DEPARTMENT OF PHILOSOPHY,
LINGUISTICS AND THEORY OF SCIENCE

GUESSWHAT? - FROM WHAT WE DISCUSSED BEFORE

Improving the VQA task in goal-oriented conversational games using the context of the preceding dialogue

Yousuf Ali Mohammed

Master's Thesis:	30 credits
Programme:	Master's Programme in Language Technology
Level:	Advanced level
Semester and year:	Spring, 2020
Supervisor	Simon Dobnik, Mohammed Mehdi Ghanimifard
Examiner	Staffan Larsson
Report number	(number will be provided by the administrators)
Keywords	Visual question answering, history module, human-like natural reasoning, grounding, computer vision and natural language processing

Abstract

Visual Question Answering (VQA) models have struggled combining the natural language and computer vision modalities to extract the question-relevant information from the image. Many state-of-the-art approaches show that language features stand-alone in VQA task in predicting the outcome for the given question. GuessWhat game is a two-player (*guesser and oracle*) question answering game based on the images from MS COCO with a limited answer vocabulary for the oracle player. A competitive neural network model with this limitation needs more detailed understanding of the image as well as the objects in the image and complex semantics in natural language.

To overcome these problems in the oracle player, we propose a sophisticated history encoded model using Long Short-Term Memory (LSTM) network that allows the current question to extract human-like natural reasoning from the previous context of the dialogue in the game to predict the correct answer. Our proposed method introduce a new feature (*history input*) where each tuple consists of a question and the previous context of the dialogue. History encoded models are trained and evaluated on the GuessWhat dataset using different combinations of feature variables to achieve the lowest error-rate. Notably, our models perform significantly better on both the language and vision inputs compared to the baseline models in the GuessWhat dataset. These findings from using previous dialogue history encoded models also provide a substantial improvement in learning categories as well as their properties (*size, shape and colour*) and spatial relation between the objects.

This study further investigates the previous dialogue history encoded using Bidirectional LSTM (BiLSTM) network and compares to the previous findings. Evaluation on the test dataset achieves very similar error-rate for the history encoded model (*question, category, spatial location, history input*) using LSTM and BiLSTM network. These findings are in contrast to the previous studies on LSTM and BiLSTM networks. Our work also examines the history model together with the most informative caption for the image from the set of MS COCO captions for each image. The results indicate that image captions do not have a significant impact on the VQA task in this approach. These findings are also in contrast to the previous study where image captions are proven helpful.

Preface

This thesis is submitted for the degree of Master in Language Technology at the University of Gothenburg. The aim of this study is to develop a neural network model that extract human-like natural reasoning from the previous context of the dialogue (prior conversation). The research aims to contribute to the gap in the VQA task. Before entering the world of Artificial Intelligence and Visual Dialog, I would like to thank the following people, without whom I would not have been able to complete this research.

I would like to especially thank my supervisors, Associate Professor Simon Dobnik and Mehdi Mohammed Ghaninmifard, for their endless support, enthusiasm, encouragement and advice they have provided throughout my time as their student. I have been extremely lucky to have two supervisors who cared so much about my work, and were always available to answer my questions. I would also like to thank all the faculty, staff and colleagues at the Department of Philosophy, Linguistics and Theory of Science and the Department of Swedish. In particular I would like to thank my colleagues in the Language Technology programme.

Finally, I would like to thank my family and friends for their constant support and love. Thank you Anna for always believing in me. This research work would have not been possible without your encouragement.

Contents

1 Introduction

Natural language processing using deep neural networks shows a considerable success in various topics such as machine translation (Johnson et al., 2016), information extraction (Mitra & Craswell, 2017), co-reference resolution (Zhang et al., 2018b), sentence generating (Hashimoto & Tsuruoka, 2018) and topic modelling (Zhang et al., 2018a). Another topic that showed notable success in deep neural networks is Question Answering (QA) based on knowledge dataset of facts (Yin et al., 2015) and memory networks (Weston et al., 2015; Kumar et al., 2015; Sukhbaatar et al., 2015). In these approaches, memory network models achieved notable improvement in answer prediction where text sentences were used as memory input. These QA networks are complicated to model as answer predictions are not only based on the given question and text sentences, but also on many natural language aspects in the questions such as language understanding, semantics interpretation and word-sense disambiguation.

Object recognition (He et al., 2015) and object detection (Ren et al., 2015) are two of the most successful image classification techniques in machine learning based on natural language processing and computer vision. In these techniques, natural language is used both in the form of question to ask about objects in the image as well as classifying objects based on the semantic categories and instance segmentation. Another interesting topic where computer vision and natural language are combined is to automatically generate image captions. One previous study shows that only 34% of the machine generated captions proved to be as good as the human captions (Fang et al., 2014). These shows that generative caption does not only depend on the object categories in the image but also on complex visual features, spatial relation in the objects as well as understanding the entire scene. Dense image captioning technique (Sriram et al., 2018) shows a considerable improvement in generating more informative captions using both image and a set of regions in the images as input.

Visual Question Answering (VQA) is simply defined as *"A neural network model that predicts the correct answer given an image and a question related to that image*". The main objective in VQA task is to extract question relevant information from the image combining the visual features, language understanding and logical reasoning. A fully-functional VQA model depends on object detection (He et al., 2015), object recognition (Ren et al., 2015), question answering (Yin et al., 2015), referring expressions (Krahmer & van Deemter, 2012), spatial features (de Vries et al., 2016) and attention mechanism (Andreas et al., 2015). To develop a neural network model that fills up the gap between computer vision, natural language understanding and AI complete is one of the most challenging task in artificial intelligence.

One of the most significant development on this topic was the release of VQA dataset (Antol et al., 2015) that consist of a rich and balanced multi-modal dialogues. This dataset was used in many innovative machine learning models such as to understand the role of image in VQA (Goyal et al., 2016), counting objects in the image (Zhang et al., 2018c) and neural modular networks (Andreas et al., 2015) that use attention mechanism. A recent study based on VQA dataset (Antol et al., 2015) developed a knowledge based OK-VQA dataset (Marino et al., 2019) where external knowledge resource was modelled together with image and question in predicting the correct answer. The knowledge categories for a selection of images from MS COCO dataset (Lin et al., 2014) were collected from Amazon Mechanical Turk[1]. The above-mentioned neural network models using natural language and computer vision uses image and textual input to solve the VQA task.

The conversation goal in the VQA task is to communicate and ground the information in an image using natural language. A successful coordination between two or more people is only possible if they share the same common ground about the current situation. The common ground to coordinate is usually depends on

[1] https://www.mturk.com

the content of the language, prior conversation and the perceptual understanding (Clark, 1991, 1996). The focus of our research is on the previous dialogue history (prior conversation) as it plays an important role in coordinating on the basis of current common ground to answer the given question. With this aim in mind, this thesis present a new neural network method (*History encoder model*) using the multi-modal dialogues from GuessWhat dataset (de Vries et al., 2016).

The remainder of this thesis is structured as follows: Section 2 introduces previous research in the area of computer vision and natural language processing, followed by the state-of-the-art approaches in QA and VQA task. After this follows the problem formulation in Section 3, the research questions and the hypotheses of our work. In the next Section 4, we will describe the material used in this work as well as a detail explanation of the propose models (*history encoder models*). Results and discussion are given in Section 5 to get a better understanding of the findings of this work. Finally, the last Section 6 concludes our work followed by possible future research on this topic.

2 Previous Research

2.1 Memory networks and natural language

One of the most complex tasks in machine learning is modelling the previous context of a dialogue to predict the correct answer to the given question. Previous studies based on memory networks (Weston et al., 2015; Kumar et al., 2015; Sukhbaatar et al., 2015) shows a significant success in natural language understanding and human-like natural reasoning. These memory networks take question and text sentences to predict the answer in QA. Weston et al. (2015) develop one of the earliest memory networks that reads and writes from a memory component. Their implementation of QA shows that the model predicts the outcome based on the current question, memory component and semantic inference. The memory component search for the most relevant sentence in the given text and saves it in a memory module and extends the search for the next most relevant apart of the text based on the result of the first search. One of the limitations of this memory component is that the search stops after 2 iterations.

Dynamic memory networks (Kumar et al., 2015) (DMN) is an another approach similar to the memory network developed using gated recurrent networks (Cho et al., 2014) (GRU). In the network, DMN first computes the question together with each of the input sentences and produces a vector counts of the relevant information using an attention mechanism. Later this memory vector is used in the answer module to generate the answer. (Sukhbaatar et al., 2015) presents single layer and multiple layer memory network to understand the natural language in the question. In their approach, output prediction is based on the input question embedding and sum of output vector generated by the memory module. The approaches mentioned above are relevant to map questions and answers as well as association between texts. Whereas, we propose a different approach by modelling the previous context of the dialogue (*question/answer pairs*) together with computer vision and natural language in VQA task. Hence, our approach, while modelling memory, captures much larger units of memory than what is proposed for memory networks where the focus is individual constituents.

2.2 Image Captions

In computer vision task, image classification and image captioning are two of the main topics where artificial intelligence achieved immense success using semantic segmentation of categories and limited natural language understanding (Donahue et al., 2014; Karpathy & Li, 2014; Vinyals et al., 2014; Wu et al., 2018). Even though these topics are very well researched, very less work is done to date in machine learning using VQA, image classification and image captioning.

One proposed model shows that generated and annotated caption together with question and image features improves the prediction accuracy in VQA task (Wu et al., 2018). An other study proposed that Dense captioning (Sriram et al., 2018) in the VQA task performs better than normal image captions. In their work, image captions as well as a set of bounding box region captions were modelled together with questions to predict the final output. Li et al. (2018) approach to the VQA task was to use generated image captions and image attributes to predict the answer. They developed a neural network based on generated image captions, image attributes and current question that use semantic reasoning.

A recent study on image captioning (Testoni et al., 2019) showed a considerable improvement in guessing the correct image from a set of candidate images in the GuessWhich task (Chattopadhyay et al., 2017). The best *Guesser* model in their work achieved an accuracy of 94.92% on combining the caption encoder and the dialogue history encoder (Testoni et al., 2019). In our work, we implement a new method that uses human annotated image captions from MS COCO (Lin et al., 2014) based on the target object in the game. The selection of caption based on the target object is briefly explained in last part of Section 4.2.2. These

captions are modelled together with a baseline module and a history module (representing the previous dialogue) as shown in Figure 4 to predict the answer for the given quesiton.

