Multimodal Dialogue Systems

A Dual Degree Project Report

submitted by

ISHU DHARMENDRA GARG

under the guidance of

DR. MITESH M. KHAPRA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS

May 2018

THESIS CERTIFICATE

This is to certify that the thesis titled **Multimodal Dialogue Systems**, submitted by **Ishu Dharmendra Garg (CS13B060)**, to the Indian Institute of Technology Madras, for the award of the degree of **Bachelors of Technology** and degree of **Masters of Technology**, is a bonafide record of the research work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Mitesh M. Khapra
Project Guide
Assistant Professor
Department of Computer Science & Engineering
Indian Institute of Technology Madras, 600036

Place: Chennai
Date: May 2018

ACKNOWLEDGEMENTS

My work is a sum total of all the love, care and efforts that people have put behind me. This incomplete acknowledgment is a small efforts to cherish the contributions of all those who have made this research work possible.

First and foremost I will like thank Prof. Mitesh for pushing me to complete the project and supporting me with any and every logistics. His undoubting trust, noble understanding, constant motivation forms some of the main supporting pillars of this work.

I believe that my interest and motivation to pursue research project in AI is because of the wonderful courses in AI that I have taken so far. I will specially like to thank Prof. Ravindran being a constant mentor and introducing and motivating me to work in the field of AI. I will also like to express my gratitude to Prof. Sutanu for introducing me to the field of Natural Language Processing and influencing me to take Computer Science as my major in my undergrads.

I will sincerely like to thank Prof. Katerina and Adam for introducing me to deep learning techniques and experimentation methods in Computer Vision and patiently attending to all my silly doubts and allowing me to make mistakes and learn from them.

A special thanks to Prof. Harish and Prof. Rahul for their support, cooperation and trust during my last semester and to Prof. Rupesh for being a friendly mentor. Prof. Siva Ram Murthy has been a great faculty advisor and I will like to thank him for his expert advice on handling my academic logistics. I also appreciate the Department of Computer Science and Engineering at Indian Institute of Technology, Madras for its kind support and assistance.

Any work that I will do or have done could not have been possible without the wonderful people I had the fortune to be with. Undoubtedly, it was my family that kept me moving during the worst of the times and needless to say that they were the powerhouse for me when I was exhausted. I don't think I would have developed into the human being I am today without the miraculous and prodigious friends I have. Among all, I will specially like to thank (going alphabetically) Abhishek, Gurneet, Mohan, Sahil, Sanchit, Santhoshini, Stack, Susanna, Suvadeep, Ujjawal, Varun, ... A magical thanks to the Star for unknowingly helping me to understand myself better. It will be a shame if I do acknowledge the movie "Bhag Milkha Bhag" for giving me the constant motivation and enthusiasm to fight the odds of life.

- Ishu D. Garg

ABSTRACT

Key Words: Visual Dialogue, Visual Question Answering, “ GuessWhat?! ” Game, Multimodal Deep Neural Networks, Multimodal Dialogue Systems

An important step towards scene understanding and computer vision will be by designing computer vision system that can hold meaningful conversations in natural language with humans on visual content. There are two fundamental problems that have to be solved before solving this problem; first, understanding natural language descriptions and second grounding these descriptions in the visual world.

The Visual Dialog problem [Das *et al.* (2017*b*)] has been recently introduced where an AI agent holds a meaningful dialog in natural language about an image. There are several variants of the problems but generally in most of the problems the common theme is that given an image, a conversation dialog history; an agent has to generate a question (which is relevant to the image and consistent with the dialog history) or the agent has to correctly answer a question about the image (given a question as a part of dialog history). This task of Visual Dialog involves inferring the context from the dialog history, grounding the parts of conversation to the image and then answering the question correctly or generating the next relevant question.

Higher-level image understanding, for example, spatial reasoning along with language grounding is also required to solve the task. Being grounded in vision enough, the problem of Visual Dialog allows objective evaluation of responses and benchmark progress; while at the same time Visual Dialog is abstract enough to serve an indicator for the machine intelligence. Solving the Visual Dialog task will help us in getting a step closer towards machine intelligence!

The state of the art results due to advancements in deep learning techniques in Computer Vision, and Natural Language Processing and Understanding motivated us to use deep learning models to solve the complex problem of Visual Dialog which lies at the intersection of both Computer Vision and Natural Language Processing and Under-

standing. In this work, we propose deep learning model for Visual Dialog task and test out our models on "Guess What?!" dataset [De Vries *et al.* (2017)].

CHAPTER 1

Introduction

Motivation

Over the time, there has been great progress in the field of computer vision and artificial intelligence. With the help of deep learning models, we have been able to satisfactorily accomplish several tasks in computer vision like simple image classification task, to more complex task such as object detection. Recently we have been witnessing the successful application of deep learning methods (with other reinforcement techniques) in solving complex tasks such as playing video games, complex robotics applications, answering questions about the images etc. With progress, researchers are also trying to come up with more complex tasks which can be a challenge to existing AI techniques. The introduction of new complex challenges ultimately will lead to the advancements of AI techniques and with the advancement of AI techniques, we will slowly move towards building a stronger and hopefully a smarter AI agent.

