Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.038695
Article

Relevant Visual Semantic Context-Aware Attention-Based Dialog

Eugene Tan Boon Hong[1], Yung-Wey Chong[1,*], Tat-Chee Wan[1] and Kok-Lim Alvin Yau[2]

[1]National Advanced IPv6 Centre, Universiti Sains Malaysia, Penang, Malaysia
[2]Lee Kong Chian Faculty of Engineering and Science (LKCFES), Universiti Tunku Abdul Rahman, Sungai Long, Selangor, Malaysia
*Corresponding Author: Yung-Wey Chong. Email: chong@usm.my
Received: 25 December 2022; Accepted: 23 May 2023; Published: 30 August 2023

Abstract: The existing dataset for visual dialog comprises multiple rounds of questions and a diverse range of image contents. However, it faces challenges in overcoming visual semantic limitations, particularly in obtaining sufficient context from visual and textual aspects of images. This paper proposes a new visual dialog dataset called Diverse History-Dialog (DS-Dialog) to address the visual semantic limitations faced by the existing dataset. DS-Dialog groups relevant histories based on their respective Microsoft Common Objects in Context (MSCOCO) image categories and consolidates them for each image. Specifically, each MSCOCO image category consists of top relevant histories extracted based on their semantic relationships between the original image caption and historical context. These relevant histories are consolidated for each image, and DS-Dialog enhances the current dataset by adding new context-aware relevant history to provide more visual semantic context for each image. The new dataset is generated through several stages, including image semantic feature extraction, keyphrase extraction, relevant question extraction, and relevant history dialog generation. The DS-Dialog dataset contains about 2.6 million question-answer pairs, where 1.3 million pairs correspond to existing VisDial's question-answer pairs, and the remaining 1.3 million pairs include a maximum of 5 image features for each VisDial image, with each image comprising 10-round relevant question-answer pairs. Moreover, a novel adaptive relevant history selection is proposed to resolve missing visual semantic information for each image. DS-Dialog is used to benchmark the performance of previous visual dialog models and achieves better performance than previous models. Specifically, the proposed DS-Dialog model achieves an 8% higher mean reciprocal rank (MRR), 11% higher R@1%, 6% higher R@5%, 5% higher R@10%, and 8% higher normalized discounted cumulative gain (NDCG) compared to LF. DS-Dialog also achieves approximately 1 point improvement on R@k, mean, MRR, and NDCG compared to the original RVA, and 2 points improvement compared to LF and DualVD. These results demonstrates the importance of the relevant semantic historical context in enhancing the visual semantic relationship between textual and visual representations of the images and questions.

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Keywords: Visual dialog; context-aware; relevant history; computer vision; natural language processing

1 Introduction

In recent years, remarkable advancements have been made in the fields of computer vision (CV) and natural language processing (NLP). These developments have been applied in various Artificial Intelligence (AI) tasks such as image classification, scene recognition, question answering, object detection, image retrieval [1], cybersecurity [2], malware detection, and many more. The NLP field is mainly focused on sentence generation, semantics, sentence semantic matching [3], among other related areas. Currently, majority of works in image caption generation concentrate on developing better methods and generate more accurate captions, with only a few aimed at enhancing the understanding of the question context. In contrast, Visual Dialog seeks to integrate conversational context into Visual Question Answering by collecting conversational data via Amazon Mechanical Turk (AMT). This data is gathered by having two workers engaging in a conversation based on the MSCOCO-2014 [4] dataset with captions provided. Fig. 1 depicts a sample validation image from MSCOCO-2014 and a snapshot of VisDial conversational dialogue between two AMT workers based on the image.

Figure 1: Visual dialog's sample question and answer pairs

However, most of the questions covered in the VisDial dataset are short and fail to capture the context of the whole image. As illustrated in Fig. 1, the questions in conversational dialogue do not encompass all the latent information, including the surrounding objects such as "bench", "handbag", and "suitcase". Instead, the questions only ask vague queries such as "any animals", "any other people", and "is it in the country". Consequently, the learned model can only comprehend the image context partially, and the generated answers are also brief and straightforward. Therefore, this research aims to comprehensively comprehend the image semantic context and capture the latent context relationships between the context and semantic information of the image. This approach helps to generate more comprehensive answers based on conversational questions.

This research aims to improve the contextual understanding of multi-round visually grounded dialogs through the development of a context-aware dataset, named DS-Dialog. With DS-Dialog, more relevant conversational history context can be provided to the model, resulting in more comprehensive answers. As shown in Fig. 2, the data extraction process involves the use of Faster Region-based Convolutional Neural Network (Faster R-CNN) [5] to detect image semantic features such as "person", "bench", and "suitcase". DS-Dialog then proceeds to identify the top-n ($n = 1, 2, 3, 4, 5$) relevant history corresponding to the image features, where n represents the ratio for a feature's relevant history. This process is illustrated in Fig. 3, where each image in DS-Dialog is enhanced with context-aware relevant history, which is trained in tandem with existing conversational dialogs from VisDial. To maintain computational efficiency, DS-Dialog contains a maximum of five image semantic features for each question in a 10-round dialog. The image semantic features are determined based on the 80 MSCOCO-2014 image categories as shown in Fig. 4 below.

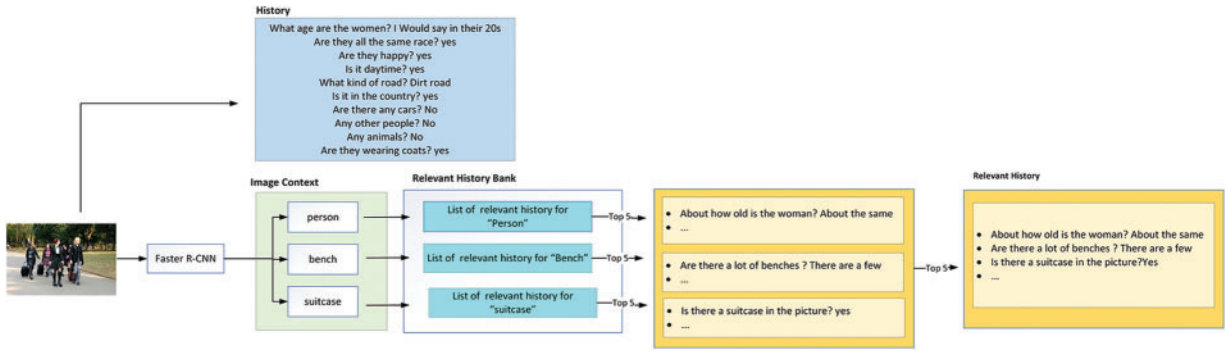

Figure 2: Illustration of the intuition of DS-Dialog

This research makes the following contributions: (1) introducing a new dataset, DS-Dialog, which is a context-aware DS-Dialog dataset that comprises relevant histories to provide more visual semantic information about an image based on other images with similar image features; and (2) proposing a novel adaptive relevant history selection, in addition to question-guided and relevant history-guided visual attentions, to resolve the missing visual semantic information for each image.

