The Art of Deep Connection - Towards Natural and Pragmatic Conversational Agent Interactions

Arijit Ray

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Engineering

Jia-Bin Huang, Chair
Devi Parikh, Co-chair
A. Lynn Abbott

April 24, 2017
Blacksburg, Virginia

Keywords: Deep Learning, Computer Vision, Natural Language Processing

The Art of Deep Connection - Towards Natural and Pragmatic Conversational Agent Interactions

Arijit Ray

ABSTRACT

As research in Artificial Intelligence (AI) advances, it is crucial to focus on having seamless communication between humans and machines in order to effectively accomplish tasks. Smooth human-machine communication requires the machine to be sensible and human-like while interacting with humans, while simultaneously being capable of extracting the maximum information it needs to accomplish the desired task. Since a lot of the tasks required to be solved by machines today involve the understanding of images, training machines to have human-like and effective image-grounded conversations with humans is one important step towards achieving this goal. Although we now have agents that can answer questions asked for images, they are prone to failure from confusing input, and cannot ask clarification questions, in turn, to extract the desired information from humans. Hence, as a first step, we direct our efforts towards making Visual Question Answering agents human-like by making them resilient to confusing inputs that otherwise do not confuse humans. Not only is it crucial for a machine to answer questions reasonably, it should also know how to ask questions sequentially to extract the desired information it needs from a human. Hence, we introduce a novel game called the Visual 20 Questions Game, where a machine tries to figure out a secret image a human has picked by having a natural language conversation with the human. Using deep learning techniques like sequence-to-sequence learning (using recurrent neural networks) and reinforcement learning, we demonstrate promise towards scalable and reasonable performances on both the tasks.

(Grant Information)

This work was supported in part by the following: National Science Foundation CAREER awards to DB and DP, Alfred P. Sloan Fellowship, Army Research Office YIP awards to DB and DP, ICTAS Junior Faculty awards to DB and DP, Army Research Lab grant W911NF-15-2-0080 to DP and DB, Office of Naval Research grant N00014-14-1- 0679 to DB, Paul G. Allen Family Foundation Allen Distinguished Investigator award to DP, Google Faculty Research award to DP and DB, AWS in Education Research grant to DB, and NVIDIA GPU donation to DB.

The Art of Deep Connection - Towards Natural and Pragmatic Conversational Agent Interactions

Arijit Ray

GENERAL AUDIENCE ABSTRACT

Research in Artificial Intelligence has reached to a point where computers can answer natural free-form questions asked to arbitrary images in a somewhat reasonable manner. These machines are called Visual Question Answering agents. However, they are prone to failure from even a slightly confusing input. For example, for an obviously irrelevant question asked to an image, they would answer something non-sensical instead of recognizing that the question is irrelevant. Furthermore, they also cannot ask questions in turn to humans for clarification or for more information. These shortcomings not only harm their efficacy, but also harm their perceived trust from human users. In order to remedy these problems, we first direct our efforts towards making Visual Question Answering agents capable of identifying an irrelevant question for an image. Next, we also try to train machines to be able to ask questions to extract more information from humans to make an informed decision. We do this by introducing a novel game called the Visual 20 Questions game, where a machine tries to figure out a secret image a human has picked by having a natural language conversation with the human. Deep learning techniques such as sequence-to-sequence learning using recurrent neural networks make it possible for machines to learn how to converse based on a series of conversational exchanges made between two humans. Techniques like reinforcement learning make it possible for machines to better themselves based on rewards it gets for accomplishing a task in a certain way. Using such algorithms, we demonstrate promise towards scalable and reasonable performances on both the tasks.

(Grant Information)

This work was supported in part by the following: National Science Foundation CAREER awards to DB and DP, Alfred P. Sloan Fellowship, Army Research Office YIP awards to DB and DP, ICTAS Junior Faculty awards to DB and DP, Army Research Lab grant W911NF-15-2-0080 to DP and DB, Office of Naval Research grant N00014-14-1- 0679 to DB, Paul G. Allen Family Foundation Allen Distinguished Investigator award to DP, Google Faculty Research award to DP and DB, AWS in Education Research grant to DB, and NVIDIA GPU donation to DB.

Contents

List of Figures

List of Tables

Preface

"Believe in the why's, the how's and what's will always follow." - modified from Simon Sinek's "Golden Circle" TED Talk.

I believe in connection - personal connection. Proper communication can arguably be stated as the panacea to a lot of the problems in the world today - personal, societal and political. My strong belief in this fact is the "why".

Like every other 12-year-old growing up with science fiction thrillers, Artificial Intelligence (AI) has fascinated me since I was a kid! This fascination established the path to finding my "how". As I paved my career into AI, I learned how deep learning has recently shown a lot of promise in common-sense and semantic understanding of concepts that were previously understood by humans alone. Hence, deep learning became my "how". I wanted to leverage the power of deep learning to make machines connect with humans.

To ground my aspirations into something concrete and achievable within the span of a Master's degree, the topic of image-grounded AI-human conversations became my "what".

Heartfelt thanks to Prof. Devi Parikh, Prof. Jia-Bin Huang, Prof. Dhruv Batra, and Prof. A. Lynn Abbott for believing in my "why", for giving me an opportunity to pursue my dream, and for being the best advisors one can hope for. This thesis is possible only because I found a community that believed in my "why" - the community at the Computer Vision Lab at Virginia Tech.

Chapter 1

Introduction

Figure 1.1: Motivation: While conversing with humans, a machine must be able to a) detect a wrong input, and b) ask clarification questions to get more information.

As Artificial Intelligence (AI) advances, it is crucial to investigate how machines can interact with humans effectively to accomplish a task. For human-machine collaboration to be effective at a task, the machine should be able to **(a)** answer questions asked by humans reasonably and **(b)** ask questions back to humans to extract further information as it needs.

Let us illustrate this with an example scenario as shown in Figure 1.1. Consider a blind woman walking down the aisle of a supermarket looking for pickles. She unknowingly points to an irrelevant object in the aisle and asks her visual assistant, "Which row are the pickles?". An ideal visual assistant would notify her of an irrelevant input saying that she is pointing to a trash can. Now say, she turns around and does point to the rows of pickles this time and asks the same question. Instead of blurting out an answer, the visual assistant should ideally ask more clarification questions to zero-down on the desired choice of pickles (e.g., Kroger brand and low-sodium) and then answer the location of them.

While the above reasoning is second-nature to humans, it is hard for robots to do. Conventional Visual Question Answering (VQA) agents (eg, [3, 30, 23]) are discriminatively trained to output a choice of answer given the question and image. This means that the VQA agent just reasons about $p(y|x_i, x_q)$ where y is the answer, x_i is the image and x_q is the question. They are not trained to

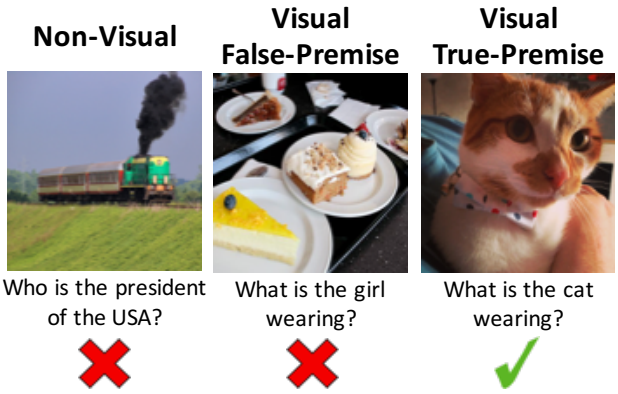

Figure 1.2: Example irrelevant (non-visual, false-premise) and relevant (visual true-premise) questions in VQA.

reason about whether x_i and x_q make sense, or to ask questions in turn to extract more information from a human.

Hence, we explore on ways to impart two key communication skills to machines in this thesis - **(a)** identifying irrelevant or confusing questions, and **(b)** asking clarification questions to extract more information as needed.

VQA Relevance Our first work is motivated by the following key observation – all current VQA systems always output an answer *regardless of whether the input question makes any sense for the given image or not*. Figure 1.2 shows examples of relevant and irrelevant questions. When VQA systems are fed irrelevant questions as input, they understandably produce nonsensical answers (Q: *"What is the capital of Argentina?"* A: *"fire hydrant"*). Humans, on the other hand, are unlikely to provide such nonsensical answers and will instead answer that this is irrelevant or use another knowledge source to answer correctly, when possible. We argue that this implicit assumption by all VQA systems – that an input question is always relevant for the input image – is simply untenable as VQA systems move beyond standard academic datasets to interacting with real users, who may be unfamiliar, or malicious. The goal of this work is to make VQA systems more human-like by providing them the capability to identify relevant questions.

While existing work has reasoned about cross-modal similarity, being able to identify whether a question is relevant to a given image is a novel problem with real-world applications. In human-robot interaction, being able to identify questions that are dissociated from the perception data available is important. The robot must decide whether to process the scene it perceives or query external world knowledge resources to provide a response.

As shown in Figure 1.2, we study three types of question-image pairs: **Non-Visual**. These questions are not questions about images at all – they do not require information from *any* image to be answered *e.g., "What is the capital of Argentina?"*. **Visual False-Premise**. While visual, these questions do not apply to the given image. For instance, the question *"What is the girl wearing?"*

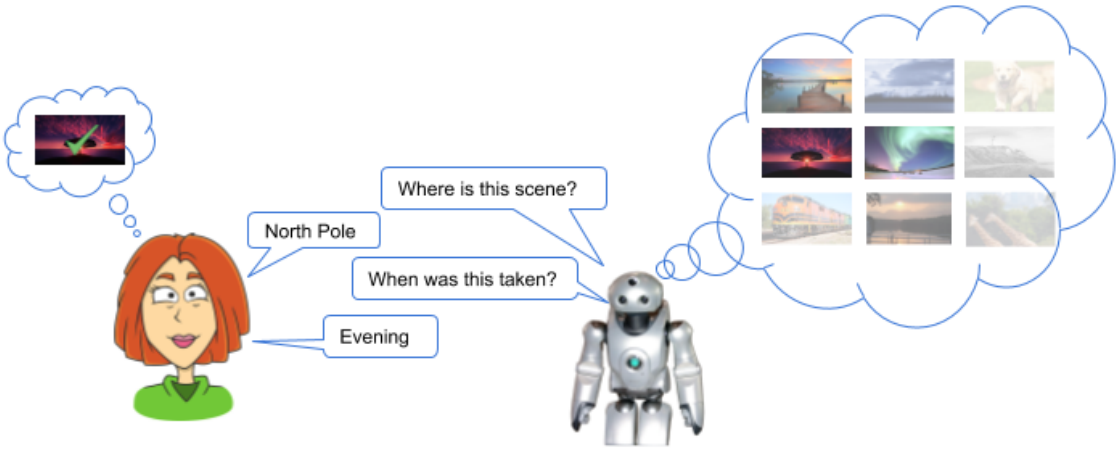

Figure 1.3: A human secretly picks an image from a pool of images, which the robot does not know. The robot asks questions sequentially to figure out what image the user secretly picked.

makes sense only for images that contain a girl in them. **Visual True-Premise**. These questions are relevant to (i.e., have a premise which is true for) the image at hand.

We introduce datasets and train models to recognize both non-visual and false-premise question-image (QI) pairs in the context of VQA. First, we identify whether a question is visual or non-visual; if visual, we identify whether the question has a true premise for the given image. For visual vs. non-visual question detection, we use a Long Short-Term Memory (LSTM) recurrent neural network (RNN) trained on part of speech (POS) tags to capture visual-specific linguistic structure. For true vs. false-premise question detection, we present one set of approaches that use the uncertainty of a VQA model and another set that use pre-trained captioning models to generate relevant captions (or questions) for the given image and then compare them to the given question to determine relevance.