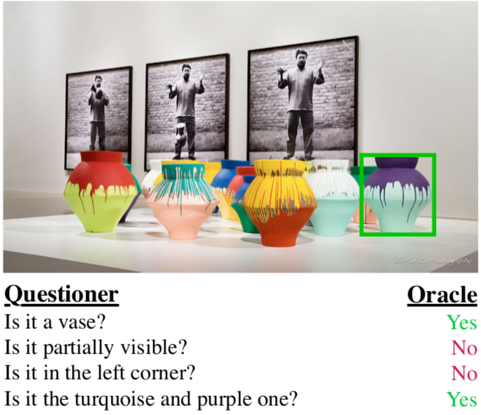
Many of the tasks in NLP (like machine translation, dialog systems, and question answering) and Vision (like scene recognition, image classification, and object detection) only take their information from single mode; many of the problems have already been solved. Other tasks which involves multi-modal input such as image captioning and visual question answering are more complex and are more closer to AI complete task. The models that have been proposed for these tasks are expected to understand both text and visual inputs, find the complex relationship between the information from all the modalities and then produce the output. But there have been observations that suggests that for these tasks, models rely more on biases and correlations in the dataset [Agrawal *et al.* (2016*a*), Zhang *et al.* (2016), Goyal *et al.* (2017), Johnson *et al.* (2017)]. For example, in the Visual Question Answering task, in most of the datasets, if there is a question "Which animal is present in the picture?"; most of the times the answer is "dog".

One can think of creating datasets that have less biases or developing the models that explicitly tries to ignore these biases. But another efficient way to overcome this problem is by introducing new task that makes it difficult for any AI model to rely on the biases (for any dataset). One such attempt in this direction is the task of Visual Dialog.

Visual Dialog: A Multimodal Dialogue System

We as humans are very good at talking to each other and describing our external world and internal emotions using words. Among different means of communications, natural language forms one of the most popular and convenient way of communications. One will also observe that most of us use natural language very effectively and efficiently. It will be wonderful if machines have a similar capability of describing the external visual world in natural language. The task of Visual Dialog was recently introduced in [Das *et al.* (2017*b*)] where an AI holds meaningful conversation with humans about an Image. Specifically, a human and AI agent talk with each other about an image. In the conversation human asks several questions about the image which the AI agent has to answer. Note that this is different from Visual Question Answering because in this task instead of having just one question being asked about the image, the human can ask several related questions on the same image. The dialog history between the AI agent and human can define the context for the next question. One can think of Visual Dialog to be a more generalized version of Visual Question Answering.

Since the introduction of Visual Dialog, tasks similar to Visual Dialog have been proposed. One such task that has been introduced is "GuessWhat?!" [De Vries *et al.* (2017)]. "GuessWhat?!" task is modeled as a two player guessing game played between players named Oracle and Questioner. In this task, the information exchange between Oracle and Questioner is limited. The limitations are that Questioner and Oracle can communicate to each other in natural language. The Questioner is only supposed to ask questions that results in Yes and No. The response of the Oracle can only be Yes/No/NA (Not Applicable).



Questioner	Oracle
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

One instance of conversation in "GuessWhat?!" Game

In this game, first the Oracle randomly picks an image from a large set of images. One can assume that the image is that it has large number of objects in it and is rich in visual scene. Oracle shares this image with the Questioner; now both Oracle and Questioner has access to the image. The Oracle selects an object in the image; which Questioner does not know about. The job of the Questioner is to guess the unknown object that is selected by the Oracle. In our case, we allow the Questioner to ask 5 questions from the Oracle, after which Oracle presents Questioner with a list of objects from the image and the Questioner then makes a guess from among the list. The game ends successfully when Questioner makes the correct guess, otherwise we say that the game has ended unsuccessfully.

Application of Multimodal Conversation Systems

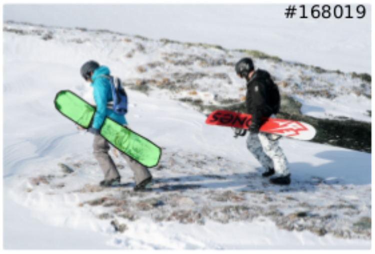
If we can come up with good models for the Visual Dialog Task, there are several real life application where such a model can be deployed. Few examples include:

- Conversation with human AI assistant
  - Mom: Can you see someone at the door?
  - Model: Yes, there is very happy boy at the door!
  - Mom: Is it Ishu?
  - Model: Yes, should I welcome him?
  - Mom: Stupid AI, you should have done that long before ...

- Aiding visually impaired users
  - Model: Your friend Sanchit has uploaded some pictures on fb!
  - Ishu: Oh, wow!! What's he doing?
  - Model: Looks like he is chilling at Bhutan.
  - Ishu: Can you like his fb post for me?
  - Model: Yes, sure.
- Security
  - Boss: Did anyone other than me entered my office yesterday?
  - Model: Yes.
  - Boss: Did anyone carry anything outside?
  - Model: No.
  - Boss: Are you sure?

Challenges

There are several challenges that should be addressed in order to solve the problem effectively. Consider the image which is one of the instance of the "GuessWhat?!" game play. We will highlight few of the major challenges with reference to the given image. For each of the challenges, we have provided a sample question which further tries to clarify the point.



Is it a person?	*No*
Is it an item being worn or held?	*Yes*
Is it a snowboard?	*Yes*
Is it the red one?	*No*
Is it the one being held by the person in blue?	*Yes*

Major Challenges

- **Memory and Reasoning:** Since question answering is an integral part of the task, the models should be able to remember the dialog history, should possess the capability of reasoning in order to give correct answers to the question.
  Example Question : *"Is it the one being held by the person in blue?"* - The model not only needs to locate and understand that there is a person in the image, but also should understand the image to locate the person in blue clothing.

- **Visual Grounding:** We need to ground the words from the dialog to the image.
  Example Question: "Is it a person?" - The model should be able to relate the word person and the parts of the image where the two persons are located.

- **Co-reference Resolution:** The model should be able to de-reference using the dialog context.
  Example Question: "Is it the red one?" - From the dialog history, the model should be able to figure out that the "it" refers to the skateboard here.

Higher-level image understanding, like spatial reasoning and language grounding is required to solve the proposed task. The good part about this setting of Visual Dialog is that it has a objective evaluation mechanism which greatly helps us in quantitatively evaluating the model. We hope that solving this task effectively will get us a step closer towards machine intelligence.

