using an autoregressive decoder model for visual dialog to generate responses that are consistent with the multi-modal context?

This paper presents a generative visual dialog system following the “pretrain then transfer” technique to build powerful and efficient conversational agents that can generate human-like responses. To achieve this objective and address the challenges, powerful transformer-based models such as ViLBERT (Lu et al., 2019) and GPT-2 (Radford et al., 2019) are utilized to enhance the model’s generation ability. BERT has a bidirectional processing ability for tokens which is eagerly needed in the encoder to help the decoder coherently generate responses. Consequently, ViLBERT is employed as our cross-modal encoder due to its compatibility with BERT and its specific design for visual-language conjunction. It effectively extracts visual and textual features, establishing connections and grounding between them to facilitate a comprehensive contextual understanding of the input modalities. GPT-2 is selected as the decoder due to its ability to generate coherent and contextually rich sentences. In the context of visual dialog, GPT-2 demonstrates improved reasoning capabilities by effectively leveraging the encoded understanding from ViLBERT to produce responses that are relevant to the image, question, and dialog history. Moreover, it generates responses with a more natural tone compared to the factual responses produced by BERT. Additionally, GPT-2 is less susceptible to dataset bias than BERT, as it is pre-trained on a larger dataset, which enhances its generalization ability and overall performance. The proposed model achieves promising results across normalized discounted cumulative gain (NDCG), rank@5, and rank@10 and the mean metrics, with scores of 64.05, 62.67, 70.17, and 15.37, respectively. The experimental results on the VisDial dataset demonstrated that the generated answers exhibited greater realism and coherence when compared to the ground truth answers.

The findings highlight the crucial role of autoregressive decoders, specifically GPT-2, in enhancing the quality of generated responses within conversational systems. This improvement leads to more engaging and effective interactions. These findings have significant implications for the development of interactive visual chatbots and can inform future research in a variety of related applications, including education, healthcare, customer service, and assistive technologies.

This paper presents a comprehensive literature review of prominent visual dialog research, focusing on key contributions, limitations, and potential avenues for advancement in section 2. Section 3 introduces the proposed model as a solution to the identified research challenges. Implementation details are provided in section 4. Experimental results and a detailed description of the dataset used are presented in section 5. Finally, section 6 summarizes our findings, draws conclusions, and outlines potential directions for future research.

2. RELATED WORKS

Recent achievements in the AI field have significantly propelled the development of multi-modal tasks, where deep models leverage a combination of text, images, video, audio, and other modalities. Vision-language (VL) models have emerged as powerful tools for a wide range of complex AI applications, owing to their ability to integrate visual and linguistic information into a shared semantic space. Various VL tasks have been addressed using deep neural modules, including captioning systems (Dai & Zhang, 2022; Xin et al., 2023), video processing (Cui et al., 2023; Zhang et al., 2024), text-to-image synthesis (Yu et al., 2024), visual question-answering (VQA) systems (Cadène et al., 2019; Yu et al., 2019), and visual dialog systems.

Visual dialog, an extension of VQA, requires an agent to engage in a series of questions and answers about a specific image aligned with the previous dialog exchanges. In contrast, VQA typically involves answering a single question about an image. Visual dialog systems have been formulated using two primary approaches: free-form conversation introduced by (Das et al., 2016), and goal-oriented conversation in the form of a guessing game introduced by de Vries et al. (2016). Each approach is associated with a specific dataset: VisDial (Das et al., 2016), and GuessWhat! (de Vries et al., 2016) datasets, both of which were collected using Amazon Mechanical Turk (AMT).

Previous studies have employed deep neural modules to train visual dialog systems on the VisDial dataset in an end-to-end fashion, utilizing an encoder-decoder architecture. The encoder processes the input tuple, consisting of an image, dialog history, and a follow-up question, to generate a contextual vector representation. This encoded context is then integrated by the decoder to infer information and produce the correct answer.

In addition to the context, the decoder network receives a list of one hundred candidate answers, including the ground-truth answer. The answering process within the decoder can be implemented either discriminatively or generatively.

A core component that has been widely adopted in previous research is the attention mechanism, which effectively aligns the question with the dialog history while grounding it on the given image, thereby enhancing the agent's multimodal comprehension capabilities. Based on this foundation, subsequent approaches have increasingly favored transformer-based models for generating coherent responses to the specified question. Transformers (Vaswani et al., 2017), employ a codec (encoder-decoder) structure with a multi-head attention mechanism, where each head focuses on distinct relationships within the input sequence. This enables them to handle long-sequence dependencies more effectively, leading to superior performance in various natural language processing (NLP) tasks such as machine translation, question-answering, text generation, and summarization.

Bidirectional Encoder Representations from Transformers (BERT) is a powerful encoder-only transformer network known for its contextual understanding capabilities. Its bidirectional nature enables deep contextualization of the input sequence, making it well-suited for various NLP tasks such as sentiment analysis, text summarization, and question-answering. BERT is pre-trained on a massive dataset using two objectives: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). MLM randomly masks tokens and trains BERT to predict them, while NSP trains BERT to predict whether a given sentence follows another. ViLBERT is another transformer-based model that extends BERT by incorporating two encoder streams, one for the visual input and the other for the textual input through applying cross-attention layers between the two streams. ViLBERT is particularly well-suited for the visual dialog task, as it naturally addresses the challenges of multimodal reasoning.

Subsequent studies in visual dialog have frequently built their models upon the BERT architecture. Murahari et al. (2020) proposed the VisDial-BERT model, aiming to leverage pre-trained checkpoints from related vision-language tasks to enhance visual dialog system performance. To achieve their objective, they developed a discriminative visual dialog agent based on the ViLBERT model. By applying transfer learning to the VisDial dataset, they adapted ViLBERT and achieved competitive results in the discriminative setting. While their model set a new state-of-the-art at the time, it did not introduce a generative agent, which is often considered more desirable in this task, as generative models can simulate more human-like conversations.

Wang et al. (2020) proposed a simpler yet more effective architecture than VisDial-BERT called (VD-BERT), addressing both the discriminative and generative settings. This unified model for visual dialog effectively integrates multimodal features through visually grounded training, leading to improved results compared to VisDial-BERT in the discriminative setting and promising results in the generative setting. While VD-BERT offers a simple and effective approach, it may be limited in its reasoning ability when faced with complex tasks that require in-depth analysis of visual and textual relationships. Additionally, training solely on the VisDial dataset can pose significant generalization challenges.

C. Chen et al. (2022) also utilized ViLBERT as their encoder and used a single transformer-based decoder to serve both discriminative and generative settings. They proposed a unified contrastive learning approach called (UTC) and they achieved comparative results compared to the prior works. However, training a unified decoder simultaneously to handle both settings causes the agent to be confused at some time to perform which task.