Our proposed models achieve accuracies of 92% for detecting non-visual, and 75% for detecting false-premise questions, which significantly outperform strong baselines. We also show through human studies that a VQA system that reasons about question relevance is picked *significantly more often* as being more intelligent, human-like and reasonable than a baseline VQA system which does not. Our code and datasets are made publicly available.

Visual 20 Questions Game In VQA, the task is for AI to answer a question for the image. However, in many applications as illustrated in Figure 1.1, it would be useful for machines to be able to ask questions to humans to figure out what exactly they have in mind.

Our second work is motivated towards training conversational agents that not only ask questions that are relevant to the context, but also are high in the amount of potential information they can extract. We explore this avenue by introducing the task of "Visual 20 Questions".

As shown in Figure 1.3, "The game of Visual 20 Questions" is an image-based version of the

classic 20q.net game. A user secretly imagines an image, and our machine learning model tries to guess the scene by asking the user a limited set of informative questions in sequence, in turn filtering the set of images based on the user answer at each step. Such a game has many useful applications such as interactive image retrieval and generation, deeper image understanding, and multimodal dialogue. As a first approach, we formalize this game by letting the user pick an image from a fixed pool of images, and evaluate the model on the retrieval accuracy of that particular image. We demonstrate a simple re-ranking-based approach that works reasonably well on this game. We experiment with ways in which we choose the best question and filter our image pool. We show that a semi-soft information gain with soft image filtering and enforcing question diversity helps improve the image retrieval rate. Next, we look at the interesting technical problem of how to learn to generate questions that are most effective for completing such a human-AI collaborative task. We explore a combination of reinforcement learning and supervised learning (based on human dialog data) to train a conversational agent to ask questions sequentially for most effectively converging down to the desired image.

Chapter 2

Related Work

Relevance. There is a large body of existing work that reasons about cross-modal similarity: how well an image matches a query tag [22] in text-based image retrieval, how well an image matches a caption [15, 36, 28, 17, 14], and how well a video matches a description [13, 20].

In our work, if a question is deemed irrelevant, the VQA model says so, as opposed to answering the question anyway. This is related to perception systems that do not respond to an input where the system is likely to fail. Such failure prediction systems have been explored in vision [37, 11] and speech [38, 31, 7, 35]. The related idea is to avoid a highly specific prediction if there is a chance of being wrong, instead of making a more generic prediction that is more likely to be right [10].

To the best of our knowledge, our work is the first to study the relevance of questions in VQA. [5] classify users' intention of questions for community question answering services. Most related to our work is [12] that extracts visual text from within Flickr photo captions to be used as supervisory signals for training image captioning systems. Our motivation is to endow VQA systems the ability to detect non-visual questions to respond in a human-like fashion. Moreover, we also detect a more fine-grained notion of question relevance via true- and false premise.

Twenty question game. There have been extensive efforts on studying games with a similar setting of the visual 20 question problem. The most relevant one is the classic 20q.net game. In one session of the game, the computer asks a participant a series of questions in an effort to figure out the specific object/topic/character in his/her mind. In this work, we extend the single concept (e.g., a person or an object) to the much richer domain (i.e., an image). While the space of commonly imagined single objects/topics might not be complicated, the space of images is virtually infinite because a single image encapsulates the complex interaction of numerous objects and topics.

Some works looked at a similar setup for improving the classification decision of an image [4]. They improve classification based on user responses for certain attributes of the image on top of

a prior belief of a classifier. Here, only one image is involved at a time, and both the user and the computer can see the image. The user's answers to the questions posed by the machine improve the machine's classification decision. In our case, the computer does not know which image user picked but can see all images. The user answers questions posed by the machine to improve the retrieval of the correct image by the machine.

Question generation. There has also been a lot of effort into generating natural language questions about images ([27] [29]). In [27], the motivation is that engaging questions are not just about the image content, but rather more about the inference of the event given the image content. e.g., in a picture of a motorcycle accident, an interesting question would be "is the motorcyclist alive?" rather than "Is there a motorcycle in the picture?" We try to generate a pragmatic question that solves the purpose rather than focusing on what's interesting. In our case, the task is the correct retrieval of the desired image. We use [29] to calculate the relevance of a question for an image pool and also use their question generator as a baseline.

Some works have also looked at making pragmatic image captioning models that describe an image to fulfill a task at hand [33] [2]. They focus on generating a caption that helps in choosing an image from a choice of two. In our case, we have a pool of images, and a computer has to ask informative questions to discriminate the desired image from the pool of images.

Interactive image-grounded dialog. There has also been significant interest in making AI agents have natural image-grounded conversations [26] [8], but most of these are not towards solving a task. We aim at having a pragmatic conversation towards effectively solving a task - the task of "Visual 20 Questions".

Our work is probably most similar to the work on Visual Dialog by [8]. They introduce a task where there are two people talking about an image. One of them cannot see the image and tries to figure out the image by asking a bunch of questions. However, the person is seeded with the caption of the image which gives him/her some idea of the image. In our case, the robot knows nothing about the picture. Hence, our idea is to ask widely relevant and informative questions to figure out the image.

There has also been interest in AI-AI conversation using reinforcement learning [9] [19] [18]. In [9], they train a questioner and answerer bot, where the questioner bot asks questions to the answerer in sequence to figure out an unknown image. However, the questioner bot is shown a caption about the image, to begin with. For our case, there is no caption/information, to begin with. The questioner bot must learn to start with widely relevant questions (like "what is in the picture?") and slowly make questions more specific to converge onto the desired image.

Chapter 3

Question Relevance in VQA- Identifying Non-Visual and False-Premise Questions

Imagine a visually impaired person asking a visual assistance system if it is safe to cross the road while unknowingly pointing at a fire hydrant by the side of the road. Visual Question Answering (VQA) models today will go ahead and answer, "yes" as shown in Figure 3.1. Ideally, we would want the VQA model to notify the user of an irrelevant input. This work [29] focuses on identifying irrelevant Question-Image (QI) pairs.

3.1 Approach

We break the detection of relevant vs. irrelevant pairs into two subparts: (1) visual vs. non-visual QI pairs, and (2) true- vs. false-premise QI pairs.

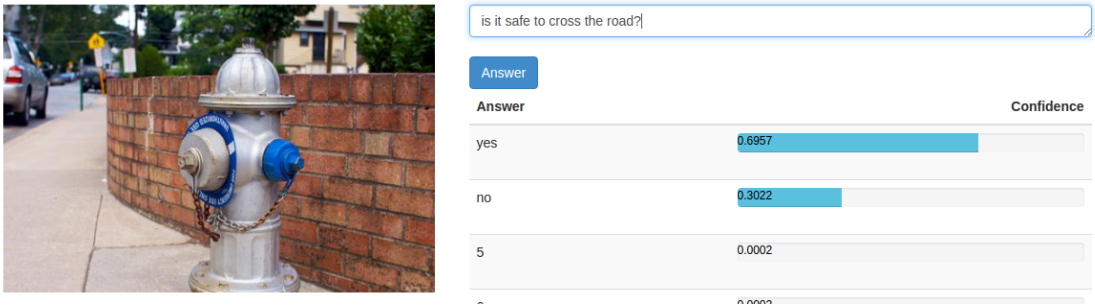

Figure 3.1: Problem with discriminative VQA models today. VQA answers an irrelevant question with a potentially wrong answer. Courtesy: CloudCV [1] demo for VQA [3], https://cloudcv.org/vqa/

3.1.1 Visual vs. Non-Visual Detection

Recall that the task here is to detect visual questions from non-visual ones. Non-visual questions, such as *"Do dogs fly?"* or *"Who is the president of the USA?"*, often tend to have a difference in the linguistic structure from that of visual questions, such as *"Does this bird fly?"* or *"What is this man doing?"*. We compare our approach (**LSTM**) with a baseline (**RULE-BASED**):

1. **RULE-BASED**. A rule-based approach to detect non-visual questions based on the part of speech (POS)[1] tags and dependencies of the words in the question. For example, if a question has a plural noun with no determiner before it and is followed by a singular verb (*"Do dogs fly?"*), it is a non-visual question.
   Various hand-coded rules based on observation and experimentation were added to make this baseline as strong as possible. We list a few examples:
   - If there is a plural noun, without a determiner before it, followed by a verb (e.g., "Do dogs fly?"), the question is non-visual.
   - If there is a determiner followed by a noun (e.g., "Do dogs fly in this picture?"), the question is visual.
   - If there is a personal or possessive pronoun before a noun (e.g., "What color is his umbrella?"), the question is visual.
   - We use a list of words that frequently occur in the non-visual questions but infrequently in visual questions. These include words such as: 'God,' 'Life,' 'meaning,' and ''universe.' If any words from this list are present in the question, the question is classified as non-visual.
2. **LSTM**. We train an LSTM with 100-dimensional hidden vectors to embed the question into a vector and predict visual vs. not. Instead of feeding question words (['what', 'is', 'the', 'man', 'doing', '?']), the input to our LSTM is embeddings of POS tags of the words (['pronoun', 'verb', 'determiner', 'noun', 'verb']). Embeddings of the POS tags are learned end-to-end. This captures the structure of image-grounded questions, rather than visual vs. non-visual topics. The latter are less likely to generalize across domains.

3.1.2 True- vs. False-Premise Detection

Our second task is to detect whether a question Q entails a false premise for an image I. We present two families of approaches to measuring this QI 'compatibility': (i) using uncertainty in VQA models, and (ii) using pre-trained captioning models.

Using VQA Uncertainty. Here we work with the hypothesis that if a VQA model is uncertain about the answer to a QI pair, the question may be irrelevant for the given image since the uncertainty may mean it has not seen similar QI pairs in the training data. We test two approaches:

[1] We use spaCy POS tagger [16].

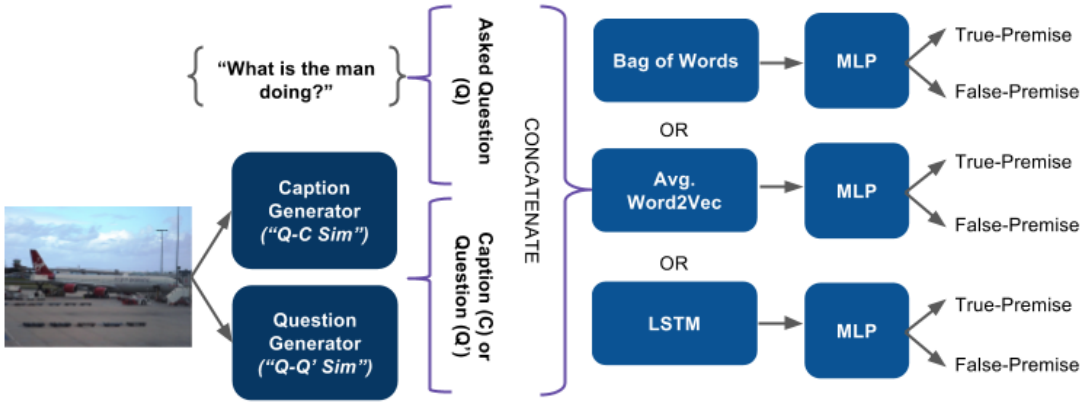

Figure 3.2: Summary of Caption Similarity Approaches. We try Bag of Words, Averaged Word2vec and an LSTM-based word2vec for concatenating question (Q) and generated caption/question(C or Q') features to learn their semantic similarity for identifying relevance.

1. **ENTROPY**. We compute the entropy of the softmax output from a state-of-the-art VQA model [3, 23] for a given QI pair and train a decision stump classifier.
2. **VQA-MLP**. We feed in the softmax output to a two-layer multilayer perceptron (MLP) with 30 hidden units in each layer, and train it as a binary classifier to predict whether a question has a true- or false-premise for the given image.