CHAPTER 2

Literature Review

Convolutional Neural Networks (CNN)

CNN are neural networks that work very well with images. CNN uses parameter sharing and also exploits the spatial information of the image. There are several models that uses CNN for the task of visual recognition (eg. image classification). They usually take raw image as input. Some of the popular work includes Krizhevsky *et al.* (2012), LeCun *et al.* (1998), Russakovsky *et al.* (2015), Simonyan and Zisserman (2014) and Szegedy *et al.* (2015).

Recurrent Neural Network (RNN)

RNN are special class of neural networks that are used to model varying sized sequential data. A special class of RNN are Long Short Term Memory (LSTM) [Hochreiter and Schmidhuber (1997)] which are more stable and easy to train. LSTMs have been used in many NLP tasks most famous of which is Machine Translation [Cho *et al.* (2014), Sutskever *et al.* (2014)].

Vision and Language

There are many problems proposed at the intersection of vision and language. Usually in such models a CNN first encodes a visual input and then a RNN model decodes this encoded information to the desired sequence of words. Some of the popular tasks involve visual storytelling [Agrawal *et al.* (2016*b*), Huang *et al.* (2016)], video/movie description [Rohrbach *et al.* (2015), Venugopalan *et al.* (2015), Venugopalan *et al.* (2014)], image captioning [Donahue *et al.* (2015), Fang *et al.* (2015), Karpathy and Fei-Fei (2015), Vinyals *et al.* (2015)].

Visual Question Answering

Some of the work in visual question answering (VQA) include Agrawal *et al.* (2016*a*), Agrawal *et al.* (2017), Das *et al.* (2017*a*), Gao *et al.* (2015), Lu *et al.* (2016), Malinowski *et al.* (2015), Ren *et al.* (2015*a*), Zhang *et al.* (2016). None of the work addressed dialog and focused on single-shot natural language interaction.

Visual Dialog

Das *et al.* (2017*b*) introduced the task of visual dialog for the first time. Specifically, given an image, a dialog history, and a question about the image, the AI agent tries ground the question in image, infer context from history, and answer the question accurately. But they do not generate questions, as it is left to the user to ask question.

GuessWhat?!

Here we list down the baseline models that has been proposed in the "GuessWhat?!" [De Vries *et al.* (2017)] paper.

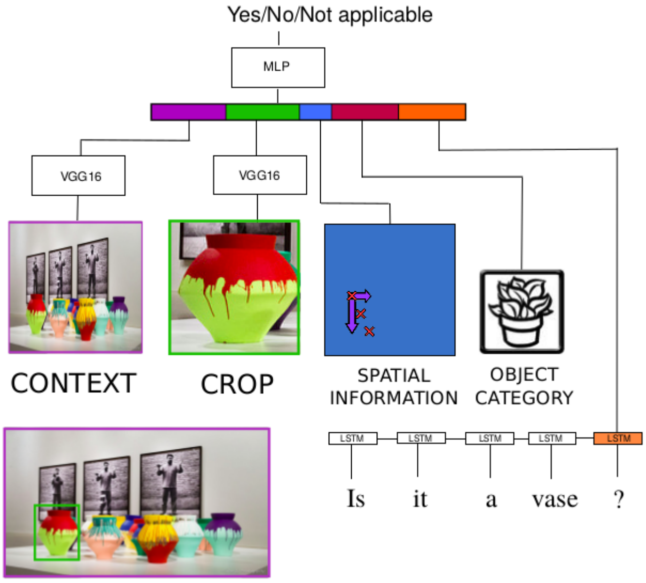
Oracle

Given the the image, the oracle answers the natural language question in yes, no and NA. The inputs of the model being:

- The image I
- the cropped object from S
- its spatial information
- its category c
- the current question q

They concatenate the embedding of all the inputs, and feed them to a single hidden layer MLP that outputs the final answer distribution using a softmax layer. They minimize the cross-entropy error during the training and report the classification error at evaluation time.

The model for Oracle is as follows:



Oracle Baseline Model

Questioner

Given an image, the questioner must ask a series of questions and guess the correct object. They separate the questioner task into two different sub-tasks that are trained independently:

1. Guesser: Given an image and a sequence of questions and answers, it predicts the correct object from the list.
2. Questioning Generator: Given an image and a sequence of questions and answers , it produces a new question.

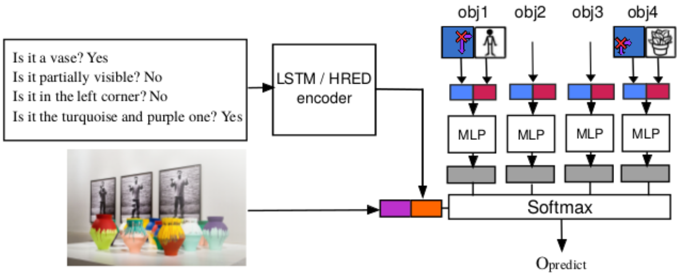
Guesser

The inputs of the model being:

- The image I

- the cropped object from S
- its spatial information
- its category c
- the current question q

They concatenate the image and dialogue features and do a dot-product with the embedding for all the objects in the image, followed by a softmax to obtain a prediction distribution over the objects.



