VISUAL SEMANTIC CONTEXT-AWARE ATTENTION-BASED DIALOG MODEL

EUGENE TAN BOON HONG

UNIVERSITI SAINS MALAYSIA

2024

VISUAL SEMANTIC CONTEXT-AWARE ATTENTION-BASED DIALOG MODEL

by

EUGENE TAN BOON HONG

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosphy

September 2024

ACKNOWLEDGEMENT

First and foremost, I would like to express my deep appreciation and gratitude to my supervisor, Dr. Chong Yung Wey, for her invaluable mentorship, encouragement, and unwavering commitment to my academic and research growth. Her guidance has been instrumental in shaping the direction of this thesis. She went above and beyond to secure the essential resources required for this research, including her relentless pursuit of funding and access to cloud computing resources, which are key factors in the success of this research.

In addition, I want to express my sincere appreciation to Mr. Thomas Ooi Wei Min for granting me the invaluable opportunity and computing resources to explore the fascinating realm of deep learning research. This opportunity has been a transformative experience in my academic and professional journey, allowing me to delve into cutting-edge technology, develop critical research skills, and contribute to the field of deep learning.

Lastly, the unwavering support, continuous encouragement, and deep understanding of my family members and friends have been my source of strength throughout these years of research.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
DS-Dialog	Diverse History-Dialog
CV	Computer Vision
NLP	Natural Language Processing
RL	Reinforcement Learning
QA	Question Answering
VQA	Visual Question Answering
AMT	Amazon Mechanical Turk
MSCOCO	Microsoft Common Objects in Context
ANN	Artificial Neural Network
SVM	Support Vector Machines
kNN	k-nearest neighbors
CNN	Convolutional Neural Network
HVLM	Human Like Visual Cognitive and Language Memory Network for Visual Dialog
LF	Late Fusion

LR	learning rate
RVA	Recursive Visual Attention
AdaGrad	Adaptive Gradient
SGDR	Stochastic Gradient Descent with Restarts
SGD	Stochastic Gradient Descent
RMSProp	Root Mean Square Propagation
Adam	Adaptive Optimizer
ReLU	Rectified Linear Unit
MLP	Multilayer Perceptron
RGB	red, green and blue
RNN	Recurrent Neural Network
ResNet	Residual Network
R-CNN	Region-based with CNN
YOLO	You Only Look Once
RPN	Region Proposal Network
LSTM	Long Short-Term Memory Networks
SPP	Spatial Pyramid Pooling
ZFNet	Zeiler Fergus Network
RoI	Region of Interest

VGG-16	Visual Geometry Group
FC	Fully Connected
R-FCN	Region-Based Fully Convolutional Networks
FPN	Feature Pyramid Networks
Bi-LSTM	Bidirectional Long Short-Term Memory Networks
GRU	Gated Recurrent Unit
GAN	Generative Adversarial Networks
MAGAN	Multi-Attention Generative Adversarial Networks
FFN	Feed-forward neural network
ELMo	Embeddings from Language Model
RAKE	Rapid Automatic Keyword Extraction
YAKE	Yet Another Keyword Extractor
RVA2	Reference Vector Algorithm
BOW	Bag-of-Words
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
PLM	Pre-trained language models
GloVe	Global Vectors
CBOW	Continous Bag-of-Words

BERT	Bidirectional Encoder Representations from Transformers
STS	Semantic Textual Similarity
NLU	Natural Language Understanding
MLM	Masked Language Modelling
RoBERTa	Robustly Optimized BERT-Pretraining Approach
CC	CommonCrawl
BPE	Byte-Pair Encoding
SBO	Span Boundary Objective
SNLI	Stanford Natural Language Inference
MNLI	Multi-Genre Natural Language Inference
ALBERT	A Lite BERT
STS-B	Semantic Textual Similarity benchmark
GLUE	General Language Understanding Evaluation
KD	knowledge distillation
ViLBERT	Vision-and-Language BERT
GPT	Generative Pre-trained Transformer
VIVO	Visual Vocabulary Pre-Training for novel object captioning
OSCAR	Object-Semantics Aligned Pre-training
FCLN	Fully Convolutional Localization Network

CLEVR	Compositional Language and Elementary Visual Reasoning
MNIST	Modified National Institute of Standards and Technology
AMEM	Attention Memory
RAA-Net	Reference-Aware Attention Network
HACAN	History-Aware Co-Attention Network
DualVD	Dual Encoding Visual Dialogue
VD-BERT	Visual Dialogue BERT
NoCaps	Novel image captioning
CAQT	Co-Attention Network with Question Type
SBE-GRU	Semantic Bi-embedded Gated Recurrent Unit
IoU	Intersection-over-Union
BLEU	Bilingual Evaluation Understudy
METEOR	Metric for Evaluation of Translation with Explicit Ordering
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
NLTK	Natural Language Toolkit
LLaMa	Large Language Model Meta AI
NDCG	Normalised Discounted Cumulative Gain
DCG	Discounted Cumulative Gain
MRR	Mean Reciprocal Rank