2 Related Work

In the field of computer vision, Visual Dialog is an extension of Visual Question Answering (VQA) [6]. VQA is a deep learning model that provides answers based on both images and their accompanying questions. It can be considered as an enhancement of image captioning, which has been limited to generating textual descriptions solely based on an image.

2.1 Image Captioning

Image captioning refers to the process of generating textual descriptions of an image. Recent research such as Visual Vocabulary Pre-Training for novel object captioning (VIVO) [7] and Object Semantics Aligned pre-training (OSCAR) [8], is leveraging the latest advances of NLP to provide better image descriptions. VIVO is trained based on visual-text alignments using image-text pairs, while OSCAR achieved poorer image captioning results by using a simplified alignment method that uses the image's object tags. Prior to VIVO and OSCAR, many studies have focused on novel image captioning [9], attention-based, and text summarization [10]. Reference [11] utilize a phrase-based

image captioning that encodes sequences of phrases and words. Densecap [12] has been widely used as it can provide region localization and image descriptions. Attention-based image captioning has gained popularity among these frameworks as it has been shown to achieve promising results.

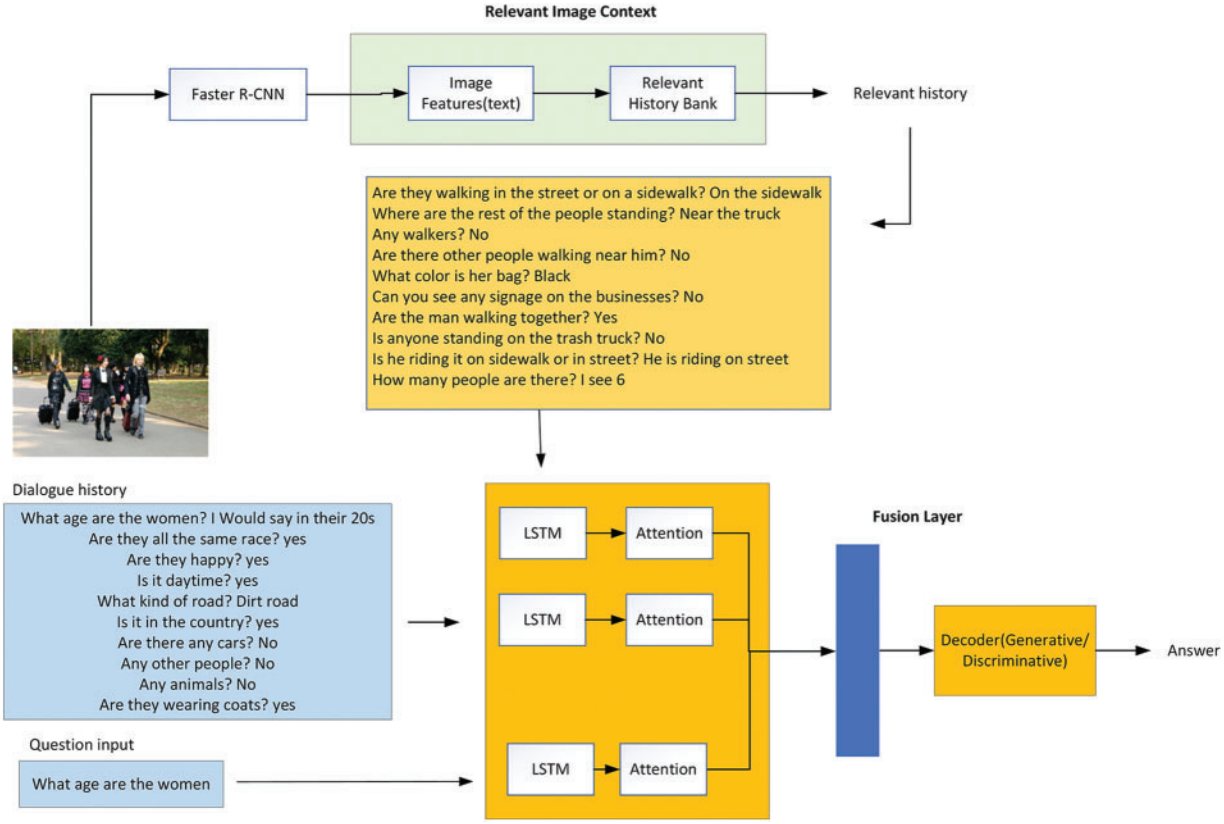

Figure 3: The context-aware DS-Dialog framework

Attention-based image captioning has been proposed by [13] to assist models in selecting the most relevant region for generating words during sentence generation by paying attention to salient objects. Reference [14] has developed a semantic attention model to attend to semantic concepts and incorporate them via top-down and bottom-up combinations. Reference [15] has proposed an adaptive attention approach that automatically determines on when and where to look respectively. Similarly, reference [16] has proposed an adaptive attention approach implemented using DenseNet. Multi-attention Generative Adversarial Network (MAGAN) [17] utilises both local and non-local attention modules for more effective feature representations.

2.2 Visual Question Answering

Unlike image captioning, which generates image descriptions, VQA provide answers based on a given question and image. VQA has the capability of cross-modal understanding and reasoning of vision and language, which sets it apart from image captioning. Recent VQA works focus on visual attention [18,19], adversarial approach [20], and handling open-ended question answering task [21]. To further enhance text representations, [22] have introduced external large-scale knowledge bases such as

DBpedia [23] to enrich the combination of both image captioning and VQA. References [24,25] utilise the attention model to retrieve region context intelligently.

COCO categories							
Food	Human	vehicle	road	animal	safety	bag	cloth
banana apple sandwich orange broccoli carrot hot dog pizza donut cake	Person	bicycle car motorcycle airplane bus Train truck boat	Traffic light Stop sign Parking meter	Bird Cat Dog Horse sheep cow elephant bear zebra giraffe	Fire hydrant	Handbag Suitcase backpack	tie
Outdoor games	utensil	toilet	gadgets	kitchen	Others	Furniture	
frisbee skis snowboard sports ball kite baseball bat baseball glove skateboard surfboard tennis racket	bottle wine glass cup fork knife spoon bowl	Toilet toothbrush	Tv Mouse Keyboard Laptop Remote Cell phone	Microwave Toaster Refrigerator Oven sink	Umbrella Vase Potted plant Hair drier Teddy bear Scissors book clock	Bench chair couch Dining table bed	

Figure 4: MSCOCO-2014 categories

2.3 Visual Dialog

However, previous works on image captioning and VQA do not include conversational context. Visual Dialog, unlike VQA, learns from multiple contexts such as multi-round dialogues, images, and questions. Visual Dialog was initially introduced by [26], which was later extended with deep reinforcement learning [27] but previous works have led to repetitive dialogues. To address this issue, [28] have proposed a method to penalize the question-bot that generates duplicated questions using a smooth-L1 penalty over questions with a high similarity score, which has improved the model's image-guessing ability. GuessWhat [29] focused on object discovery with yes or no questions. Reference [30] is used to transfer knowledge from discriminative learning to generative learning. It uses the current question to attend to the exchanges in the history, and then uses the question and attended history to attend to the image to get the final encoding. The attention model in this work can help the discriminator in paraphrasing answers. However, Visual Dialog still ignores the semantic feature of the images. To address this limitation, reference [31] has proposed a method involving object feature extraction and selection to extract relevant visual information from images and filter irrelevant visual information, which is assisted by of semantic guidance from both question and dialog history.