Is it a vase? Yes
Is it partially visible? No
Is it in the left corner? No
Is it the turquoise and purple one? Yes

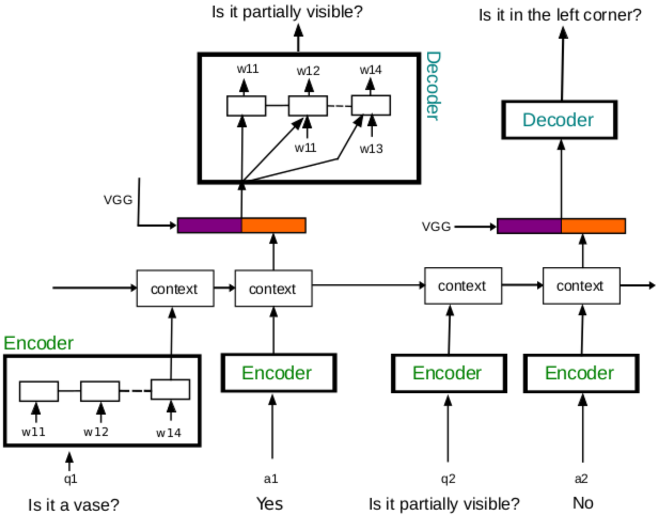
Guesser Baseline Model

Question Generator

Each of the inputs is embedded using the encoder as specified before. The inputs of the model being:

- The image I
- Summarized Question Answer

The model is a Hierarchical Recurrent Encoder Decoder model conditioned on the image. Finally, they train their proposed model by maximizing the conditional log-likelihood.



Question Generator Baseline Model

CHAPTER 3

Model

We believe that under the "GuessWhat?!" game setting, the task of generating question is harder than the task of answering questions. If we can generate a good model for the Questioner (which generates intelligent questions), a similar model with minor tweaks can be used for the Oracle part. Additionally, there has been a lot of work already done in the field of Visual Question Answering. Oracle Model can be borrowed from any one of those models with little modifications.

Unlike the baseline models, instead of having a separate guesser and a question generator, we have a single model for the Questioner. The baseline model for question generator is trained by maximizing the conditional log-likelihood. In this training objective, the question generator does not get any kind of feedback about if the questions that it generated were useful in making the right guess about the object. It just learns to produce human-like questions related to the image. Hence instead of separating the two task of guessing and question generation, we use a single model to do perform both the tasks and allow the question generation part of the model to take feedback from the guesser part and learn to generate questions that help in making right guesses.

Important Features of the Model

Apart from defining a single model for Questioner, there are three important features that has been incorporated in our model that were not the part of the baseline model:

1. **Better Fusion of Multimodal Features**: Usually whenever multimodal data is involved, features from individual mode is extracted using some form of neural network. All these features are concatenated together to form a feature space representing all the modalities. A similar approach is used in the baseline model where the visual features are extracted using a CNN and textual features are extracted using a LSTM and then concatenated together.

There has been several new models proposed in the field of Visual Question Answering which does a better job at merging features from visual and text modalities. For our task, we use FiLM [Perez *et al.* (2017)] mechanism, which is one of the state of the art models for the task of Visual Question Answering for CLEVER dataset [Johnson *et al.* (2017)] and other visual reasoning tasks.

FiLM stands for Feature-wise Linear Modulation. FiLM is a generalized Conditional Normalization mechanism which acts on any intermediate features of a neural network. It applies a affine transformation on the intermediate feature conditioned on an abitrary input. Mathematically, given an arbitrary input x_i, FiLM calculates γ_i and β_i using two functions f and h acting on input x_i:

$$\gamma_i = f(x_i)$$

$$\beta_i = h(x_i)$$

This γ_i and β_i is used to do a affine transformation of intermediate features of the neural network in the following way. Assuming F_i to be the intermediate feature corresponding to x_i,

$$FiLM(F_i|\gamma_i, \beta_i) = \gamma_i F_i + \beta_i$$

In our case, f and h are fully connected neural networks with single hidden layer and x_i is the last state of lstm (summarizing all the questions and answers) after the end of all the dialog.

2. **Regularization**: In general regularization helps to counter overfitting. In general, models which are very complex and has large number of parameters tend to overfit on the training data. Neural networks are complex models which overfit the data easily. In our case, to counter overfitting using the dropout regularization [Srivastava *et al.* (2014)] technique.

3. **Attention**: Attention mechanisms have been very successful in Dialog Systems and Text-based Question Answering. The baseline models introduced in "GuessWhat?!" do not use attention mechanism. Attention mechanism will help us en-

sure that we focus only on the relevant information conditioned on the various inputs.

In our model we use Glimpse attention mechanism [Kim *et al.* (2016)]. Given the filmed image features, we apply Glimpse attention on the image features conditioned on the last state of lstm (summarizing all the questions and answers). We think that the attention mechanism will help better visual grounding.

Like the baseline model, we use an encoder-decoder model. Using the encoder we convert the Image, object features and dialog history to a vector space. Later using a decoder we convert this vector embedding back to the desired output.

Encoding

At any stage the following are the ways of encoding which we use to encode the input:

- **Image:** To encode a image, we pass it through RESNET [He *et al.* (2016)] pretrained on ImageNet [Deng *et al.* (2009)] to obtain the feature maps at the end of block4. We then use max pooling over these features and then flatten it to get the image features. If the RESNET if FiLMed, we do max pooling over the FiLMed features (Note that Filmed features are of the exact same shape and size as that of features that come out of block4 layer of RESNET). At the end, we apply a single layer of fully connected layer to get the visual features.
- **Dialog History:** To encode the text (which is the dialog history in our case), we use RNN LSTM [Hochreiter and Schmidhuber (1997)]. The initial state of the RNN LSTM is always set to zero state.
- **Object features:** We use object spatial information and object category to represent object in our model. We do not use the visual features as of now as it has been observed in the baseline model that the visual features degrades the model performance.
  - *Objects category:* We represent the object category using one hot vector of the class.

- *Object spatial information:* Let w_{box} and h_{box} denote the width and height of the bounding box. We use the same form for representing the spatial information for locating the object as done in the baseline model. The spatial location is represented by a vector of the form [x_{min}, y_{min}, x_{max}, y_{max}, x_{center}, y_{center}, w_{box}, h_{box}]. We pass this vector through a shared layer of MLP to get the object features. Like in the baseline model, image height and width are normalized such that coordinates range from -1 to 1 with center of image being the origin of the normalized coordinates.