Using Pre-trained Captioning Models. Here we utilize (a) an image captioning model, and (b) an image question-generation model – to measure QI compatibility. Note that both these models generate natural language capturing the semantics of an image – one in the form of statement, the other in the form of a question. Our hypothesis is that a given question is relevant to the given image if it is similar to the language generated by these models for that image. Specifically:

1. Question-Caption Similarity (**Q-C SIM**). We use NeuralTalk2 [17] pre-trained on the MSCOCO dataset [21] (images and associated captions) to generate a caption C for the given image, and then compute a learned similarity between Q and C (details below).
2. Question-Question Similarity (**Q-Q' SIM**). We use NeuralTalk2 re-trained (from scratch) on the questions in the VQA dataset to generate a question Q' for the image. Then, we compute a learned similarity between Q and Q'.

We now describe our learned Q-C similarity function (the Q-Q' similarity is analogous). Our Q-C similarity model is a 2-channel LSTM+MLP (one channel for Q, another for C). Each channel sequentially reads word2vec embeddings of the corresponding language via an LSTM. The last hidden state vectors (40-dim) from the 2 LSTMs are concatenated and fed as inputs to the MLP, which outputs a 2-class (relevant vs. not) softmax. The entire model is learned end-to-end on the VTFQ dataset with a frozen captioning model. We also experimented with other representations (*e.g.* bag of words) for Q, Q', C, which is included in the supplement for completeness. A summary of our caption similarity approaches is visualized in Figure 3.2

Finally, we also compare our proposed models above to a simpler baseline (**Q-GEN SCORE**), where we compute the probability of the input question Q under the learned question-generation model. The intuition here is that since the question generation model has been trained only on relevant questions (from the VQA dataset), it will assign a high probability to Q if it is relevant.

3.1.3 Training Implementation Details

For training **BOW**, **AVG. W2V**, **LSTM W2V** and **VQA-MLP**, we use the Keras Deep learning Library [6] for Python. For pre-training the question and caption generation models from scratch, we use the Torch Deep Learning Library [**torch**]. We use *rmsprop* as the optimization algorithm for **LSTM W2V**, and *adadelta* for **BOW** and **AVG. W2V**. For all our models, we use a gaussian random weights initialization, a learning rate of 0.001, and no momentum.

3.2 Experiments

Table 3.1: Normalized accuracy results for visual vs. non-visual detection and true- vs. false-premise detection.

Visual vs. Non-Visual		**True- vs. False-Premise**				
RULE-BASED	**LSTM**	**ENTROPY**	**VQA-MLP**	**Q-GEN SCORE**	**Q-C SIM**	**Q-Q' SIM**
75.68	**92.27**	59.66	64.19	57.41	74.48	**74.58**

Visual vs. Non-Visual Detection. We use a random set of 100,000 questions from the VNQ dataset for training, and the remaining 31,464 for testing. The results are shown in Table 3.1. We see that **LSTM** performs 18.59% better than **RULE-BASED**.

True- vs. False-Premise Detection. We use a random set of 7,195 (66%) QI pairs from the VTFQ dataset to train and the remaining 3597 (33%) to test. Table 3.1 shows the results. While the VQA model uncertainty based approaches (**ENTROPY**, **VQA-MLP**) perform reasonably well (with the MLP helping over raw entropy), the learned similarity approaches perform much better (10% gain in normalized accuracy). High uncertainty of the model may suggest that a similar QI pair was not seen during training; however, that does not seem to translate to detecting irrelevance. The language generation models (**Q-C SIM**, **Q-Q' SIM**) seem to work significantly better at modeling the semantic interaction between the question and the image. The generative approach (**Q-GEN SCORE**) is outperformed by the discriminative approaches (**VQA-MLP**, **Q-C SIM**, **Q-Q' SIM**) that are trained explicitly for the task at hand.

Table 3.2: True- vs. false-premise question detection.

		True-Premise		False-Premise		
		Recall	Precision	Recall	Precision	**Norm Acc.**
ENTROPY		68.07	28.28	51.25	85.05	59.66
Q-GEN SCORE		64.73	25.23	50.09	84.51	57.41
VQA-MLP		57.38	36.13	71.01	85.62	64.19
Q-C SIM	**BOW**	70.48	40.19	69.91	90.46	70.19
	AVG. W2V	69.88	**48.81**	**78.35**	91.24	74.12
	LSTM W2V	72.37	46.08	76.60	91.55	74.48
Q-Q' SIM	**BOW**	68.05	44.00	75.79	90.28	71.92
	AVG. W2V	**74.62**	46.51	74.77	**92.27**	74.69
	LSTM W2V	74.25	44.78	74.90	91.93	**74.58**

For completeness, we also list the results of the three choices for feature extraction of the questions and captions that we explored:

1. **BOW**. We test a bag-of-words approach with a vocabulary size of 9,952 words to represent questions and captions, where we train an MLP to predict whether the question is relevant or not. The representation is built by setting a value of 1 in the features at the words that are present in either the question or the caption and a 2 when the word is present in both. This means each question-caption pair is represented by a 9,952-dim (vocab length) vector. The MLP used on top of **BOW** is a 5-layer MLP with 30, 20 and 10 hidden units respectively.
2. **AVG. W2V**. We extract word2vec [25] features for the question and captions words, compute the average of the features separately for the question and caption and then concatenate them. Similar to **BOW**, we train a 5-layer MLP with 200, 150 and 80 hidden units respectively.
3. **LSTM W2V**. These are the features we used in the main paper. The LSTM has 40 hidden units using a 4-layer MLP with 40 and 20 hidden units respectively.

Table 3.2 shows a comparison of the performance in the recall, precision, and normalized accuracy.

3.2.1 Human Qualitative Evaluation

We also perform human studies where we compare two agents: (1) **AGENT-BASELINE**– always answers every question. (2) **AGENT-OURS**– reasons about question relevance before responding. If question is classified as visual true-premise, **AGENT-OURS** answers the question using the same VQA model as **AGENT-BASELINE** (using [23]). Otherwise, it responds with a prompt indicating that the question does not seem meaningful for the image. A total of 120 questions (18.33% relevant, 81.67% irrelevant, mimicking the distribution of the VTFQ dataset) were used. Of the relevant questions, 54% were answered correctly by the VQA model. Human subjects on AMT were shown the response of both agents and asked to pick the agent that sounded more intelligent, more

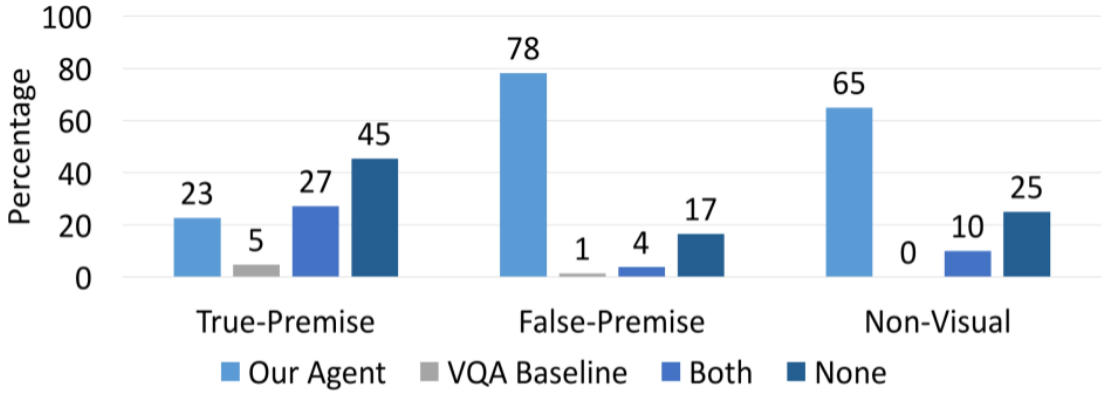

Figure 3.3: Human Study for perceived smartness for our agent vs. a baseline agent. Our agent reasons about relevance while answering questions, while baseline agent does not. Our agent is always perceived as smarter than the baseline agent. Both or none are chosen when humans thinks both or none of the two agents are smart.

reasonable, and more human-like after every observed QI pair. Each QI pair was assessed by 5 different subjects. Not all pairs were rated by the same 5 subjects. In total, 28 unique AMT workers participated in the study. AGENT-OURS was picked 65.8% of the time as the winner, AGENT-BASELINE was picked only 1.6% of the time, and both considered equally (un)reasonable in the remaining cases.

Figure 3.3 shows the percentage of times each robot gets picked by the workers for true-premise, false-premise, and non-visual questions.

Interestingly, humans often prefer **AGENT-OURS** over **AGENT-BASELINE** even when *both models are wrong* – **AGENT-BASELINE** answers the question incorrectly and **AGENT-OURS** incorrectly predicts that the question is irrelevant and refuses to answer a legitimate question. Users seem more tolerant to mistakes in relevance prediction than VQA.

3.2.2 Qualitative Results

Here we provide qualitative results for our visual vs. non-visual question detection experiment, and our true- vs. false-premise question detection experiment.

Visual vs. Non-visual detection

Here are some examples of non-visual questions correctly detected by **LSTM**:

- "Who is the president of the United States?"
- "If God exists, why is there so much evil in the world?"
- "What is the national anthem of Great Britain?"

- "Is soccer popular in the United States?"

Here are some example questions that **RULE-BASED** failed on, but that were correctly identified as non-visual by **LSTM**:

- "What color is Spock's blood?"
- "Who was the first person to fly across the channel?"

Here are some visual questions correctly classified by **LSTM**, but incorrectly classified by **RULE-BASED**:

- "Where is the body of water?"
- "What color are the glass items?"
- "What is there to sit on?"
- "How many pillows are pictured?"

True- vs False- Premise Detection

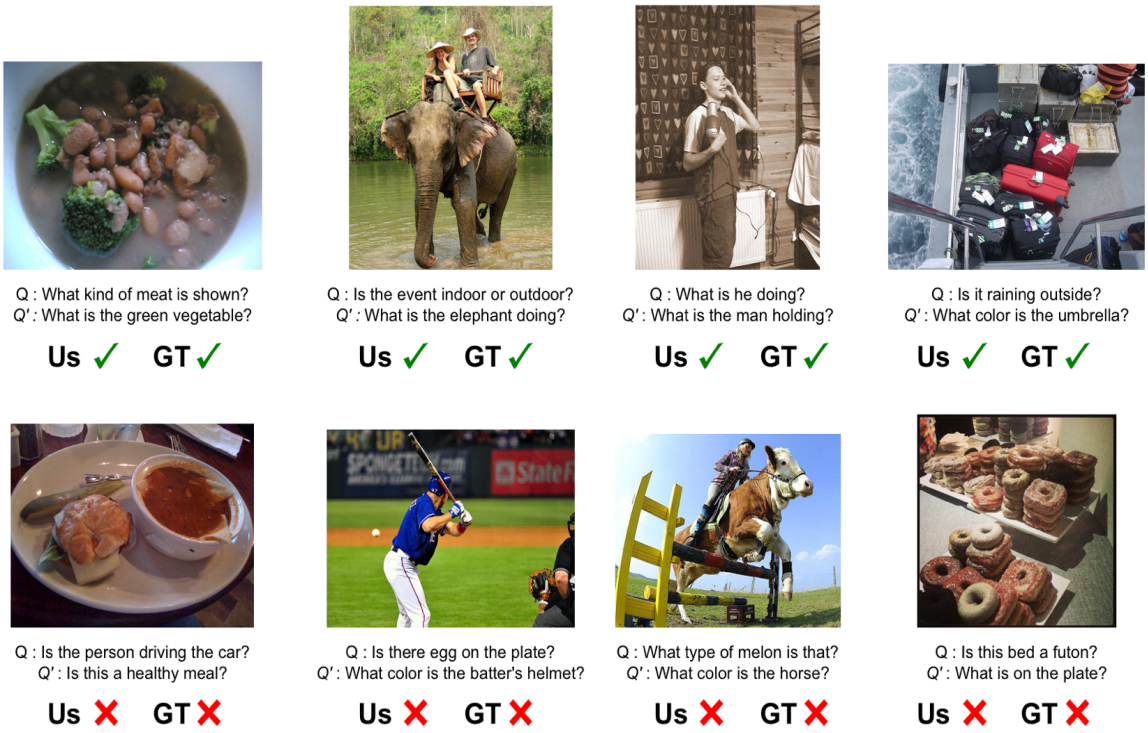

Figure 3.4: Success Cases: The first row illustrates examples that our model thought were True-Premise, and were also labeled so by humans. The second row shows success cases for False-Premise detection.

