Min-Max Inverse Reinforcement Learning for learning bi-modal dialogue policies.

Gandharv Patil

Department of Electrical & Computer Engineering
McGill University
Montréal, Québec, Canada

April 2020

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Engineering.

2020/04/27

Abstract

In this thesis, we explore the use of Inverse Reinforcement Learning (IRL) in image-based dialogue systems. Visually-grounded dialogue systems are still in their infancy. Their success hinges on an efficient feature extraction method and the training mechanism used for learning the language model. At present, dialogue systems trained using maximum likelihood estimation tend to produce deflective or uninformative responses. We propose to tackle this issue by formulating the dialogue generation problem in the IRL setup. To that end, we adapt the IRL framework for processing complex high-dimensional inputs by leveraging a connection between deep learning, min-max optimisation and energy-based models. We also propose a neural network architecture for generating state representation by combining signals from text and image modalities. Finally, we show that subject to the availability "enough" training data; our proposed method can mitigate the issue of bland or deflective responses.

Résumé Scientifique

Dans cette thèse, nous explorons l'utilisation de l'apprentissage par reinforcement inverse (IRL) dans les systèmes de dialogue basés sur les images. Les systèmes de dialogue visuellement fondés n'en sont encore qu'à leurs débuts. Leur succès repose sur une méthode efficace d'extraction des caractéristiques et sur le mécanisme de formation utilisé pour apprendre le modèle linguistique. De plus, les systèmes de dialogue formés à l'aide de l'estimation du maximum de vraisemblance tendent à produire des réponses défléchies ou non informatives. Nous proposons d'aborder cette question en formulant le problème de la génération de dialogue dans la configuration IRL. Pour ce faire, nous adaptons le cadre IRL pour le traitement d'entrées complexes de grande dimension en tirant parti d'une connexion entre l'apprentissage approfondi, l'optimisation min-max et les modèles énergétiques. Nous proposons également une architecture de réseau neuronal pour générer une représentation d'état en combinant des signaux provenant de modalités texte et image. Enfin, nous montrons que, sous réserve de la disponibilité de données " suffisantes " sur la formation, la méthode que nous proposons peut atténuer le problème des réponses fades ou défectives.

Acknowledgements

I will be forever grateful to my supervisors Jeremy Cooperstock and Doina Precup, for guiding me in my research and encouraging me to work on ambitious ideas. I thank Jeremy for giving me the intellectual freedom of pursuing my research interests even when they drifted away from his overall research agendas. Doina's patience and insight have been invaluable in helping me discover my research path, and I thank her for that. I also want to thank Jayakumar Subramanian for helping me appreciate the intricacies of Reinforcement Learning. I have thoroughly enjoyed our countless research discussions which made me rediscover my love for mathematics.

A significant part of the work in the thesis was carried out at Microsoft Research with Layla El Asri. I want to thank Layla for her guidance on dialogue systems and for giving me the opportunity of working in an established research lab in the early stages of my research career.

All the members of Shared Reality Lab have been wonderful colleagues. I especially want to thank Antoine Weill-Duflos and Jeff Blum for putting up with my incessant requests for fixing the lab workstations. I also want to thank Pascal Fortin for helping me with the french translation of the thesis abstract. Special thanks are owed to Preeti Vyas for being a fun office mate.

Finally, I would like to thank my parents for their love, care and support at every step of my life.

To my parents.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Dialogue systems have a long legacy of research within the Linguistics, Artificial Intelligence (AI) and Human-Computer Interaction (HCI) communities. Traditionally, these systems have been implemented using a modularised pipeline, coarsely consisting of three key modules: Language Understanding (LU), Dialogue Manager (DM) and Natural Language Generation (NLG) [1]. Each of these modules specialises in operations necessary for coherent synthesis and generation of language. Lack of adaptability and narrow scope of applicability are significant disadvantages of such complex modular pipelines. Both these shortcomings can be attributed to two leading causes: (1) The credit assignment problem [1] where the user feedback cannot be propagated to upstream modules and (2) Lack of modularity: To ensure global optimisation, changing one module requires modifying all the other connected modules. This strong interdependence makes modular dialogue systems less adaptable.