We concatenate both the object object category and object spatial information to encode information about the object.

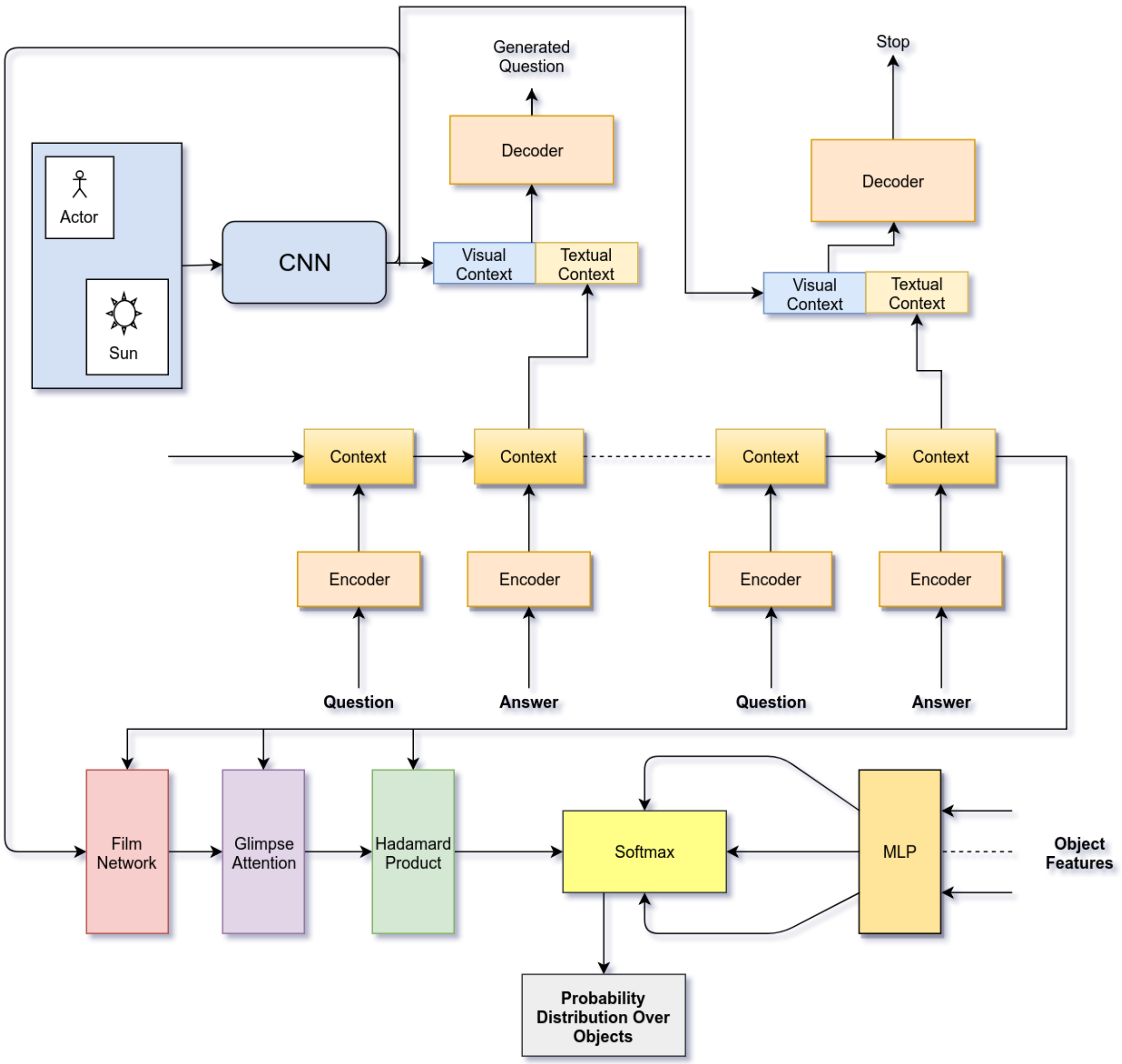
Decoding

Decoding has to be done for two of the following tasks:

- **Generating Questions**: A standard RNN LSTM is used for the task of decoding and generating the questions. Visual context (output of CNN) and text context (output of encoder LSTM) are concatenated and set as initial state for the decoder LSTM.

- **Guessing Object:** First we take the visual features from a CNN (in our case a RESNET). We apply FiLM layer on top of the visual features and then apply attention over it conditioned on text features; then we take a hadamard product between the visual features and the text features. At the end, we take a dot product between the output of hadamard operation and features of each of the object and then apply a softmax layer to obtain a distribution over the objects.

The output of the encoder is set an a initial state of the LSTM RNN language model. The training is done by maximizing the log-likelihood of the ground truth answer (encoded as vector). This model is used in the questioner part, where the questioner generates questions in natural language.


Generated Question
Stop
Decoder
Decoder
Actor
Sun
CNN
Visual Context
Textual Context
Visual Context
Textual Context
Context
Context
Context
Context
Encoder
Encoder
Encoder
Encoder
Question
Answer
Question
Answer
Film Network
Glimpse Attention
Hadamard Product
Softmax
MLP
Object Features
Probability Distribution Over Objects

CHAPTER 4

Experiments and Results

Dataset Statistics

We experiment our model on "GuessWhat?!" dataset. This dataset is a collection of 155,280 dialogues over 66,537 images taken from the MS COCO dataset [Lin *et al.* (2014)] with 134,073 unique objects. There can be varied number of question/answer payer per game and in total there are 831,889 question/answer pairs in the dataset. On an average, there are 2.3 games played on a single image. On average, each dialogue has 5.2 question/answer pair.