The attention mechanism has a significant impact on improving Visual Dialog, particularly with the introduction of self-attention mechanism proposed in [32]. In addition to penalizing approach, the attention mechanism has been used to improve the Visual Dialog performance. For example, in [33], Attention Memory (AMEM) has created a new synthetic visual dialog dataset called MNIST-Dialog, which is the combination of MNIST (Mixed National Institute of Standards and Technology) and VisDial datasets to resolve the Visual Dialog's sequential dependencies through an attention memory and a dynamic attention combination process. Reference-Aware Attention Network (RAA-Net) [34]

has proposed multi-head textual attention and visual-two-step reasoning to overcome latent semantic and semantic correlation issues, respectively, for generating better answers. Recursive Visual Attention (RVA) [35] aims to overcome existing soft attention that is unable to predict the discrete attentions over topic-related history by introducing recursive visual attention. RVA can make discrete decisions as a response to the input content by recursively browsing the dialog history and computes the visual attentions until it meets an unambiguous description. The synergistic model [36] has been introduced to generate more comprehensive answers rather than just "yes" and "no". Recently, research has attempted to resolve the visual co-reference using the neural networks at the word level [37]. Further, Visual Dialog does not emphasize the conversation history and only exploits ground-truth history. Wrong answers are imposed in a conversational context and collected measurement based on the adverse critics [38]. In [39], low-level information in both image and text are covered via three low-level attention modules such as H2H attention that focuses on connections between words, H2Q attention, and R2R attention which focuses on the relationship between spatial feature and object feature.

Meanwhile, Dual Encoding Visual Dialogue (Dual VD) has been proposed in [40] to extract objects and their relationships from the visual module and feed them into the semantic module. Multi-level image captions, which combine both image captions and dense captions, have been utilized to provide a more comprehensive description of the image by localizing and describing image regions in a natural language.

The Dialog Network, which has been introduced in [41], aims to improve existing Visual Dialogs by accurately understanding questions. This enhancement to the Visual Dialog encoder allows for better representation and focus on the intended region of interest.

2.4 Transformers and BERT

As Recurrent Neural Network (RNN) encountered bottlenecks at the end of computations with sequential text processing, transformers with its self-attention module have been introduced to compute the attention score for each sentence with parallelism [42]. Transformers have made a significant impact and have been widely adapted by works in NLP, image captioning [43,44], scene segmentation [45], etc. Bidirectional Encoder Representations from Transformers (BERT) [46] has been introduced by Google with 340 M parameters, and it is trained on 3.3 billion words. BERT has set new state-of-the-art performance of various NLP tasks such as sentence classification, sentence-pair regression tasks like semantic textual similarity (STS), question-answering(QA) [47], text-classification (TC) [48], natural language understanding(NLU) [49,50] and keyphrase extraction [51]. BERT is trained using the masked language modelling (MLM) task that randomly masks some tokens in a text sequence, and then independently recovers the masked tokens by conditioning on the encoding vectors obtained by a bidirectional Transformer. Numerous works have been carried out to improve BERT, such as SentenceBERT [52], VisualBERT [53], VD-BERT [54], VU-BERT [55], RoBERTa [56], VisDial-BERT [57], ALBERT [58], DistillBERT [59], SpanBERT [60], and KeyBERT [61]. RoBERTa is more robust than BERT and it is trained using much more training data. ALBERT reduces memory consumption and increases the training speed of BERT. DistillBERT utilizes knowledge distillation during pre-training to reduce the size of BERT by 40% while retaining 99% of its original capabilities and making the inference 60% faster. SpanBERT extends BERT to better represent and predict text spans. SentenceBERT is introduced to overcome the drawbacks of BERT embeddings, which produces bad sentence embeddings. SentenceBERT is using Siamese and triplet network structures to derive sentence embeddings with semantic meanings. Meanwhile, KeyBERT aimed to augment the quality of extracted keyphrases.

2.5 Visual Dialog Datasets

Various works such as CLEVR-Dialog [62] and MNIST-Dialog, propose new visual dialogues for new test cases. CLEVR-Dialog focuses on visual reasoning using images from diagnostic datasets such as Compositional Language and Elementary Visual Reasoning (CLEVR) [63]. Meanwhile, MNIST-Dialog consists of images of MNIST digits, and it uses attention memory to resolve visual co-reference. Attention memory helps the neural network to learn by storing an image attention map in each round.

However, most frameworks focus on the correlation between the image context and the conversation context, without emphasizing the image semantic context, which helps to generate more comprehensive answers alongside conversational dialog. Although RVA can infer visual co-reference between questions and history, it does not contribute to a better image context. For example, this research makes general deduction based on what has been observed from the image, including identified objects such as people, cars, and so on. Since images can have overlapping objects, it is a possible that more than one image contains a similar context.

3 Methodology

In this section, the generation process for DS-Dialog, which is an enhanced dataset containing contextual conversational history related to the image semantic context, is explained in Section 3.1. This is followed by the explanation of Context-Aware DS-Dialog in Section 3.2.

3.1 DS-Dialog Dataset

The DS-Dialog dataset generation process consists of three stages: (a) extraction of image semantic feature (b) identification of relevant question for each image (c) creation of relevant history bank.

Fig. 5 illustrates the formation of the relevant history bank, which includes all 80 MSCOCO image feature categories. Each category in the MSCOCO image feature set comprises of a list of relevant questions that can be utilized as relevant history during the training of the model.

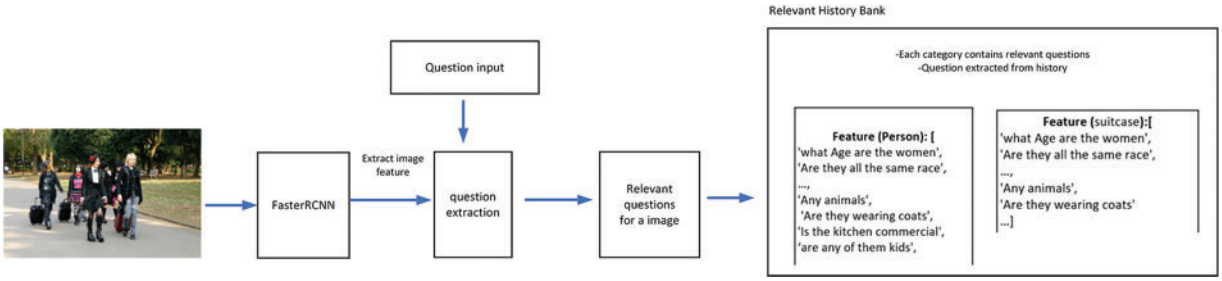

Figure 5: Relevant history bank

3.1.1 Image Semantic Feature Extraction

The image contains detailed information about the image context. Image semantic feature extraction is used to extract important visual context about the image using Faster R-CNN. Faster R-CNN determines the coordinates and categories for each detected object in the bounding box. The categories of each detected object are determined based on the 80 MSCOCO image feature categories. The extracted image features are denoted as $F = \{f_1, f_2, \ldots, f_n\}$, where n is the number of extracted image features.