Figures 3.4 and 3.5 show success and failure cases for true- vs. false- premise question detection using **Q-Q' SIM**. Note that in the success cases, the contextual and semantic similarity was learned

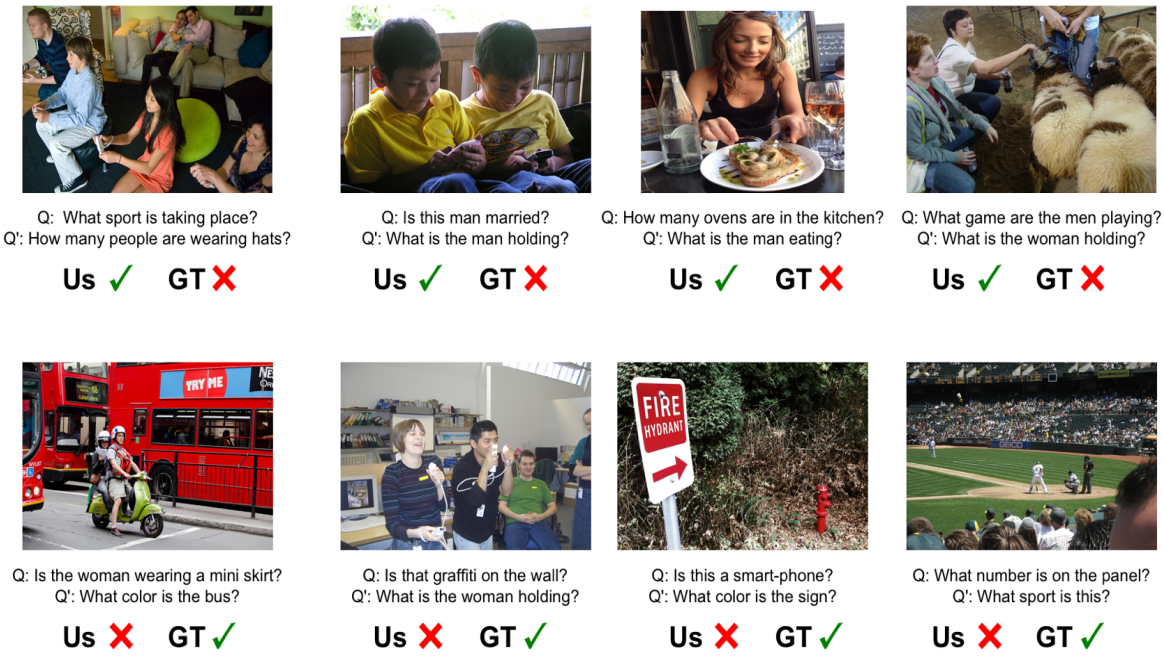

Figure 3.5: Failure Cases: The first row illustrates examples that our model thought was False-Premise, but were actually labeled as True-Premise by humans. Vice versa in the second row

even when the words in the question generated by the captioning model (Q') were different from the input question (Q).

3.3 Conclusion

We introduced the novel problem of identifying irrelevant (i.e., non-visual or visual false-premise) questions for VQA. Our proposed models significantly outperform strong baselines on both tasks. A VQA agent that utilizes our detector and refuses to answer certain questions significantly outperforms a baseline (that answers all questions) in human studies. Such an agent is is perceived as more intelligent, reasonable, and human-like. Our system can be further augmented to communicate to users what the assumed premise of the question is that is not satisfied by the image, e.g., respond to *"What is the woman wearing?"* for an image of a cat by saying *"There is no woman."*

Chapter 4

What are you thinking? The Visual 20 Questions Game

4.1 Motivation

Answering questions about images reasonably is definitely a desired skill a machine must have. However, as was illustrated in the motivating example in the introduction (Figure 1.1), a machine must also know how to ask questions to humans in order to extract the desired information it needs to make an informed decision when there are multiple ambiguous choices.

We explore this avenue by proposing a game called Visual 20 Questions (V20Q).

Visual 20 Questions (V20Q) is a game where an AI agent tries to imagine the kind of scene/image a human is imagining. It is similar to the famous 20q.net game, where an AI tries to guess the object/topic a human is imagining. Here, we try to guess the image/scene. A single image encapsulates numerous interactions between various objects/topics/concepts. It is certainly not possible to annotate every kind of image with its objects and their interactions. So, figuring out the desired image via questions and answers requires the understanding of image and language semantics without any human annotation on the images. Hence, this is definitely a more interesting task than figuring out a standalone object/topic.

Since the space of images is virtually infinite, in order to formalize the game, we frame it as an image retrieval task: a human secretly picks an image from a pool of fixed images, and an AI agent tries to guess which image the human picked. In the end, the human player decides whether the returned image is the exact image s/he picked, or whether it is a similar enough image or neither. Note that while we use a fixed pool of images for computation purposes, we do not use any ground truth annotations of the images. Our pool of images can be replaced with any set of random images.

4.2 Basic Pipelined Approach

4.2.1 Approach

As a strong baseline, we first implement a simple re-ranking based baseline for the game.

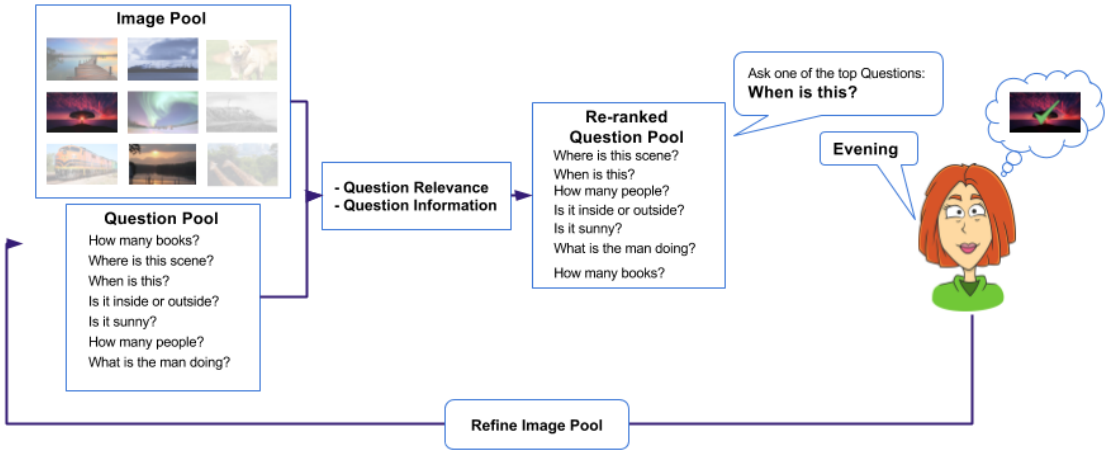

Figure 4.1: A basic approach where a predefined set of questions are re-ranked to choose the best question to ask at each time-step. The image pool is also refined in either a hard or soft manner based on the answer received at each time step.

A basic high-level sketch for image retrieval using the V20Q setup is as follows:

1. Start with X images and Y questions.
2. Get the top r most relevant questions. The top question is the most relevant question for all the X images. Relevance is determined using the best Q-Q' model in [29].
3. From the top r relevant questions, pick the most informative question.
4. Ask that question to the user.
5. Score/Prune the X images based on answer received from a human on the above question.
6. Repeat from step 2.

Figure 4.1 shows the basic pipeline that we implement and post publicly for the V20Q game.

We experiment with steps 2 and 3 for **question choice** and step 5 for **image pool pruning** to come up with the best possible pipeline that has the highest retrieval rate.

Question Choice:

The best question to ask at each step is a question that is not only relevant to the image the user is possibly imagining, but also has the ability to extract the most amount of information from the user in order to converge to the desired image with high confidence.

Global Question Relevance Let there be an image pool, I, consisting of N images. We define the global relevance of a question by a score ranging from 0 to 1 that denotes how relevant the question is to all N images in the pool- 0 indicating low relevance for all images, and 1 indicating high relevance for all images. Let $p(q, i)$ represent the relevance probability of the question, q for the image, i. The global relevance of q is then simply defined as:

$$r = \frac{1}{N} \sum_{i \in I} p(q, i) \tag{4.1}$$

We use the best performing model (Avg.W2V) in [29] for computing $p(q, i)$.

The above approach works well for a hard set of images. If we are dealing with a set of softly scored images, we have two ways by which we can compute the specificity.

- Compute a weighted average instead of a simple average based on the scores of the image pool. This way, the relevance of "important" images matter more than unimportant images.
- Simply use Equation 4.1 on the top k percentile of images.

Question Information: There can be three ways of calculating the information gain of a question for a pool of images.

- Hard Approach: For each of the top K relevant questions, compute the entropy of the image pool lengths for the answer choices for that question. If N denotes the total number of images, and n_{ia} denotes the number of images that have the answer a for question, q, the entropy of the q is given by:

$$H_q = -\sum_a \frac{n_a}{N} log \frac{n_a}{N} \tag{4.2}$$

  We pick the question with the highest entropy

- Soft Approach: Let IG_q denote the information gain for a question q. $H()$ denotes the entropy function. S_i denotes the scores of the image pool at iteration i. Let S_{i+a} denote the scores of the image pool if answer a is chosen after the i^{th} iteration.

$$IG_q = \frac{1}{n_a} \sum_a H(S_i) - H(S_{i+a}) \tag{4.3}$$

n_a represents the number of answer choices, which in our case is 1000.

We use the VQA model in [24] for the answer predictions for both the hard and soft approaches.

For an intuitive explanation of Equation 4.3, we are basically trying to pick a question that reduces the entropy of the scores over the image pool the most on average over all possible answer choices. The lower the entropy of the scores over the image, the more confident we become for a few specific images to be the result.

The $H(S_i)$ is the current entropy of the scores over the image. The $H(S_{i+a})$ are the entropies of the image score if question q is asked and answer a is chosen by the user. The question that has low entropy scores for all the n_a answer choices will have the second term as low, and hence the IG for that question q will be high. Hence, it means that this question q has the potential to raise the confidence over a few specific images the most.

- Semi-Soft Information Gain:

  The soft approach relies on finding a question that reduces the entropy of scores on the image pool. Often so, such a question does not exist because a question that might reduce the entropy of scores over the images we care about increase the entropy over other images. Hence, such a question never gets chosen, and the soft approach resorts to asking repetitive questions because asking the same question again at least keep the entropy the same. However, if we hard prune some of the images we surely do not care about, it is more likely to find a question that reduces the entropy of the scores over the images we care about. Hence, in this approach, we prune the image pool to the top kth percentile of image scores at each time step, and the information gain is computed in a soft manner over these top k percentile images.

Question Diversity:

We also observe that the soft and semi-soft approaches tend to ask similar questions over and over again. To remedy this, we add in a diversity score while re-ranking the questions. The diversity score is simply the negative of the number of words in the present question that matches words of all previous questions.

Final Question choice:

Using the tools above to determine global relevance and information gain, we try the following approaches for coming up with the best question:

VQA Re-ranking: We take a random set of Q questions from the VQA dataset and re-rank them according to information gain and relevance. The calculation of information gain and relevance

can be done in a hard, soft or semi-soft manner.

Diverse Questions Re-ranking: We generate k diverse questions for all the images in the pool using the diverse captioning model in [34] and re-rank them according to information gain and relevance. Once again, this can be done in a hard, soft and semi-soft manner.

We find that VQA Re-ranking and Diverse Questions Re-ranking perform comparably for a certain fixed pool of images. Hence, we choose VQA Re-ranking for all our experiments with other parameters since the type of questions made by humans in the VQA dataset is more diverse and diverse generated questions.