2.3 Visual Question Answering

Goal-directed dialogue is more informative in solving the VQA task as the question/answer pairs consist of both human-like natural reasoning and natural language understanding. The challenging problem is to extract the semantic information and attribute relation from the image using this multi-modal dialogue. The state-of-the-art method for the VQA task (Antol et al., 2015) use free-form and open-ended questions about the image to predict a natural language answer. The baseline results in their work for the language models shows better performance compare to the vision model. de Vries et al. (2016) shows similar outcome in VQA task where the baseline model achieves an error-rate of 21.5% for yes-no answers .

Many interesting neural network methods are developed on the VQA task such as Neural Modular Networks (Andreas et al., 2015), counting objects (Zhang et al., 2018c), co-attention (Yang et al., 2019) and external knowledge resources (Marino et al., 2019). Neural modular networks (Andreas et al., 2015) (NMN) is an interesting approach to VQA task as some of the questions need multiple steps of reasoning to generate a correct answer. NMN uses an external parser to parse the question together with image features using an attention mechanism. The question is turned into a formal representation that can be matched with neural models, sub-networks specialised to do a particular task. Then reinforcement learning is used to train the sequences of such modules and their parameters to produce the correct answer to a question.

Zhang et al. (2018c) implemented a new and unique neural component that learns and count the object instances in the image using attention that improves the prediction accuracy of VQA task. One of the most recent development in VQA task was using co-attention (Yang et al., 2019) mechanism in deep neural networks where question embedding and question attention is concatenated with image embedding and visual attention to predict the outcome. We present a different approach compared to the previously studies, where our model extracts question related semantic information from the image using visual features and previous context of the dialogue.

2.4 The GuessWhat game

Guesser	Oracle
Is the object a person?	*No*
Is the object a train?	*No*
Is the object a building?	*No*
Is the object on right of the photo?	*No*
Is the object a bag?	*Yes*
Is the bag blue?	*No*
Is the bag red with a white logo?	*Yes*

FIGURE 1: *Example dialogue from GuessWhat dataset (target object - red bag)*

Is a two-player game (de Vries et al., 2016) based on the images from MS COCO dataset (Lin et al., 2014). In this game, player one (*the oracle*) has access to both the image and a set of objects in that image, whereas the second player (*the guesser*) only has access to the image and not to the objects. The oracle's role in the game is to select an object in the image and answer the guessers questions. To identify the target object in the image, guesser ask a series of yes-no questions to the oracle as shown in Figure 1. The game ends when the guesser identifies the target object in the image. This dataset is one of the largest multi-modal dialogue datasets relevant to the goal-directed dialogue research (de Vries et al., 2016). It is also very similar to other goal-directed dialogue such as ReferIt (Kazemzadeh et al., 2014) and GuessWhich (Chattopadhyay et al., 2017). The task is pushing the limits of vision and language models beyond image captioning approaches since the answer is not the most salient object in the scene and conversational context also needs to be taken into account. This is very nicely shown in the Figure 1 above.

Oracle baseline model. The oracle's main role in this game is to understand and ground the meaning of the question related to the image and correctly answer it with a limited vocabulary of *Yes, No* and *N/A*. de Vries et al. (2016) also proposes a list of oracle baseline models based on different combination of image and natural language features for the VQA task using the GuessWhat dataset. The features used in their work are image (*I*), target object in the image (*O*), category of the target object (*C*), spatial location of the target object (*S*) and question asked by the guesser (*Q*) to predcit the correct answer. A combination of *question, category (target object)* and *spatial location* with an error-rate of 21.5% on the test dataset was the best baseline model from their work. This shows that the semantic segmentation of category and spatial location plays an important role in predicting the answer (de Vries et al., 2016). Our work builds on top of this oracle baseline model using GuessWhat dataset (de Vries et al., 2016).

3 Problem formulation

Our main aim is to develop a more sophisticated neural network model to minimise the gap between natural language understanding and computer vision. Here, we are interested in modelling the previous context of the multi-modal dialogue in the VQA task to achieve a better answer prediction rate for the current question. An improvement in the prediction rate will also result in learning new linguistic features and semantic reasoning. Many previous studies have shown a significant improvement in the VQA task using natural language processing, semantic segmentation of categories in image and spatial location.

In this work, we will investigate the following research questions:

Q1 Does encoding the previous dialogue history (prior conversation) in the baseline model perform better than the state-of-the-art model?

Q2 If it does, What type of linguistic features show a considerable improvement in answering questions?

Q3 Does the encoded history module show better performance using BiLSTM compare to the LSTM?

Q4 Does encoding the image caption together with the history module show any improvement in learning?

Q5 Does encoding the visual features (*image and objects*) in the history module has an advantage over the research question **Q1** and **Q3**?

To answer these overarching research questions, we formulate the following hypotheses:

A History feature (previous question/answer pair in the dialogue) shows similar or better results compared to the state-of-the-art method

B Encoding the visual image features together with the history module will improve the image classification problem.

C Encoding the history module will improve the prediction accuracy of the target object.

D Previous dialogue history encoded as BiLSTM performs better than LSTMs.

E History module with added image caption encoded improves the prediction accuracy.

Our first Hypothesis **A** is that the previous context of the multi-modal dialogue will improve the VQA task in predicting the output. Initially, we will preprocess feature variables for both image and dialogue and develop a baseline model similar to the GuessWhat baseline model (i.e. the Oracle model). Later, we will preprocess the dialogue dataset and create a history input feature that we use in training the history encoded model. Each input of sequences from the history input (*sequence of the question/answer pairs*) feature is computed using Long short-term memory (Hochreiter & Schmidhuber, 1997) (LSTM) and Bi-directional long short-term memory (Schuster & Paliwal, 1997) (BiLSTM) in this history encoded model. We will also evaluate and compare the answer predictions for the history encoded model and the baseline model to examine the semantic similarity and linguistic features to answer the second research question.

Previous studies in deep neural network showed that BiLSTM based models have a better learning rate compare to the LSTM models (Zhang et al., 2017; Zhou et al., 2018). To investigate these findings and also to see if our fourth Hypothesis **D** holds, we will implement and compare our proposed history model using LSTM and BiLSTM. In the latter part of this thesis, we will implement and investigate image captions in the history model. Human annotated captions for an image are more descriptive and informative compare

to the generative captions (Fang et al., 2014). However, it is not clear whether these image captions will help in VQA task as the caption describers normally focus on the most salient situation in the scene. This problem is also shown in the previous example Figure 1 where the target object in the game was the red bag. To overcome this problem as well as to answer our research question **Q4**, we will use the most informative and linguistically correct caption related to the unknown object in the image. We will also extend the history model by modelling the visual features of image and object to see if the context based model outperform the baseline model.

The main contributions of this thesis are as follows:

1. We show the role of the previous dialogue history (prior conversation and current common ground) in the visual question answering task.
2. Evaluation result on target object categories shows a significant improvement in previous dialogue history encoded model compared to the baseline model.
3. History feature with length 10 and below (previous dialogues of question/answer pairs) shows a decrease in error-rate compare to no history feature.
4. Our history encoded model correctly identifies the tiny patterns and the complex visual features in the image using the vision features of the image that are hard to detect by human-eyes.
5. Similarity in the error-rates for the LSTM and BiLSTM networks are contrast to the previous findings.
6. Our findings also show that image captions can be more informative in the VQA task if a separate caption dataset is created for the target objects in the games.

4 Material and Methods

4.1 Material

The goal-oriented dialogue dataset used in our work is from GuessWhat dataset (de Vries et al., 2016) and the images are from MS COCO dataset (Lin et al., 2014). We also use MS COCO caption dataset (Chen et al., 2015) in our work which is not a part of GuessWhat dataset (de Vries et al., 2016). As the main goal our work was to develop a VQA model that extract question relevant information using the current common ground from the encoded previous dialogue history, we introduced a new feature (*history feature*) as shown in Section 4.1.2 using the dialogue dataset (de Vries et al., 2016). The preprocessed data in our work is different compared to the de Vries et al. (2016) as we used our own preprocessing parameters. The visual features (FC8) of the images and the objects are extracted using the pre-trained VGG16 model (Zhang et al., 2015). We preprocessed the caption dataset (Chen et al., 2015) and created a caption feature to implement in the Model 3..

Our baseline model is similar to (de Vries et al., 2016) as the main focus of our work was to develop a VQA model based on the prior conversation (*current common ground*). The images and the dialogues used in the report were qualitative analysed and considered to avoid any kind of ethical issues. The implications on the quality of the dataset and the feature variables are further discussed in Section 5 and Section 6.

4.1.1 GuessWhat Dataset

GuessWhat dataset (de Vries et al., 2016) was created by a crowdsourcing experiment on *Amazon Mechanical Turk* (AMT)[1] and this two-player guessing game is based on the images from *MS COCO* dataset (Lin et al., 2014). Only, a subset of the original training and validation images and objects are used in creating this dialogue dataset. Initially, the objects that are very small and ambiguous are discarded from the image dataset. To create a quality dataset with meaningful dialogues, images that contains three to twenty objects are only used in this experiment. Total number of unique images in GuessWhat dataset is *66,537* that contain *134,073* objects. The words of the dialogue dataset that do not occur in English dictionary are manually corrected by GuessWhat team to minimise the spelling mistakes as the participants were asked to type the questions manually in the game (de Vries et al., 2016).

The GuessWhat dataset (de Vries et al., 2016) is a well balanced multi-modal dialogue dataset and also one of the largest and most relevant goal-directed dialogue datasets. This dataset comprises of *84.6%* successful, *8.4%* unsuccessful and *7.2%* incomplete games. The data is randomly split into 3 parts where each game is a pair of an image and its corresponding dialogue.

- *Training set - 70%*
- *Validation set - 15%*
- *Testing set -15%*

There are total of 160,745 dialogues and 821,889 question/answer pairs with an average of 5.2% questions per dialogue. In this two-player game, the oracle's vocabulary is limited to 3 tokens in the dataset with *Yes, No* and *N/A*. The total tokens in the GuessWhat dataset is about 3,986,192. In our work, the vocabulary of tokens with at least 3 occurrence are used and the remaining tokens are marked as *UNK*. The maximum

[1] https://www.mturk.com

number of question/answer pairs in a dialogue is 54 and the minimum is 2. The detailed statistics of the dataset used in our work are shown in Table 1.

	Training	Validation	Testing	Total
Dialogues	113221	23739	23785	160745
Images	46794	9844	9899	66537
Objects	976459	202720	206023	1385197
Questions	579633	120318	121938	821889
Success	95429	19977	19994	135400
Failure	11592	2506	2502	16600
Incomplete	6200	1256	1289	8745

TABLE 1: *Dataset statistics*

4.1.2 Feature variables

Features play an important role in machine learning models as these variables are the building blocks for the dataset from which the algorithm can learn. This makes the preprocessing of feature variables one of the most challenging parts in developing neural models. In this sub-section, we will focus on data preprocessing and representing the feature input shapes for the feature variables that are used for training and evaluating neural network models.