Kang et al. (2023) proposed a Generative Self-Training approach (GST) based on ViLBERT and BERT. They employed the same encoder architecture proposed by Murahari et al. (2020) and they adapted BERT to act as an autoregressive decoder for answer generation. They applied the GST algorithm to enable the model to generate multi-turn visual question-answer data to leverage unlabeled Web images effectively. The GST approach achieved state-of-the-art performance on the VisDial dataset. However, training GST required creating a synthetic dataset by generating questions and answers for given images and their captions. This significantly increased the training dataset size from 1.2 million to 12.9 million which required substantial computational power. In addition, relying on the generated data by their agent to perform the GST approach leads to limited domain knowledge for the agent which in turn could raise the dataset bias problem.

Regarding these issues, the authors believe that this task requires an auto-regressive decoder model by nature so that it can form better, realistic, and creative responses. Auto-regressive decoders have proven their ability to cover this problem in other similar tasks as mentioned in Radford et al. (2019), as well as their ability to be fine-tuned for the task objective. In this paper, the primary objective is to examine the advantages and limitations of creating an agent based on the GPT-2 (Radford et al., 2019) model as the decoder. The objective is to improve the quality, creativity, uniqueness, and relevance of the generated responses, which are essential for conversational tasks. Additionally, these goals are pursued while maintaining a lower computational cost compared to the state-of-the-art approaches.

An evaluation framework was developed with a set of criteria metrics to facilitate a comprehensive comparison with existing methods. The evaluation framework incorporates the same retrieval metrics established by Das et al. (2016), NDCG, Rank@k, Mean, and MRR. Additionally, it involves a comparative analysis of the model complexity across different methods, regarding hardware resource requirements (GPUs), dataset size, and decoder parameter configurations.

A small version of GPT-2 (117M parameters) is used, containing approximately one-third the number of parameters compared to BERT-base (340M parameters), resulting in reduced computational cost.

Additionally, the model is evaluated on the VisDial dataset, where it outperforms the UTC model C. Chen et al. (2022) on NDCG, R@5, R@10, and mean rank. Promising results are achieved compared to the model by Kang et al. (2023) in evaluation metrics, with lower computational cost, utilizing a smaller dataset and fewer decoder parameters.

3. METHOD

The proposed model consists of two main components as illustrated in Fig. 1. A cross-modal encoder for encoding the triplet input (i.e. visual features, question (Q_t), and history ($H_{1\to t-1}$)) all together forming the context. In addition to an autoregressive decoder to generate an answer $\hat{A}_i^t$ for the current question Q_t at round i conditioned on the given triplet context. In this architecture, ViLBERT is utilized as the cross-modal encoder and GPT-2 as the autoregressive decoder. As illustrated in Fig. 1, the ViLBERT encoder takes both text and vision input and then fuses them to output the encoded context (E_h). Next, this context is fed into the GPT-2 decoder to generate the desired answer. This section formally describes the visual dialog task and then presents the proposed approach.

3.1. Problem formulation

This section formally describes the visual dialog problem and presents the details of the proposed codec model.

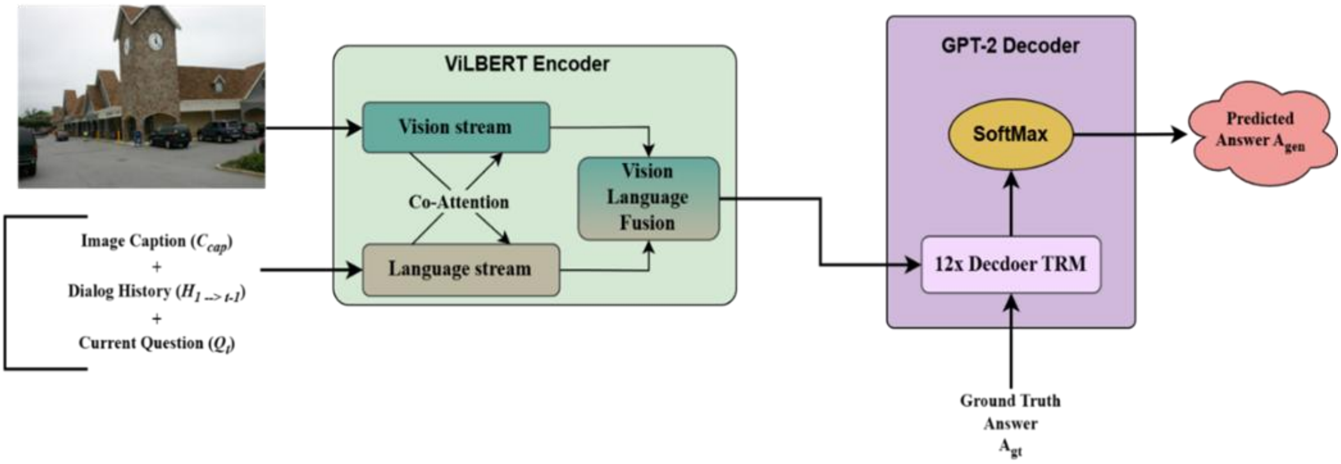


Fig. 1. The overall architecture of the proposed model

The definition proposed by Das et al. (2016) is followed for the visual dialog task. For a given question Q_t about an image I with the previous dialog history rounds (i.e. from round 0 to round $t-1$) formulated as $H = \{C_{cap}, h_{1,}\ h_{2,} \dots ,\ h_{t-1}\}$ where C_{cap} is the image caption treated as history round 0, h_i is a question/answer pair for each round i. In addition to a list of 100-candidate answers for each history round given as $A_t = \{A_1^t, \dots, A_{100}^t\}$, it contains the ground-truth answer A_{gt}^t as an option from the 100 options. The visual dialog agent should predict an answer either by discrimination or generation. In the discriminative setting, it sorts the candidate list so that the A_{gt}^t ranked on top of the list. The ranking is done by calculating the posterior probability of each candidate's answer by computing the dot product between the candidate and the encoded context. The objective is to maximize the log-likelihood of the correct answer candidate within the given list. Conversely, in the generative setting, the decoder aims to generate an answer sequence that is as accurate as possible by maximizing the log-likelihood of the generated answer's encoded representation A_{gen}^t relative to the ground-truth answer's encoded representation A_{gt}^t. In common sense, the generative setting is more practical in real life as it doesn’t require pre-defined answers.

3.2. Cross modality encoding

This section provides a brief introduction to the ViLBERT model in Section 3.2.1, followed by an explanation of its application to the visual dialog task in the proposed approach in Section 3.2.2.

3.2.1. Preliminary on ViLBERT Model

ViLBERT was initially pretrained on the Conceptual Captions dataset (Sharma et al., 2018) which consists of nearly 3 million images paired with corresponding captions. The pretraining process involved two objective tasks: masked language modeling (MLM) and masked image region (MIR). Subsequently, the pretrained model can be finetuned for various multi-modal tasks such as VQA. As shown in Fig. 2, the visual features are embedded and then fed into the vision-stream transformer blocks. At the same time in the second stream, the textual features are embedded and fed into the transformer blocks. At a certain level, co-attention is applied between both streams to establish a mapping between the region of interest in the image and their corresponding tokens in the text. Thereafter, the outputs of the two streams are fused together into a unified hidden representation which can be leveraged in various downstream vision-language tasks.