RR	Reciprocal Rank
R@K	Recall@k

MODEL DIALOG BERASASKAN PERHATIAN KONTEKS SEMANTIK VISUAL

ABSTRAK

Dialog visual merangkumi konteks imej MSCOCO yang luas dan mengumpul soalan melalui platform AMT. Penggunaan sejarah soalan dan imej yang sedia ada tidak lagi memberi sumbangan kepada pemahaman konteks imej secara keseluruhan. Oleh itu, penyelidikan ini mencadangkan DsDial set data, satu konteks-sedar dialog visual yang menggabungkan semua sejarah dialog yang relevan berdasarkan kategori imej MSCOCO masing-masing. Penyelidikan ini juga mengeksploitasikan semantik visual yang bertindih antara imej-imej melalui penyesuaian pilihan sejarah dialog relevan berdasarkan semua sejarah dialog relevan. Ini adalah separuh daripada 2.6 juta pasangan soalan-jawapan. Pada masa yang sama, penyelidikan ini juga mencadangkan DS-Dialog menyelesaikan isu semantik visual yang hilang bagi setiap imej melalui perhatian visual konteks-sedar. Perhatian visual konteks sedar ini merangkumi perhatian visual soalan terbimbing dan sejarah dialog relevan terbimbing yang membolehkan model untuk mendapatkan konteks visual yang relevan selepas mencapai konfiden yang tinggi. Keputusan kualitatif dan kuantitatif terhadap set data VisDial v1.0 dan DsDial menunjukkan bahawa DS-Dialog bukan sahaja dapat mengatasi prestasi bagi kaedah yang sedia-ada, DS-Dialog juga dapat mencapai keputusan yang kompetitif dengan menyumbangkan pengekstrakan semantik visual yang lebih baik. Set data DsDial telah membuktikan kepentingan konteks relevan apabila dibandingkan dengan dataset VisDial v1.0. Set data DsDial dapat menunjukkan signifikasi terhadap model LF berbanding dengan VisDial v1.0. Keputusan kuantitatif secara keseluruhan menunjukkan bahawa DS-Dialog dengan set data DsDial dapat mencapai skor uji yang terbaik bagi

recall@1, recall@5, recall@10, pangkat purata, MRR, dan NDCG."

VISUAL SEMANTIC CONTEXT-AWARE ATTENTION-BASED DIALOG MODEL

ABSTRACT

Visual dialogue dataset, i.e. VisDial v1.0 includes a wide range of Microsoft Common Objects in Context (MSCOCO) image contents and collected questions via a crowdsourcing marketplace platform (i.e. Amazon Mechanical Turk). The use of existing question history and images no longer contributes to a better understanding of the image context as they do not cover the entire image semantic context. This research proposes the DsDial dataset, which is a context-aware visual dialogue that groups all relevant dialogue histories extracted based on their respective MSCOCO image categories. This research also exploits the overlapping visual context between images via adaptive relevant dialogue history selection during new dataset generation based on the groups of all relevant dialogue histories. It is half of 2.6 million question-answer pairs. Meanwhile, this research proposes Diverse History-Dialog (DS-Dialog) to resolve the missing visual semantic information for each image via context-aware visual attention. The context-aware visual attention includes the question-guided and relevant-dialogue-history-guided visual attention modules to get the relevant visual context when both have achieved great confidence. The qualitative and quantitative experimental results on the VisDial v1.0 and DsDial datasets demonstrate that the proposed DS-Dialog not only outperforms the existing methods, but also achieves a competitive results by contributing to a better visual semantic extraction. DsDial dataset has proven its significance on LF model as compared to VisDal v1.0. Overall quantitative results show that DS-Dialog with DsDial dataset has achieved the best test scores for recall@1, recall@5, recall@10, mean rank, MRR, and NDCG respectively.

CHAPTER 1

INTRODUCTION

In recent years, significant progress has been achieved in various AI tasks, encompassing image classification (Maurício, Domingues, & Bernardino, 2023), scene recognition (Xie, Lee, Liu, Kotani, & Chen, 2020), image captioning (Deng et al., 2021; Elbedwehy, Medhat, Hamza, & Alrahmawy, 2022), question answering, object detection (Zou, Chen, Shi, Guo, & Ye, 2023), image retrieval (X. Wei, Qi, Liu, & Liu, 2017), and visual question answering (Antol et al., 2015; Salaberria, Azkune, de Lacalle, Soroa, & Agirre, 2023; Z. Shao, Yu, Wang, & Yu, 2023). These advancements have been driven by deep learning models that excel in pattern recognition and data processing, enabling machines to understand and interact with visual data more effectively. For example, in image classification, models like Residual Network (ResNet) have revolutionized the ability to categorize images into predefined classes with high accuracy. Scene recognition has similarly benefited from these advancements, allowing AI systems to comprehend complex environments and context within images, improving applications like autonomous driving and robotics.

Simultaneously, the field of Natural Language Processing (NLP) has seen substantial growth, with tasks such as sentence generation, semantics, and sentence semantic matching (X. Zhang, Lu, Li, Peng, & Zhang, 2019) receiving considerable attention. NLP models have advanced in their ability to understand and generate human language, leading to more natural and context-aware interactions between humans and machines. Techniques like Transformer-based architectures, including BERT and

GPT, have been instrumental in this progress, providing powerful tools for tasks like language translation, sentiment analysis, and conversational AI.

The intersection of these fields has led to more complex tasks like image caption generation, Visual Question Answering (VQA) and Visual Dialog, where both visual and textual information are integrated to enable AI systems to answer questions about images or engage in a dialog about visual content. Much of the research in image caption generation is directed toward improving methods for generating more accurate captions, with limited emphasis on enhancing contextual understanding. Unlike image captioning, VQA can generate answers based on both a given question and the associated image. Visual Dialog, in particular, extends the capabilities of VQA by incorporating multiple rounds of questioning, requiring the system to maintain context and understand the evolving nature of the conversation. In a Visual Dialog system, the human user typically interacts by asking a series of questions about a given image. The system must process the image, understand the question in the context of previous dialog, and generate a coherent, contextually relevant response. For instance, given a picture of a few people walking down the street as shown in Figure 1.1, a user might start by asking, "Is it day time?" followed by, "Are they wearing coats?" and then, "Are there any car?" The system not only needs to provide accurate answers but also keep track of the ongoing conversation to maintain relevance and context across multiple turns. This task not only requires sophisticated visual understanding, but also demands a deep integration of NLP techniques to handle the linguistic aspects of the dialog, making it one of the more challenging and promising areas of research in AI today.