- *Image* - Each game in GuessWhat datasets (de Vries et al., 2016) is based on an image from MS COCO dataset (Lin et al., 2014). Each image *I* is represented as an $M \times N$ matrix of real number $\mathbb{R}$ as shown in Equation 4.1. We used an input shape 224×224, so that images can be applied to a pre-trained VGG model (Zhang et al., 2015) which gives us the output shape of $7 \times 7 \times 512$, whereby 224 units can be mapped to 7 units. The transformation from the original size to the image feature size is shown in Equation 4.2, where *?* is a value between $\{0, \infty\}$.

$$I \in \mathbb{R}^{M \times N} \tag{4.1}$$

$$I^{?\times?} \Rightarrow I^{224\times224} \Rightarrow I^{7\times7\times512} \tag{4.2}$$

- *Object* - The selection of images used in the dataset are based on the number of objects in the image. Each image in the dataset consists of 3 to 20 segmented objects and each object is of size $M \times N$. Given a set of objects $\{O_1, O_2.....O_k\}$ in an image and bounding box co-ordinates $O_i = (x, y, w, h)$ for each object, we re-calculated the new co-ordinates $O_i^{'}$, for the new image size 224×224 where *x, y, w, h* are the $x_{coordinate}, y_{cordinate}, width, height$. Using these new co-ordinates, each object in the image is resized to 224×224 and processed through a pre-trained VGG16 (Zhang et al., 2015) to encode the visual features of objects. The final input shape for each of the object in the image is $7 \times 7 \times 512$ as shown in Equation 4.3.

$$\{{O_i}^{7\times7\times512}, {O_2}^{7\times7\times512}, ..., {O_k}^{7\times7\times512}\} \tag{4.3}$$

- *Spatial co-ordinates* - The bounding box co-ordinates of the objects that are re-calculated in the previous step are used to generate spatial features for each object in the image. We initially calculated

the upper, lower and centre co-ordinates for *x* and *y* before normalising. The co-ordinates *width* and *height* are normalised using Equation 4.5 and the other co-ordinates using Equation 4.4. The input shape of spatial features for the target object in our model was of shape (8,) as shown in Equation 4.6.

$$C_{coordinate} = round((1. \times x_i/224) \times 2 - 1, 4) \tag{4.4}$$
$$C_{coordinate} = round((1. \times w_i/224) \times 2, 4) \tag{4.5}$$
$$S_i = [x_{left}, x_{right}, x_{centre}, y_{lower}, y_{upper}, y_{centre}, x_{width}, y_{height}] \tag{4.6}$$

- *Category* - Category refers to the target object category in the game. The total list of target object categories are shown in Appendix 7.5. There are total of 80 categories of objects in MS COCO dataset (Lin et al., 2014) and each of the segmented object in the dataset belongs to one of these 80 categories. The category for the target object in the game is represented by C_i, where $i \in \{1, ..., 80\}$. As every object belongs to a particular category, the input shape of this feature is (1,) in our models.

- *Question/answer pair* - Each game in the dataset consists of a series of question-answer pairs that represent simple dialogue. The minimum length of dialogues in the dataset is 2 whereas the maximum length of dialogues is 54. A dialogue is represented as $QA = \{q_1a_1, q_2a_2, ..., q_ja_j\}$, where q_i is the question asked by the guesser and a_i is the answer given by the oracle in *A*. Question q_i is a sequence of word tokens W such that,

$$q_i = \{w_1, w_2, ..., w_l\} \tag{4.7}$$

  where *l* is maximum length of the question (*i.e. 45*) in the dataset. We added zero padding in front of the question word sequences that were below the maximum length of the question in the dataset. Equations below show the original and the formatted shape of the question input.

$$q_i = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7\} \tag{4.8}$$
$$q_i = \{0, 0, ...0, w_1, w_2, w_3, w_4, w_5, w_6, w_7\} \tag{4.9}$$

  The length of the answer vocabulary in the game is limited to 3 (*yes, no and n/a*) and the maximum length of the answer is 1 as there are yes-no questions. The final inputs used in the neural model after the preprocessing and padding are of shape (1,45,) for the question and (1,) for the answer.

- *History input* - The goal of this work is to develop a network that predicts the answer for the given question using the previous context of the dialogue and to show that context based answering performs better than the state-of-the-art approach. The main challenge of the model is to learn the current common ground in the image using the prior conversation. Therefore, we propose a new feature (*history input*) in this work.

  From the previous step, we know that the dialogue QA is set of question/answer pairs where Q_i (q_ia_i) is one such pair.

$$QA = \{Q_1, Q_2, ..., Q_k\} \tag{4.10}$$

  To create the history input from the QA dialogue, we first represent QA as a list of lists as shown

below.

$$\begin{aligned} H_{input} \quad = \quad & [[Q_1], \\ & [Q_1, Q_2], \\ & [Q_1, Q_2, Q_3], ...[Q_1, Q_2, Q_3, ..., Q_k]] \end{aligned}$$

In the above list of lists, $list_1$ consists of the first question/answer pair in the game and the $list_2$ consists of the first two question/answer pairs and so on. The lengths of these lists are uneven as each game consists of a different number of question/answer pairs. To format these lists into the same shape as the list QA, we add zero padding sequences Z_0. In neural networks, zero padding is added to have the input data acceptable by the model during training and evaluation. This padding doesn't have any influence on the learning of the network. We extract the last question and answer from this sequence list and generate H_{input}, Q_k, A_k as shown in Equation 4.11 for a game with 6 question/answer pairs. Question Q_1 has only zero sequences as H_{input} as it is the first question in the dialogue, whereas the second question Q_2 has zero padding sequences and Q_1 as H_{input}. The input shape of the H_{input} is (1,54,46) where 54 is the maximum length of a dialogue game and 46 is the maximum length of a question + answer (45+1) in the dataset

$$\begin{matrix} & & H_{input} & & & Q_k & A_k \\ \begin{pmatrix} Z_0 & Z_0 & Z_0 & Z_0 & Z_0 \\ Z_0 & Z_0 & Z_0 & Z_0 & Q_1 \\ Z_0 & Z_0 & Z_0 & Q_1 & Q_2 \\ Z_0 & Z_0 & Q_1 & Q_2 & Q_3 \\ Z_0 & Q_1 & Q_2 & Q_3 & Q_4 \\ Q_1 & Q_2 & Q_3 & Q_4 & Q_5 \end{pmatrix} & \begin{pmatrix} Q_1 \\ Q_2 \\ Q_3 \\ Q_4 \\ Q_5 \\ Q_6 \end{pmatrix} & \begin{pmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \end{pmatrix} \end{matrix} \tag{4.11}$$

4.1.3 Implementation and settings

We preprocessed the GuessWhat dataset (de Vries et al., 2016) and used 70% for training, 15% for validation and 15% for testing. We implemented the neural network models in Keras[2] with TenserFlow[3] backend and we trained the models on the full training dataset (*successful, unsuccessful and incomplete games*). Below is a list of parameters and settings used. The code for this work can be found on Github[4].

- Vocabulary size for question: 5875, answers: 3
- Maximum length of the question is 45 and the longest dialogue consists of 54 question/answer pairs
- Pre-trained VGG16 is used for image and object features
- Batchsize
  - Training: 64
  - Validation and Testing: 128
- Learning rate: $1e^{-4}$, optimiser: *ADAM*, loss function: *sparse_categorical_crossentropy*

[2]https://keras.io
[3]https://www.tensorflow.org
[4]https://github.com/SamirYousuf/Thesis

- Epochs: 15, EarlyStopping: 7 epochs on *validation_loss*
- ModelCheckpoint saves the best model based on *validation_accuracy*

4.2 Methods

The main purpose of this work is to investigate whether visual question answering models based on the previous context of the dialogue perform better than the baseline models. To answer this research question, we implement a baseline model and three new history encoded models as shown below:

1. *Baseline Model*
2. *Baseline Model with added history*
   2.1. *Previous dialogue history encoded as LSTM*
   2.2. *Previous dialogue history encoded as bi-directional LSTM*
   2.3. *Previous dialogue history with encoded Image caption*

4.2.1 Baseline Model

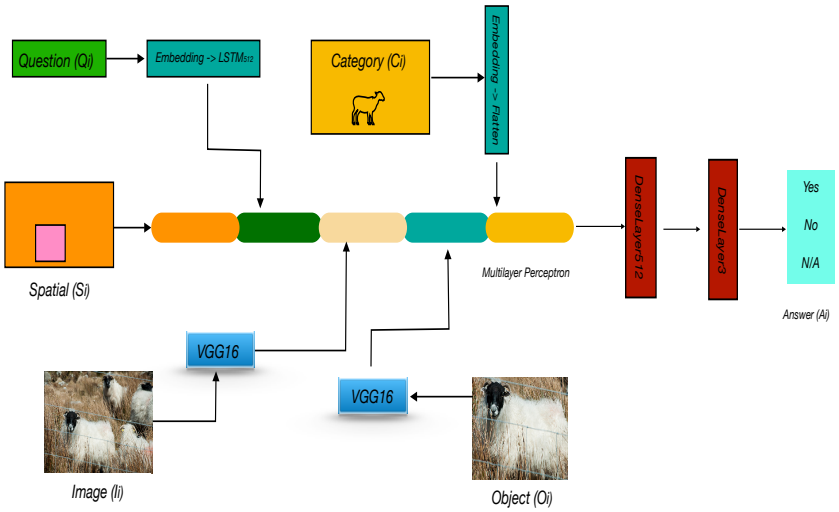


FIGURE 2: *Neural network architecture for Baseline Model*

We re-implemented a simple version of the oracle model proposed by the de Vries et al. (2016) and apply it on our own pre-processed GuessWhat dataset. The model shown in Figure 2 is used as the baseline model throughout our work. The inputs for the model are an image I_i, a question Q_i, a category (*target object*) C_i, a spatial location (*target object*) S_i, the selected target object O_i and an answer A_i. We use pre-trained VGG16 (Zhang et al., 2015) to extract features for an image I_i and a set of objects $O_1, O_2, ..., O_k$ in the image. GlobalAveragePooling2D is used on the image and the target object visual features to generalise over features while training as shown in Equation 4.12 and Equation 4.13 . A Dense layer with 512 hidden units and the *ReLu* activation computes the vectorised output of GlobalAveragePooling2D before it is concatenated into a fully-connected layer. Only the visual features of the target object O_s in the game is used during training from the set of objects $O_1, O_2, ..., O_k$ in the image. The question Q_i is initially used as input sequence for an embedding layer with vocabulary length of 5875 and 300-dimensions. The vectorised output is passed through an LSTM with 512 hidden units before it is concatenated to the fully-connected

layer. The target object category C_i is processed through an embedding layer with the length of category (1,) and 512-dimensions. This embedded output is flattened to get into same shape as others inputs as shown in Equation 4.15. The spatial location S_i of the target object is concatenated to the fully-connected layer with the same embeddings.