Image Convergence:

Hard approach We just prune the image pool based on the answer choice provided by the user. The matching of the answer to the images is done using the top prediction of the VQA model [24] for the question asked.

Soft Approach The basic idea here is to score the images based on the answer provided by the user. The scoring is done using the $p(I|Q, A)$ value obtained using the VQA model by [24]. The scores are accumulated as the user answers more and more questions. There are various ways of accumulating the score - we tried simply adding, averaging and exponential averaging and found that exponential averaging tends to be the most stable. In simple adding and averaging, the later answers do not affect the image pool as much, and hence, its becomes harder to recover from a wrong answer in the beginning. Exponential average, on the other hand, weighs the earlier questions lesser than the current question, and hence, it is easier to recover.

$$S_{i+a} = \frac{S_i + S_a}{1.3} \tag{4.4}$$

$$S_i = S_{i+a} \tag{4.5}$$

Empirically, we found that a 1.3 exponential average tends to work the best.

The anecdotal results for retrieval based on plays by experts at this game using various methods of question selection and image pruning are noted in Table 4.1.

4.2.2 Experiments

In the sections below, we experiment with various parameters on the individual modules of the V20Q game. We then pick a reasonable choice of implementation of each of the modules, imple-

ment the entire pipeline, and run user experiments to see the efficacy of our algorithm in retrieving the desired image.

Global Question Relevance

To start with, we took a set of 1500 images from MSCOCO Validation dataset and a random subset of 5000 VQA Val questions. We removed the images from 1500 that were seen by the training of the relevance module used in this experiment.

We use Equation 4.1 to rank the questions in order of global relevance. We observe that it mostly picks out questions starting with "is this...", "is it...", "is there...", "what is..." and so on.

It scores "Is it indoors?", "Is it daytime?", "are there people in this picture?" , "are there trees?" higher (meaning these are more relevant) as opposed to "Is this an image of a pedestal sink?" or "Is this a good store to hide something in?" or "Is the truck a fire truck?" (meaning these are more specific questions)

To investigate further, we took 1500 COCO images that all have some object in them (like microwaves, so that we know that they are all likely to be kitchens or bedroom and such) and noted if the list of questions became a little more specific (but still relevant to all images in the pool).

Specifically, we tried: microwave, book, umbrella and kites.

- **Microwave:** Was this taken inside? Is this a home scene? Where was this picture taken? Is it daytime or nighttime? is there a tablecloth? Are there people in the photo?
- **Book:** Is it inside? are there people? is there a patio? where was this picture taken? when is this picture taken? is it evening? is there a tablecloth?
- **Kite:** Are there many people? is this a camping scene? is it daytime or nighttime? is the picture outside? is there sunshine? are there trees?
- **Umbrella:** where is this picture taken? Is the background of the picture blurry? Is the picture blurry? is the background blurry? is this a romantic scene? is it outside? is it daytime or nighttime?

A lot of the questions like "Is it daytime?", "are there people?", "where is this picture taken?", "it is inside?", "is it outside?" appear in all the cases because they are always widely relevant. The more indoor-oriented items like books, microwave have "is it inside?", "is there a tablecloth?", "is there a patio?", "is this a home scene?" ranked higher than the outdoor-oriented items like kite and umbrella. They have "is this a camping site?", "Are there many people?" ranked higher. The differences are very subtle, but it does seem like the relevance model rank questions that would make sense to a broad spectrum of images higher.

Table 4.1: Anecdotal Performances of various implementations of the V20Q game.

Ques. Sel	Img Pruning	No. Imgs	No. Ques	Success	Success Sim	Failure	Freq
HardIG + Rel	hard	1000	1000	19.23%	57.69%	23.07%	52
HardIG + Rel	hard	3000	500	13.79%	70.68%	15.51%	58
SoftIG + Rel + QuesDiv	soft	1000	1000	45.23%	38.09%	16.66%	42
SemiSoftIG + Rel + QuesDiv	soft	1000	1000	64.15%	20.75%	15.09%	53

V20Q Pipeline

We implement the V20Q pipeline as shown in Figure 4.1 in Python3. As evaluation criteria, we asked users to choose one of three choices at the end of the game.

1. SUCCESS : the exact chosen image was returned.
2. SUCCESS SIMILAR : The exact image was not returned, but a very similar image was returned.
3. FAILURE : Returned image was completely off.

The results of different versions of the game are summarized in Table 4.1

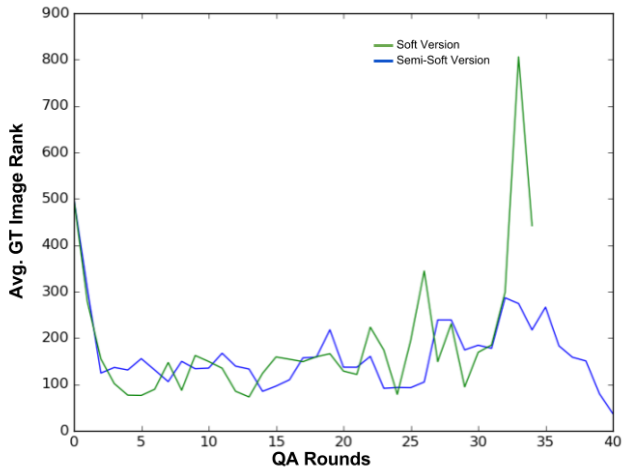

Figure 4.2: X axis is the number of questions asked and y-axis is the rank of the desired image across 42 runs for both, semi-soft (blue) and soft (green)

For the soft approaches where the image is scored based on the answers supplied by the user, we also plot the average rank of the desired image across rounds of dialog as shown in Figure 4.2. Each question asked has a certain amount of information associated with it. For example, a binary

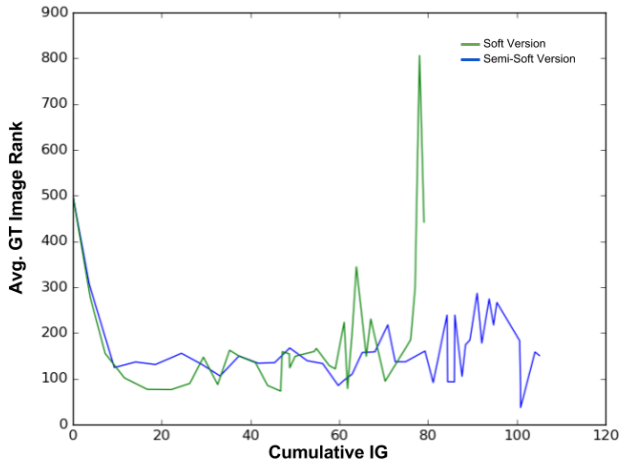

Figure 4.3: X axis is the cumulative information gain of the questions asked and y-axis is the rank of the desired image averaged across 42 runs for both, semi-soft (blue) and soft (green).

question has two answer choices and hence adds 1 bit of information to the system. We also plot the rank of the desired image with the cumulative information bits received so far as shown in Figure 4.3. For a pool of say, 1024 images, ideally, the desired image should be found by asking questions worth of 10 bits of information. Note how the biggest dip occurs between 0 to 10 bits of information gain.

Hard Version In this setting, we return the pruned image pool once it is pruned to below 5 images. We experiment with pool sizes of:

- 1000 Images and 1000 Questions - We see that it often fails to return the desired image. We observe that most choices converge to one or two images right after the first question asked. This intensifies the effect of inconsistencies in the VQA model. If there a few beach images in the pool, there is a good chance none of them would be selected when a user answers beach to "where is this?" simply because the answer "beach" did not feature as the top answer for those few images. So, we test to see whether an increased image pool improves the result.

- 3000 Images and 500 Questions- We observe that as the image pool increases and the question pool decreases, it gets harder to reach the desired image as the there is not enough freedom in the questions to retrieve the desired information needed to discriminate the desired image from the others. However, we do note that we are better at returning a similar image here. This is because, with a larger image pool, inconsistencies in the VQA system get ironed out as it is more likely that some similar images would still survive in the pool (statistical law of large numbers). For example, if there are a lot of beach images in the pool, there is a higher chance of some them being included as it is more likely at least some of them would have "beach" as the top answer.

To make things run fast enough in real time, we pre-compute the answer probabilities for all pos-

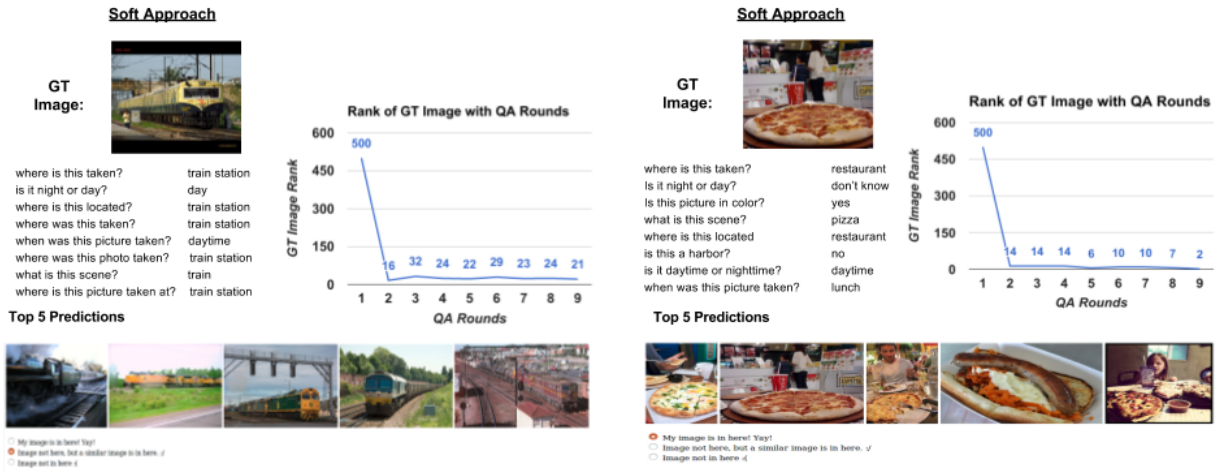

Figure 4.4: Example runs of the Soft Version of the V20Q game. Note how the soft version asks repetitive questions in-spite of enforcing question diversity.

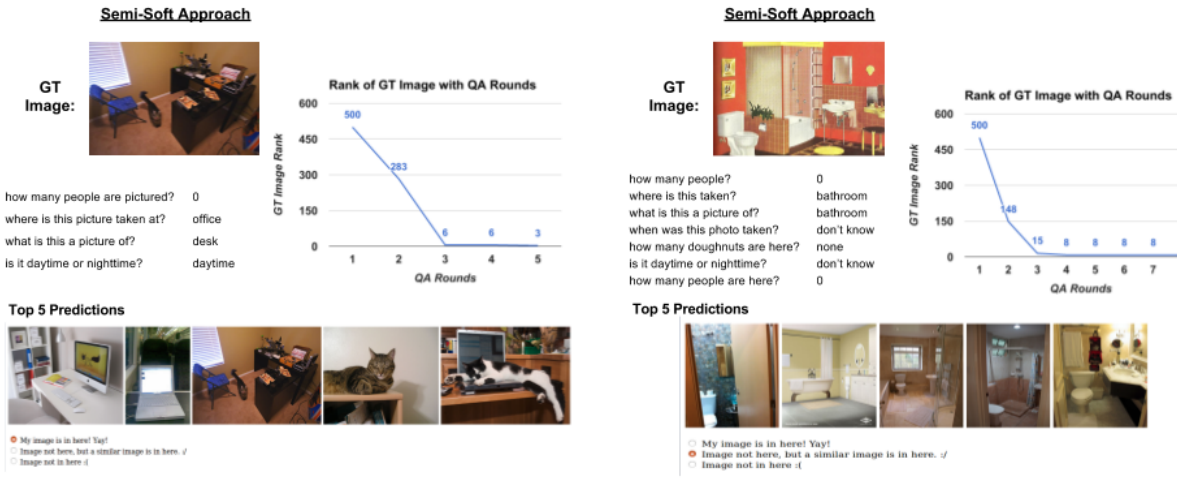

Figure 4.5: Example runs of the Semi-Soft version of the V20Q game.

sible combinations of QI pairs in the pool. We release our game on a public domain to collect data for analyzing the efficacy of the game.