Figure 1.1: Visual Dialog's Sample Question and Answer Pairs

1.1 Problem Statement

As depicted in Figure 1.1, the questions posed in conversational dialogues based on VisDial v1.0 dataset often fail to encompass all latent information, including details about surrounding objects such as "benches", "handbags", and "suitcases". Instead, these questions tend to be vague, asking about the presence of "animals", "other people", or the location being "in the country". Consequently, the learned models can only partially grasp the image context. None of the mentioned architectural approaches adequately addresses the global historical context. The existing VisDial v1.0 dataset did not support global dialogue history such as relevant dialogue history. This research endeavor aims to comprehensively understand the image's semantic context by leveraging globally captured historical information that is semantically relevant based on similar image contexts observed in other images. Moreover, some queries in the VisDial v1.0 dataset are vague and unrelated to the image's content, such as "*Are there any animals?*", "*Are there any other people?*" and "*Is it in the countryside?*". These irrelevant questions can mislead trained models, causing them to rely solely on the image's spatial structure rather than incorporating visual-textual inputs. Additionally, existing research primarily focuses on visual-textual representations and often over-

looks the broader historical context. Unclear or imprecise inquiries and the absence of concealed details in models lacking a comprehensive historical context can limit the comprehension of events within the image. As a result, the output of the Visual Dialog models are irrelevant to the corresponding visual context(L. Zhao, Lyu, Song, & Gao, 2021). Therefore, the development of a well-designed mechanism for incorporating relevant semantic information is necessary to address the absence of crucial latent information and promote a deeper understanding of both the question and image context.

The majority of Visual Dialog frameworks primarily focus on establishing correlations between the image context and the ongoing conversation, often overlooking the crucial aspect of the image's semantic context. This semantic context plays a vital role in enabling models to generate more comprehensive responses within conversational dialogues.

For instance, Recursive Visual Attention (RVA)(Y. Niu et al., 2019), excels at inferring visual co-references within the question and dialogue history but does not significantly contribute to enhancing the overall understanding of the image context. It remains focused on the existing dialogue history context rather than embracing a broader historical context that could lead to more generalized deductions based on common elements observed in images, such as people or cars. Images often contain overlapping objects, and multiple images may share similar contextual features. Unfortunately, Visual Dialog models typically do not explore keyphrase extraction techniques to deduce relevant keywords based on the semantic features extracted from the image.

In the realm of Visual Dialog, there are several challenges, one of which involves effectively conveying comprehensive visual information from the image. Although there have been efforts to address semantic feature representation (Kang, Park, Lee, Zhang, & Kim, 2021; Schwartz, Yu, Hazan, & Schwing, 2019) and dialogue history co-reference feature representation(Kang, Lim, & Zhang, 2019; Seo, Lehrmann, Han, & Sigal, 2017), many existing approaches still rely on simply extracting visual features from images using Convolutional Neural Network (CNN). However, this approach falls short in capturing the global semantic context and historical information (L. Zhao et al., 2021), which are crucial for a holistic understanding of the image context.

For example, approaches like Dual Encoding Visual Dialogue (DualVD) (Jiang et al., 2020; Yu et al., 2020) primarily focus on global semantics based on image captions but tend to overlook historical information. In contrast, Human Like Visual Cognitive and Language Memory Network for Visual Dialog (HVLM) (K. Sun, Guo, Zhang, & Li, 2022) aims to enrich the visual context by considering both global and local perspectives, establishing relationships between objects by incorporating external knowledge. Visual Dialog questions often lack relational semantics (Yu et al., 2020), which can impact the overall context of the dialogue history. Unfortunately, none of these approaches comprehensively address the global historical perspective concerning visual content.

Some approaches concentrate on analyzing dialogue history by recovering the dialogue's relational structure (Zheng, Wang, Qi, & Zhu, 2019), addressing issues related to imperfect dialogue history(T. Yang, Zha, & Zhang, 2019), ensuring dialogue consistency(Q. Wu, Wang, Shen, Reid, & Van Den Hengel, 2018), enhancing visual-

semantic information between the image and dialogue history(F. Chen et al., 2020), and concentrate on enhancing the reference entity within the encoder(X. Chen, Lao, & Duan, 2020), rather than exploring keyphrase extraction to deduce relevant keywords based on the image's semantic features.

Objects serve as integral components of the image by providing detailed visual attributes and semantic category concepts. In images, there can be overlapping spatial information, where attributes and objects contribute to the overall visual context. Hence, these absent questions could offer valuable insights by addressing objects like benches, handbags, or suitcase details.

1.2 Objective

The objective of this thesis are listed as follows:-

- To propose a context-aware DS-Dialog model with context-aware visual attention, designed to address the missing global visual semantic information in images.
- To enhance the VisDial v1.0 dataset(referred to as DsDial dataset) by providing additional visual semantic information about an image by leveraging the characteristics of other images with similar features.
- To evaluate the performance of the proposed dataset and model in comparison to existing models and datasets.