$$i_i = relu(Dense_{512}(GlobalAveragePooling2D(I_i))) \tag{4.12}$$
$$o_i = relu(Dense_{512}(GlobalAveragePooling2D(O_i))) \tag{4.13}$$
$$q_i = LSTM_{512}(Embedding\_layer_{(vocab_{len}, emb_{unit})}(Q_i)) \tag{4.14}$$
$$c_i = Flatten()(Embedding\_layer_{(cat\_len, no\_of\_units)}(C_i)) \tag{4.15}$$
$$s_i = Input(S_i) \tag{4.16}$$
$$concat = relu(Dense_{512}(i_i + o_i + q_i + c_i + s_i)) \tag{4.17}$$
$$output = softmax(Dense_{512}(concat)) \tag{4.18}$$

As shown in the Equation 4.17, the fully-connected layer is processed with a dense layer with 512 hidden units and the *ReLu* activation. This fully-connected layer is followed by a softmax layer to predict the output layer (*answer*) for the given question. This Multilayer Perceptron (MLP) (*the softmax layer*) is used for multi-class softmax classification task in machine learning and contains a dense layer with 512 hidden units and the *softmax* activation . Several models were trained with different combinations of feature but constant parameters to achieve a high accuracy and low error-rate.

4.2.2 Baseline model with added history

In this section, we extend the baseline model described previously by adding a history component. This addresses the research questions described in Section 3.

1. ***Previous dialogue history encoded as LSTM*** - The model is shown in Figure 3. The Previous dialogue history is a sequence of question answer pairs as shown in Equation 4.11 in Section 4.1.2. Modelling the previous dialogue history to understand and respond to the present situation is one of the most challenging tasks in artificial neural networks. The model needs to learn the current common ground in the image using the prior knowledge. These type of models depends on various aspects of linguistics such as natural language understanding, referential expressions, common-sense reasoning and linguistic relations. The other challenge is to know what kind of information is relevant and useful in making decisions and predicting answers.

   In this model, each sequence is a question and answer pair fed into the TimeDistributed and LSTM layer. TimeDistributed handles sequences and the LSTM layer learns the weights to predict the next item in the sequence, based on the previous dialogue history and the current input. Vectorised data generated by the Embedding layer with 300-dimensions and vocabulary length of 5875 is passed to the first Time Distributed layer as shown in the Equation 4.22. These weights are later processed through the LSTM layer with 512 hidden units.The output of the LSTM layer is a list of sequences that is passed through another Time Distributed layer to learn another layer of weights. We have two layers to learn sufficient generalisation of sequence representations. The output of the Time Distributed layer is passed again to an LSTM representing with 512 hidden units. This embedded LSTM layer is concatenated to a fully-connected layer representing different combinations of feature variables. Figure 3 shows the architecture of the entire system with added history.

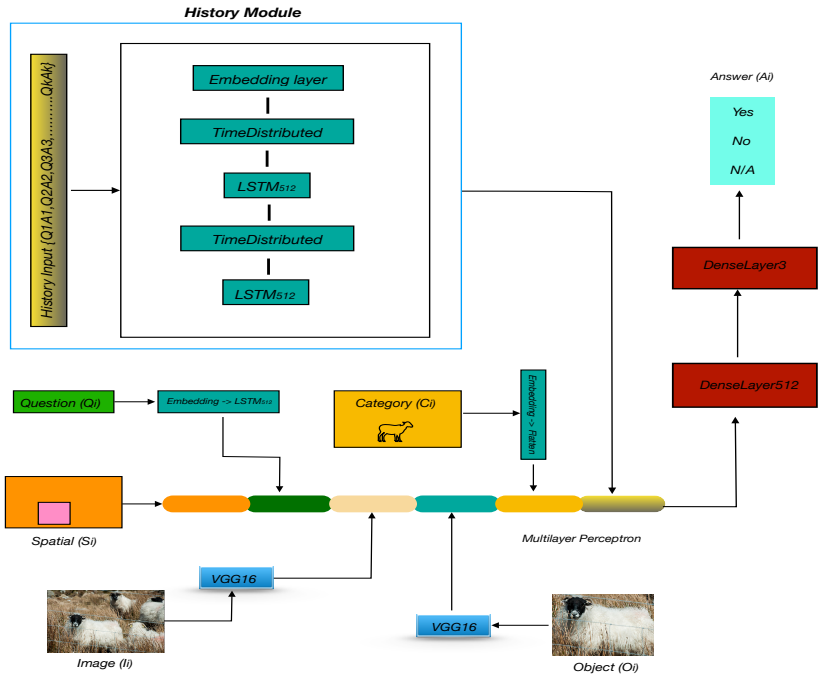


FIGURE 3: *Neural network architecture for History Model*

$$h_i = TimeDistributed(Embedding\_layer_{(voc,unit)}(H_{input})) \tag{4.19}$$
$$h_i = LSTM_{512}(h_i) \tag{4.20}$$
$$h_i = TimeDistributed(h_i) \tag{4.21}$$
$$h_i = LSTM_{512}(h_i) \tag{4.22}$$

2. ***Previous dialogue history encoded as bi-directional LSTM*** - Previous studies shows that BiLSTM networks are more successful in natural language processing tasks compared to LSTMs(Zhang et al., 2017; Zhou et al., 2018) as they learn more features as well as the weights of the sequence in both the directions. The history feature used in our work is a sequence of question/answer pairs of the previous dialogue where each pair is an informative event (question/answer pair) related to the goal-oriented dialogue game. To learn the weights and new features from this set of previous events, we re-implemented the previous model in Figure 3 with a bi-directional LSTM layer as shown below. The goal of this approach is to increase the accuracy of the language model.

$$h_i = TimeDistributed(Embedding\_layer_{(voc,unit)}(H_{input})) \tag{4.23}$$
$$h_i = BiLSTM_{512}(h_i) \tag{4.24}$$
$$h_i = TimeDistributed(h_i) \tag{4.25}$$
$$h_i = BiLSTM_{512}(h_i) \tag{4.26}$$

3. ***Previous dialogue history with encoded image caption*** - The encoded model is shown in Figure 4. We propose another variation of the model where we also include the image caption with the conversation history. Previous studies on this topic showed a significant improvement in GuessWhich task

(Chattopadhyay et al., 2017) where the *Guesser* model was given an image caption and the dialogue history as input. The prediction accuracy of their model was 94.92% compared to the model using only the *caption* (49.99%) or only the *dialogue* (49.99%) as input. The guessers task in their work was to correctly identify the image from a set of candidate images (10k) grounding the linguistic features in the caption and the dialogue history (Testoni et al., 2019). The goal of the oracle model in our task is to answer about the target object using previous dialogue history encoder which is quite similar to the Testoni et al. (2019). To increase the prediction accuracy and to investigate whether the caption encoder improves the VQA task, we incremented our previous model in Figure 3 by encoding a caption module as shown in Figure 4.

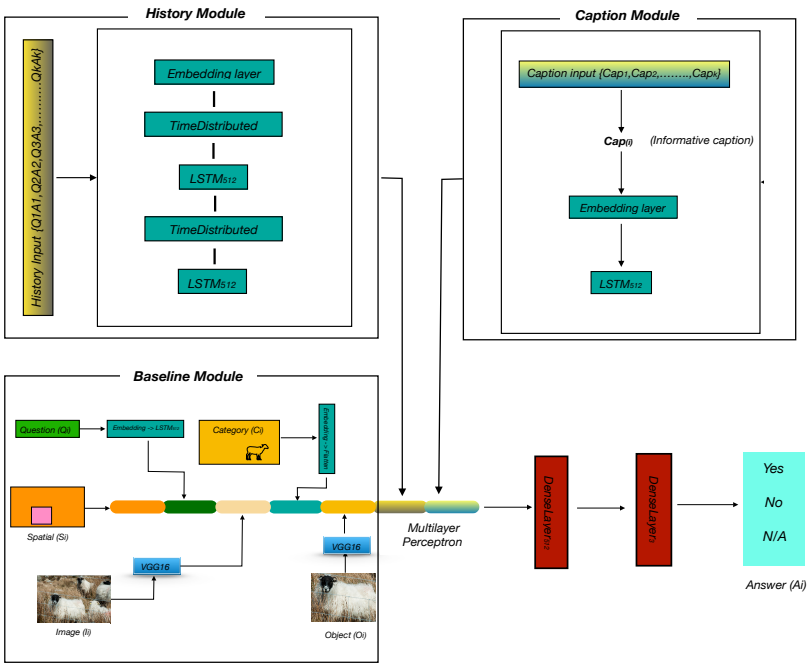


FIGURE 4: *Neural network architecture for History Model together with encoded Image Captions*

Image captions in MS COCO caption dataset (Chen et al., 2015) are human annotated descriptions of the image that were collected before the GuessWhat conversations. In our work, we encoded the image captions as an additional knowledge for the oracle player to correctly answer the given question. Every image in MS COCO caption dataset (Chen et al., 2015) consist of captions between 1 to 7. We represent them as a set $Cap_1, Cap_2, ..., Cap_k$ where $k \in \{1, 7\}$ that describes the scene. As the goal-directed dialogue is always about finding the target object that the oracle selected in the image, we processed the captions and extracted the most relevant caption that describe the target object. Selection of image caption for each game is done based on the conditions below.

- *Condition 1.* We extracted the captions that describes the target object in an image. From this set of captions we selected the caption that describes most of the object categories (MS COCO category dataset (Lin et al., 2014) as shown in Appendices 7.5). E.g. *The **man** is standing behind the **car** with a **laptop** and an **umbrella**.*
- *Condition 2.* If there is only one caption that describes the target object from the set of captions for an image, then we just selected that caption. E.g. *The **man** is walking in the park.*
- *Condition 3.* If none of the captions describes the target object ($O_i \notin \{Cap_1, Cap_2, ..., Cap_k\}$), then we extracted the first caption Cap_1 from the set. E.g. *Two trains on tracks near one another.*(This is caption for Figure 1)

The extracted captions share the same word embedding as the question feature in Section 4.1.2. This vectorised data is processed through an embedding layer with a vocabulary length 5875 and 300-

dimensions before it is passed through an LSTM with 512 hidden units. Caption module shown in Figure 4 describes the encoded architecture for each game. It is later concatenated to the fully-connected layer together with the history module and the baseline module. History module from Figure 3 is used in this model. In this approach, we used only the first 10 question/answer pairs of each dialogue as history features. There are two reason for limiting the length of the question/answer pairs: (1) Dialogue games are linguistically meaningful. (2) Minimise the training time for each model.