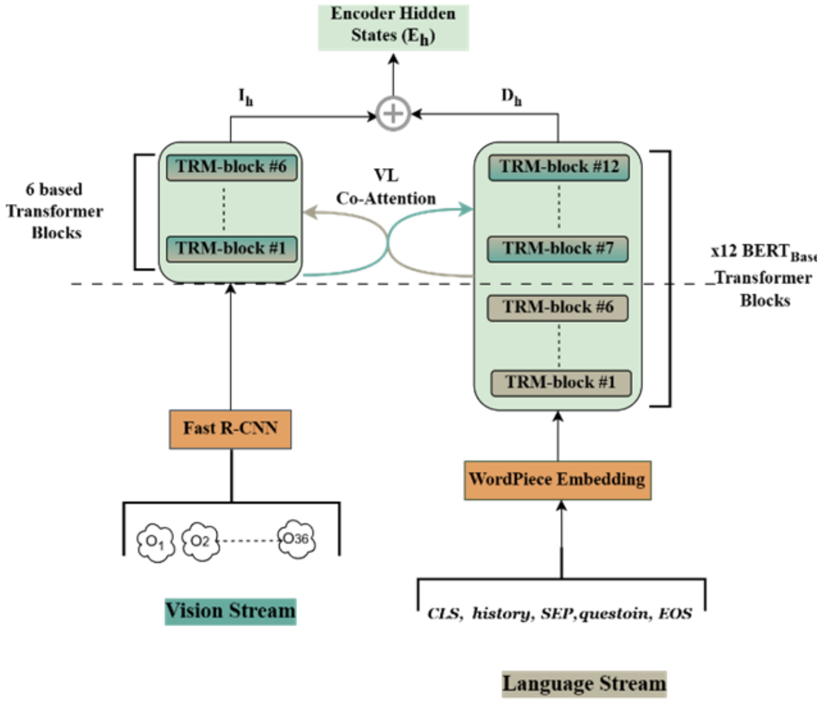


Fig. 2. ViLBERT encoder network used in the proposed architecture

3.2.2. VILBERT for visual dialog

To apply ViLBERT to this task, the same procedure as in Murahari et al. (2020) is followed. It is firstly pretrained on the Conceptual Captions dataset with two training objectives: MLM and Masked Image Regions (MIR). This pretraining step aims to optimize the summation of MLM and MIR losses to efficiently predict the masked token conditioned on the context and predict the semantic category of the masked visual object. Secondly, it is finetuned on VQA before the final finetuning on the visual dialog task due to the large similarity between the two tasks. VQA is similar to our task in generating an answer sequence for a given question grounded on a given image. On the other hand, the visual dialog task has multiple question/answer rounds (i.e. history) which makes it more complex than VQA. Following Lu et al. (2019) VQA finetuning is done by training two multi-layer perceptron as a simple decoder on top of the visual-language fusion block where it takes the fused hidden states and predicts the answer representation. The final pretrained model is now ready to be finetuned on the visual dialog task.

For visual dialog fine-tuning, MLM loss and MIR loss are used to train the model, where 15% of word tokens and image features are randomly masked out. The masked token is replaced with the [MASK] token, and the image features are zeroed out. The NSP loss is not utilized, as it is primarily designed for discriminative settings to enhance ground-truth answer prediction and exact ranking given a candidate list, whereas the focus here is on the generative setting.

For image features representation, Murahari et al. (2020) were followed, where the image undergoes an object extraction process via a Faster R-CNN (Ren et al., 2017) (with a ResNet-101 backbone (He et al., 2016) network) pretrained on Visual Genome dataset (Krishna et al., 2017). This process extracts the top 36 objects with corresponding bounding boxes and visual features. The visual feature embeddings along with their spatial information are then concatenated and passed through encoding where masking is applied as shown in Eq.(1).

$$V = [v_1, \dots v_{36}] = fast\ R - CNN(I) \tag{1}$$

where V represents the resNet101 visual features for the 36 detected objects by Fast R-CNN.
For text feature representation, recall this input format $Q_t, H = \{C_{cap}, h_{1,}\ h_{2,} \dots ,\ h_{t-1}\}$ where h_i is a question/ answer pair for each round i. The question and dialog history (including the caption) are concatenated together as one vector D with additional special tokens for masking and padding (e.g. CLS, PAD, MASK) thus the final input format for round i is $[CLS, C_{cap}, SEP, Q_1, SEP, A_1, SEP, \dots, Q_i, SEP, A_i, SEP, Q_t, PAD]$ Then this vector passes through an encoding process where masking is applied as shown in Eq.(2).

$$T = [w_0, \dots, w_n] = WordPieceEmbedding(D) \tag{2}$$

where T is the language features, $[w_0, \dots, w_n]$ is the embedded tokens resulted from WordPiece embedding.
As ViLBERT is an extension from BERT, it uses WordPiece embedding (Wu et al., 2016) where the text embedding is formulated as a summation of three vectors which are token embeddings, positional embeddings, and segment embeddings. Subsequently, the embedded text features T is passed through 6 transformer blocks to focus on the most important parts in the text. Subsequently, cross attention is applied to get the final hidden states which are fed into the decoder network. Eq.(3).

$$V_{cross-att} = attention(Q_v, K_t, V_t) * V$$

$$T_{cross-att} = attention(Q_t, K_v, V_v) * T$$

$$E_h = CONCAT(V_{cross-att}, T_{cross-att}) \tag{3}$$

where Q, K, V are the query, key, and value matrices for self- attention mechanism for both modalities. E_h are the encoder's final hidden states that are fed into the decoder network.

3.3. Autoregressive decoder

Autoregression is a type of statistical model that predicts the current value in a time series given all previous values in the same time series. Autoregressive language models (AR) provide the capability to predict the next word y_t in a given sequence at time-step t depending on the previously predicted words $\{y_{1\rightarrow t-1}\}$ as shown in Eq.(4). Consequently, it has been adopted in various NLP tasks such as machine translation, text generation, etc. Fig. 3 illustrates the autoregressive process with an example.

$$y_t = b + \sum_{i=1}^{t-1} w_i y_i + \epsilon_t \tag{4}$$

where b is the bias term, y_i are the previously predicted tokens, ϵ_t is the time-step error and w_i are the random weights.

The proposed generative decoder is based on deep AR models which are feed-forward networks used for sequence generation tasks. These AR models have powerful capabilities that make them SOTA language models because of their training simplicity (i.e. just feed-forward not recurrent) and stability (i.e. supervised generative nature). This stability offers easy hyperparameter tuning and inexpensive inference computations. Therefore, they are compelling alternatives for RNNs for sequential data and GANs for generation tasks. One of the AR-based generative models is the GPT-family networks. In our proposed model we employ GPT-2 for answer generation where it generates the required text autoregressively given the context as shown in Fig. 3.