Experimentation Details

In the prepossessing step, we rescale all the images to a size of 224×224 and then normalize them. We also ignore all words that appear less than 4 times in the all of the dialogues. Our code [link] is implemented in Tensorflow [Abadi *et al.* (2015)]. We use a minibatch of size 2 and train our model with Adam optimizer [Kingma and Ba (2014)] with learning rate of 1e-4. We train our model for 30 epochs. We use a pretrained RESNET and keep its parameters fixed during training. All other parameters of the model like the word embeddings, parameters of the MLP layers etc are trainable.

There are two learning objectives of our model. Hence there are two kinds of losses which are as follows:

- Question Generation Loss: For generating dialogues we try to maximize the conditional log likelihood (log P(Question | Dialog History, Image)) of the ground truth question.

- Guesser Loss: The guesser part of our models gives us a distribution over the objects. We try to maximize the negative log-likelihood of the correct answer.

At the time of training, we appropriately weight both the types of losses such that the scaled losses are of similar magnitudes. After each epoch we find model performance on the validation set and save the best model and report results for that model on test set. Note that for evaluation purposes we need pre-trained Oracle. In our experiments we use the same pre-trained Oracle as used for bench marking the baseline models of "GuessWhat?!" [De Vries *et al.* (2017)] paper.

Results

Similar to "GuessWhat?!" [De Vries *et al.* (2017)], we measure the performance of model by measuring the percentage error that the Questioner makes in guessing the object selected by the Oracle. We compare our model performance against the "GuessWhat?!" baselines, which to our knowledge is the only paper that reports results for the Questioner part of the "GuessWhat?!" game. Since in the original paper, the questioner was divided into two parts namely Guesser and the Question Generator, we compare our results with Guesser of them.

Model	Train err %	Val err %	Test err %
Baseline Model	27.9	37.9	38.7
Our Model	27.8	35.3	36.2

Comparing results with baseline model

CHAPTER 5

Directions for Future work and Conclusion

There are two promising ideas that can further improve the models which are following:

- **More world knowledge and explicit Information:** One of the main reasons why humans are very good in Visual Dialog task is because the huge world knowledge we possess, collected over the many years. Moreover, there are existing solutions which can give us caption for the image by only taking raw image as an input. We can use this information and provide it explicitly to our model making the task of decoding image easier.
- **Getting better visual features for objects:** Lets analyze the classification errors for the oracle baseline. The reported results are for different models of oracle for different input settings.

	Model	Train err %	Val err %	Test err %
1.	Dominant class (no)	47.4	46.2	50.9
2.	Question	40.2	41.7	41.2
3.	Image	45.7	46.7	46.7
4.	Crop	40.9	42.7	43.0
5.	Question + Crop	22.3	29.1	29.2
6.	Question + Image	37.9	40.2	39.8
7.	Question + Category	23.1	25.8	25.7
8.	Question + Spatial	28.0	31.2	31.3
9.	Question + Category + Spatial	17.2	21.1	**21.5**
10.	Question + Category + Crop	20.4	24.4	24.7
11.	Question + Spatial + Crop	19.4	26.0	26.2
12.	Question + Category + Spatial + Crop	16.1	21.7	22.1
13.	Question + Spatial + Crop + Image	20.7	27.7	27.9
14.	Question + Category + Spatial + Image	19.2	23.2	23.5

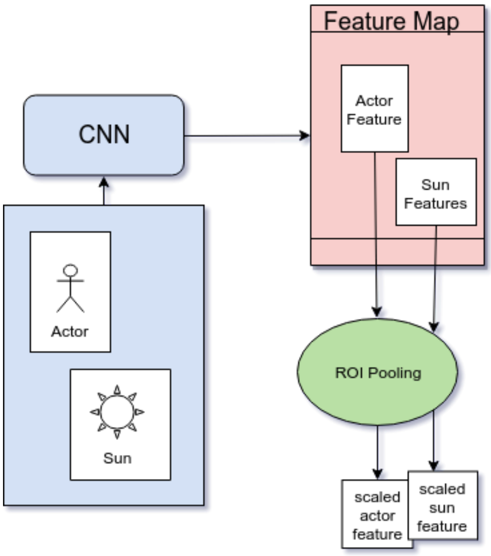
Table 5.1: Oracle Baseline Results

Lets compare row 14, and 9; 12, and 9. We see that the Oracle performs better when it is not given any visual input to the Oracle. This is rather contrary to the general intuition. In general, having extra information about the visual features of the object and image must help in generating better informed questions for the Oracle and thus should result in more number of successful games. But we see that this extra information deteriorate the model correctness. One logical explanation one can come up for this observation is that the visual features must be imperfect. These visual features can be noisy and hence does not improve the model performance.

So one promising improvement that can be made to improve the Oracle model is to feed improved visual features to the Oracle. As of now, visual features for images are obtained by passing image through a pre-trained RESNET. So one can train this RESNET on the images of "GuessWhat?!" dataset and make RESNET a trainable part of the Oracle model and train it in an end to end fashion. Currently, for getting the visual features of the objects, the cropped images of the objects are *re-scaled* and are passed through the *pre-trained VGG net*. This pre-trained VGG net is trained on *full scale images* for the task of *image classification*.

One can use features extracted from object detection models such as Faster-RCNN [Ren *et al.* (2015*b*)]. Since we have the bounding box for each object, we can take the part of the feature map corresponding the bounding box of the image. Over this part of the feature map, we can apply ROI pooling and scale it to get a fixed size feature map.

One can expect to get better visual features for objects as unlike image classification objective which extracts feature from entire image operating on the entire image, the training objectives of object detection models is much more fine grained operating on section of image and is focused towards finding objects in different section of image.