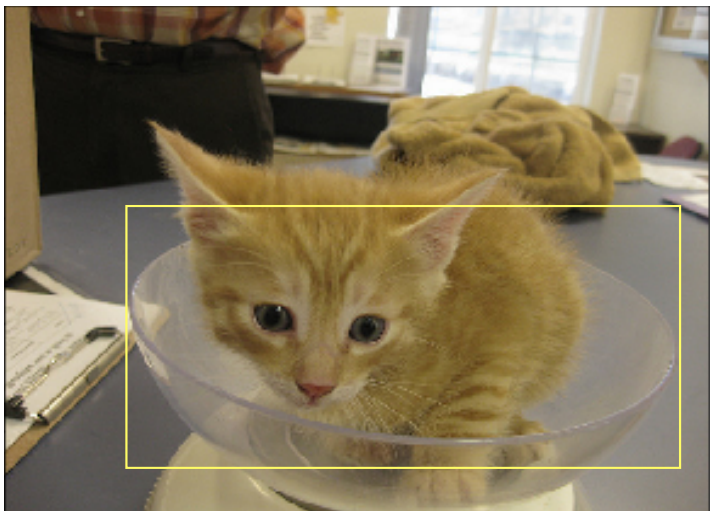
5 Results and Discussion

In this section, we will illustrate and compare results from the baseline model and the proposed models with integrated history and discuss how these results answer our hypotheses.

5.1 Baseline Model

Results. The evaluation result for the baseline models on test dataset are shown in Table 2 and answer predictions for the two baseline models are shown in Figure 5. We implemented a baseline model based on (de Vries et al., 2016) to compare the results using our pre-processed data. The baseline model with only question as an input achieved an error-rate of 40.5% in predicting the correct answer compared to the de Vries et al. (2016) model (*41.2%*) reported in the paper. The error-rate using the most common answer (*No*) would be 51%. An increase of more than 8% in the error-rate for the model with only question as input shows that natural language understanding plays an important role in answering the question. We further examined the baseline model with question by adding other information (*image, target object, category, spatial features*) mentioned in Section 4.1.2. The error-rate decreased down to 26.4% for the model with the input question+category whereas question+spatial features achieved an error-rate of 31.1%. The differences in the error-rates shows a significant improvement in learning when target object category is used compared to using spatial relation between target object and other objects.

We concatenated one more feature variable to the baseline model to see if any of these combinations will achieve a better result compared to the question + category model. The best baseline model in our work achieved an error-rate of 21.9% for the input question + category + spatial location. To understand the baseline models, we qualitatively evaluate and compare two baseline models (question, question + category + spatial features) as shown in the Figure 5. The main reason for selecting these two models is to examine the differences in answer prediction with and without using target object's *category* and *spatial features*. The predicted answers as shown in Figure 5 indicate that the best model (Q+C+S) learns semantic segmentation of the target object category and spatial location of the objects in the image.



Guesser	Human	Q	Q+C+S
Is it a cat?	*No*	*No*	*No*
Is it a person?	*No*	***Yes***	***No***
Is it bowl?	*Yes*	***No***	***Yes***
Is the cat in it?	*Yes*	*No*	*Yes*

FIGURE 5: *A comparison of answer predictions on a successful game from the test dataset, Human vs BaselineModel with inputs (Question, Question+Category+Spatial)*

The dataset used in this work is a well-balanced multi-modal dialogue dataset which is very large. The run-time for each training model is between 3 to 48 hours. Due to the long training times, we trained more models based on natural language inputs and a few on vision. The evaluation results for models using image and objects features together with a question is not so promising as shown in Table 2. Another issue that we encounter is over-fitting of the data when we concatenate 4 or more inputs for training. We do compress the image and the object features obtained by the pre-trained VGG model (Zhang et al., 2015) using a compress function to overcome this problem but the results are not so favourable as most of the models are over-fitting after a single epochs.

Model name	de Vries et al. (2016)	Ours
Question	41.2%	**40.5%**
Question + Image	39.8%	41.2%
Question + Object	29.2%	39.0%
Question + Category	25.7%	26.4%
Question + Spatial	31.3%	31.1%
Question + Category + Spatial (*Q+C+S*)	**21.5%**	**21.9%**
Question + Category + Spatial + Object	22.1%	30.0%

TABLE 2: *Error-rates of the baseline models on the test dataset*

Discussion. The results from our baseline models show that understanding the question and correctly grounding the objects in the image plays an important role in the VQA task. Our best model with question+category+spatial location achieved an error-rate of 21.9% whereas the model with only a question as an input achieved 40.5%. We did a qualitative evaluation of these two models on a set of games from the test dataset. The main finding is that the baseline model learns semantic segmentation of the category in predicting the answer to the given question. The qualitative evaluation of two baseline models (question, question+category+spatial features) as shown in Figure 5 signify that the model *Q+C+S* learn to correctly answer about the category of the objects in the image.

The model *Q+C+S* also shows a significant improvement in learning spatial relation of the objects depending on the number of objects in the image. The last question (*Is the cat in it?*) in Figure 5 correctly ground the categories (*cat* and the *bowl*) and the physical relation between these objects. The baseline results show that the language models perform better without the visual features of the image and the target object. This concludes that the model is learning the natural language in the text and also the physical and statistical relations between the objects in the image.

The error-rate for our baseline model achieves 40.5% compared to 41.2% from de Vries et al. (2016) with only question feature as input. The model with inputs question+category+spatial location shows an increase of 0.4% compared to the their results. We did investigate the difference and we found that our pre-processing data and the training parameters are different compared to de Vries et al. (2016). We could have further tested different configurations with the aim of improving on these results by changing the implementation as well as preprocessing the data but our main purpose of developing this baseline model is to compare the results and findings with our new proposed model including an integration of conversational history.

5.2 Baseline model with added history

In this section we present and discuss the results for our proposed history module using LSTM, BiLSTM and informative image captions. We also evaluate and discuss the results based on the categories of the object and the length of the previous dialogue. We further train and evaluate the previous dialogue history encoded as LSTM with image visual features. Finally, we evaluate and compare the main findings of our work with the baseline model.

5.2.1 Previous dialogue history encoded as LSTM

Results. We pre-processed the question/answer pairs in each dialogue of the dataset and created a new feature (*history feature*) as shown in Section 4.1.2 Equation 4.11. The history features are trained using TimeDistributed and LSTM layers to learn the weights from the input sequences of question/answer pairs using the dense function. Later these weights are used as a history encoder in predicting the final output as shown in Section 4.2.2 Figure 3.

Guesser	Human	Q	Q+C+S	Q+C+S+H
Is it in the kitchen area?	*No*	*Yes*	*No*	*No*
Is it the tv?	***No***	*No*	*Yes*	***Yes***
Is it on the wall?	***Yes***	*No*	*No*	***No***
Is it the one on top?	***No***	***No***	***No***	***Yes***
The black box like fireplace?	*Yes*	*No*	*No*	*Yes*

FIGURE 6: *Evaluation on the test dataset, Human-generate answers, Baseline model (Question, Question+Category+Spatial) and the same model with history feature*

The evaluation results for the previous dialogue history encoded model using various combinations of input features are shown in Table 3. The qualitative evaluation and comparison to previous findings are shown in Figure 6. The first history encoded model using question and history feature achieved an error-rate of 37.5%. Compared with the previous finding for the baseline model using only question input (40.5%), our previous dialogue history encoded model shows a decrease of 3.3% in the error-rate. Several different models using different combinations of features together with history feature were trained. The model with question+category+history feature minimises the error-rate to 25.8% compared to 26.4% from the baseline model without the history information for question + category. The difference between the baseline model and the history encoded model shows that the previous context of the dialogue does have an effect answering the current question in the game. Hypothesis **C** is partially confirmed from these findings.

Model name	de Vries et al. (2016)	Baseline model	History model
Question	41.2%	**40.5%**	**37.2%**
Question + Image	39.8%	41.2%	38.7%
Question + Category	25.7%	26.4%	**25.8%**
Question + Category + Spatial	**21.5%**	**21.9%**	**21.3%**
Question + Category + Image	27.4%		27.2%
Question + Category + Spatial + Image	23.5%		**23.1%**

TABLE 3: *Error-rate comparison History model with de Vries et al. (2016) and Baseline model on test dataset*

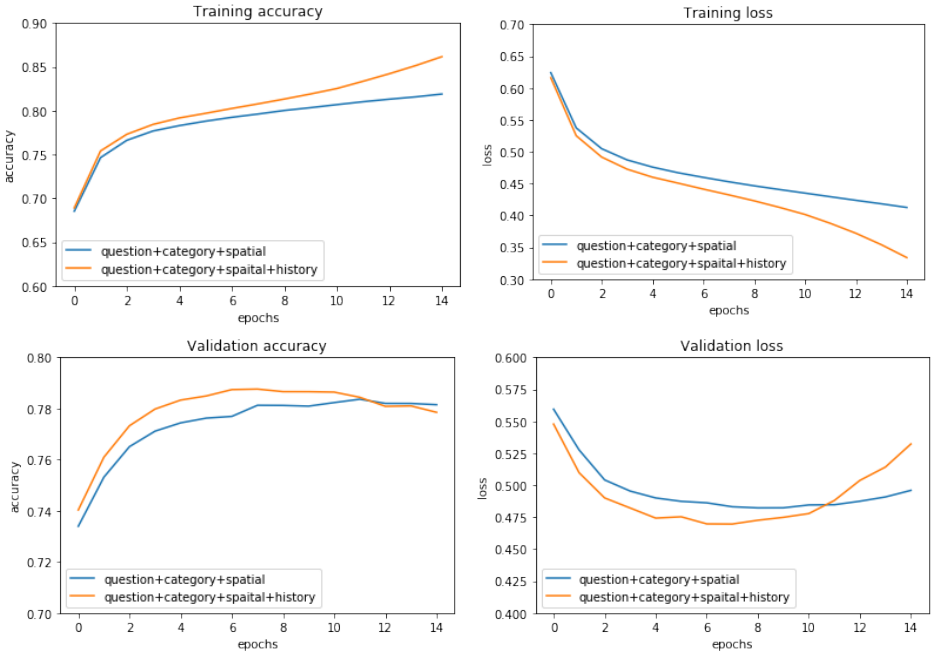
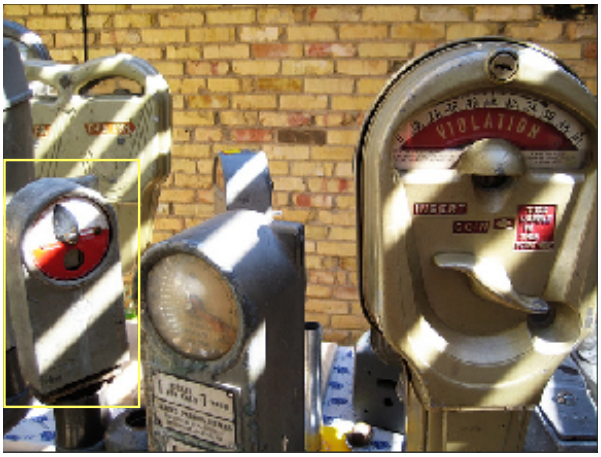