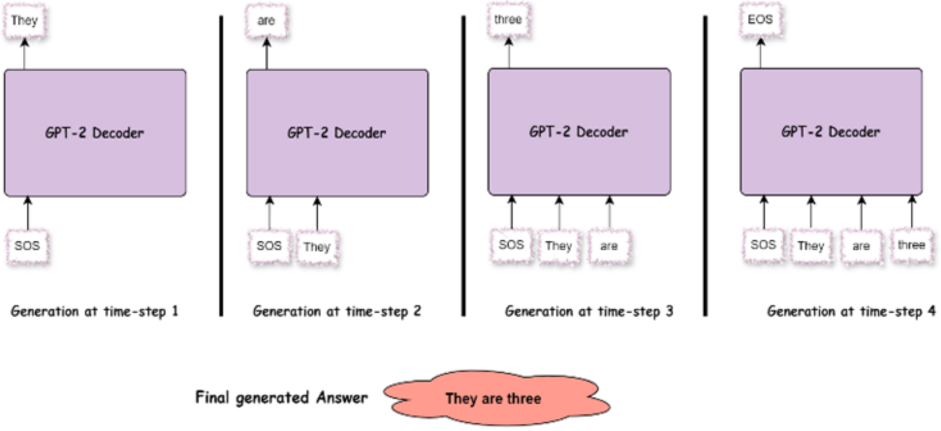


Fig. 3. Illustrated autoregressive process for sequence generation

In what follows, we present the main contribution to the visual dialog task which is employing the GPT-2 (Radford et al., 2019) model as the autoregressive decoder. First, we will give a brief preliminary study about the GPT-2 model (section 3.3.1). Second, we will dive into our finetuned GPT-2 decoder on the VisDial dataset (Section 3.3.2).

3.3.1. Preliminary on GPT-2 model

Radford and Narasimhan (2018) were the first to introduce GPT-family (i.e. Generative Pre-training Transformers) models (Brown et al., 2020; OpenAI, 2023; Radford et al., 2019; Radford & Narasimhan, 2018) for improving natural language understanding and generation tasks. GPTs have shown significant breakthroughs in NLP tasks, allowing machines to understand and generate language with unprecedented quality. GPT-2 is a successor of the GPT model which is in turn an extension of decoder-only transformer models.

GPT-2 is a large-scale unsupervised language model that generates coherent text conditioned on the given task. It is pre-trained on a massive amount of dataset (created by Radford and Narasimhan (2018)) consisting of more than 8 million web pages (combining BookCorpus (Zhu et al., 2015), Common Crawl, and Web Text). It consists of 10 times the parameters and 10 times the dataset amount of its predecessor GPT. It is trained with a casual language modeling objective to efficiently predict the next word given all previous words within a sequence. This massive amount of data helps this objective to naturally boost the model generalization ability for various tasks across diverse domains (i.e. multitasking ability). GPT-2 exhibits a strong capability to synthesize coherent and realistic sequences of text that are often indistinguishable from human text, making it a valuable tool for NLP tasks, such as summarization, question-answering, and translation.

3.3.2. GPT-2 decoder for visual dialog

To effectively adapt GPT-2 to the visual dialog domain, its pre-trained checkpoints are strategically utilized as a warm start for the decoder network, relying on a fine-tuning process. This approach leveraged the model's existing proficiency in natural language processing and realistic text generation. However, given the absence of visual context understanding in the pre-trained checkpoints, it was necessary to incorporate transfer learning alongside fine-tuning to adapt the model to our specific objectives.

While the underlying GPT-2 architecture remained unchanged, we modified the pre-trained weights during the training loop by passing the integrated visual and linguistic features (i.e. encoder hidden states) as the hidden state to GPT-2. Our decoder was subsequently trained in an autoregressive manner to reconstruct the answer sequence word by word, as illustrated in Figure 4. By conditioning the model's training on the context triplet (i.e. I, H, Q_t), we enable it to generate contextually relevant responses, aligned with the visual and textual information provided, and demonstrate a comprehensive understanding of the multi-modal context.

Figure 4 represents GPT-2 decoder architecture. The input embeddings represent how the decoder deals with the encoder's hidden states as well as the target answer embeddings. The final layer of the GPT-2 language model is a SoftMax activation layer which selects the word to be generated. It normalizes the input vector and then outputs a vector of probabilities that sums up to 1. The value with the highest probability is the desired word.

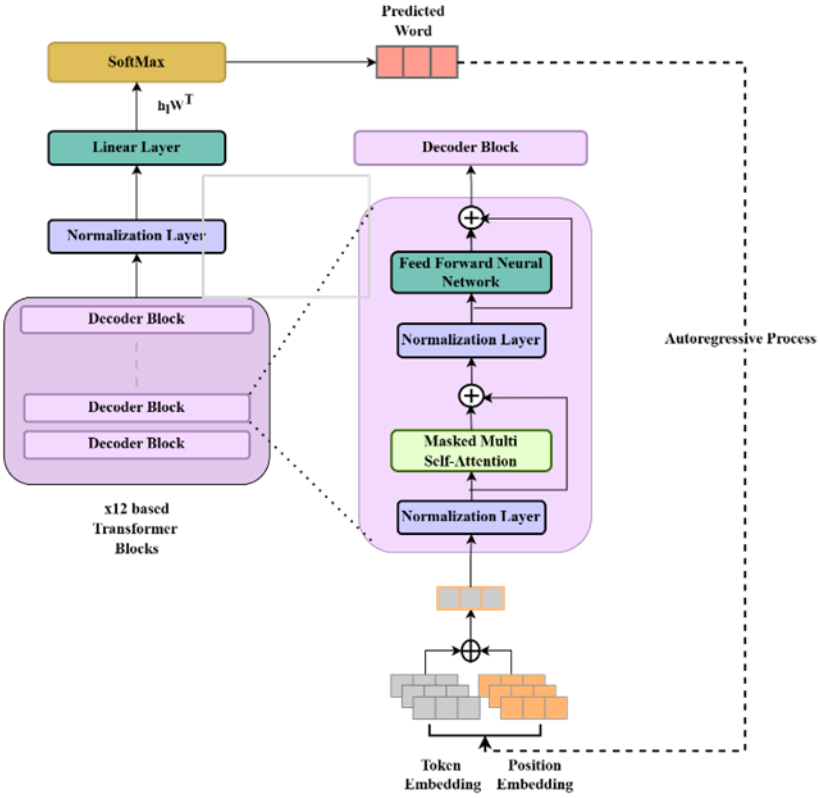


Fig. 4. GPT-2 decoder architecture for *v*isual *d*ialog

The objective function for GPT-2 in this context is to minimize the negative log-likelihood of the generated response given the input image and dialog history (E_h). This can be formulated as illustrated in Eq.(5):

$$Loss = -\log P(\hat{y}\ |E_{h,} y_{gt}) = -\log \prod_{i=1}^{t} P(\hat{y}_i\ |E_h, y_1, \ldots y_{i-1}) \tag{5}$$

where y is the target sequence, $\hat{y}$ is the generated sequence, E_h represents the context triplet (I, Q, H) hidden states from the encoder. $\hat{y}_i$ is the generated token at time step *i* given all previous tokens. $P(\hat{y}_i\ |E_h, y_1, \ldots y_{i-1})$ calculates the joint probability of the generated sequence using the softmax activation function followed by cross-entropy loss for each generated token.