Some of the qualitative results from a few example run of the game are shown in Figure 4.4 and Figure 4.5.

Soft Version The problem with the hard version is that it is impossible to recover a pruned image. So, if the desired image was thrown off due to the slight inconsistency of the VQA model/user answer, it is impossible to get it back. Using the soft version, all images are kept in the pool and are just scores, and hence, we do see a rise in the retrieval rate of the desired image.

Semi-soft Version In this setting, the image pool is pruned to the top 25 percentile just for purposes of question selection. All images are kept in the pool for consideration. This tends to work the best in terms of retrieval rate of the desired image.

We also include some qualitative runs of the soft and semi-soft runs of the game in Figures 4.4 and 4.5 respectively.

4.3 Learned Approach

4.3.1 Approach

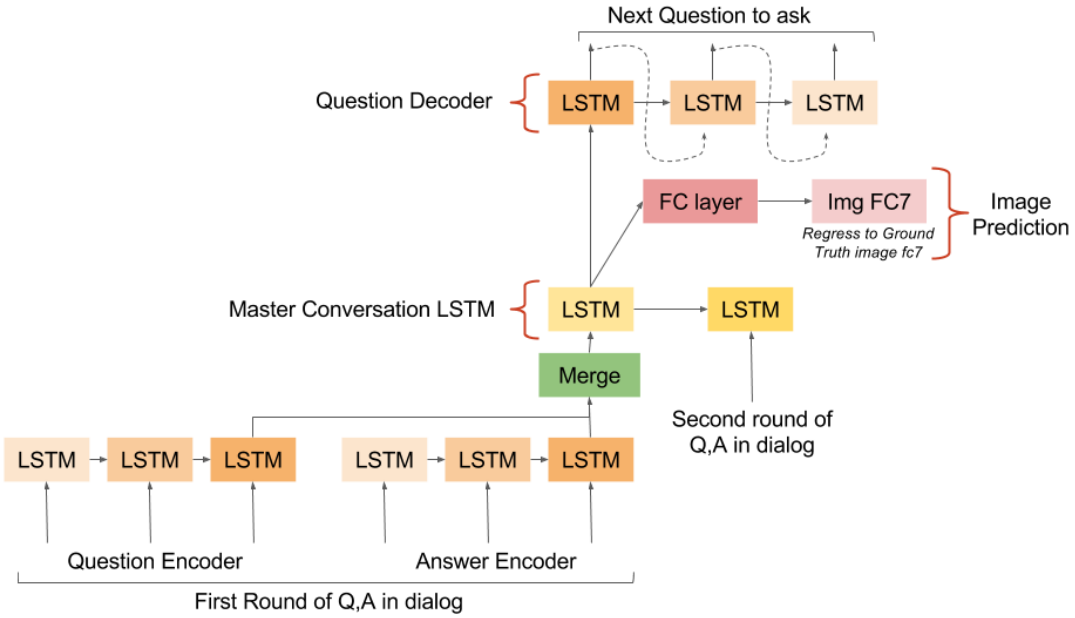

Figure 4.6: The model for an end-to-end learned approach. the model takes in the current question and answer received and predicts the next question to ask and the current belief of the image at each time step.

Needless to say, the pipelined approach is not scalable. If the whole pipeline is to be run real-time, it has a time complexity of $O(n_q n_i)$ where n_q is the number of questions in the pool and n_i is the number of images in the pool. Otherwise, it requires pre-computation of the answer probabilities of all possible image question pairs in the pool. This rises to intractable numbers even with a small increase in the number of images and questions and hence, is either slow, or requires a gigantic amount of hard-drive space.

Furthermore, re-ranking questions do not give us much flexibility in the type of questions we can ask. We are bound to the questions in the VQA dataset, and cannot generate discriminative enough questions that help to distinguish the target image from the other images.

Hence, we explore a combination of supervised and reinforcement learning to train a conversational agent to ask informative, yet relevant questions that most effectively solves the task. Instead of re-ranking a pool of fixed images, we train the model to predict the VGG16 [32] FC7 features of the Ground Truth (GT) image. Now, the pool of images can basically be replaced by an arbitrarily large pool, and the top K closest images according to VGG16 [32] FC7 distances will be the predicted images.

The model is shown in Figure 4.6 takes as input the currently asked question and the answer to that question and predicts the next question to be asked along with its image belief based on the questions and answers it has seen so far in the dialog. The idea is that the model will learn to ask the next best question given all previous questions and answers in dialog in order to predict the desired image accurately. The image predictions should also get closer to the desired image with progressing rounds of dialog.

Dataset The Visual Dialog Dataset [8] contains 10 rounds of dialog of two people talking about an image. Person A only sees a caption of the image and person B sees the full image. Person A asks questions sequentially to person B for extracting information about the image. This is slightly different from our task since person A sees the caption and hence, does not have to ask widely relevant questions at first since he/she already has some idea about the image. However, we can leverage this dataset to pre-train our model.

Supervised Learning (SL) for pre-training At each time step, we use question and answer pairs from the Visual Dialog dataset to train the model to predict the next question in dialog and the ground truth image.

We see that after about a day of training, the model learns to ask sensible questions, but does not quite learn how to strategically choose questions that lower the rank after the first question asked while playing with a VQA model.

Reinforcement Learning (RL)

Training using reinforcement learning requires playing the game live. Since playing with a human numerous times is time-consuming and expensive, we first play with a frozen VQA model, which can be thought of as a noisy human. We show that by just playing with a noisy VQA model, the RL trains the model reasonably well to play with humans as well. As a preliminary anecdotal result, experts at the game (internal creators) play multiple times with the model to evaluate it's efficacy. As further work, we plan to conduct a systematic human play evaluation as well as training.

With VQA We play 500 rounds of the game with VQA. Here is our outline for training using reinforcement learning:

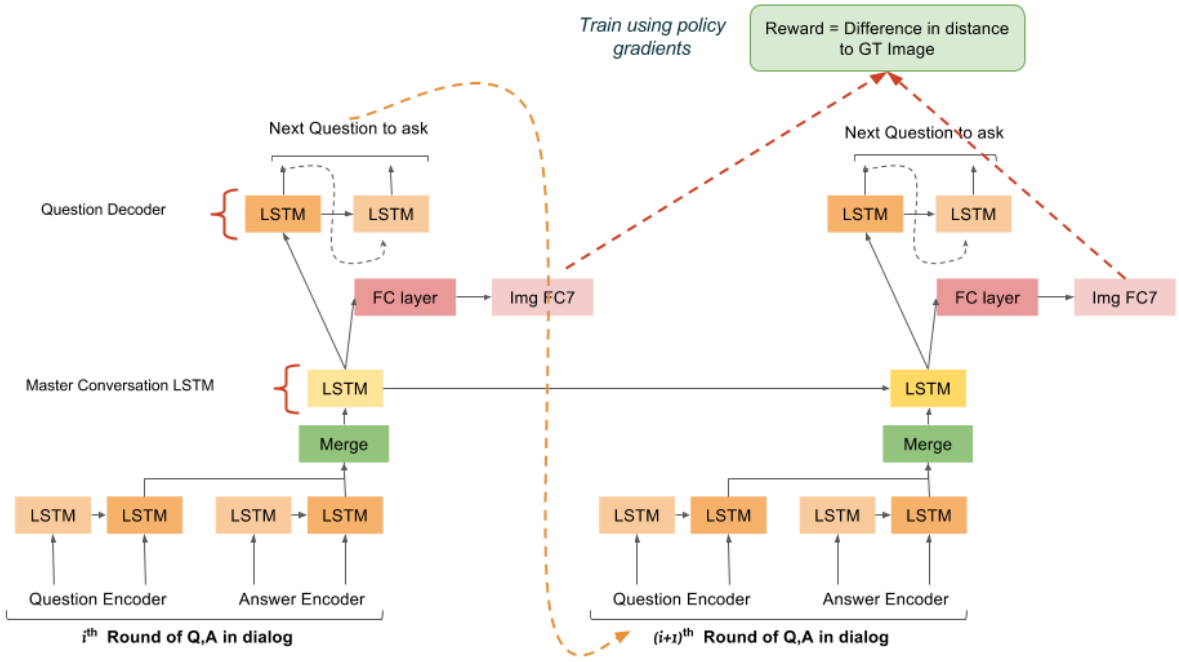

Figure 4.7: Reinforcement Learning Approach for the training our model. We either play with a VQA robot or human. The model plays 10 rounds of QA dialog and the reward at each time step is defined by whether the image distance got closer to the desired ground truth image or not. The model is trained using policy gradients based on the reward.

1. Choose a random ground truth image from the Visual Dialog Training Dataset.
2. Predict the first question to ask.
3. Get the VQA answer.
4. Pass it through the one-time step of the LSTM model and get the next predicted question and current image prediction. Keep track of ground truth (GT) image rank
5. Ask Question
6. Repeat from step 3 for 10 rounds.
7. The question and answer pairs are the policy that was chosen. We get the reward for this policy using the difference in GT image ranks at each time step. So, if the image rank goes down, it is a positive reward and vice versa. This is done at each time step.
8. Scale, the learning rate, using the reward and train the network using the policy chosen with the scaled learning rate.
9. So, if the reward was positive, the network is positively reinforced for the question it just asked and vice versa.

While evaluating, we evaluate based on 50 plays with the VQA model on validation pool of 40,000 images.

With Human When playing with a human, we just replace the VQA answer with a live human answer. Since this is expensive; we only play 20 rounds with a human for evaluating efficacy of the game. We show that our model is able to play reasonably well with a human even when trained with a noisy VQA model.

4.3.2 Experiments

Only Supervised Training We train the model on the Visual Dialog Training dataset on around 80,000 visual conversation examples. This is done to align the model utterances with natural English. At first the model word utterances are random, and hence playing with a human from a cold start would very frustrating and would also take a long time for the RL to learn good utterances just from reward feedback. Also, had we done Reinforcement Learning (RL) right from the beginning with a VQA, our LSTM model and the VQA would have cooperatively discovered their own language which might not be understandable with humans [18] [9]. Hence, we first give the model supervised labels to align itself to English utterances. We experiment with beam search vs. random sampling for generating the next question. In Beam search, top k words are kept at each time step. From the top k, we branch out to the next k words for each of the previous k. Now, we prune the k^2 to the top k beams based on their joint probability. Random Sampling simply samples (based on probability) a random word from the top k probable words. We note that random sampling tends to generate a lot more diverse questions than beam search. Beam search tends to generate very repetitive and safe questions.

Reinforcement Learning

With VQA Using reinforcement learning, we see that the model learns to adapt to the noisy VQA input better. We reinforce learn only 500 times with the HieCoAtten VQA model [23]. Fig 4.11 shows an example play with VQA.

The average ground truth image ranks over 50 plays with the VQA model for SL and RL are shown in Figure 4.8 (using beam search to generate questions) and 4.9 (using random sampling to generate questions). Note that the number of game plays is very small, and hence, these trends are not statistically conclusive. However, based on this preliminary evaluation, RL shows promise in making the model learn reasonably relevant and informative questions to ask.

With Human Since playing numerous times with a human enough to be able to significantly train a deep network is expensive, we train with VQA plays (which can be thought of as a noisy human). RL training with VQA also works reasonably while playing with a human. Figure 4.12 an

example run of the model playing with a human. Figure 4.10 shows the ranks when a human plays with a VQA-RL trained model. Almost perfect answers by humans lead to better performance. However, the model is adapted to VQA's noisy answers, and hence, does not perform drastically better with perfect human answers. Although the evaluations are on samples that are too small to make strong claims on, we do seem to see trends that show promise in the direction of training with humans on a large scale.