FIGURE 7: *Accuracy and loss for the baseline model and the history model on training and validation dataset*

The baseline model with question+category+spatial location described in the previous section has the lowest error-rate of 21.9%. We trained and evaluated a new model using these features together with the history features and obtained an error-rate of 21.3%. This decline in the error-rate by 0.6% shows that history encoder based on the previous context support our Hypothesis **A** and Hypothesis **C**. A qualitative comparison of the predicted answers for different models in Figure 6 shows that the model using history encoder is more accurate in answering question than a human as the unknown object in the image is *TV* and it is not on the wall. The prediction results outperforms the baseline model also in answering about the spatial relation between the objects and correctly grounding the referring expression - *The black box like fireplace*. These findings confirms our Hypothesis **A** and **C** as well as answer our research question **Q1** and **Q2**. Appendix 7.1 shows comparison results for more games in the test dataset.

In neural networks, the model with the highest accuracy and the lowest loss value is considered as the best model. The graphs in Figure 7 shows the accuracy and loss for the two best model in our work (baseline model and the previous dialogue history encoded model). The validation accuracy curve shows that the previous dialogue history encoded model achieves the best accuracy after 6 epochs compared to the baseline model (after 11 epochs). The decrease in the loss for the validation dataset shows that the model with history module is learning more linguistic features compared to the baseline model. The contrast between the accuracy and loss curves in Figure 7 supports our Hypothesis **A**.

Image visual features. Eventhough the results for the baseline model using visual features (*image, object*) were not so promising, we implemented and trained a history encoded model combining visual features of the image with the question, category, spatial relation and history features. The result of this model using vision and language achieved an error-rate of 23.1% compared to the baseline model (23.5%) in de Vries et al. (2016) using the same features but excluding the history features. The difference in these results shows that the history model does extract visual knowledge and common-sense reasoning from the image scene when answering a question as shown in Figure 8. One interesting example where the system outperforms the human was the question - *"Is there writing on it?"* with answer *Yes*. This finding also raises an interesting question on the quality of the dataset as the human answers in the dataset are not perfectly correct.

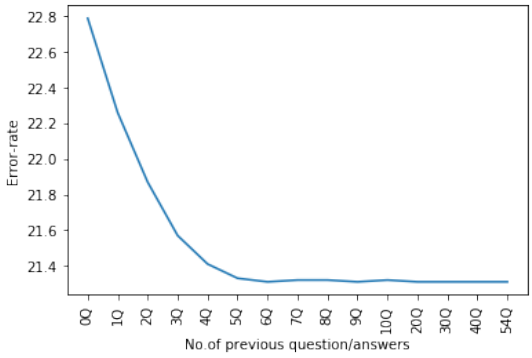


Guesser	Human	Q+C+S	Q+C+S+H	Q+C+S+H+I
*Is it **gold**?*	*No*	*No*	*No*	*No*
*Is it **silver**?*	*Yes*	*Yes*	*Yes*	*Yes*
*Does it have **red** on it?*	*Yes*	*Yes*	*Yes*	*Yes*
*Is there **writing** on it?*	***No***	***No***	***No***	***Yes***
*Is it closest to the **bricks**?*	*No*	*No*	*Yes*	*No*
*Is it on the **left**?*	*Yes*	*Yes*	*Yes*	*Yes*

FIGURE 8: *Evaluation and comparison of the baseline model(Q+C+S), the history model (Q+C+S+H), history model with vision (Q+C+S+H+I) and human baseline(Human)*

We qualitatively evaluated and compared these models on a set of failure games as shown in Appendix 7.3. These findings show that the previous dialogue history encoded model with image visual features outperforms the human answering in recognising the visual features in the image. Even though the accuracy of answer prediction is not as good as the history encoded model based on language (Q+C+S+H), this indicates that the visual model can recognise tiny objects and text that are hard to notice by human eyes. The evaluation results confirm our Hypothesis **B** that the visual features improve the image classification task in VQA.

Dialogue length. We also evaluate the best model (previous dialogue history encoded as LSTM) by changing the length of the dialogue history. The graph in Figure 9 shows that the error-rate gradually decreases as the length of the dialogue increases in the history input. 0Q (*no previous context of the dialogue*) achieves an error-rate of 22.8% whereas 10Q (*previous 10 question/answer pairs of the dialogue*) gives an error-rate of 21.3%. The improvement of more than 1.5% shows that context based history encoded models are more competitive compared to the baseline models in solving the VQA task. The difference of 1.5% in the error-rate for 0Q and 10Q question/answer pairs support our Hypothesis **A**.

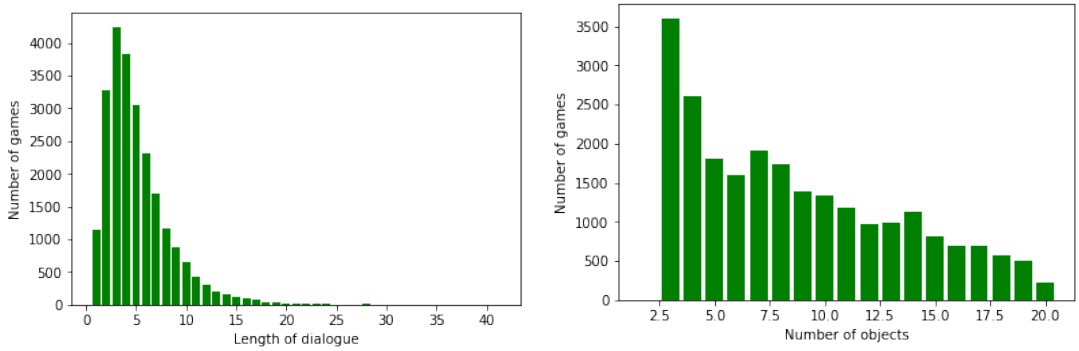


(A)

No.of previous QA	Accuracy
0Q	**22.79**
1Q	22.26
2Q	21.87
3Q	21.57
4Q	21.41
5Q	21.33
10Q	**21.31**
20Q	21.31
54Q	21.31

(B)

FIGURE 9: *Error-rate for the history model using different input lengths of the previous QA pairs*

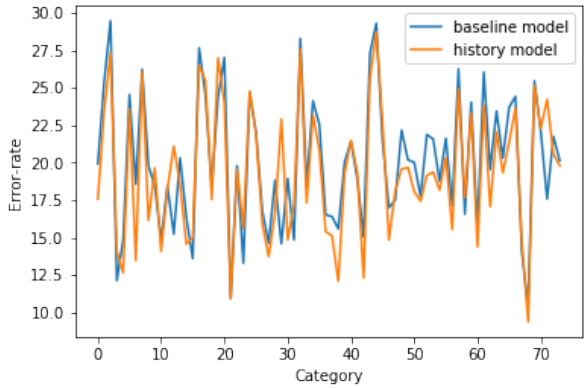


(A) (B)

FIGURE 10: *(A) Histogram showing the number of games by dialogue length (B) Histogram showing the number of games by number of objects in the image*

The evaluation results become constant on further increasing the length of the history input as shown in the Figure 9b. Further investigation on the length of the dialogue in the games as shown in Figure 10a confirms that most of the dialogue are of length 10 and below. This is probably because such is a natural length of a dialogue required for a human to get the necessary information and find the target object. These results also confirms that the dialogue length of 10 and below in the dataset are linguistically informative compare to the longer dialogues.

Category of the target object. We have further evaluated the baseline model and the previous dialogue history encoded model based on the category of the target object. The graph and the table in Figure 11 show a significant variance in error-rate for some of the categories. The category with the least error-rate in both models is *tennis racket* with 10.35% and 9.37% respectively. The category with the maximum variance between the models is *knife* with 18.55% and 13.47% respectively. Further investigation shows that more than 80% of the games where the target object is *knife* are having objects 5 to 19. The variance of more than 5% in the error-rate shows that the history encoded models is learning the target object category (*knife*) based on the salient features in the image scene.



(A)

Category	BaseL	History
knife	18.55	13.47
bench	18.93	14.84
baseball bat	19.79	16.14
dining table	15.58	12.09
bowl	27.02	23.97
parking meter	21.87	19.14
person	26.24	**24.89**
tennis racket	10.35	**9.37**
chair	18.81	16.61

(B)

FIGURE 11: *Error-rate comparison of different categories using baseline model and history model*

The evaluation results on the models also demonstrate that the category *person* differs with more than 1.5%. This is an interesting finding as 30% of the games in the test dataset are related to persons. The other categories with considerable variance in the error-rates are *bench* and *baseball bat*. Some of the categories where the baseline model outperforms the history model are *vase, toothbrush* and *refrigerator*. Appendix 7.4 shows the error-rate comparison for all the target object categories in the dataset.

Discussion. The previous results demonstrate that the previous dialogue history encoded models improve the answer prediction using the current common ground in the prior conversations. The results of the history encoded model using an LSTM layer show a considerable decrease in error-rate to 21.3%. Qualitatively analysing and comparing the outcome of the games in the test dataset to the predictions of the system shows that the model learn about the objects in the image, common-sense reasoning and spatial location of the objects as shown in Figure 6. The model correctly differentiate between the *living room* and the *kitchen* in an image as well as the physical relation between the objects (*Is it on the wall?*). The model outperforms the human answering in correctly locating the target object. The results also confirms our Hypothesis **A**and **C**.