1.3 Scope of Research and Limitation

The scope of this research is as follows:

- This research employs an encoder-decoder approach with the following encoders: DualVD, Late Fusion (LF), RVA, and the proposed DS-Dialog. The decoder is a discriminative model.
- Due to hardware limitations, the framework will be tested with three existing works, including DualVD, LF, and RVA.
- The new framework will be tested using only the VisDial v1.0 dataset and the newly proposed DsDial dataset.

1.4 Contributions

The contributions of this research are as follows:

- DsDial dataset improves the visual comprehension(visual semantic context) with relevant dialogue history.
- DS-Dialog
  - Quantitative results indicate that DS-Dialog model shows higher retrieval score as compared to other models.
  - Qualitative results indicate that DS-Dialog model ranked answer candidates with great relevance to the corresponding visual context.

1.5 Thesis Structure

Chapter 1 provides an introduction to visual dialogue and discusses the issues found in existing works. Based on the problem statement, the objectives, scope of the research, and contributions are outlined.

Chapter 2 summarizes the multimodal tasks such as image captioning, and visual question answering. It also details about visual dialogue and the gaps between visual dialogue models.

Chapter 3 describes the methodology and provides detailed descriptions of the proposed DsDial dataset, which offer more visual semantic context for the image through a global historical perspective related to image content.

Chapter 4 covers the implementation of the proposed DS-Dialog model. Detail explanation, including the neural network designs, defined modules, for the DS-Dialog's encoder and the decoder.

Chapter 5 outlines the dataset preprocessing, experimentation designs, and the test cases. Detailed analysis between the proposed DS-Dialog model and the existing models using both VisDial v1.0 and DsDial dataset are also evaluated and analyzed.

Chapter 6 presents the conclusion and outlines future work for this thesis.

CHAPTER 2

RELATED WORK

This chapter presents the existing literature related to multimodal approach such as image captioning, VQA, and visual dialogue.

As computer vision techniques, like image recognition, mature, there is growing interest in expanding research to encompass full scene understanding (Malinowski & Fritz, 2014). Images contain high-level semantic concepts that are relatively unexplored by both NLP and computer vision. Vision-to-Language tasks, including image captioning, VQA, and visual dialogue, aim to bridge the semantic gaps between visual context and natural language information (X. Li, Yuan, & Lu, 2019; Q. Wu, Wang, Shen, Dick, & Van Den Hengel, 2016). Table 2.1 summarizes the multimodal vision-to-language tasks.

Table 2.1: Multimodal Focus.

Vision-to Language Tasks	**Focus**	**Models involved**
Image captioning	Provide descriptions on image	Image, text
VQA	Text conversation	Text
Visual dialog	Conversation with visual context	Image, text

2.1 Image Captioning

Image captioning involves generating a description of an image. Recent works like Visual Vocabulary Pre-Training for novel object captioning (VIVO) (X. Hu et al., 2021) and Object-Semantics Aligned Pre-training (OSCAR) (X. Li et al., 2020)

use state-of-the-art NLP to provide image descriptions. Unlike other image captioning models, VIVO is trained using image-text pairs and multi-layer Transformers for visual-text alignment. This process is followed by a linear layer and softmax function, producing results in the joint embedding space of tags and image region features. There are many research have been done prior to VIVO and OSCAR, such as novel image captioning (Agrawal, Harsh and Desai, Karan and Wang, Yufei and Chen, Xinlei and Jain, Rishabh and Johnson, Mark and Batra, Dhruv and Parikh, Devi and Lee, Stefan and Anderson, Peter, 2019; Johnson, Karpathy, & Fei-Fei, 2016; Tan & Chan, 2019; Xiao, Wang, Ding, Xiang, & Pan, 2019), semantic-concept-based and attention-based. Novel image captioning (NoCaps) provides a benchmark with images from the Open Images dataset to test models' capability of describing novel objects that are not found in the training corpus. The benchmark consists of 166,100 human-generated captions describing 15,100 images from the Open Images validation and test sets. Densecap by Johnson et al. (2016) is widely used as it can provide region localization and description of the image to identify and describe important areas in images with natural language.

Semantic-concept-based methods selectively attend to a set of semantic concept proposals extracted from the image (Hossain, Sohel, Shiratuddin, & Laga, 2019). It also ensures detailed and coherent description of semantically vital objects (Sharma, Dhiman, & Kumar, 2023). The extracted features are then fed into language generation model while semantic features are fed into various hidden states of language model to enhance image description with semantic information. Wanyan, Yang, Ma, and Xu (2023) reduces the semantic gap between graphs obtained . Unlike previous work, X. Liu and Xu (2020) proposes adaptive attention by fusing the image features and

high-level semantics, with the assistance of a language generation model. Shi, Zhou, Qiu, and Zhu (2020) presented a caption generation model that consists of caption-guided visual relationship graphs and later words can be predicted based on the visual relationship.