This research then adopts KeyBERT to perform keyphrase extraction from the data itself to find the top two highest pairwise cosine similarity scores between the image features and the MSCOCO category group as depicted in Fig. 6. The purpose of this research is to gain insights into the closest semantic relationship between the detected image semantic information and relevant MSCOCO keywords. This leads to the generation of relevant histories for the image.

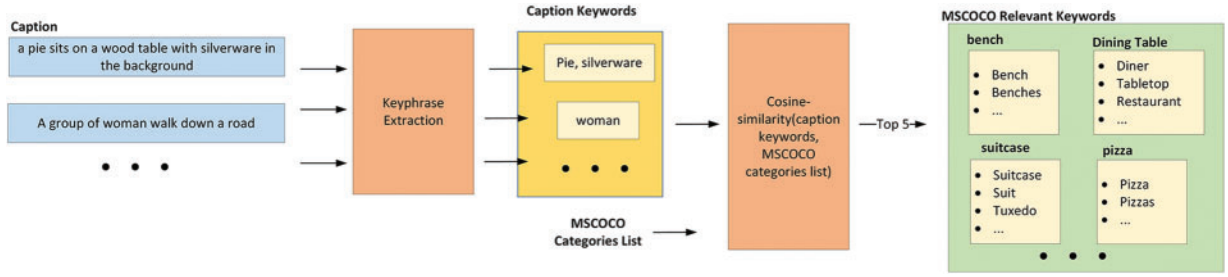

Figure 6: Keyphrase extraction based on image caption

3.1.2 Relevant Question Extraction

The top relevant questions for the detected image feature in the MSCOCO category are extracted using cosine similarity, as shown in Fig. 7. Eq. (1) represents the cosine similarity equation between two sentences, where S_1 is the vector of the source sentence, and S_2 is the vector of the target sentence.

$$\cos\theta = \frac{S_1 \times S_2}{|S_1| \times |S_2|} \tag{1}$$

This research adapted SentenceBERT to perform relevant question extraction, using fine-tuned STS-benchmark (STSb) RoBERTa model. This research will also feed the VisDial history and respective image caption into SentenceBERT to extract the top five relevant questions based on the top five semantic features detected based on image input. SentenceBert will calculate the cosine similarity between the image caption and dialog history context. Pairwise cosine similarity is calculated between each question in the original VisDial history and the respective image caption. The sentence are then sorted in descending order based on their similarity score, and the top two sentences are extracted from each sorted sentences based on each semantic feature. This process is repeated for each detected semantic feature, resulting in the final relevant question history bank, which consists of 10 relevant question histories as shown in Fig. 7.

3.1.3 Dialog Generation

Before proceeding with the neural network training, this research generates a relevant history bank by grouping all questions that pertain to the same feature, as illustrated in Fig. 5. The Faster R-CNN is used to extract image features based on the MSCOCO categories, as depicted in Fig. 4, and the original question history is tagged along with the features. For example, the image features extracted from the sample image input in Fig. 2 are "person", "suitcase", and "bench". Then, the original question history from the VisDial dataset that mainly consists of ten questions, as shown in Fig. 1, will be added to each of the feature question lists. In other words, the question history for image in Fig. 1 is saved into the "person", "suitcase", and "bench" feature question lists respectively. This step is repeated for all the images in the dataset. Consequently, the relevant question dataset has a total of 80 features, and each feature contains a list of questions that are relevant to the image semantic feature.

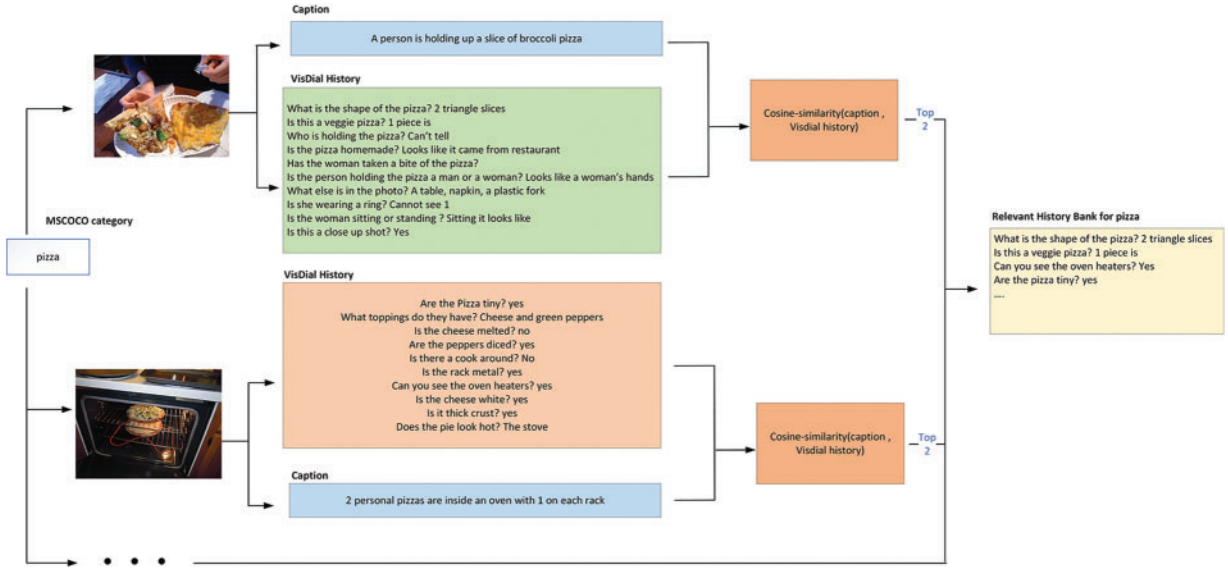

Figure 7: Relevant question extraction process for a MSCOCO feature

Fig. 8 shows the total number of the relevant questions, that can be retrieved for training and validation datasets to form the relevant history bank, respectively, based on MSCOCO categories.

MSCOCO category	relevant question history (train)	relevant question history (val)
person	3188	43
bicycle	2495	66
car	1643	33
motorcycle	2594	60
airplane	2509	64
bus	3086	92
train	3581	110
truck	2393	53
boat	2123	39
Traffic light	1186	17
Fire hydrant	1216	37
Stop sign	2266	46
Parking meter	2554	66
bench	1970	40
bird	2887	65
cat	4194	94
dog	4195	91
horse	3159	88
sheep	1171	32
cow	1667	21

MSCOCO category	relevant question history (train)	relevant question history (val)
elephant	2554	70
bear	2102	49
zebra	1986	53
giraffe	2652	69
backpack	373	14
umbrella	2339	58
handbag	1301	27
tie	196	10
suitcase	1289	34
frisbee	1726	35
skis	2687	66
snowboard	2712	57
Sports ball	3512	53
kite	1843	57
Baseball bat	3499	63
Baseball glove	2791	34
skateboard	3473	76
surfboard	2260	87
Tennis racket	3219	116
bottle	1197	34