Visual features of the image in the history encoded model have proven to be more informative in recognising patterns in complex images compared to the Q+C+S baseline model and the best history encoded model. The model also shows a significant improvement in learning the properties of objects (*gold, silver and red*) in images using the previous context of question/answer pairs as shown in the Figure 8. The image visual features model also outperforms the human in classifying the image scene and correctly recognising the tiny patterns (*Is there writing on it?*) as shown in Figure 8 (Hypothesis **B** is confirmed). The results also confirms that the computer vision can further improve the prediction accuracy in the VQA task.

The qualitatively evaluation and comparison results shown in Figure 8 and Appendix 7.3 raises an interesting question on the quality of the dataset and the human answering. The answer predictions confirm that the human answers are not perfectly correct especially in the failure games. The incorrect information provided by the *Oracle* player can be the reason for the *Guesser* player to guess the wrong object in the image. Our model shows consistency in correctly answering these questions as it learns and generalise from a set of training samples. To overcome this implication in the dataset, one can trained the models only on the successful games in the dataset and evaluate on both the successful and the unsuccessful games.

The evaluation results based on the length of the previous context of the dialogue show a considerable decrease in the error-rate of 1.5%. These findings indicates that dialogues with the length of 10 and shorter are more informative compared to longer dialogues in the dataset. The results also supports our Hypothesis **A**. This is probably because such is a natural length of a dialogue required for a human to get the necessary information and find the target object. The histogram in Figure 10a shows that most of the dialogues in the dataset are below 10. The previous dialogue history encoded model performs better than the Q+C+S model in correctly predicting the target object category in the game. The table in Appendix 7.4 shows that more than 75% of the categories achieved a lower error-rate for history encoded model compared to the Q+C+S model (Hypothesis **C** is confirmed).

The reasons for not training more combinations of the input in the history encoded as LSTM model are over-fitting, large requirements on preprocessing data and long training time. Most of the models with visual features of objects over-fit after one or two epochs. Over-fitting was also a problem for the history feature as the data was too large after preprocessing. The run-time for each epoch of the history encoded model was on average 4 hours while the total run-time for a history encoded model was between 48 to 72 hours.

5.2.2 Previous dialogue history encoded as bi-directional LSTM

Results. We modified our previous dialogue history encoded model by replacing the LSTM layer with a BiLSTM layer as previous studies in natural language processing using BiLSTM show a significant improvement (Zhang et al., 2017; Zhou et al., 2018). We trained only the two best history encoded models (*Question+Category, Question+Category+Spatial*) from our previous experiments using BiLSTM. The run-time for training these models was 72+ hours. The results using BiLSTM in the history model where very similar to the previous results using LSTM as shown in the Table 4.

Model name	de Vries et al. (2016)	Baseline model	LSTM	BiLSTM
Question + Category	25.7%	26.4%	**25.8%**	**25.8%**
Question + Category + Spatial	21.5%	21.9%	**21.3%**	**21.2%**

TABLE 4: *Error-rate comparison of de Vries' model, our baseline model, the history model using LSTM and BiLSTM on the test dataset*

Discussion. The main finding from using BiLSTM networks is that its performance is similar to the history encoded model using an LSTM network, even though BiLSTM recurrent neural networks train on context from both forward and backward directions. This result also contrast with the previous studies on BiLSTM models as the language model shows a significant decrease in the history model using LSTM. We further investigate these results by training the BiLSTM models using a dropout layer(0.2). The evaluation results show a decrease of 0.1% compared to the model without dropout layer. These findings partially reject our Hypothesis **D** as the bi-directional LSTM models perform the same as the LSTMs.

5.2.3 Previous dialogue history with encoded image caption

Results. Table 5 shows comparison of different models in this work and their error-rates. From our previous findings, we know that the previous dialogue history encoded model as LSTM shows a considerable improvement in learning the categories (Figure 11) as well as the previous question/answer pairs in the dialogue with length of 10 are most informative (Figure 9). Here we examine whether a separate image caption that is not a part of the dialogue and was collected in a different task is also informative. As described in the last part of Section 4.2.2, we selected only those image captions that are potentially referring to the target object or the image scene. The results for the baseline model (*Q+C+S*) with an added module encoding the caption achieved an error-rate of 22.9% which is an increase of 1% as shown in Table 5 compared to the best baseline mode (*Q+C+S*). When we use encoded caption with previous dialogue history encoded models we achieved an error-rate of 22.3%.

Model name	Error-rate
Question (Q)	40.5%
Question + Category + Spatial (Q+C+S)	21.9%
Q+C+S+Ca(Caption)	**22.9%**
Q+C+S+H(History)	21.3%
Q+C+S+H+Ca	**22.3%**

TABLE 5: *Error-rate comparison of model with and without captions*

Discussion. The findings indicate that image captioning together with the previous dialogue history encoded model are not informative in the VQA task (Hypothesis **E** is rejected). The results are in contrast to the previous study on this topic which showed that image captions are informative in correctly identifying the image form a set of candidate images in the VQA task (Testoni et al., 2019). Further investigation of the preprocessed caption dataset indicates that only 56% of the captions are potentially referring to the target object. One extension of this work could be to train and evaluate models using only the 56% of the captions that describe the target object.

6 Conclusion

The VQA task is one of the most challenging problems in the domain of natural language processing and computer vision as it requires common-sense reasoning, language and scene understanding, object recognition and object detection to extract question-relevant information from the image. In our work, we proposed an approach that uses previous context of the dialogue to understand and interpret the correct answer to the given question modelled as the history module using LSTM and bi-directional LSTM. The preprocessed history feature consists of the previous question/answer pairs of the dialogue. We evaluated and compared our proposed models to the baseline models from de Vries et al. (2016). A considerable difference in error-rates between these models and a baseline model that selects the most frequent answer shows that natural language plays an important role in the VQA task. This relates to experimental studies in linguistics which have demonstrated that conversational participants coordinate their believes in conversation in order to optimise the task of referring and therefore reach a common ground (Clark, 1996). The dialogue history therefore provides an important information what these mutual believes are which give clues about the future potential strategies to answer questions.

The evaluation results on the test dataset of the previous dialogue history encoded model as LSTM supports our Hypothesis **A** that the previous context of the dialogue does lead to a better performance compares to the baseline model. These findings also confirm that there is a significant relation between the previous context of the dialogue, the current question and visual features of the image. Moreover, we also qualitatively analysed a set of successful games in the test dataset as predicted by the baseline model and the history model. The answer predictions of these systems shows that the history encoded model learns the target object categories (Hypothesis **C** is confirmed) and spatial relations between the objects in the image scene. The proposed context based history model also shows a considerable improvement in correctly answering questions about the properties of the category as well as understanding some of the referring expressions in the question. Furthermore, the evaluation results for the history model based on target object categories and the length of the previous dialogue confirm our Hypothesis **A** that previous context of the dialogue improves the prediction accuracy.

The proposed history encoded model together with the visual features of the image proved to be more informative in correctly recognising the tiny and complex patterns in the image scene. The model also outperforms the human answering in recognising the visual features in the image. These findings confirm that the computer vision plays an important role in answering questions about the image scene in the VQA task. Even though the error-rate was 2% higher than the best language model (Q+C+S together with history encoder), it learns new features from the whole image. The results also approve our Hypothesis **B** that the image visual features are informative. The evaluation results from this model raises an important question regarding the quality of the dataset and human answering.

Surprisingly, the history model using BiLSTM shows very similar results to the model using LSTM. This result is in contrast to the previous studies (Zhang et al., 2017; Zhou et al., 2018) where BiLSTM models outperform the LSTM models on natural language processing task (Hypothesis **D** is rejected). We also used additional image caption for each image as knowledge input together with the history module to further improve the learning rates of the model. The results of these models do not seem to be satisfactory since related work (Testoni et al., 2019) shows that image captions are informative (Hypothesis **E** is rejected). Our explanation of this result is that image captions are not informative for this task of the GuessWhat dialogue as it describe the most salient objects in the image, whereas the GuessWhat dialogue is to identify any object in the image. Further research is still required to test how image captions can be useful in the VQA task.

Future Research. The main findings from our experiments show a significant improvement in answering question using previous dialogue history approach compared to the baseline approach. Therefore, the

proposed models are very much the key component in future attempts towards a more natural, human-like visual question answering. One interesting approach based on this goal-directed dialogue dataset could examine by concatenating the previous question/answer pairs in the dialogue as a single sequence. This approach could be extended by implementing an attention mechanism to ground entities in the image over the history input sequence that can later be used as memory feature in the output prediction. There are many challenges in this ensuring approach such as GPU memory, processing speed and the maximum length of the previous context of the dialogues that one needs to consider.

One interesting question for future research that can be derived from our findings is whether vision could be more informative for the VQA task as the best history encoded model together with the vision input performed better than the baseline model with the vision input. Due to computational complexity, over-fitting and training time for complex inputs, we were not able to further investigate the visual features in the VQA task. One can explore this area by using different machine learning techniques and also by pre-processing and compressing the image and object visual features by applying convolution neural networks to avoid the over-fitting problem. Future research could also examine the role of visual information in the VQA model by using visual features of all the objects in the image together with natural language features.

Another interesting topic that can be further investigated is natural language in the VQA task. The Guess-What dataset is very rich in referring expressions and spatial relations between objects in images. The best language models (*the baseline model and the history encoded model*) in our work show an error-rate of nearly 21% which confirms that linguistic features such as vocabulary, referring expressions and co-reference resolution could be helpful. Successfully modelling these linguistics features could solve many problems in the VQA task. Obviously, the use of other feature variables such as the category of the target object and spatial location of objects could be further investigated since they showed a considerable influence on the prediction of results.

References

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2015). Deep compositional question answering with neural module networks. *CoRR*, abs/1511.02799.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: visual question answering. *CoRR*, abs/1505.00468.

Chattopadhyay, P., Yadav, D., Prabhu, V., Chandrasekaran, A., Das, A., Lee, S., Batra, D., & Parikh, D. (2017). Evaluating visual conversational agents via cooperative human-ai games. *CoRR*, abs/1708.05122.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, P., & Zitnick, C. (2015). Microsoft coco captions: Data collection and evaluation server.

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Clark, H. H. (1991). Words, the world, and their possibilities. In *The perception of structure* chapter 17. Washington DC: APA.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

de Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., & Courville, A. C. (2016). Guesswhat?! visual object discovery through multi-modal dialogue. *CoRR*, abs/1611.08481.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2014). Long-term recurrent convolutional networks for visual recognition and description. *CoRR*, abs/1411.4389.

Fang, H., Gupta, S., Iandola, F. N., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J. C., Zitnick, C. L., & Zweig, G. (2014). From captions to visual concepts and back. *CoRR*, abs/1411.4952.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2016). Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *CoRR*, abs/1612.00837.