Attention-based image captioning was first proposed by K. Xu et al. (2015) to assist the model to select the most relevant region for generating words during sentence generation by paying attention to salient objects. It can be trained using standard back-propagation techniques or stochastically by maximizing a variational lower bound. You, Jin, Wang, Fang, and Luo (2016) developed a semantic attention model to attend to semantic concept and incorporate them via the top-down and bottom-up combinations. The algorithm learns to focus on semantic concept proposals and integrates them into the hidden states and outputs of recurrent neural networks. This selection and integration create feedback between top-down and bottom-up computations. Lu, Xiong, Parikh, and Socher (2017) proposes an attention-based neural encoder-decoder frameworks that is able to determine automatically on when to look and where to look respectively. This model can decide whether to attend to the image or to the visual sentinel at each time step, allowing it to extract meaningful information for sequential word generation without relying on visual information for non-visual words or words that can be predicted from the language model alone. To address the challenge of extracting global features from images for image captioning and the limitations of attention methods that force each word to correspond to an image region, Deng, Jiang, Lan, Huang, and Luo (2020) also propose an adaptive attention that is implemented using DenseNet (G. Huang, Liu, Van Der Maaten, & Weinberger, 2017) to extract global image features and an adaptive attention mechanism with a sentinel gate to de-

cide whether to use image feature information for word generation. L. Zhou, Zhang, Jiang, Zhang, and Fan (2019) proposed two-phase learning learning image captioning model which both phases would take place in decoder. By combining top-down and bottom-up attention, it would help in identifying the salient image regions. Z. Zhang, Wu, Wang, and Chen (2021) highlights the salient parts of the image and encrypts the interactions between objects and the scene. Generative Adversarial Networks (GAN)s has been adapted into image captioning tasks but there are limitations of GANs-based methods that only capture local information. Therefore, Multi-Attention Generative Adversarial Networks (MAGAN) (Y. Wei, Wang, Cao, Shao, & Wu, 2020) was introduced to utilizing both local and non-local attention modules for more effective feature representation. The generator generates more accurate sentences, while the discriminator determines if generated sentences are human-described or machine-generated.

Transformer-based image captioning (Deng et al., 2021; Elbedwehy et al., 2022) is leveraging transformer models to generate image captioning. Deng et al. (2021) is utilizing transfomer model to extracts multi-level image features before fusing those image features with scaled-dot product. Later it need to get the relative position between the image features in order to generate image caption. (Elbedwehy et al., 2022) uses attention-based transformer to perform image feature extraction before fed into LSTM-based decoder for caption generations.

2.2 Visual Question Answering

Unlike image captioning, VQA is able to provide answer based on given question and image. VQA has the capability of cross-modal understanding and reasoning

of vision and language as compared to image captioning. Recent VQA works focuses on visual attention (C. Yang, Jiang, Jiang, Zhou, & Li, 2019; Zeng, Zhou, & Wang, 2019), adversarial approach (Y. Liu, Zhang, Huang, Cheng, & Li, 2020) and handling open-ended question answering task(J. Hu & Shu, 2019). Question-guided visual attention uses the whole question feature which might mislead attention and image features extracted by image-guided visual attention might not closely related to keypoints of question. Therefore, C. Yang et al. (2019) proposed Co-Attention Network with Question Type (CAQT) to further divide the VQA question datasets into several categories and also fuse question types by concatenating with multimodal join representation. Meanwhile, Zeng et al. (2019) introduces residual self-attention models to increase convergence and improve accuracy of the model. The attention module consists of multiple stages, including bottom-up attention, residual self-attention and top-down attention. Y. Liu et al. (2020) argued that existing VQA models are ineffective to reflect the answer information. Therefore, they proposed the adversarial models that include question-image and question-answer representations. J. Hu and Shu (2019) propose Semantic Bi-embedded Gated Recurrent Unit (SBE-GRU) to handle issue with open-ended visual question answering task. It feeds the question and image into the stacked GRU and CNN respectively to generate a list of answers. The best answer will be chosen from the answer list based on the top cosine similarity between word2vec's generated answer and each candidate answer.

To further enhance text representations, Q. Wu, Shen, Wang, Dick, and Van Den Hengel (2017) add external Large-scale Knowledge Bases such as DBpedia(Auer et al., 2007) on top of the combination for both image captioning and VQA. The external knowledge base provides the text-based information for the model to improve the an-

swer generation with the help of question-guided knowledge selection scheme. With the advancement in pre-trained large language model such as Generative Pre-trained Transformer (GPT)-3, recent knowledge-based VQA work such as Prophet (Z. Shao et al., 2023) combined vanilla VQA model and GPT-3. Vanilla VQA is responsible for answer heuristics generation while integrated GPT-3 is responsible for heuristics-enhanced prompting. However, Ravi, Chinchure, Sigal, Liao, and Shwartz (2023) highlights that current VQA models are either factual(Marino, Rastegari, Farhadi, & Mottaghi, 2019; P. Wang, Wu, Shen, Dick, & Van Den Hengel, 2017) or common-sense knowledge(Schwenk, Khandelwal, Clark, Marino, & Mottaghi, 2022; Zellers, Bisk, Farhadi, & Choi, 2019), which leads to facts retrieval only appropriate in a certain contexts.

Anderson et al. (2018) and Z. Yang, He, Gao, Deng, and Smola (2016) use attention model to retrieve region context intelligently. Z. Yang et al. (2016) highlighted the importance of repetitive reasoning in order to get the accurate answer. It can be achieved with the implementation of multiple-layer stacked attention network in which query can be made to the image multiple times to infer the answer progressively. Anderson et al. (2018) combines both bottom-up and top-bottom attention to improve the relationship between salient objects detection and the image region generation.

2.3 Visual Dialog

However, none of the previous works include conversational context. Unlike VQA, Visual Dialog learns from multiple contexts such as multi-round dialogues, image and questions. The Visual Dialog dataset, i.e. VisDial v1.0 dataset (X. Chen et al., 2020)

consists of 133000 images, whereby about 123387 and 10064 images are MSCOCO and Flickr respectively. Each image has a caption and ten rounds of question-answer pairs. Each question also paired with 100 candidate answers with one ground-truth human response, 50 answers to similar questions, 30 commonly used answers, and some randomly selected answers from the data set. VisDial v1.0 dataset was formed by collecting conversational data through Amazon Mechanical Turk (AMT). This data is gathered by having two workers engage in a conversation based on the MSCOCO-2014 (T.-Y. Lin et al., 2014) dataset with provided captions. In Figure 1.1 a sample image from MSCOCO-2014 and a snapshot of a VisDial conversational dialogue between two AMT workers based on the image are depicted.