MSCOCO category	relevant question history (train)	relevant question history (val)
Wine glass	2083	46
cup	235	4
knife	1206	38
spoon	245	8
bowl	1224	24
banana	1914	49
apple	471	17
sandwich	3762	91
orange	774	25
broccoli	1154	20
carrot	450	18
Hot dog	689	36
pizza	4419	120
donut	1247	24
cake	2126	46
chair	3997	88
couch	2463	52
Potted plant	208	4
bed	4060	110
Dining table	3275	79

MSCOCO category	relevant question history (train)	relevant question history (val)
toilet	3699	121
tv	1133	35
laptop	3188	81
mouse	605	17
remote	1526	51
keyboard	2776	69
Cell phone	1711	34
microwave	1293	19
oven	3877	113
toaster	777	28
sink	1643	62
refrigerator	1479	29
book	300	16
clock	4050	72
vase	1361	24
scissors	1206	38
Teddy bear	2646	56
Hair drier	619	35
toothbrush	586	22

Figure 8: DS-Dialog relevant history based on MSCOCO categories

As each image contains various visual semantic information, it may have multiple MSCOCO categories assigned to it. For example, the sample image in Fig. 2 was assigned to three MSCOCO categories: “person”, “suitcase”, and “bench”. Using Eq. (1), relevant question history can be further fine-tuned to obtain the top ten relevant history questions as depicted in Fig. 9. The final result of

relevant history will be 10 relevant questions for all images, which in total, amounts to 2.6 million dialogs as depicted in Table 1.

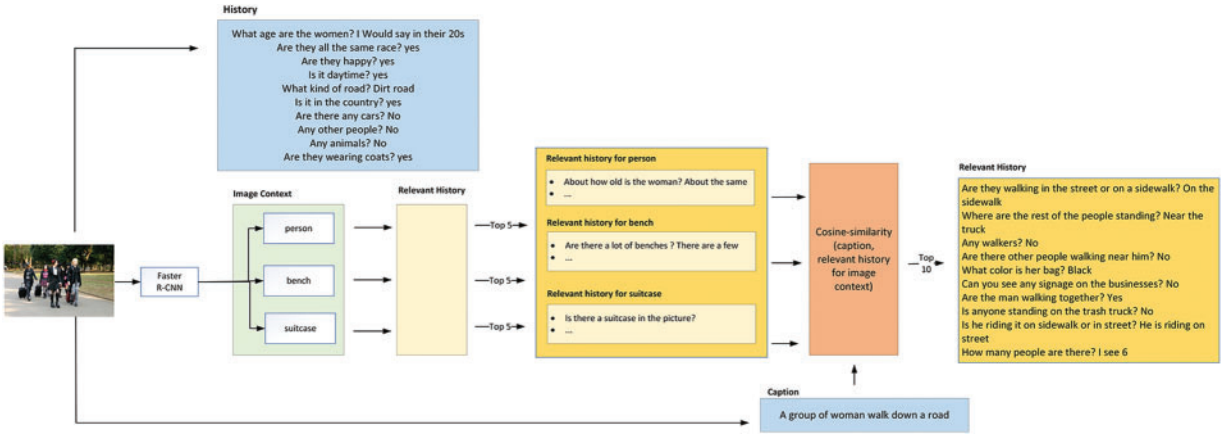

Figure 9: Relevant history generation for an image in DS-Dialog

Table 1: Dataset statistics comparing VisDial to DS-Dialog

Name	DS-Dialog (ours)	VisDial
Images	133 K	133 K
Dialogs	2.6 M	1.3 M
a) History	1.3 M	1.3 M
b) Relevant history	1.3 M	-

3.2 Context-Aware DS-Dialog

3.2.1 Feature Representation

To avoid introducing irrelevant visual information, this research proposes incorporating relevant history into the model training with relevant semantic guidance. The relevant history is collected based on the image semantic information for each question Q_t in round t, with each word in the questions embedded using the Global Vectors for Word Representation (GLoVE) [64] embedding matrix.

Visual Feature. Object features are extracted using Faster R-CNN which contains visual information for attributes and semantic concepts.

Text Feature. Word embedding of the questions, history, and relevant history are defined as $W = \{W_1, W_2, \ldots, W_n\}$, $H = \{H_1, H_2, \ldots, H_n\}$, and $R = \{R_1, R_2, \ldots, R_n\}$, respectively. These word embeddings are passed through a bidirectional Long Short-Term Memory (LSTM). The matching score is calculated between question and history, question, and relevant history to provide better image context.

3.2.2 Adaptive Relevant History Selection

In order to incorporate both visual and semantic image representations, SentenceBERT is used to selectively extract relevant history information based on image and question input, thereby providing

more semantic context in addition to the existing history, question, and visual input. As illustrated in Fig. 10, the extracted image caption keywords are matched against MSCOCO-relevant keywords, and if a match is found, common keywords such as "ovens" and "pizza" are used to locate their corresponding history from the newly generated relevant history bank as described in Section 3.1.2. The process of adaptively selecting relevant history is also depicted in Fig. 10, where the top ten relevant histories from the corpus of relevant history for the "oven" and "pizza" features. The minimum cosine similarity score used in SentenceBERT is set to 0.5.