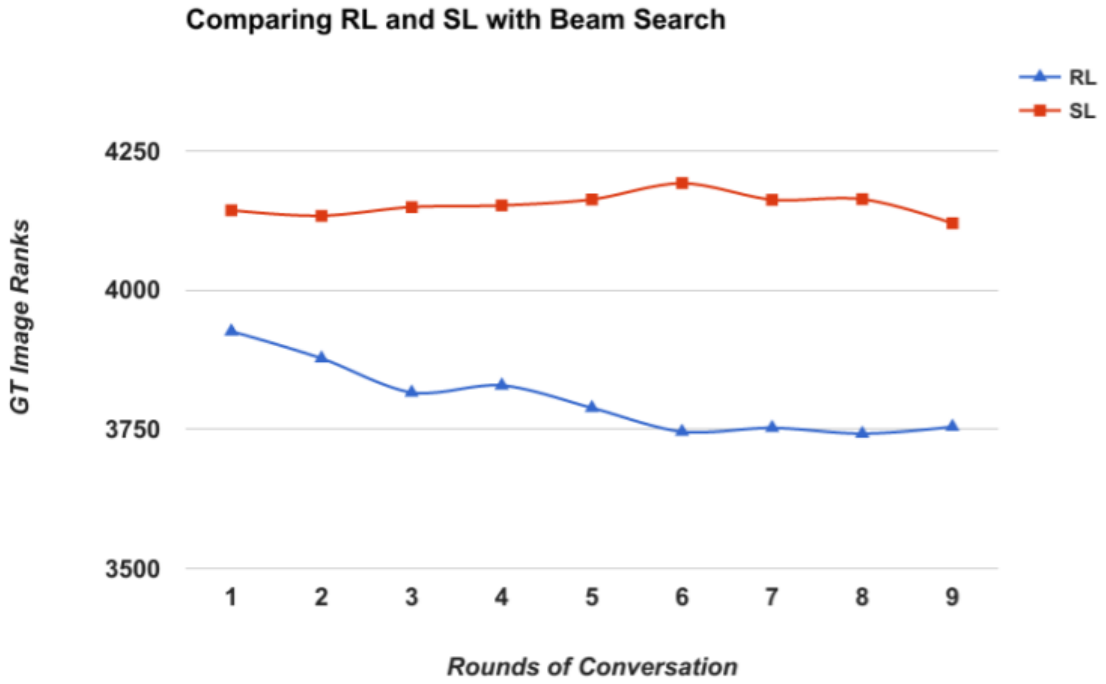

Figure 4.8: Ground Truth Image Ranks for 50 random plays with the VQA after trained using SL vs training using SL+RL. RL plays were done only for 500 times. Questions here were generated by beam search

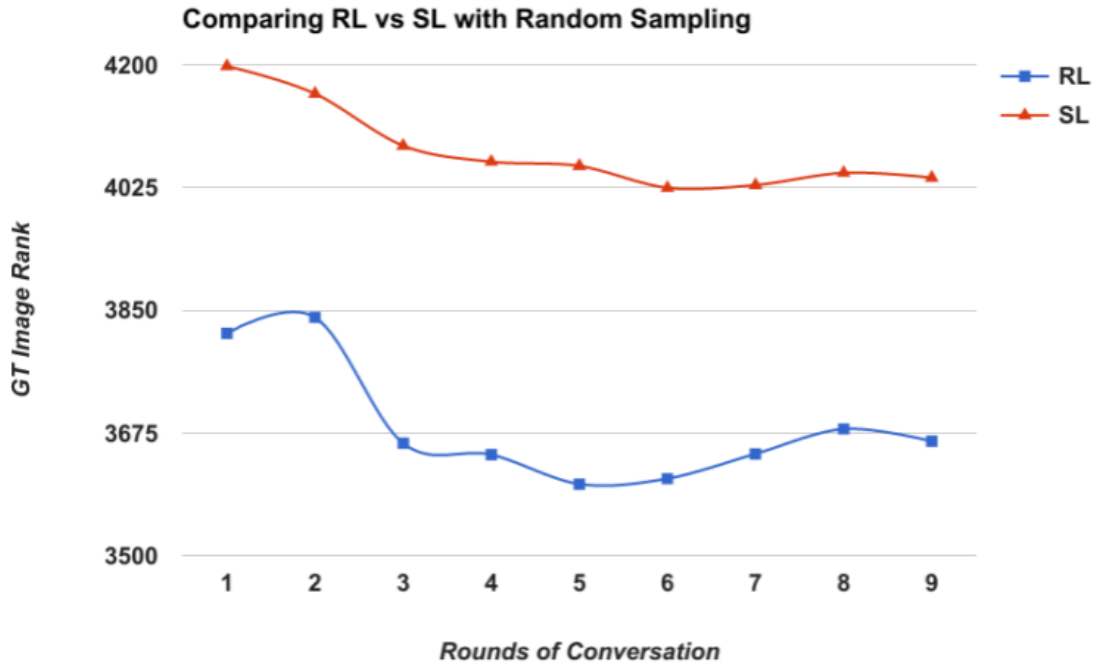

Figure 4.9: Ground Truth Image Ranks for 50 random plays with the VQA after trained using SL vs training using SL+RL. RL plays were done only for 500 times. Questions here generated by random sampling. Random sampling tends to generate more diverse questions and hence, performs better than beam search.

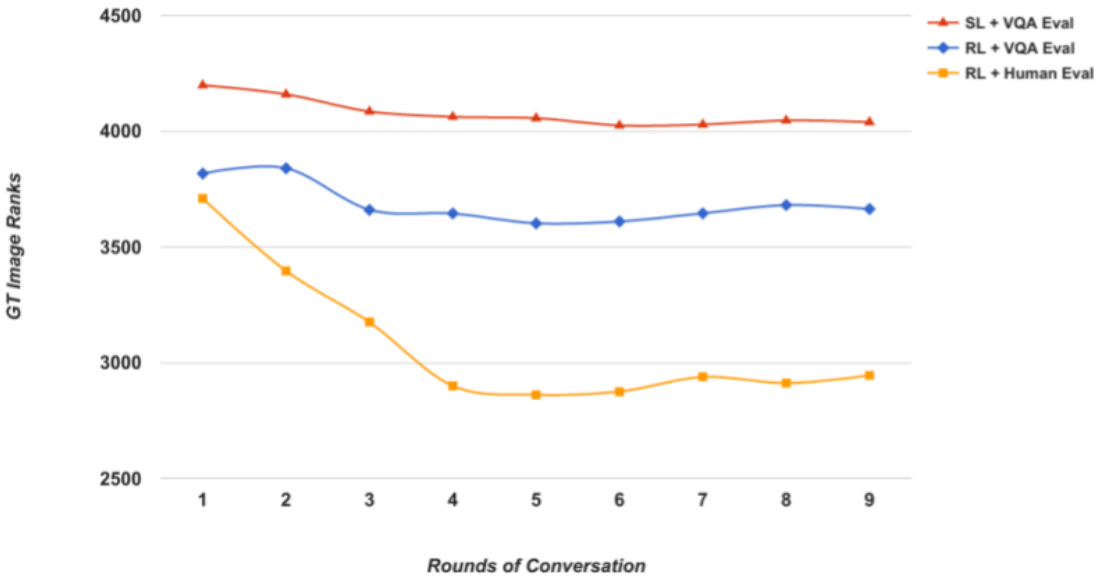

Figure 4.10: Anecdotal comparison of performance of a human playing with model trained using RL with VQA. The blue and red lines are the same as Figure 4.9. Note: Human plays were averaged only over 20 runs.

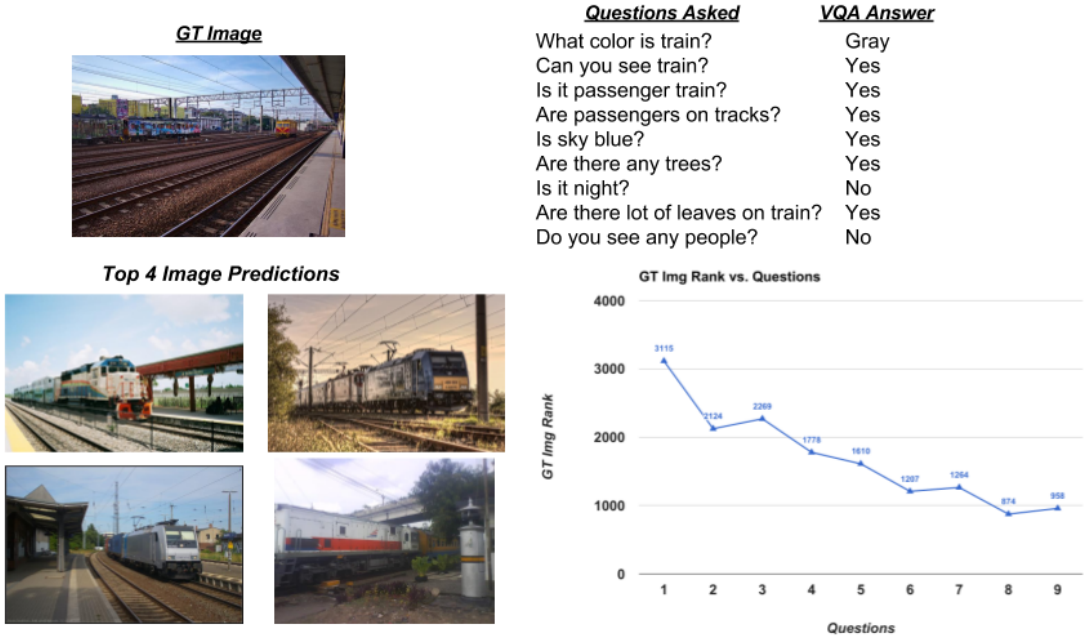

Figure 4.11: Example testing with VQA after training with Reinforcement Learning with VQA

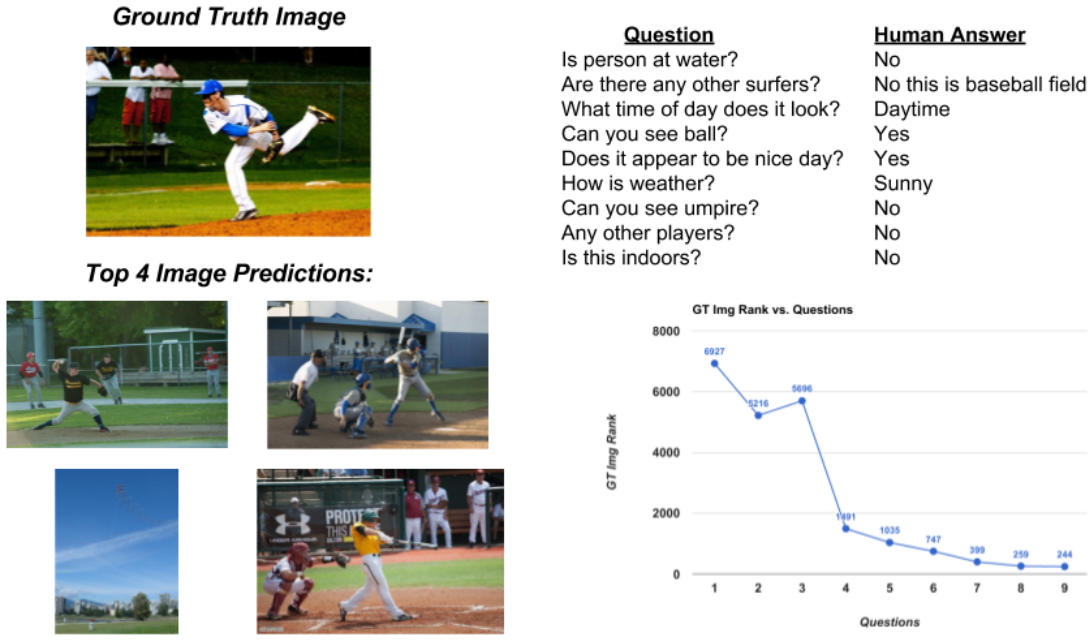

Figure 4.12: Example testing with human after training with Reinforcement Learning with VQA

Chapter 5

Future Work and Conclusion

5.1 Custom Dataset Collection

Our approach does not explicitly need a custom dataset because RL explores and learns the best possible questions to ask. However, if one were to collect a custom dataset for the "Visual 20 Questions" game, the performance could potentially be improved. This is because the model will then have more signals to learn a widely relevant yet highly informative first few question. Our qualitative preliminary results show that some of the first question asked are crucial to the success of identifying the desired image. The Visual Dialog dataset [8] has one person seeded with the caption. As a result, training on the Visual Dialog dataset doesn't teach the model to ask a widely relevant question at first.