Visual Dialog was initially introduced by Das, Kottur, Gupta, et al. (2017), with LF. It is later extends the Visual Dialog with deep reinforcement learning (Das, Kottur, Moura, Lee, & Batra, 2017) as RL enhances the capability of models to handle tasks based on action-rewards-policy concept (J. Li et al., 2016; Mousavi, Schukat, & Howley, 2018). However, previous work leads to repetitive dialogues. Murahari, Chattopadhyay, Batra, Parikh, and Das (2019) enabled question-bot to ask diverse question by introducing smooth-L1 penalty over questions with high similarity score. The model will penalise the bot that have generated duplicated questions. Meanwhile, Fan, Zhu, Yang, and Wu (2020) introduced Dialog Network to enhance visual dialog encoder for understanding the question accurately by focusing on the intended region of interest. GuessWhat(de Vries et al., 2017) focused on object discovery with yes or no questions. Lu, Kannan, Yang, Parikh, and Batra (2017) transfer knowledge from discriminative learning to generative learning. It uses the current question to attend to the exchanges in the dialogue history, and then use the question and attended dialogue

history to attend to the image to get final encoding. The attention model in it can help the discriminator on paraphrasing answers. Visual Dialog also ignored the semantic feature of the images. Q. Wang and Han (2019) involves object feature extraction and selection in order to extract relevant visual information from the image and filter irrelevant visual information with assistance of semantic guidance from both question and dialogue history.

Several works such as CLEVR-Dialog (Kottur, Moura, Parikh, Batra, & Rohrbach, 2019) and MNIST-Dialog (Seo et al., 2017), proposing new visual dialogues for new test cases. CLEVR-Dialog focuses on visual reasoning using images from diagnostic dataset such as Compositional Language and Elementary Visual Reasoning (CLEVR) (Johnson et al., 2017), focusing on grounding objects based on a natural language expression, and deals with additional visual and linguistic challenges that require multi-round reasoning in visual dialog. Meanwhile, MNIST-Dialog consists of images of Modified National Institute of Standards and Technology (MNIST) digits, used attention memory to resolve visual co-reference. Attention memory helps the neural network to learn by storing image attention map at each round. CLEVR-Ref+ (R. Liu, Liu, Bai, & Yuille, 2019) is a diagnostic dataset based on CLEVR images for visual reasoning in referring expressions.

2.3.1 Attention-based Visual Dialog

The idea of attention is inspired by the human understanding of an object or text by focusing only on certain parts. For example, instead of paying attention to all parts when looking at an image, a person will concentrate only on specific details to better

understand that image. Similarly, it is possible to allow a model to focus only on specific kinds of information that are considered most important in achieving a better understanding.

In sequence-to-sequence modelling, encoding the entire source text into a fixed-length vector requires large memories and leads to the problem of long-term dependencies that negatively affects the performance of the model. Alternatively, the model can utilize the attention mechanism which dynamically searches for the most relevant parts by using a dynamically changing context in the decoding process (Gu, Lu, Li, & Li, 2016). Therefore, before generating a word, the attention mechanism is used to compute word weights to determine how much attention should be paid to each input word. This idea began with Bahdanau, Cho, and Bengio (2014) for English to French statement translations by means of automatic alignment, followed by image caption generation (K. Xu et al., 2015), short text conversation by (Shang, Lu, & Li, 2015), and many more. To improve results by better handling name-entities and long sentences, Luong, Pham, and Manning (2015) suggested global and local attentions for the machine training task. The global attention pays attention to all the words of the source input, regardless of its length, while the local attention focuses only on a selective subset of input positions at each time step.

Besides that, attention modules also gave huge impact on improving Visual Dialog. Seo et al. (2017) proposed Attention Memory (AMEM) and created new synthetic visual dialogue dataset called MNIST-Dialog, which is the combination of MNIST and VisDial datasets to resolve Visual Dialog's sequential dependencies through an attention memory and a dynamic attention combination process. Visual Dialog also has

issue in determining the latent semantic co-reference between question and history. Thus, Guo, Wang, Wang, and Wang (2020) has proposed Reference-Aware Attention Network (RAA-Net) to overcome latent semantic and semantic correlation issues respectively. RAA-Net contains two stage, i.e. multi-head textual attention and visual-two-step visual reasoning. In multi-head textual attention, semantic concentration is determined via attention concentration of words between input and dialogue history, one hot encoding, and word embedding. Guided attention is used to extract relevant textual semantic from dialogue history while both question features and new dialogue history features are concatenated to generate textual reference aware vector. The first stage of visual-two-step visual reasoning is to use Faster R-CNN to focus on self visual, especially the man's face ; visual grounding of related objects and salient relevant regions given textual query, visual key and value. Second stage of visual reasoning is using VGG19 to focus on the cross-visual that covers whole body of man. RAA-net's attention modules contains guided attention and co-attention. Guided attention use do-product attention to learn new embedding of sequence. Co-attention combines other learnable parameters into same feature dimension RVA is trying to overcome the existing soft attention that is unable to predicts discrete attention over topic-related dialogue history by introducing recursive visual attention. It can make discrete decision on replying input content by recursively browses the dialogue history and computes visual attention until it meets unambiguous description. Synergistic model (Guo, Xu, & Tao, 2019) was introduced to generate more comprehensive answer rather than just "yes" and "no". Recently, there are researches attempted to resolve the visual co-reference using neural network at word level (Kottur, Moura, Parikh, Batra, & Rohrbach, 2018). Further, Visual Dialog does not emphasize on the conversation history and only exploit