4. IMPLEMENTATION DETAILS

The ViLBERT model is integrated as the encoder with GPT-2 as the decoder, and both models are trained end-to-end for sequence generation. The ViLBERT encoder consists of two transformer-based networks: vision and language streams, as illustrated in Fig. 2. The vision stream is composed of 6 layers of transformer blocks with 8 attention heads and a hidden state of size 1024. While for the language stream, it is based on $BERT_{base}$ architecture. It is composed of 12 layers of transformer block with 12 attention heads and a hidden state of size 768. For the co-attention layer, the vision stream transformer blocks are co-attended to the last 6 transformer layers of the language stream with 8 attention heads and a hidden state of size 1024. For tokenization, we use a maximum sequence length of 256. We add an encoder-decoder cross-attention mechanism to generate the answer grounded on the encoded context. We ran the training for 70 epochs using a single A100 GPU.

5. EXPERIMENTAL RESULTS AND DISCUSSION

The dataset is described in Section 5.1, followed by the evaluation metrics in Section 5.2. Finally, the experiment and results are presented in Section 5.3.

5.1. Dataset

The authors evaluate their proposed approach on the VisDial v1.0 dataset (Das et al., 2016), collected by the AMT two-person chat about MS-COCO (Lin et al., 2014) images. The VisDial v1.0 dataset contains 123,287 images for the training split, 2,064 images for the validation split, and 8,000 images for the testing split. Each image is associated with a caption sentence from COCO and one dialog (i.e. 10 rounds of question-answer pairs).

5.2. Evaluation metrics

The same metrics introduced by Das et al. (2016) are used for evaluating visual dialog models. Both the generative and the discriminative tasks are evaluated by retrieval-based evaluation metrics which are Mean Reciprocal Rank (MRR), Recall @k (R @k, k = {1, 5, 10}), Mean Rank (Mean), and finally the Normalized Discounted Cumulative Gain (NDCG). Mean reciprocal rank is the average of 1/rank of the ground truth answer option. 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. NDCG penalizes the low-ranked answer options with high relevance. The lower the value for mean is better and the higher value for the other three metrics is better. For the generative setting, the model uses its calculated log-likelihood scores for ranking the 100-candidate answers then the model is evaluated using the mentioned four metrics.

5.3. Experiments

GPT models have demonstrated their efficacy in conversational chatbot applications, as exemplified by the ChatGPT agent's ability to provide comprehensive and informative responses across a wide range of topics. In this paper, we applied different GPT models to investigate their efficacy in visual-based conversational systems. To our knowledge, we are the first to explore GPT models on the VD task for answer generation. We conduct two experiments to support our hypothesis using different GPT models as our decoder. In the two experiments, the ViLBERT model is used as the encoder with the same architectural details mentioned above in section 4.

The proposed model integrates an autoencoder (ViLBERT) and an autoregressive model (GPT-2). However, the disparate tokenization schemes employed by these models WordPiece for ViLBERT and byte-pair encoding (BPE) for GPT-2 presented a challenge. During training, we observed that the BPE tokenizer negatively impacted the language modeling loss. To mitigate this, we implemented transfer learning and finetuning techniques to adapt GPT-2 to generate text using the WordPiece tokenizer, resulting in improved performance.

Experiment 1 aims to investigate the extent to which the complexity of the model's architecture affects its performance on the visual dialog task. To apply this objective DistilGPT-2 and GPT-2 models were used. GPT-2 and DistilGPT-2 are both language models developed by OpenAI, but they differ significantly in their size and training methodology. DistilGPT-2 is a smaller, distilled version of GPT-2. It is trained on a subset of the GPT-2 training data using knowledge distillation techniques (Sanh et al., 2019). A student model (DistilGPT-2) is trained to mimic the behavior of a teacher model (GPT-2). This allows DistilGPT-2 to learn from GPT-2's knowledge, even though it is smaller. This makes it more efficient to run, especially on devices with limited computational resources. Table 1 illustrates the architectural details of both decoder models used in our experiment.

Tab. 1. Architectural details of the conducted experiments

Experiment No.	Decoder network	No. of transformer blocks	No. of attention heads	Hidden state size
1	DistilGPT-2	6	6	768
2	GPT-2	12	12	768

To compare the performance of both models, we train each one for 70 epochs (3079 iterations/epoch) with a batch size of 40 using a single A100 GPU on the VisDial dataset v1.0 training set. The learning rate is 2e-5 and linearly decays to 1e-5 after 10k iterations.

Experiment 2 aims to evaluate the performance of the GPT-2 model with and without fine-tuning on the visual dialog task. GPT-2, a pretrained transformer generative model, served as the foundation for our decoder architecture. To assess the model's capacity for the VD task without fine-tuning, we employed a transfer learning approach, freezing the GPT-2 decoder checkpoints while training the ViLBERT encoder for 25 epochs. Subsequently, we conducted a comparative experiment, fine-tuning the GPT-2 decoder and training the entire model end-to-end. The convergence behavior of the fine-tuned model was evaluated across varying numbers of epochs (25, 40, 50, 70), maintaining consistent hyperparameters throughout these experiments.

5.4. Results and discussion

In Experiment 1, it was observed that the GPT-2 decoder converges faster than the DistilGPT-2 decoder during training. Subsequently, GPT-2 outperforms DistilGPT-2 on the VisDial validation set across all evaluation metrics by approximately 3%, as shown in Table 2. Although knowledge distillation techniques have been proven to be effective in transferring knowledge from larger models to smaller ones (Z. Chen et al., 2023; Sanh et al., 2019) there are some limitations in preserving the full complexity and nuances of the original model. In the current case, the VD task has a very complex nature which requires a strong model with deep understanding capabilities to achieve the desired objective, thus GPT-2 has outperformed DistilGPT-2 here.

As shown in Table 2, as training continues the model performance diverges away from the global minima. GPT-2's larger capacity and exposure to a wider range of data make it a strong opponent for achieving superior results in visual dialog systems.

Tab. 2. Evaluation results between GPT-2 And DistilGPT-2 decoders

Decoder	**NDCG**↑	**MRR**↑	**R@1**↑	**R@5**↑	**R@10**↑	**Mean**↓
DistilGPT-2	61.01	49.52	39.64	59.57	66.77	17.37
GPT-2	64.05	52.22	42.34	62.67	70.17	15.37

In experiment 2, it was observed that the GPT-2 decoder, when used without fine-tuning, exhibited suboptimal performance on the visual dialog task. This was primarily attributed to the failure of the LM loss to converge. This can be explained by two key factors: first, GPT-2's pre-training solely on text data does not adequately account for the additional visual features present in the current task, impacting contextual comprehension. Second, the GPT-2 model, utilizing BPE tokenization, is incompatible with the tokenization scheme of the encoder hidden states passed to the decoder, leading to misunderstandings and erroneous sequence generation. Figure 5 shows the LM convergence during training with and without finetuning of the GPT-2 model.