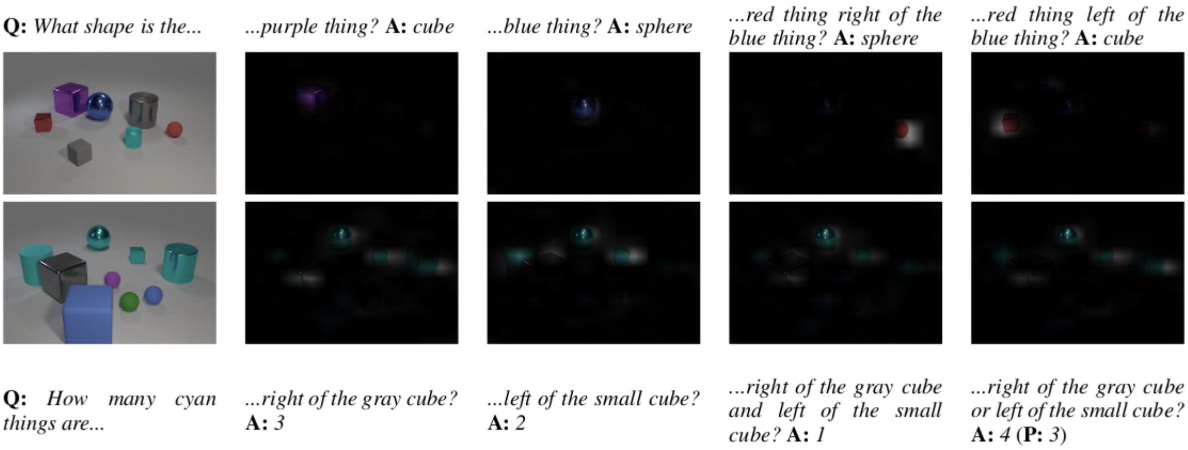


Object Features using ROI pooling

- **More supervision using available information:** At the end of the game, the Questioner knows the object that was selected by the Oracle, including the location of the object in the image. None of the models have used this information so far. We can define a probability distribution over each pixel of the image. Let $\mathcal{P}_i$ denote the probability that pixel i is a part of the selected object. Let y_i be 1 when pixel i is a part of the selected object else 0. Using this, we can define an additional supervised cross entropy loss. We can add this (weighted if necessary) loss on top of all the other losses and help train our model with more supervised information.

$$l_i = y_i \; log(\mathcal{P}_i) + (1 - y_i) \; log(1 - \mathcal{P}_i)$$

There are several ways in which one can define P_i and extend our model. For CLEVER dataset [Johnson *et al.* (2017)], which is one of the popular Visual Question Answering dataset, it has been observed that whenever the model proposed in [Perez *et al.* (2017)] predicts the correct answer, the visualizations of the last globally max-pooled features of FiLM layer correctly localizes objects referred in the answer. Moreover the localizations are not correct when the answer of the model is not correct. This is a clear indication that we can use these features to define our probability distribution. Since we already use the FiLM layer in our model, we can we can easily extend the current architecture to get P_i. First we can use a sequence of de-convolution layers on the last globally max-pooled features of FiLM layer to match the size of the image. We can then apply a neural network with only one layer with logistic activation function to get P_i.


Q: What shape is the...
...purple thing? A: cube
...blue thing? A: sphere
...red thing right of the blue thing? A: sphere
...red thing left of the blue thing? A: cube
Q: How many cyan things are...
...right of the gray cube? A: 3
...left of the small cube? A: 2
...right of the gray cube and left of the small cube? A: 1
...right of the gray cube or left of the small cube? A: 4 (P: 3)

Visualization of last max-pooled features of FiLM layer for CLEVER dataset

It also will give us a visualization tool to monitor the training of the model. This probability distribution can be thought of as a heat map that tells us the location of the images where our model is paying more attention. Analyzing such a heat map can bring more insights, which can further be used to improve the model.

REFERENCES

1. **Abadi, M.**, **A. Agarwal**, **P. Barham**, **E. Brevdo**, **Z. Chen**, **C. Citro**, **G. S. Corrado**, **A. Davis**, **J. Dean**, **M. Devin**, **S. Ghemawat**, **I. Goodfellow**, **A. Harp**, **G. Irving**, **M. Isard**, **Y. Jia**, **R. Jozefowicz**, **L. Kaiser**, **M. Kudlur**, **J. Levenberg**, **D. Mané**, **R. Monga**, **S. Moore**, **D. Murray**, **C. Olah**, **M. Schuster**, **J. Shlens**, **B. Steiner**, **I. Sutskever**, **K. Talwar**, **P. Tucker**, **V. Vanhoucke**, **V. Vasudevan**, **F. Viégas**, **O. Vinyals**, **P. Warden**, **M. Wattenberg**, **M. Wicke**, **Y. Yu**, and **X. Zheng** (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

2. **Agrawal, A.**, **D. Batra**, and **D. Parikh** (2016*a*). Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*.

3. **Agrawal, A.**, **J. Lu**, **S. Antol**, **M. Mitchell**, **C. L. Zitnick**, **D. Parikh**, and **D. Batra** (2017). Vqa: Visual question answering. *International Journal of Computer Vision*, **123**(1), 4–31.

4. **Agrawal, H.**, **A. Chandrasekaran**, **D. Batra**, **D. Parikh**, and **M. Bansal** (2016*b*). Sort story: Sorting jumbled images and captions into stories. *arXiv preprint arXiv:1606.07493*.

5. **Cho, K.**, **B. Van Merriënboer**, **C. Gulcehre**, **D. Bahdanau**, **F. Bougares**, **H. Schwenk**, and **Y. Bengio** (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

6. **Das, A.**, **H. Agrawal**, **L. Zitnick**, **D. Parikh**, and **D. Batra** (2017*a*). Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, **163**, 90–100.

7. **Das, A.**, **S. Kottur**, **K. Gupta**, **A. Singh**, **D. Yadav**, **J. M. Moura**, **D. Parikh**, and **D. Batra**, Visual dialog. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2. 2017*b*.

8. **De Vries, H.**, **F. Strub**, **S. Chandar**, **O. Pietquin**, **H. Larochelle**, and **A. Courville**, Guesswhat?! visual object discovery through multi-modal dialogue. *In Proc. of CVPR*. 2017.