ground-truth history. History-Aware Co-Attention Network (HACAN)(T. Yang et al., 2019) imposes the wrong answers in conversational context and collect measurement on adverse critic. J. Zhang, Wang, and Han (2020) was aimed to cover low-level information in both image and text via three low-level attention modules such as History-to-History attention that focuses on connections between words, History-to-Question attention, and Relevant History-to-Relevant History attention that focuses on relationship between spatial feature and object feature. Meanwhile, Yu et al. (2020) proposed DualVD which is able to extract the objects and their relationships from visual module and then feed into the semantic module. With the help of multi-level image captions that combines both image captions and dense captions. Dense captions localize and describe image regions in natural language by providing more comprehensive description on the image itself.

2.3.2 Visual Dialog's Dialogue History

There are approaches focusing on analyzing dialogue history by recovering the dialogue relational structure(Zheng et al., 2019), imperfect dialogue history (T. Yang et al., 2019), dialogue consistency(Q. Wu et al., 2018), and leveraging the learned dialogue state(Pang, 2023). (F. Chen et al., 2020) is only focusing on enriching visual-semantic information between image and dialogue history. Park, Whang, Yoon, and Lim (2021) is trying to overcome the missing image features highlighted the missing question intent.

2.3.3 Transformer-based Visual Dialog

Prior to training visual dialogue solely on deep neural network, there are works extending visual dialog to integrate Transformer-based model such as BERT. They are Visual Dialogue BERT (VD-BERT) (Y. Wang et al., 2020), VU-BERT (Ye et al., 2022), and VisDial-BERT (Murahari, Batra, Parikh, & Das, 2020),

VisDial-BERT adapted ViLBERT and pretrained on conceptual captions and VQA dataset, before fine-tuning the VisDial dataset. VisDial-BERT discovered that dense annotations from VisDial v1.0 dataset does not correlates well with original ground-truth dialogue answers. VisDial-BERT is mainly focused on discriminative decoder, rather than both generative and discriminative decoder. Similarly to ViLBERT, there will be two streams in VisDial-BERT, i.e. visual stream and language stream. Visual stream has a total of six layers, with a hidden size of 1024 and eight attention heads, whereas language stream has a total of 12 layers, with a hidden size of 768 and 12 attention heads.

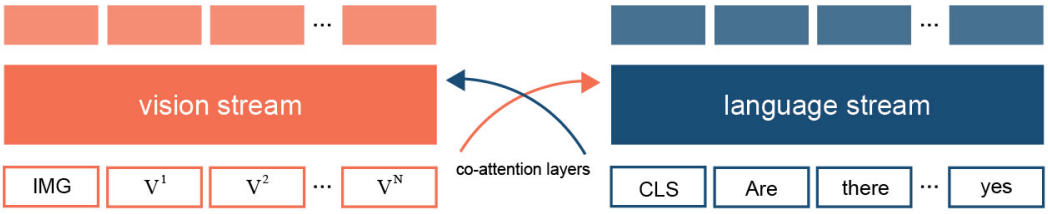

Figure 2.1: Visdial-BERT.

VD-BERT is the BERT-based Visual Dialog model. VD-BERT is a unified vision-

dialog Transformer that leverages on the pre-trained BERT language models for Visual Dialog tasks. It captures all interactions between the image and multi-turn dialog using a single-stream Transformer encoder and supports both answer ranking and generation through the same architecture. VD-BERT adapts BERT for effective fusion of vision and dialog contents via visually grounded training.

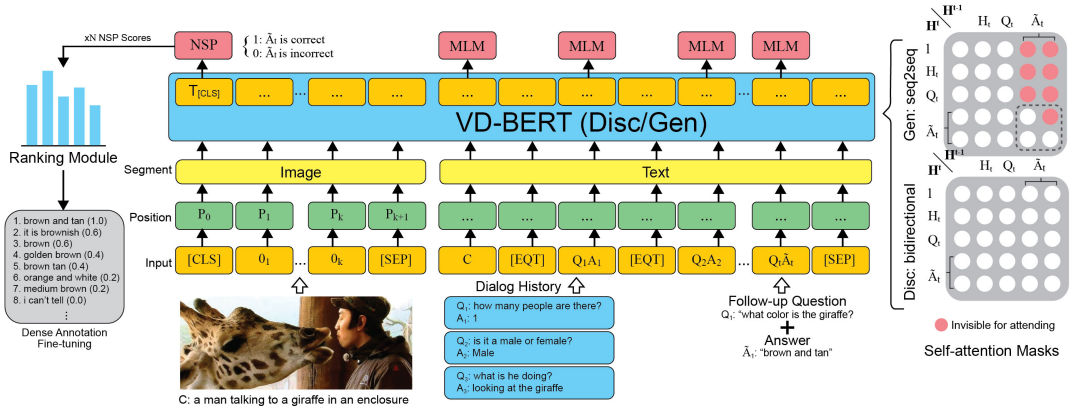

Figure 2.2: VD-BERT.