In the past decade, researchers have started to explore fully data-driven end-to-end (e2e) approaches for uni/multi-modal conversational response generation. These models are trained entirely on data without resorting to any expert knowledge. Statistical language models with expressive parameterisations supplant the complex hand-engineered pipelines and broaden the application domain of these systems. The broader scope of applicability, however, comes at the cost of accuracy. The learning mechanism used for e2e models makes them susceptible to producing incoherent, bland and inappropriate responses. Another key shortcoming for contemporary e2e dialogue models is their lack of grounding in the real world. For instance, the question *How is the weather today?*, would be followed by responses like *rainy*, or *sunny*. As such, the algorithms do not have a basis for differentiating between several plausible responses. Data-driven dialogue systems have a remarkable ability to learn statistical correlations in the data, but they cannot connect the correlations to real-world facts [2]. In other words, the responses generated by e2e models are often pragmatically correct but the semantic content of the response can often be incorrect. Recent research in e2e dialogue has increasingly focused on designing grounded neural conversational models. In this thesis we explore the avenue of grounding a dialogue system in the visual domain. Naturally, exploration of this avenue calls for innovations in representation learning techniques which can efficiently combine different information modalities.

Alternatively, given the sequential nature of the dialogue generation process, it can be easily cast into the paradigm of reinforcement learning (RL). RL-based methods can be used to learn a dialogue generation policy that is optimal for a long-term user-designed objective. One significant hurdle in the applicability of RL for dialogue policy learning is the lack of a well-defined reward function for the optimisation process. Inverse Optimal Control (IOC) [3, 4] is one solution for policy learning which can bypass the need for a well defined reward function. In spite of being an attractive alternative to RL, IOC is restricted in its applicability due to its inability to handle complex, high-dimensional systems. Recent research has demonstrated that this challenge can be addressed by using expressive nonlinear function approximators like neural networks. Of specific interest to us is Ho *et al.*'s Generative Adversarial Imitation Learning (GAIL) [5], where the authors establish a resemblance between IOC and Generative Adversarial Networks (GAN) [6, 7], a neural network model designed to model multi-dimensional distributions.

In this thesis, we extend the GAIL framework for learning dialogue policy from human-human dialogues grounded on real-world images. The the key innovation in our method is to use the equivalence relation between GANs, Maximum Entropy RL (MaxEntRL) [8] and Energy-Based Models (EBMs) [9] established by Finn *et al.* [10]. As opposed to Maximum Likelihood Estimation we show that this alternative formulation can help prevent the model from generating bland and repetitive responses, albeit at a computational cost. The contributions of this work can thus be summarised as follows:

1. In the context of bi-modal dialogue systems, we develop a novel neural network architecture to capture sensory information from both image and text modalities.

2. We experimentally show that unifying EBMs and GANs can help in learning "better" dialogue policies which are less prone to producing deflective and bland responses.

Chapter 2

Literature Survey

The contributions of this thesis build on innovations in dialogue systems, image-based question answering and adversarial inverse reinforcement learning. This warrants that the literature related to all these innovations be discussed separately.

It is also important to note that the work presented in this thesis has been carried out in the second half of 2017 and predates works of both Li *et al*. [11] and Wu *et al*. [12] mentioned in the following section. We include a discussion of these works in section 2.3.