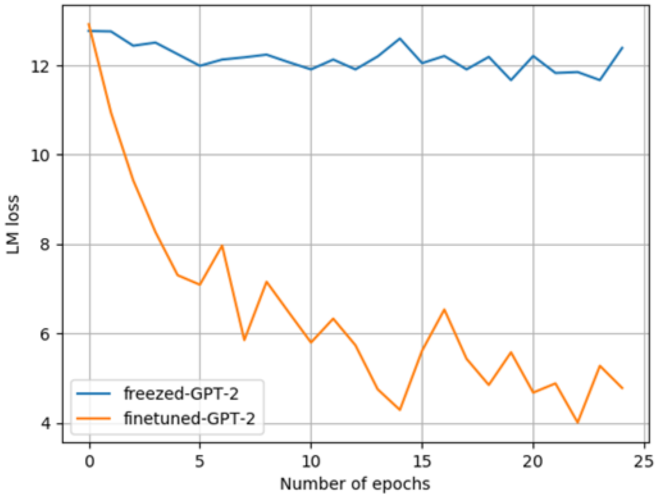


Fig. 5. LM loss convergence during training for 25 epochs for both the freezed and finetuned GPT-2

In contrast to the GPT-2 model with frozen parameters, which exhibited LM loss oscillatory convergence between 12 and 11, the end-to-end trained GPT-2 model demonstrated a positive impact on language modeling performance, converging to a value of 1.2 after training for 70 epochs.

From previous experiments, we observed that our best model is the finetuned GPT-2 trained end-to-end with ViLBERT. We compare our approach with the state-of-the-art models regarding the generative setting as shown in Table 3.

Tab. 3. Comparison with the state-of-the-art model on val. Split VisDial dataset V1.0. $^{\uparrow}$ Indicates higher is better and $^{\downarrow}$ indicates lower is better

Model	**NDCG**$^{\uparrow}$	**MRR**$^{\uparrow}$	**R@1**$^{\uparrow}$	**R@5**$^{\uparrow}$	**R@10**$^{\uparrow}$	**Mean**$^{\downarrow}$
MN (Das et al., 2016)	51.86	47.99	38.18	57.54	64.32	18.60
HCIAE (Lu et al., 2017)	59.70	49.07	39.72	58.23	64.73	18.43
CoAtt (Wu et al., 2018)	59.24	49.64	40.09	59.37	65.92	17.86
Primary (Guo et al., 2019)	-	49.01	38.54	59.82	66.94	16.60
DMRM (F. Chen et al., 2020)	-	50.16	40.15	60.02	67.21	15.19
ReDAN (Gan et al., 2019)	60.47	50.02	40.27	59.93	66.78	17.40
DAM (Jiang et al., 2020c)	60.93	50.51	40.53	60.84	67.94	16.65
KBGN (Jiang et al., 2020a)	60.42	50.05	40.40	60.11	66.82	17.54
LTMI (Nguyen et al., 2020).	63.58	50.74	40.44	61.61	69.71	14.93
MITVG (F. Chen et al., 2021)	61.47	51.14	41.03	61.25	68.49	14.37
UTC (C. Chen et al., 2022)	63.86	52.22	42.56	62.40	69.51	15.67
GST (Kang et al., 2023)	65.47	53.19	43.08	64.09	71.51	14.34
Ours	**64.05**	**52.22**	**42.34**	**62.67**	**70.17**	**15.37**

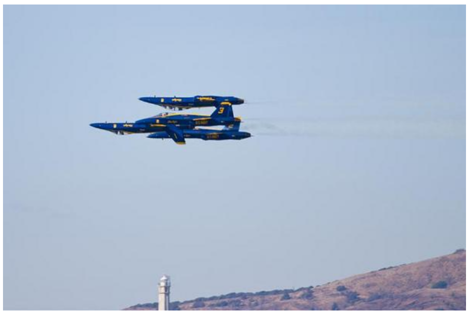
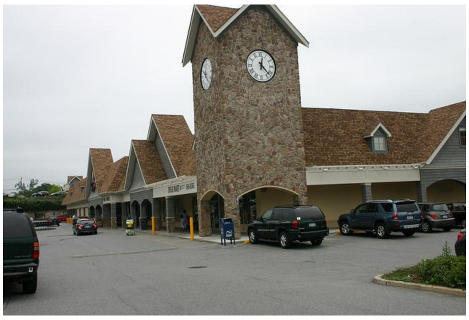
Some of the state-of-the-art models are not transformer-based which are: MN (Das et al., 2016), HCIAE (Lu et al., 2017), CoAtt (Wu et al., 2018), Primary (Guo et al., 2019), DMRM (F. Chen et al., 2020), ReDAN (Gan et al., 2019), DAM (Jiang et al., 2020c), KBGN (Jiang et al., 2020a), and LTMI (Nguyen et al., 2020). While the rest are transformer-based approaches which are: MITVG (F. Chen et al., 2021), UTC (C. Chen et al., 2022), and GST (Kang et al., 2023). The proposed approach is evaluated on the validation split of VisDial v1.0, following the methodology of previous works. The model demonstrates comparable performance on the VisDial v1.0 dataset, surpassing most prior approaches by a significant margin. Specifically, it achieves improvements of 0.19 in NDCG, 0.27 in R@5, 0.66 in R@10, and 0.3 in Mean, falling short only to Kang et al. (2023).

The findings highlight the promising potential of GPT-2 for multi-modal tasks, as evidenced by its strong performance with basic training. In comparison, Kang et al. (2023) (GST) approach, which leverages a substantially larger dataset (12.9 M), required more training time and computational resources. Our model, trained solely on the VisDial dataset (1.2 M), closely approached their best results, demonstrating the effectiveness of GPT-2. In addition, the complexity of our proposed models is one-third less than that of their proposed architecture because of the decoder parameters (117 M vs. 340 M). Moreover, the UTC (C. Chen et al., 2022) algorithm demanded significant computational power (8 A100 GPUs) yet our model outperforms it across NDCG, R@5, R@10, and Mean metrics. Furthermore, MITVG (F. Chen et al., 2021) relied on a single transformer-based encoder, without a decoder, for understanding and generation tasks which hindered their results.

Furthermore, the generated text exhibited a greater degree of realism and informativeness compared to the ground truth responses, as illustrated in Table 4. In certain instances, the model's generated responses are more aligned with the visual context and provide more specific details (highlighted in blue text), surpassing the ground truth in terms of relevance and accuracy. Examining the first sample, the generated answer for the initial question (*Yes, with jet planes*) provides more comprehensive and informative information compared to the ground truth answer (*Yes*). In a real-world scenario, such as an AI copilot, a simple '*Yes*' response would be insufficient. However, the model exhibited an error in the fifth question, likely due to confusion between 'jet steam' discussed in the fourth question and the presence of clouds. Furthermore, in the ninth question, while our model produced a factually correct answer (i.e., 2), it did not align with the intended correct answer (i.e., 3). This discrepancy may be attributed to the influence of the eighth question, which focused on identifying pilots clearly visible in jet #3. Consequently, when prompted to identify the closest jet in the subsequent question, our model prioritized proximity to jet #3, potentially overlooking other relevant factors.