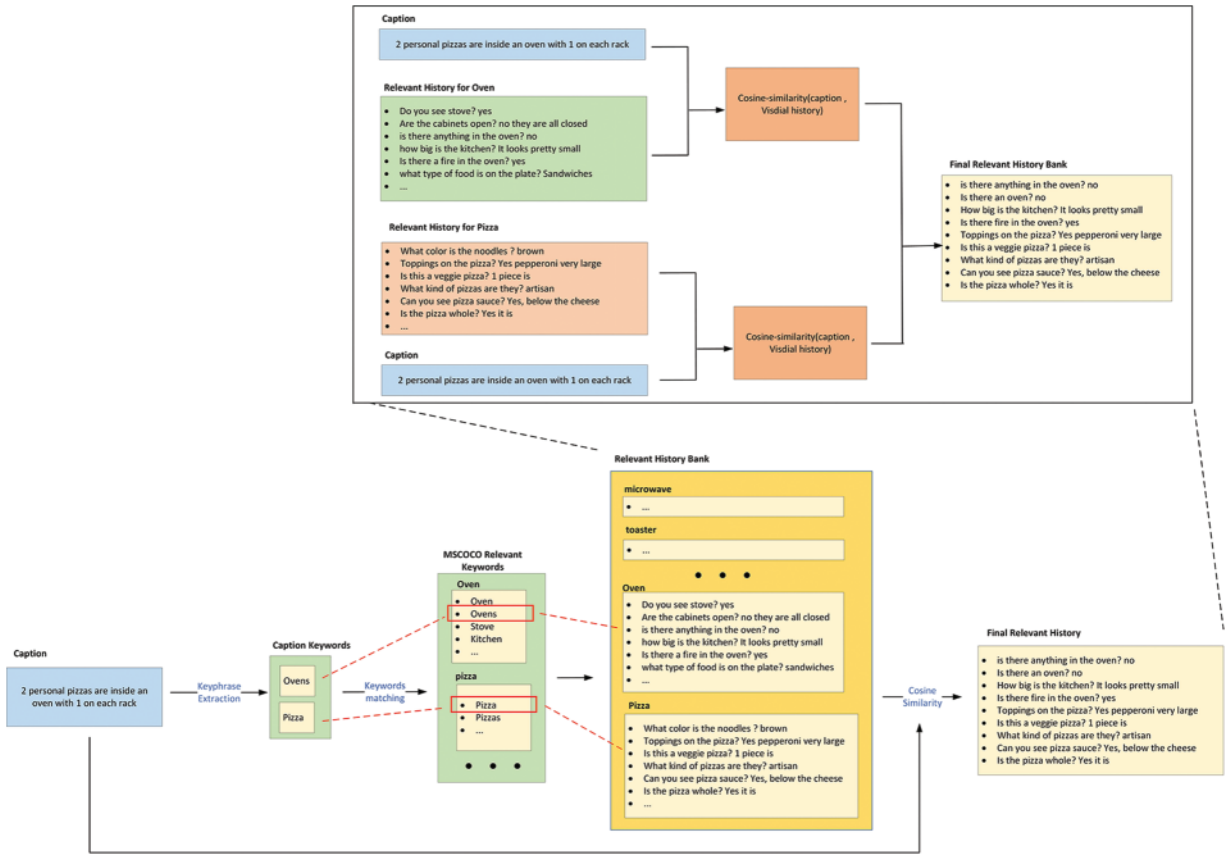

Figure 10: Adaptive relevant history selection in DS-Dialog

3.2.3 Attention Module

The attention module is divided into two parts, namely question-guided visual attention and relevant history-guided visual attention. Given that the model inputs, which consists of visual feature $V_t = \{v_1, \ldots, v_k\}$, question Q_t, history $H_t = \{h_1, \ldots, h_t\}$, and relevant history $R_t = \{r, \ldots, r_t\}$, the attention weights α_x and α_y are calculated for question-guided and relevant-history-guided visual attention, respectively, using the following equations:

$$x_t^a = \text{L2Norm}\left(f_q^a Q_t * f_v^a V_t\right) \tag{2}$$

$$\alpha_x^a = \text{softmax}\left(W^a x_t^a\right) \tag{3}$$

$$y_t^a = \text{L2Norm}\left(f_r^a R_t * f_v^a V_t\right) \tag{4}$$

$$\alpha_y^a = \text{softmax}\left(W^a y_t^a\right) \tag{5}$$

where f_q^a , f_r^a, and f_v^a represent non-linear transformation to visual and semantic feature embeddings, while $*$ denotes an element-wise operation between text semantic and visual.

3.3 Multimodal Fusion

The same method in [26] is used to calculate joint embedding, E_t^J by fusing the visual feature V_t, the question feature Q_t, the history feature H_t, and the relevant history feature R_t with tangent activation.

$$E_t^J = \tanh\left(W^B \left| V_t \cdot Q_t \cdot H_t \cdot R_t \right|\right) \tag{6}$$

where $\cdot$ denotes the concatenation operation.

4 Implementation

4.1 Datasets Preparation and Setup

The experiment utilises the VisDial v1.0 dataset which contains 123 K images for training, 2 K images for validation, and 8 K images for testing. Each image in the dataset is accompanied by a caption, and a text dialog that comprises of ten rounds of questions and answers, except the testing dataset. The history list is formed by combining ten rounds of questions and answers. Moreover, the proposed model extends the VisDial v1.0 dataset by integrating relevant questions based on the detected image semantic features and generating the relevant history. As a result, for each visual input, there is a relevant history comprising 10 round question-answers pairs in addition to the existing history.

4.2 Implementation Details

The proposed model in this research is implemented in PyTorch. A pretrained Faster R-CNN is employed as the object-detector to extract the 2048-dim visual context from input images. All LSTMs have 2 layers with 512-dim hidden states. All words and images are represented with 300-dim embeddings. During training, the proposed model is optimized by minimizing cross-entropy loss using the Adam optimizer with a learning rate of 0.01. Training for discriminative models is carried out with 3 epochs and a batch size of 24. The proposed model is trained with the combination of the original warm-up strategy and the cosine annealing learning strategy together to learn the model parameters. The parameters used in this research include the warm-up factor of 0.3 and the cosine annealing learning strategy, with an initial learning rate of 1×10^{-3} and a termination learning rate of 3.4×10^{-4}. All experiments are conducted on four NVIDIA Tesla T4 with 16 GB memory. Since too many relevant question histories can reduce the computation efficiency, only the top ten relevant histories are selected for each of the image features extracted.

Evaluation Metrics: The generated answer accuracy is evaluated by retrieving the ground truth answer from a 100-option answer list [24]. This research adopts a retrieval-based evaluation metrics set which includes (a) the mean rank of human response; (b) Recall@K (R@K), which is the percentage of human response in top-k ranked responses; (c) the mean reciprocal rank of the human response (MRR); and (d) the Normalised Cumulative Gain (NDCG), which penalizes answers with the a lower rank of the high relevance. The model has a better accuracy if it has a lower mean value and higher values in R@k and MRR.

4.3 Baselines

This research evaluates the proposed DS-Dialog model by comparing its performance with other Visual Dialog models. Specifically, the discriminative visual dialogs from Late Fusion (LF), RVA, and DualVD are adopted for benchmarking. LF is a model that initially encodes image, history, and questions individually, and then combines them through concatenation. RVA is an attention-based model that iteratively refines the visual attention by incorporating the visual co-reference solution when the model obtains sufficient confidence. DualVD, on the other hand, is a dual encoding model that extracts information from an image.

4.4 Results and Discussion

Comparison of model variants: In this research, the newly generated DS-Dialog dataset is used to compare the performance of different variants of the existing approaches, including: (a) without SentenceBERT to extract a maximum of 400 words of the relevant histories; (b) without Sentence-BERT to extract a maximum of 500 words for the relevant histories; (c) without SentenceBert and accepts a maximum number of 800 words for the relevant histories; (d) with SentenceBert and accepts a maximum number of 400 words for the relevant histories; (e) with SentenceBert and accepts a maximum number of 500 words for the relevant histories; and (f) with SentenceBert and accepts a maximum number of 800 words for the relevant histories to achieve the highest accuracy. Table 2 shows that DS-Dialog with SentenceBERT outperforms other variants without SentenceBERT. Specifically, the DS-Dialog with SentenceBERT, which takes a maximum of 400 words for relevant history input achieves 1 point for recall@k, with k being 5 and 10, a lower mean, 2 points higher for NDCG, and so on.