Hashimoto, K. & Tsuruoka, Y. (2018). Accelerated reinforcement learning for sentence generation by vocabulary prediction. *CoRR*, abs/1809.01694.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9, 1735–80.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Karpathy, A. & Li, F. (2014). Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306.

Kazemzadeh, S., Ordonez, V., Matten, M., & Berg, T. L. (2014). Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*.

Krahmer, E. & van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Comput. Linguist.*, 38(1), 173–218.

Kumar, A., Irsoy, O., Su, J., Bradbury, J., English, R., Pierce, B., Ondruska, P., Gulrajani, I., & Socher, R. (2015). Ask me anything: Dynamic memory networks for natural language processing. *CoRR*, abs/1506.07285.

Li, Q., Fu, J., Yu, D., Mei, T., & Luo, J. (2018). Tell-and-answer: Towards explainable visual question answering using attributes and captions. *CoRR*, abs/1801.09041.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Marino, K., Rastegari, M., Farhadi, A., & Mottaghi, R. (2019). OK-VQA: A visual question answering benchmark requiring external knowledge. *CoRR*, abs/1906.00067.

Mitra, B. & Craswell, N. (2017). Neural models for information retrieval. *CoRR*, abs/1705.01509.

Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497.

Schuster, M. & Paliwal, K. (1997). Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45, 2673 – 2681.

Sriram, V., Ramachandran, N., & Kehoe, E. (2018). Visual question answering via dense captioning.

Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). Weakly supervised memory networks. *CoRR*, abs/1503.08895.

Testoni, A., Shekhar, R., Fernández, R., & Bernardi, R. (2019). The devil is in the detail: A magnifying glass for the guesswhich visual dialogue game. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers* London, United Kingdom: SEMDIAL.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2014). Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555.

Weston, J., Chopra, S., & Bordes, A. (2015). Memory networks. *CoRR*, abs/1410.3916.

Wu, J., Hu, Z., & Mooney, R. J. (2018). Joint image captioning and question answering. *CoRR*, abs/1805.08389.

Yang, C., Jiang, M., Jiang, B., Zhou, W., & Li, K. (2019). Co-attention network with question type for visual question answering. *IEEE Access*, 7, 40771–40781.

Yin, J., Jiang, X., Lu, Z., Shang, L., Li, H., & Li, X. (2015). Neural generative question answering. *CoRR*, abs/1512.01337.

Zhang, C., Tao, F., Chen, X., Shen, J., Jiang, M., Sadler, B. M., Vanni, M., & Han, J. (2018a). Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. *CoRR*, abs/1812.09551.

Zhang, R., dos Santos, C. N., Yasunaga, M., Xiang, B., & Radev, D. R. (2018b). Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. *CoRR*, abs/1805.04893.

Zhang, X., Zou, J., He, K., & Sun, J. (2015). Accelerating very deep convolutional networks for classification and detection. *CoRR*, abs/1505.06798.

Zhang, Y., Hare, J. S., & Prügel-Bennett, A. (2018c). Learning to count objects in natural images for visual question answering. *CoRR*, abs/1802.05766.

Zhang, Y., Sun, X., & Yang, Y. (2017). Does higher order LSTM have better accuracy in chunking and named entity recognition? *CoRR*, abs/1711.08231.

Zhou, Q., Zhang, Z., & Wu, H. (2018). Nlp at iest 2018: Bilstm-attention and lstm-attention via soft voting in emotion classification.

7 Appendices

7.1 Baseline model vs History model

Guesser	Human	Q+C+S	Q+C+S+H
Is the object human?	*No*	*No*	*No*
Is the object attached to human?	*Yes*	*Yes*	*Yes*
Do you hit with it?	*No*	*No*	*No*
Does you catch things with it?	*Yes*	*Yes*	*Yes*

Guesser	Human	Q+C+S	Q+C+S+H
Is it a plate?	*No*	*No*	*No*
Is it a cake?	*No*	*No*	*No*
Is it a red collar?	*No*	*No*	*No*
Is it a spoon?	*Yes*	*No*	*Yes*

Guesser	Human	Q+C+S	Q+C+S+H
Is it an animal?	*Yes*	*Yes*	*Yes*
Is it facing forward?	*No*	*Yes*	*No*
Is it in the front?	*Yes*	*Yes*	*No*

FIGURE 12: *A comparison of answers generated by a human, the baseline model and the history model on the test dataset*

7.2 History model (with and without vision features of the image)

Guesser	Human	Q+C+S+H	Q+C+S+H+I
Is it a mammal?	*Yes*	*Yes*	*Yes*
Is it laying down?	*No*	*Yes*	*No*
Is it eating grass?	*No*	*No*	*No*
Is it facing away?	*No*	*Yes*	*No*

Guesser	Human	Q+C+S+H	Q+C+S+H+I
Is it a sign?	*No*	*No*	*No*
Is it black?	*Yes*	*No*	*Yes*
Can u sit on it?	*No*	*No*	*No*
Is it square?	*No*	*Yes*	*No*
Does it have light?	*Yes*	*Yes*	*Yes*
Is it towards the right?	*Yes*	*Yes*	*Yes*

Guesser	Human	Q+C+S+H	Q+C+S+H+I
Is it a bus?	*Yes*	*Yes*	*Yes*
It is blue?	*No*	*No*	*No*
Does it have a driver inside?	*Yes*	*No*	*No*

FIGURE 13: *A comparison of answers generated by a human, the history model (with and without visual features of the image) on the test dataset*

7.3 Answer predictions on failed games

Guesser	Human	Q+C+S	Q+C+S+H	Q+C+S+H+I
Is it food?	*No*	*No*	*No*	*No*
Is it bottle?	*No*	*No*	*No*	*No*
Is it a laptop?	*No*	*No*	*No*	*No*
Is it the mouse?	*No*	*No*	*No*	*No*
Is it on the left?	*Yes*	*No*	*No*	*No*
Is it the plate?	*No*	*No*	*No*	*No*
Is it blue?	*No*	*No*	*No*	*No*

Guesser	Human	Q+C+S	Q+C+S+H	Q+C+S+H+I
Is it a light?	*No*	*No*	*No*	*No*
A signal board?	*No*	*No*	*No*	*No*
Is a vehicle?	*Yes*	*Yes*	*Yes*	*Yes*
A car in the left?	*Yes*	*No*	*No*	*No*

Guesser	Human	Q+C+S	Q+C+S+H	Q+C+S+H+I
Is it a vehicle?	*No*	*No*	*Yes*	*Yes*
Does it have stairs?	*Yes*	*No*	*No*	*Yes*
Can you use it to get off a plane?	*Yes*	*Yes*	*Yes*	*Yes*

FIGURE 14: *A comparison of answers generated by a human, the baseline model and the history model (with and without visual features of the image) on failed games in the test dataset*

7.4 Category based evaluation

Category	Baseline Model	History Model
tennis racket	10.35	9.37
fork	10.93	10.93
dining table	15.58	12.09
dog	15.03	12.3
skateboard	14.84	12.65
knife	18.55	13.47
cell phone	14.62	13.72
frisbee	12.12	14.06
laptop	14.84	14.06
sports ball	15.62	14.37
couch	16.3	14.55
spoon	14.06	14.73
bench	18.93	14.84
keyboard	16.99	14.84
baseball glove	13.59	15.0
toilet	16.4	15.1
tie	16.51	15.4
backpack	17.12	15.55
mouse	13.28	15.62
tv	16.75	15.88
baseball bat	19.79	16.14
chair	18.81	16.61
horse	19.53	17.07
cat	14.84	17.18
umbrella	18.48	17.31
handbag	17.73	17.42
cup	18.44	17.54
microwave	19.92	17.57
wine glass	16.56	17.65
remote	20.01	18.06
bus	17.5	18.12
surfboard	18.75	18.16
bicycle	18.43	18.28
hot dog	20.31	18.48
parking meter	21.87	19.14
truck	20.05	19.27
oven	18.94	19.33

Category	Baseline Model	History Model
train	20.31	19.33
elephant	21.56	19.37
snowboard	19.79	19.53
car	22.16	19.56
sink	18.42	19.64
motorcycle	20.18	19.66
potted plant	20.12	19.79
pizza	21.61	20.31
zebra	21.74	20.57
broccoli	22.56	20.83
clock	15.23	21.09
bed	23.69	21.35
skis	21.38	21.48
carrot	22.01	21.87
sandwich	23.43	22.03
airplane	22.07	22.26
giraffe	21.41	22.65
refrigerator	14.58	22.91
traffic light	24.11	23.17
boat	24.02	23.3
orange	25.39	23.43
teddy bear	24.53	23.59
suitcase	24.41	23.63
bottle	26.03	23.81
bowl	27.02	23.97
toothbrush	17.57	24.21
kite	24.66	24.77
person	26.24	24.89
book	25.44	25.13
cake	24.51	25.48
bird	27.28	25.6
cow	26.22	26.06
sheep	27.66	26.49
vase	24.36	26.98
apple	29.46	27.34
banana	28.28	27.57
donut	29.29	28.71

TABLE 6: *Error-rates comparison of the baseline models and the history model on the test dataset*

7.5 Object categories in GuessWhat dataset

Category(*target object*)	Number of games
person	47389
bowl	2668
bottle	3993
banana	1458
oven	797
tennis racket	1244
sandwich	1030
baseball glove	837
refrigerator	657
motorcycle	1341
bus	1366
bench	1955
dining table	3189
knife	1338
laptop	1219
cat	1241
bed	1116
tie	1142
umbrella	1384
donut	1342
horse	1537
truck	2243
snowboard	512
sports ball	1390
apple	955
vase	1400
bicycle	1160
train	934
hair drier	57
scissors	289
chair	5403
cow	1855
clock	1076
book	3954
kite	1151
surfboard	1120
skateboard	1138
car	8102
sink	1381
couch	1613

Category(*target object*)	Number of games
sheep	1893
cell phone	1495
elephant	1408
boat	2183
mouse	621
bird	1798
zebra	1356
skis	1178
giraffe	1285
backpack	1624
carrot	1523
potted plant	1656
suitcase	1193
spoon	967
keyboard	750
cup	3438
wine glass	884
handbag	1909
teddy bear	1060
tv	1480
frisbee	709
orange	1075
hot dog	662
pizza	1245
traffic light	1947
airplane	1154
baseball bat	703
fire hydrant	363
remote	1354
cake	1245
dog	1457
toilet	837
fork	1013
broccoli	1531
stop sign	350
parking meter	309
microwave	389
toothbrush	542
bear	127
toaster	56

TABLE 7: *Target object categories in GuessWhat dataset (de Vries et al., 2016)*