The study suggests that autoregressive models, such as GPT-2, are superior to autoencoder models for visual dialog tasks. While GPT-2 demonstrates a remarkable capacity for generating realistic text, it exhibited limitations in achieving the desired level of contextual understanding when applied to the visual dialog task, compared to the GST model. This discrepancy represents a potential avenue for future research.

Tab. 4. Samples from the generated answers. Text in blue indicates that the generated answer is not the same as the ground truth answer but it logically can be a correct answer. Text in red indicates a wrong-generated answer.

Image & caption	Questions	Ground truth	GPT-2 generated answers
 3 blue colored jets performing barrel rolls through the sky	Q1: is the sky clear?	Yes	Yes, with jet planes.
	Q2: are there other colors on the jets?	Blue and yellow	Blue and orange and black tip.
	Q3: do the jets have numbers?	Yes	Yes
	Q4: is there steam coming out of the jets?	Yes	Yes
	Q5: are there any clouds?	No	Yes
	Q6: can you see the pilots?	Yes	Yes
	Q7: are there any words on the jets?	Yes	Yes, and on the tail propeller.
	Q8: are the pilots wearing helmets	Yes	Yes
	Q9: what number is the closest jet?	3	2
	Q10: are any of the jets upside down?	Yes	Yes
 The shopping center has multiple cars that are pulled up to it	Q1: is this in color?	Yes	Yes
	Q2: are the cars parked?	Yes	Yes, on the sides of the street
	Q3: is this a strip mall?	Yes	Yes
	Q4: is there a restaurant?	I can't see 1	Yes
	Q5: are the cars parked right in front or in a parking lot?	Yes	Yes
	Q6: are any of the cars white?	No	No
	Q7: are any of the vehicles suv's?	Yes	Yes
	Q8: are there any black vehicles?	Yes	Yes
	Q9: is there a health store?	Yes	Yes
	Q10: are there people?	I only see 1	Yes 2 in front of the store

6. CONCLUSION AND FUTURE WORK

This paper introduces a generative model for visual dialog tasks, employing a transformer-based encoder-decoder architecture. ViLBERT, a transformer-based model encodes the input triplet (dialog history, image, and current question) into a contextual vector representation. GPT-2, another transformer-based model, is employed as the decoder, generating an answer grounded in the established context. The proposed model, trained and evaluated on the VisDial dataset v1.0, demonstrates comparable performance to state-of-the-art models. Notably, top-tier scores were achieved in NDCG (64.05), R@5 (62.67), and R@10 (70.17), surpassing most prior works.

Addressing the research question, 'What are the advantages and limitations of using an autoregressive decoder model for visual dialog to generate responses that are consistent with the multi-modal context?', it can be concluded that autoregressive models, particularly transformer-based AR models, are highly effective for sequence generation tasks, especially in conversational systems. These models demonstrate a strong capability to enhance the quality, realism, and engagement of generated responses with minimal effort. However, their application to complex tasks such as visual dialog requires significant computational resources, posing a potential challenge in this domain.

Future work directions will investigate the integration of advanced techniques such as contrastive learning to enhance the model's understanding and response quality. Furthermore, few-shot and zero-shot learning techniques will be explored to enable the model to adapt to datasets with limited or no labels. Additionally, the continual evolution of transformer-based language models motivates an investigation into their applicability within the context of the visual dialog task. Moreover, the practical implications of the model in various real-world applications will be examined to develop more impactful systems. These potential applications can include navigation tasks where an agent must comprehend instructions and utilize visual observations to navigate an environment. Robotic systems enable the dialog agent to interact with the environment, execute actions, and provide feedback based on visual observations. Education by integrating the model into educational systems to facilitate the explanation of complex scientific concepts to students through interactive simulations and visualizations within a conversational framework. Assistive technologies which employ the model to assist visually impaired individuals in interacting with their surroundings.

Author Contributions

Author Contribution All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Ghada M. Elshamy and Marco Alfonse. The first draft of the manuscript was written by Ghada M. Elshamy and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

REFERENCES

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv, abs/2005.14165*. https://doi.org/10.48550/ARXIV.2005.14165

Cadène, R., Dancette, C., Ben-younes, H., Cord, M., & Parikh, D. (2019). RUBi: Reducing unimodal biases in visual question answering. *ArXiv, abs/1906.10169*. https://doi.org/10.48550/arXiv.1906.10169

Chen, C., Tan, Z., Cheng, Q., Jiang, X., Liu, Q., Zhu, Y., & Gu, X. (2022). UTC: A unified transformer with inter-task contrastive learning for visual dialog. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 18082–18091). IEEE. https://doi.org/10.1109/CVPR52688.2022.01757

Chen, F., Meng, F., Chen, X., Li, P., & Zhou, J. (2021). Multimodal incremental transformer with visual grounding for visual dialogue generation. *Findings of the Association for Computational Linguistics*, 436–446. https://doi.org/10.18653/v1/2021.findings-acl.38

Chen, F., Meng, F., Xu, J., Li, P., Xu, B., & Zhou, J. (2020). DMRM: A dual-channel multi-hop reasoning model for visual dialog. *AAAI Conference on Artificial Intelligence* (pp. 7504-7511). https://doi.org/10.1609/aaai.v34i05.6248

Chen, Z., Qiu, G., Li, P., Zhu, L., Yang, X., & Sheng, B. (2023). MNGNAS: Distilling adaptive combination of multiple searched networks for one-shot neural architecture search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(11), 13489-13508. https://doi.org/10.1109/TPAMI.2023.3293885

Cui, X., Khan, D., He, Z., & Cheng, Z. (2023). Fusing surveillance videos and three-dimensional scene: A mixed reality system. *Computer Animation and Virtual Worlds*, *34*(1), e2129. https://doi.org/10.1002/cav.2129

Dai, J., & Zhang, X. (2022). Automatic image caption generation using deep learning and multimodal attention. *Computer Animation and Virtual Worlds*, *33*(3–4), e2072. https://doi.org/10.1002/cav.2072

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M. F., Parikh, D., & Batra, D. (2016). Visual dialog. *ArXiv, abs/1611.08669*. https://doi.org/10.48550/ARXIV.1611.08669

Das, A., Kottur., S., Moura, J. M. F., Lee, S., & Batra, D. (2017). Learning cooperative visual dialog agents with deep reinforcement learning. *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 2970–2979). IEEE. https://doi.org/10.1109/ICCV.2017.321

de Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., & Courville, A. (2016). GuessWhat?! Visual object discovery through multi-modal dialogue. *ArXiv, abs/1611.08481*. https://doi.org/10.48550/ARXIV.1611.08481

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *2019 Conference of (NAACL-HLT*) (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Fan, H., Zhu, L., Yang, Y., & Wu, F. (2020). Recurrent attention network with reinforced generator for visual dialog. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *16*(3), 78. https://doi.org/10.1145/3390891

Gan, Z., Cheng, Y., Kholy, A. E., Li, L., Liu, J., & Gao, J. (2019). Multi-step reasoning via recurrent dual attention for visual dialog. *57th Annual Meeting of the Association for Computational Linguistics* (pp. 6463–6474). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1648