VU-BERT, another Visual Dialog that attempts on training by using BERT model. It is a unified framework for image-text joint embedding that simplifies the model by using patch projection to obtain vision embedding in visual dialog tasks. The visual dialog task trains an agent with VisDial v1.0 dataset to answer multi-turn questions given an image, requiring a deep understanding of interactions between the image and dialogue history. VU-BERT is trained over two tasks: masked language modeling and next utterance retrieval, which help in learning visual concepts, utterances dependence, and the relationships between these two modalities. Based on Figure 2.3, the image is divided into smaller segments called patches, which are then linearly projected to create patch embeddings. The input is made up of both image and text embeddings, which are calculated by adding the position and segment embeddings together.

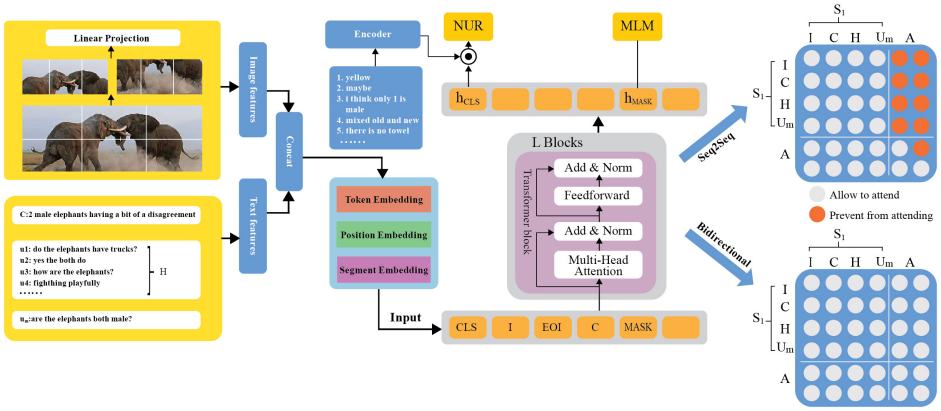

Figure 2.3: VU-BERT.

Unlike VU-BERT, VD-BERT requires high computing resources and cannot scale to a large number of candidates because it must concatenate every answer candidate with the input and go through a forward pass of the entire model.

Table 2.2: A Summary of Transformer-based Visual Dialog.

Model	Layers	Hidden Size	Attention heads	Total parameters	Transformer type
VDBert	12	768	12	110M	Encoder
VisDial-Bert	12	768	12	110M	Encoder
VU-Bert	12	768	12	110M	Encoder

2.3.4 Discussion on Visual Dialog

However, most of the Visual Dialog frameworks are focusing on the co-relation between image context and the conversation context. They did not emphasize on the image semantic context, where it helps the model to generate more comprehensive answers, alongside with conversational dialog. Although RVA is able to infer visual co-reference between question and dialogue history, it does not help to contribute to a better image context. RVA only focus on existing dialogue history context rather than a

global historical context. For example, this thesis can have general deduction based on what has been observed from the image such as people, car, and so on. Images can be overlapping objects and there is a possibility that more than one image contain similar context.

Visual Dialog has plenty of challenges especially expressing more comprehensive visual information of the image. There is some pre-work trying to address semantic feature representation (Kang et al., 2021; Schwartz et al., 2019), dialogue history co-referential relationship feature representation (Kang et al., 2019; Seo et al., 2017), and local historical contextPang (2023). Most of the existing approaches simply extract visual features from an image with CNN (F. Chen et al., 2020; Shukla et al., 2019; L. Zhao et al., 2021).

Merely enriching visual-semantic information between image and dialogue history is not sufficient to understand the whole image context as the global semantic context is not captured (K. Sun et al., 2022) and lack of historical information (L. Zhao et al., 2021). Existing works such as DualVD (Jiang et al., 2020; Yu et al., 2020) only focuses on global semantics based on image captions rather than historical information. HVLM (K. Sun et al., 2022) focuses on enriching visual context from global and local perspectives. Further, questions in visual dialogue have limited relational semantics as it covers wider contents (Yu et al., 2020), which in turns affects the overall context of the dialogue history.

Although there are approaches trying to resolve the missing question semantic intent such as Park et al. (2021), it is only focusing on the model implementation but the

data that provides the missing question semantic intent were not supplied, and thus is lacking of global historical context.

Most of the Visual Dialog models do not explore keyphrase extraction (Giarelis, Kanakaris, & Karacapilidis, 2021) to deduce the relevant keywords based on the image semantic features extracted. None of those covers the global historical perspective with respect to visual content.

Existing Visual Dialog works that enhances the VisDial v1.0 datasets or proposing new datasets are lacking of relevant dialogue histories. Both CLEVR-Dialog (Kottur et al., 2019) and CLEVR-Ref+ (R. Liu et al., 2019) are using CLEVR dataset which only focuses on visual reasoning; MNIST-Dialog by Seo et al. (2017) is using MNIST's hand-written digits to resolve visual co-reference. Even the recent proposed knowledge-based models(A.-A. Liu et al., 2023; S. Zhang, Jiang, Yang, Wan, & Qin, 2022; Z. Zhang, Ji, & Liu, 2023) are merely focusing on common-sense knowledge instead of global historical context that is relevant to the image. Although VQA models such as VLC-BERTRavi et al. (2023) discusses the importance of sentence similarity over its proposed model, there is no existing works in visual dialog that covers the similar approach. Any question-answering model would require a more comprehensive visual-textual context support to understand the enquired context. Additionally, existing visual dialogue models do not utilize semantic textual similarity and keyword extraction models to enhance context understanding by exploring potentially relevant historical contexts, thereby obtaining relevant visual-textual representations in the same embedding space.