9. **Deng, J.**, **W. Dong**, **R. Socher**, **L.-J. Li**, **K. Li**, and **L. Fei-Fei**, Imagenet: A large-scale hierarchical image database. *In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009.

10. **Donahue, J.**, **L. Anne Hendricks**, **S. Guadarrama**, **M. Rohrbach**, **S. Venugopalan**, **K. Saenko**, and **T. Darrell**, Long-term recurrent convolutional networks for visual recognition and description. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

11. **Fang, H.**, **S. Gupta**, **F. Iandola**, **R. K. Srivastava**, **L. Deng**, **P. Dollár**, **J. Gao**, **X. He**, **M. Mitchell**, **J. C. Platt**, *et al.*, From captions to visual concepts and back. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

12. **Gao, H.**, **J. Mao**, **J. Zhou**, **Z. Huang**, **L. Wang**, and **W. Xu**, Are you talking to a machine? dataset and methods for multilingual image question. *In Advances in neural information processing systems*. 2015.

13. **Goyal, Y.**, **T. Khot**, **D. Summers-Stay**, **D. Batra**, and **D. Parikh**, Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *In CVPR*, volume 1. 2017.

14. **He, K.**, **X. Zhang**, **S. Ren**, and **J. Sun**, Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

15. **Hochreiter, S.** and **J. Schmidhuber** (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

16. **Huang, T.-H. K.**, **F. Ferraro**, **N. Mostafazadeh**, **I. Misra**, **A. Agrawal**, **J. Devlin**, **R. Girshick**, **X. He**, **P. Kohli**, **D. Batra**, *et al.*, Visual storytelling. *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016.

17. **Johnson, J.**, **B. Hariharan**, **L. van der Maaten**, **L. Fei-Fei**, **C. L. Zitnick**, and **R. Girshick**, Clevr: A diagnostic dataset for compositional language and elementary visual

reasoning. *In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017.

18. **Karpathy, A.** and **L. Fei-Fei**, Deep visual-semantic alignments for generating image descriptions. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

19. **Kim, J.-H.**, **K.-W. On**, **W. Lim**, **J. Kim**, **J.-W. Ha**, and **B.-T. Zhang** (2016). Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.

20. **Kingma, D. P.** and **J. Ba** (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

21. **Krizhevsky, A.**, **I. Sutskever**, and **G. E. Hinton**, Imagenet classification with deep convolutional neural networks. *In Advances in neural information processing systems*. 2012.

22. **LeCun, Y.**, **L. Bottou**, **Y. Bengio**, and **P. Haffner** (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**(11), 2278–2324.

23. **Lin, T.-Y.**, **M. Maire**, **S. Belongie**, **J. Hays**, **P. Perona**, **D. Ramanan**, **P. Dollár**, and **C. L. Zitnick**, Microsoft coco: Common objects in context. *In European conference on computer vision*. Springer, 2014.

24. **Lu, J.**, **J. Yang**, **D. Batra**, and **D. Parikh**, Hierarchical question-image co-attention for visual question answering. *In Advances In Neural Information Processing Systems*. 2016.

25. **Malinowski, M.**, **M. Rohrbach**, and **M. Fritz**, Ask your neurons: A neural-based approach to answering questions about images. *In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2015.

26. **Perez, E.**, **F. Strub**, **H. De Vries**, **V. Dumoulin**, and **A. Courville** (2017). Film: Visual reasoning with a general conditioning layer. *arXiv preprint arXiv:1709.07871*.

27. **Ren, M.**, **R. Kiros**, and **R. Zemel**, Exploring models and data for image question answering. *In Advances in neural information processing systems*. 2015*a*.

28. **Ren, S.**, **K. He**, **R. Girshick**, and **J. Sun**, Faster r-cnn: Towards real-time object detection with region proposal networks. *In Advances in neural information processing systems*. 2015*b*.

29. **Rohrbach, A.**, **M. Rohrbach**, **N. Tandon**, and **B. Schiele**, A dataset for movie description. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

30. **Russakovsky, O.**, **J. Deng**, **H. Su**, **J. Krause**, **S. Satheesh**, **S. Ma**, **Z. Huang**, **A. Karpathy**, **A. Khosla**, **M. Bernstein**, *et al.* (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, **115**(3), 211–252.

31. **Simonyan, K.** and **A. Zisserman** (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

32. **Srivastava, N.**, **G. Hinton**, **A. Krizhevsky**, **I. Sutskever**, and **R. Salakhutdinov** (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, **15**(1), 1929–1958.

33. **Sutskever, I.**, **O. Vinyals**, and **Q. V. Le**, Sequence to sequence learning with neural networks. *In Advances in neural information processing systems*. 2014.

34. **Szegedy, C.**, **W. Liu**, **Y. Jia**, **P. Sermanet**, **S. Reed**, **D. Anguelov**, **D. Erhan**, **V. Vanhoucke**, **A. Rabinovich**, *et al.*, Going deeper with convolutions. Cvpr, 2015.

35. **Venugopalan, S.**, **M. Rohrbach**, **J. Donahue**, **R. Mooney**, **T. Darrell**, and **K. Saenko**, Sequence to sequence-video to text. *In Proceedings of the IEEE international conference on computer vision*. 2015.

36. **Venugopalan, S.**, **H. Xu**, **J. Donahue**, **M. Rohrbach**, **R. Mooney**, and **K. Saenko** (2014). Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.

37. **Vinyals, O.**, **A. Toshev**, **S. Bengio**, and **D. Erhan**, Show and tell: A neural image caption generator. *In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015.

38. **Zhang, P.**, **Y. Goyal**, **D. Summers-Stay**, **D. Batra**, and **D. Parikh**, Yin and yang: Balancing and answering binary visual questions. *In Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016.