Guo, D., Xu, C., & Tao, D. (2019). Image-question-answer synergistic network for visual dialog. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10426–10435). IEEE. https://doi.org/10.1109/CVPR.2019.01068

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90

Jiang, X., Du, S., Qin, Z., Sun, Y., & Yu, J. (2020a). KBGN: knowledge-bridge graph network for adaptive vision-text reasoning in visual dialogue. *28th ACM International Conference on Multimedia* (pp. 1265–1273). Association for Computing Machinery. https://doi.org/10.1145/3394171.3413826

Jiang, X., Yu, J., Qin, Z., Zhuang, Y., Zhang, X., Hu, Y., & Wu, Q. (2020b). DualVD: An adaptive dual encoding model for deep visual understanding in visual dialogue. *AAAI Conference on Artificial Intelligence* (pp. 11125–11132). AAAI Technical Track: Vision. https://doi.org/10.1609/aaai.v34i07.6769

Jiang, X., Yu, J., Sun, Y., Qin, Z., Zhu, Z., Hu, Y., & Wu, Q. (2020c). DAM: Deliberation, abandon and memory networks for generating detailed and non-repetitive responses in visual dialogue. *Twenty-Ninth International Joint Conference on Artificial Intelligence* (pp. 687–693). https://doi.org/10.24963/ijcai.2020/96

Kang, G.-C., Kim, S., Kim, J.-H., Kwak, D., & Zhang, B.-T. (2023). The dialog must go on: Improving visual dialog via generative self-training. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6746–6756). IEEE. https://doi.org/10.1109/CVPR52729.2023.00652

Kottur, S., Moura, J. M. F., Parikh, D., Batra, D., & Rohrbach, M. (2018). Visual coreference resolution in visual dialog using neural module networks. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (Vol. 11219, pp. 160–178). Springer International Publishing. https://doi.org/10.1007/978-3-030-01267-0_10

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., & Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision, 123,* 32–73. https://doi.org/10.1007/s11263-016-0981-7

Li, L., Huang, T., Li, Y., & Li, P. (2023). Trajectory-BERT: Pre-training and fine-tuning bidirectional transformers for crowd trajectory enhancement. *Computer Animation and Virtual Worlds*, *34*(3–4), e2190. https://doi.org/10.1002/cav.2190

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Computer Vision - ECCV 2014* (pp. 740–755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48

Lin, X., Sun, S., Huang, W., Sheng, B., Li, P., & Feng, D. D. (2023). EAPT: Efficient attention pyramid transformer for image processing. *IEEE Transactions on Multimedia*, *25*, 50–61. https://doi.org/10.1109/TMM.2021.3120873

Liu, A.-A., Zhang, G., Xu, N., Guo, J., Jin, G., & Li, X. (2022). Closed-loop reasoning with graph-aware dense interaction for visual dialog. *Multimedia Systems*, *28*, 1823–1832. https://doi.org/10.1007/s00530-022-00947-1

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, *32*.

Lu, J., Kannan, A., Yang, J., Parikh, D., & Batra, D. (2017). Best of both worlds: transferring knowledge from discriminative learning to a generative visual dialog model. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.

Murahari, V., Batra, D., Parikh, D., & Das, A. (2020). Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (Vol. 12363, pp. 336–352). Springer International Publishing. https://doi.org/10.1007/978-3-030-58523-5_20

Nguyen, V.-Q., Suganuma, M., & Okatani, T. (2020). Efficient attention mechanism for visual dialog that Can handle all the interactions between multiple inputs. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (Vol. 12369, pp. 223–240). Springer International Publishing. https://doi.org/10.1007/978-3-030-58586-0_14

OpenAI. (2023). GPT-4 technical report. *ArXiv, abs/2303.08774*. https://doi.org/10.48550/arXiv.2303.08774

Radford, A., & Narasimhan, K. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *ArXiv, abs/1910.01108*. https://doi.org/10.48550/ARXIV.1910.01108

Schwartz, I., Yu, S., Hazan, T., & Schwing, A. G. (2019). Factor graph attention. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2039–2048). IEEE. https://doi.org/10.1109/CVPR.2019.00214

Seo, P. H., Lehrmann, A. M., Han, B., & Sigal, L. (2017). Visual reference resolution using attention memory for visual dialog. *Advances in Neural Information Processing Systems 30*, *30*, 3719–3729.

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *56th Annual Meeting of the Association for Computational Linguistics* (pp. 2556–2565). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1238

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems 30*, *30*.

Wang, A., & Cho, K. (2019). BERT has a mouth, and It must speak: BERT as a markov random field language model. *Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, 30–36. https://doi.org/10.18653/v1/W19-2304

Wang, Y., Joty, S. R., Lyu, M. R., King, I., Xiong, C., & Hoi, S. C. H. (2020). VD-BERT: A unified vision and dialog transformer with BERT. *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3325–3338). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.269

Wu, Q., Wang, P., Shen, C., Reid, I., & Hengel, A. V. D. (2018). Are you talking to me? reasoned visual dialog generation through adversarial learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6106–6115). IEEE. https://doi.org/10.1109/CVPR.2018.00639

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. *ArXiv, abs/1609.08144*. https://doi.org/10.48550/ARXIV.1609.08144

Xin, B., Xu, N., Zhai, Y., Zhang, T., Lu, Z., Liu, J., Nie, W., Li, X., & Liu, A.-A. (2023). A comprehensive survey on deep-learning-based visual captioning. *Multimedia Systems*, *29*, 3781–3804. https://doi.org/10.1007/s00530-023-01175-x

Yang, T., Zha, Z.-J., & Zhang, H. (2019). Making history matter: History-advantage sequence training for visual dialog. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 2561–2569). IEEE. https://doi.org/10.1109/ICCV.2019.00265

Yu, Y., Yang, Y., & Xing, J. (2024). PMGAN: Pretrained model-based generative adversarial network for text-to-image generation. *The Visual Computer*, *44*, 303–314. https://doi.org/10.1007/s00371-024-03326-1

Yu, Z., Yu, J., Cui, Y., Tao, D., & Tian, Q. (2019). Deep modular co-attention networks for visual question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6274–6283). IEEE. https://doi.org/10.1109/CVPR.2019.00644

Zhang, B., Ma, R., Cao, Y., & An, P. (2024). Swin-VEC: Video swin transformer-based GAN for video error concealment of VVC. *The Visual Computer*, *40*, 7335–7347. https://doi.org/10.1007/s00371-024-03518-9

Zhang, J., Zhao, T., & Yu, Z. (2018). Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog. *19th Annual SIGdial Meeting on Discourse and Dialogue* (pp. 140–150). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5015

Zhao, L., Lyu, X., Song, J., & Gao, L. (2021). GuessWhich? Visual dialog with attentive memory network. *Pattern Recognition*, *114*, 107823. https://doi.org/10.1016/j.patcog.2021.107823

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 19–27). IEEE. https://doi.org/10.1109/ICCV.2015.11