Table 2: Ablation study on proposed model variants based on the validation set of DS-Dialog

Framework	R@1	R@5	R@10	Mean	MRR	NDCG
DS-Dialog + max 400 words relevant history	49.07	79.76	88.71	4.40	62.83	53.73
DS-Dialog + max 500 words relevant history	**49.15**	79.76	88.61	4.43	62.87	52.75
DS-Dialog + max 800 words relevant history	48.84	79.86	88.65	4.41	62.77	52.98
DS-Dialog + Adaptive relevant history selection + max. 800 words relevant history	48.24	79.26	88.44	4.5	62.24	53.57
DS-Dialog + Adaptive relevant history selection + max. 500 words relevant history	48.94	79.70	88.75	4.4	62.74	53.79

(Continued)

Table 2 (continued)

Framework	R@1	R@5	R@10	Mean	MRR	NDCG
DS-Dialog + Adaptive Relevant History Selection + max. 400 words relevant history	49.04	**80.11**	**89.20**	**4.33**	**62.98**	**54.23**

Comparison with state-of-art methods: Table 3 presents a comparison between the proposed DS-Dialog, which includes relevant question histories and the previous methods based on the DS-Dialog dataset. The integration of the proposed DS-Dialog with adaptive relevant history selection using the new context-aware dataset has resulted in a significant improvement in most of the metrics. This improvement highlights the importance of relevant image context that enhances image representation by providing more relevant question context using similar image features. Compared to RVA, LF and DualVD baselines, the proposed DS-Dialog model with relevant question history contexts achieves a better performance. Results show that the proposed DS-Dialog can achieve approximately 1 point improvement on R@k, mean, MRR, and NDCG compared to the original RVA, and 2 points to LF and DualVD. DS-Dialog shows a significant improvement over the original Visual Dialog model, with 8% higher MRR, 11% higher R@1%, 6% higher R@5%, 5% higher R@10%, 8% higher NDCG as compared to LF.

Table 3: Performance of discriminative models based on the validation set of DS-Dialog

Framework	R@1	R@5	R@10	Mean	MRR	NDCG
LF	44.11	75.16	85.00	5.45	58.29	50.84
DualVD	47.03	78.42	87.84	4.62	61.20	52.69
RVA	48.74	79.57	88.93	4.43	62.53	53.37
DualVD + Adaptive relevant history selection	47.34	78.45	88.20	4.57	61.41	52.91
DS-Dialog + Adaptive relevant history selection	**49.04**	**80.11**	**89.20**	**4.33**	**62.98**	**54.23**

5 Conclusion

In this research, visual textual semantic limitations of existing visual dialog datasets are addressed. The new dataset, called DS-Dialog, overcomes these limitations by grouping relevant histories linked to the corresponding image context. DS-Dialog enriches the current dataset by providing additional context-aware relevant history, which enhances the visual semantic context for each image. In addition, a novel adaptive relevant history selection, comprising question-guided and relevant history-guided visual attention, is proposed to resolve missing visual semantic information for each image. Experimental results demonstrate that the proposed DS-Dialog model outperforms previous visual dialog models, achieving higher mean reciprocal rank (MRR), recall at rank 1 (R@1), R@5, R@10,

and normalized discounted cumulative gain (NDCG) compared to LF, RVA, and DualVD. These findings highlight the significance of incorporating relevant semantic historical context to improve the visual semantic relationship between textual and visual representations in visual dialog systems. Moving forward, this research can be extended to incorporate in-depth language parsing modules for more accurate relative question history generation.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] X. Wei, Y. Qi, J. Liu and F. Liu, "Image retrieval by dense caption reasoning," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, Florida, USA, pp. 1–4, 2017.

[2] A. S. A. AL-Ghamdi, M. Ragab, M. F. S. Sabir, A. Elhassanein and A. A. Gouda, "Optimized artificial neural network techniques to improve cybersecurity of higher education institution," *Computers, Materials & Continua*, vol. 72, no. 2, pp. 3385–3399, 2022.

[3] X. Zhang, W. Lu, F. Li, X. Peng and R. Zhang, "Deep feature fusion model for sentence semantic matching," *Computers, Materials & Continua*, vol. 61, no. 2, pp. 601–616, 2019.

[4] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona *et al.,* "Microsoft COCO: Common objects in context," in *European Conf. on Computer Vision (ECCV)*, Zurich, Switzerland, Springer, Cham, pp. 740–755, 2014.

[5] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. of the 28th Int. Conf. on Neural Information Processing Systems*, Montreal, Canada, pp. 91–99, 2015.

[6] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra *et al.,* "VQA: Visual question answering," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Cambridge, USA, pp. 2425–2433, 2015.

[7] X. Hu, X. Yin, K. Lin, L. Zhang, J. Gao *et al.,* "VIVO: Visual vocabulary pre-training for novel object captioning," in *Proc. of the AAAI Conf. on Artificial Intelligence*, Virtual Event, vol. 35, no. 2, pp. 1575–1583, 2021.

[8] X. Li, X. Yin, C. Li, P. Zhang, X. Hu *et al.,* "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European Conf. on Computer Vision*, Glasgow, United Kingdom, pp. 121–137, 2020.

[9] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain *et al.,* "NOCAPS: Novel object captioning at scale," in *Proc. of IEEE/CVF Int. Conf. Computer Vision*, Seoul, Korea, pp. 8948–8957, 2019.

[10] J. Chen and Z. Hai, "News image captioning based on text summarization using image as query," in *2019 15th Int. Conf. on Semantics, Knowledge and Grids (SKG)*, Guangzhou, China, pp. 123–126, 2019.

[11] Y. H. Tan and C. S. Chan, "Phrase-based image caption generator with hierarchical LSTM network," *Neurocomputing*, vol. 333, pp. 86–100, 2019.

[12] J. Johnson, A. Karpathy and F. F. Li, "Densecap: Fully convolutional localization networks for dense captioning," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 4565–4574, 2016.

[13] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville *et al.,* "Show, attend and tell: Neural image caption generation with visual attention," in *Int. Conf. on Machine Learning (PMLR)*, Lille, France, pp. 2048–2057, PMLR, 2015.

[14] Q. You, H. Jin, Z. Wang, C. Fang and J. Luo, "Image captioning with semantic attention," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 4651–4659, 2016.

[15] J. Lu, C. Xiong, D. Parikh and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 375–383, 2017.