This is an example of how one might collect a custom dataset for this task:

- Person A sees a pool of 20 or so images. One of them is the Ground Truth (GT) and the image pool has at least some X images very similar to GT image.
- Person B sees GT Image.
- Person A asks questions to Person B to figure out GT image.
- Person A gets paid only if s/he can figure out the correct image. This encourages them not to stop the conversation early and make a wild guess without being sure.

Using facilities in the lab, we tested out an example run of the interface as shown in Figure 5.1. Our code for the interface will be publicly released for the community to work on.

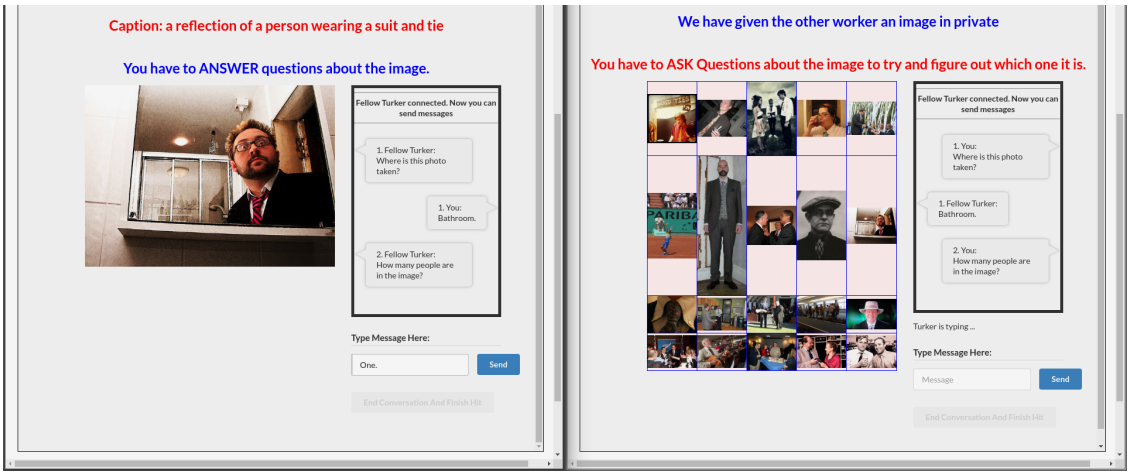

Figure 5.1: Amazon Mechanical Turk (AMT) Interface for collecting a Visual 20 Questions Dataset. Courtesy: Chris Dusold

5.2 Systematic Human Study

Conducting systematic human studies for the pipelined approach and reinforcement learned approaches to evaluate the efficacy of the games is definitely a interesting future direction to explore. Currently, the evaluation is done by playing the game within our internal development team, i.e. experts at the V20Q game and at AI. It will be interesting to evaluate the efficacy of the game when played by (a) non-experts at AI and (b) experts at AI, but non-experts of the game.

5.3 Human-Driven Reinforcement Learning

Finally, using humans for reinforcement learning on a large scale is definitely an interesting direction to pursue. This can be done in the following two ways. First, ask AMT workers to play the game numerous times. This will be a higher quality, but more expensive way to train the model. Second, one could also launch the demo on Reddit and let it train itself based on public plays. This will be cheap, but the signal it gets from public plays might be noisy if people do not take the game too seriously.

5.4 Conclusion

In this thesis, we took two humble, yet important steps towards seamless human-AI communication.

First, we made Visual Question Answering systems resilient to adversarial and confusing input. This is a crucial ability in order to be trustworthy while giving answers to a human. Second, we

attempted to equip machines to also ask questions to humans to extract the information the machine needs in order to make an informed decision. Being able to ask questions, answer questions, and ask follow-up questions effectively is a skill humans employ everyday for effective communication. Hence, an AI machine having this skill will definitely facilitate human-AI partnership. Although the visual 20 questions game is evaluated on a small sample in this thesis, we do seem to see certain trends that show promise in pursuing human-based reinforcement training.

Human-AI collaboration is essential in the days to come. Whether it is collaborating with "Siri" or "Alexa" for completing day-to-day tasks, or collaborating with AI teams that require the joint expertise of humans and AI, seamless communication is crucial for human-AI teamwork to be effective.

Will machine conversation ever be as sentient as human conversations are? Will machines ever be able to connect deeply with humans like humans are able to with one another? These are some of the questions I seek to answer in the future.

Exciting times lie ahead!

Chapter 6

References

[1] Harsh Agrawal, Clint Solomon Mathialagan, Yash Goyal, Neelima Chavali, Prakriti Banik, Akrit Mohapatra, Ahmed Osman, and Dhruv Batra. "CloudCV: Large Scale Distributed Computer Vision as a Cloud Service". In: *CoRR* abs/1506.04130 (2015). URL: http://arxiv.org/abs/1506.04130.

[2] Jacob Andreas and Dan Klein. "Reasoning About Pragmatics with Neural Listeners and Speakers". In: *Proceedings of EMNLP* abs/1604.00562 (2016). URL: http://arxiv.org/abs/1604.00562.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. "VQA: Visual question answering". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2425–2433.

[4] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. "Visual Recognition with Humans in the Loop". In: *Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*. Ed. by Kostas Daniilidis, Petros Maragos, and Nikos Paragios. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 438–451. ISBN: 978-3-642-15561-1. DOI: 10.1007/978-3-642-15561-1_32. URL: http://dx.doi.org/10.1007/978-3-642-15561-1_32.

[5] Long Chen, Dell Zhang, and Levene Mark. "Understanding User Intent in Community Question Answering". In: *Proceedings of the 21st International Conference on World Wide Web*. WWW '12 Companion. Lyon, France: ACM, 2012, pp. 823–828. ISBN: 978-1-4503-1230-1. DOI: 10.1145/2187980.2188206. URL: http://doi.acm.org/10.1145/2187980.2188206.

[6] François Chollet. *Keras*. https://github.com/fchollet/keras. 2015.

[7] Stephen Choularton. "Early Stage Detection of Speech Recognition Errors". PhD thesis. Macquarie University, 2009.

[8] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. "Visual Dialog". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[9] Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. "Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning". In: *arXiv preprint arXiv:1703.06585* (2017).

[10] Jia Deng, Jonathan Krause, Alexander C Berg, and Li Fei-Fei. "Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 3450–3457.

[11] Pandu R Devarakota, Bruno Mirbach, and Björn Ottersten. "Confidence estimation in classification decision: a method for detecting unseen patterns". In: *Proceedings of the sixth international conference on advance topics in pattern recognition (ICAPR), Kolkata, India*. Citeseer. 2007.

[12] Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alexander C. Berg, and Tamara L. Berg. "Detecting Visual Text". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL HLT '12. Montreal, Canada: Association for Computational Linguistics, 2012, pp. 762–772. ISBN: 978-1-937284-20-6. URL: `http://dl.acm.org/citation.cfm?id=2382029.2382153`.

[13] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. "Long-term Recurrent Convolutional Networks for Visual Recognition and Description". In: *CoRR* abs/1411.4389 (2014). URL: `http://arxiv.org/abs/1411.4389`.

[14] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. "From Captions to Visual Concepts and Back". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.

[15] Yansong Feng and Mirella Lapata. "Automatic caption generation for news images". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35.4 (2013), pp. 797–812.

[16] Matthew Honnibal and Mark Johnson. "An Improved Non-monotonic Transition System for Dependency Parsing". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language*. 2015.

[17] Andrej Karpathy and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3128–3137.

[18] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. "Multi-Agent Cooperation and the Emergence of (Natural) Language". In: *CoRR* abs/1612.07182 (2016). URL: http://arxiv.org/abs/1612.07182.

[19] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. "Deep Reinforcement Learning for Dialogue Generation". In: *CoRR* abs/1606.01541 (2016). URL: http://arxiv.org/abs/1606.01541.

[20] Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. "Visual semantic search: Retrieving videos via complex textual queries". In: *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE. 2014, pp. 2657–2664.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft COCO: Common objects in context". In: *Computer Vision–ECCV 2014*. Springer, 2014, pp. 740–755.

[22] Dong Liu, Xian-Sheng Hua, Meng Wang, and HongJiang Zhang. "Boost search relevance for tag-based social image retrieval". In: *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE. 2009, pp. 1636–1639.

[23] Jiasen Lu, Xiao Lin, Dhruv Batra, and Devi Parikh. *Deeper LSTM and normalized CNN Visual Question Answering model*. https://github.com/VT-vision-lab/VQA_LSTM_CNN. 2015.

[24] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. "Hierarchical Question-Image Co-Attention for Visual Question Answering". In: *CoRR* abs/1606.00061 (2016). URL: http://arxiv.org/abs/1606.00061.

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[26] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. "Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation". In: *CoRR* abs/1701.08251 (2017). URL: http://arxiv.org/abs/1701.08251.

[27] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. "Generating Natural Questions About an Image". In: (Aug. 2016), pp. 1802–1813. URL: http://www.aclweb.org/anthology/P16-1170.

[28] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. "Im2Text: Describing Images Using 1 Million Captioned Photographs". In: *Advances in Neural Information Processing Systems 24*. Ed. by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger. Curran Associates, Inc., 2011, pp. 1143–1151. URL: http://papers.nips.cc/paper/4470-im2text-describing-images-using-1-million-captioned-photographs.pdf.

[29] Arijit Ray, Gordon Christie, Mohit Bansal, Dhruv Batra, and Devi Parikh. "Question Relevance in VQA: Identifying Non-Visual And False-Premise Questions". In: *Proceedings of EMNLP*. 2016.

[30] Mengye Ren, Ryan Kiros, and Richard Zemel. "Exploring models and data for image question answering". In: *Advances in Neural Information Processing Systems*. 2015, pp. 2935–2943.

[31] Arup Sarma and David D Palmer. "Context-based speech recognition error detection and correction". In: *Proceedings of HLT-NAACL 2004: Short Papers*. Association for Computational Linguistics. 2004, pp. 85–88.

[32] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *CoRR* abs/1409.1556 (2014).

[33] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. "Context-aware Captions from Context-agnostic Supervision". In: *CoRR* abs/1701.02870 (2017). URL: http://arxiv.org/abs/1701.02870.

[34] Ashwin K. Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. "Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models". In: *Proceedings of NIPS* (2016). URL: http://arxiv.org/abs/1610.02424.

[35] Kimberly Voll, Stella Atkins, and Bruce Forster. "Improving the utility of speech recognition through error detection". In: *Journal of digital imaging* 21.4 (2008), pp. 371–377.

[36] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention". In: *arXiv preprint arXiv:1502.03044* (2015).

[37] Peng Zhang, Jiuling Wang, Ali Farhadi, Martial Hebert, and Devi Parikh. "Predicting failures of vision systems". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 3566–3573.

[38] Tongmu Zhao, Akemi Hoshino, Masayuki Suzuki, Nobuaki Minematsu, and Keikichi Hirose. "Automatic Chinese pronunciation error detection using SVM trained with structural features". In: *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE. 2012, pp. 473–478.