[16] Z. Deng, Z. Jiang, R. Lan, W. Huang and X. Luo, "Image captioning using DenseNet network and adaptive attention," *Signal Processing: Image Communication*, vol. 85, no. 1, pp. 15836, 2020.
[17] Y. Wei, L. Wang, H. Cao, M. Shao and C. Wu, "Multi-attention generative adversarial network for image captioning," *Neurocomputing*, vol. 387, pp. 91–99, 2020.
[18] C. Yang, M. Jiang, B. Jiang, W. Zhou and K. Li, "Co-attention network with question type for visual question answering," *IEEE Access*, vol. 7, pp. 40771–40781, 2019.
[19] D. Zeng, G. Zhou and J. Wang, "Residual self-attention for visual question answering," in *2019 1st Int. Conf. on Electrical, Control and Instrumentation Engineering (ICECIE)*, Kuala Lumpur, Malaysia, pp. 1–7, IEEE, 2019.
[20] I. Ilievski and J. Feng, "Generative attention model with adversarial self-learning for visual question answering," in *Proc. of the on Thematic Workshops of ACM Multimedia 2017*, Mountain View, CA, USA, pp. 415–423, 2017.
[21] J. Hu and X. Shu, "Semantic BI-embedded GRU for fill-in-the-blank image question answering," in *Proc. of the 2nd Int. Conf. on Computer Science and Software Engineering*, Xi'an, China, pp. 108–113, 2019.
[22] Q. Wu, C. Shen, P. Wang, A. Dick and A. V. D. Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1367–1381, 2017.
[23] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak *et al.,* "DBpedia: A nucleus for a web of open data," in *The Semantic web*. Berlin, Heidelberg: Springer, pp. 722–735, 2007.
[24] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson *et al.,* "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 6077–6086, 2018.
[25] Z. Yang, X. He, J. Gao, L. Deng and A. Smola, "Stacked attention networks for image question answering," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 21–29, 2016.
[26] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav *et al.,* "Visual dialog," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 326–335, 2017.
[27] A. Das, S. Kottur, J. M. F. Moura, S. Lee and D. Batra, "Learning cooperative visual dialog agents with deep reinforcement learning," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, Venice, Italy, pp. 2951–2960, 2017.
[28] V. Murahari, P. Chattopadhyay, D. Batra, D. Parikh and A. Das, "Improving generative visual dialog by answering diverse questions," in *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 1449–1454, 2019.
[29] H. D. Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle *et al.,* "Guesswhat?! Visual object discovery through multi-modal dialogue," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 5503–5512, 2017.
[30] J. Lu, A. Kannan, J. Yang, D. Parikh and D. Batra, "Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model," in *Proc. of the 31st Int. Conf. on Neural Information Processing Systems*, Long Beach, California, USA, pp. 313–323, 2017.
[31] Q. Wang and Y. Han, "Visual dialog with targeted objects," in *2019 IEEE Int. Conf. on Multimedia and Expo (ICME)*, Shanghai, China, pp. 1564–1569, 2019.
[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.,* "Attention is all you need," in *Proc. of the 31st Int. Conf. on Neural Information Processing Systems*, Long Beach, California, USA, pp. 6000–6010, 2017.
[33] P. H. Seo, A. Lehrmann, B. Han and L. Sigal, "Visual reference resolution using attention memory for visual dialog," in *Proc. of the 31st Int. Conf. on Neural Information Processing Systems*, Long Beach, California, USA, pp. 3722–3732, 2017.
[34] D. Guo, H. Wang, S. Wang and M. Wang, "Textual-visual reference-aware attention network for visual dialog," *IEEE Transactions on Image Processing*, vol. 29, pp. 6655–6666, 2020.

[35] Y. Niu, H. Zhang, M. Zhang, J. Zhang, Z. Lu *et al.,* "Recursive visual attention in visual dialog," in *Proc. of the IEEE CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 6679–6688, 2019.
[36] D. Guo, C. Xu and D. Tao, "Image-question-answer synergistic network for visual dialog," in *Proc. of the IEEE CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 10434–10443, 2019.
[37] S. Kottur, J. M. F. Moura, D. Parikh, D. Batra and M. Rohrbach, "Visual coreference resolution in visual dialog using neural module networks," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 153–169, 2018.
[38] T. Yang, Z. Zha and H. Zhang, "Making history matter: History-advantage sequence training for visual dialog," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Long Beach, CA, USA, pp. 2561–2569, 2019.
[39] J. Zhang, Q. Wang and Y. Han, "Multi-modal fusion with multi-level attention for visual dialog," *Information Processing & Management*, vol. 57, no. 4, pp. 102152, 2020.
[40] X. Jiang, J. Yu, Z. Qin, Y. Zhuang, X. Zhang *et al.,* "DuaLVD: An adaptive dual encoding model for deep visual understanding in visual dialogue," in *Proc. of the AAAI Conf. on Artificial Intelligence*, New York, USA, vol. 34, no. 7, pp. 11125–11132, 2020.
[41] 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 (TOMM)*, vol. 16, no. 3, pp. 1–16, 2020.
[42] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu *et al.,* "Deep learning-based text classification: A comprehensive review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.
[43] Z. Deng, B. Zhou, P. He, J. Huang, O. Alfarraj *et al.,* "A position-aware transformer for image captioning," *Computers, Materials & Continua*, vol. 70, no. 1, pp. 2005–2021, 2021.
[44] S. Elbedwehy, T. Medhat, T. Hamza and M. F. Alrahmawy, "Efficient image captioning based on vision transformer models," *Computers, Materials & Continua*, vol. 73, no. 1, pp. 1483–1500, 2022.
[45] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao *et al.,* "Dual attention network for scene segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 3146–3154, 2019.
[46] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv, pp. 1810.04805, 2018.
[47] S. Garg, T. Vu and A. Moschitti, "TANDA: Transfer and adapt pre-trained transformer models for answer sentence selection," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 34, no. 5, New York, USA, pp. 7780–7788, 2020.
[48] C. Sun, X. Qiu, Y. Xu and X. Huang, "How to fine-tune bert for text classification?," in *China National Conf. on Chinese Computational Linguistics*, Changsha, China, Springer, Cham, pp. 194–206, 2019.
[49] X. Liu, P. He, W. Chen and J. Gao, "Multi-task deep neural networks for natural language understanding," arXiv preprint arXiv:1901.11504, 2019.
[50] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang *et al.,* "Semantics-aware BERT for language understanding," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 34, no. 5, New York, USA, pp. 9628–9635, 2020.
[51] N. Giarelis, N. Kanakaris and N. Karacapilidis, "A comparative assessment of state-of-the-art methods for multilingual unsupervised keyphrase extraction," in *IFIP Int. Conf. on Artificial Intelligence Applications and Innovations*, Hersonissos, Crete, Greece, pp. 635–645, 2021.
[52] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese bert-networks," arXiv preprint arXiv, pp. 1908.10084, 2019.
[53] L. H. Li, M. Yatskar, D. Yin and C. J. Hsieh, "VisualBERT: A simple and performant baseline for vision and language," arXiv preprint arXiv:1908.03557, 2019.
[54] Y. Wang, S. Joty, M. R. Lyu, I. King, C. Xiong *et al.,* "VD-BERT: A unified vision and dialog transformer with bert," arXiv preprint arXiv:2004.13278, 2020.

[55] T. Ye, S. Si, J. Wang, R. Wang, N. Cheng *et al.,* "VU-BERT: A unified framework for visual dialog," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, pp. 6687–6691, 2022.
[56] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi *et al.,* "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
[57] V. Murahari, D. Batra, D. Parikh and A. Das, "Large-scale pretraining for visual dialog: A simple state-of-the-art baseline," in *ECCV 2020*, Glasglow, UK, pp. 336–352, 2020.
[58] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma *et al.,* "ALBERT: A lite bert for self-supervised learning of language representations," arXiv preprint arXiv:1909.11942, 2019.
[59] V. Sanh, L. Debut, J. Chaumond and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.
[60] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer *et al.,* "SpanBERT: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020.
[61] M. Grootendorst, "KeyBERT: Minimal keyword extraction with bert," [Online]. Available: https://maartengr.github.io/KeyBERT/index.html
[62] S. Kottur, J. M. Moura, D. Parikh, D. Batra and M. Rohrbach, "Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog," arXiv:1903.03166, 2019.
[63] J. Johnson, B. Hariharan, L. V. D. Maaten, F. F. Li, C. L. Zitnick *et al.,* "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 2901–2910, 2017.
[64] J. Pennington, R. Socher and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543, 2014.