2.1 Dialogue generation

Study of text-only dialogue generation has a long history in the Natural Language Processing (NLP) literature. It has also led to the development of commercial systems like Microsoft's Xiaoice and Facebook's Its Alive. In the past two decades the research focus of the community has shifted from rule-based dialogue systems to end-to-end learnable conversational systems, which mainly use neural networks. This shift has been due to the breakthrough in deep learning [13] and the invention of techniques like neural probabilistic models [14] and sequence-to-sequence (seq2seq) [15] which can be scaled up to handle massive data sets. Consequently, some of the early works in this line of research leverage a large amount of conversational data available on the internet. Ritter *et al*. [16] propose the idea of modelling conversations from micro-blogging websites with generative models. They formulate the response generation problem in the machine translation paradigm where the post is translated into a response. It was observed that generating answers is a difficult task as a post can have many plausible responses. To overcome this hurdle Shang *et al*. [17] use the seq2seq [15] model for summarising the semantical content of a post by a vector representation and then using it to generate responses. Subsequent works by Vinyals *et al*. [18] and Serban *et al*. [19, 20] generalise these approaches to multi-turn dialogue.

RL [21] has also been used for training neural network-based dialogue systems. Li *et al*. [22] introduce a method which uses simulation-based training for learning dialogue policies which are optimal for handcrafted reward functions encouraging informativeness, ease of answering and coherence. Asghar *et al*. [23] and Su *et al*. [24] supplant the use of handcrafted reward functions by proposing a framework where seq2seq models trained using supervised learning are improved by using RL with human feedback.

2.2 Image-Based Dialogue

Contemporary research on combining images and natural language started in 2014 with the introduction of the MS COCO dataset [25] for image captioning. Followed by image captioning, the first image-based question-answer (QA) dataset DAQUAR [26] was introduced in 2014. Release of the DAQUAR was succeeded by five other QA datasets: COCO QA [27], FM-IQA [28], VQA [29], Visual7W [30] and Visual Genome [31]. On the most popular dataset VQA, algorithms now achieve 75% accuracy [32]. Although promising, these results are fraught with concerns over multitudes of biases in the data. The biases manifest both in the type of questions asked and the type of answers to the questions. One of the most detrimental biases induced during data collection is the *visual priming bias*: as human annotators can see the image when asking a question they specifically ask questions about objects present in the image. As a result 79% of the questions starting with *Is there* ... have the answer as *yes* [33]. This bias thus benefits language-only models inflating their performance on the task. Moreover, these datasets do not have features of dialogue, *i.e.*, they do not contain multiple QA rounds on a single image.

De Vries *et al*. [34] introduce “Guess What?!” to make up for the lack of a multi-turn QA dataset. This dataset contains multiple rounds of QA pairs obtained via a two-player object guessing game. In this game, one player is tasked with guessing a randomly selected object from the image by asking questions about it, and the other player answers these questions as Yes/No/NA to help the other player. Interaction is deemed successful if the questioner finds out the assigned object. Although exciting in regards to sustained communication and clear evaluation criteria, this dataset still requires the generation of single-word responses.

Mostafazadeh *et al*. [35] introduce the Image Grounded Conversation (IGC) dataset with a task that requires a model to produce sustained natural-sounding conversations about an image. However, this dataset focuses on the style, form and context of the language as opposed to visual reasoning.

Finally, Visual Dialogue or VisDial (v0.9, v1.0) [36] is the latest in succession of vision and language datasets. In contrast to the previous datasets, VisDial requires the machine to hold a meaningful dialogue in natural language about the visual content. The data collection setup ensures it steers clear of the visual priming bias as the questioner does not have access to the image. Therefore the questions rely solely on the information provided by the image caption and the answers. Additionally, Das *et al*. evaluate several baseline models on the dataset. The follow-up work by Lu *et al*. [37] demonstrates the benefits of combining classification-based and generation-based training methods as an alternative to the maximum likelihood-based training.

2.3 Adversarial Imitation Learning

The adversarial learning framework has been successfully applied in problems of learning continuous probability distributions, particularly in the computer vision domain for the generation of realistic images. Li *et al*. [22] introduce the adversarial learning setup for dialogue generation as an alternative to human evaluation by mimicking a human user’s behaviour. They propose to jointly train a seq2seq model to produce sequences of responses and a discriminator to identify if the responses are by humans or not. Yu *et al*. [38] propose to replace the standard optimisation methods like Adam [39] by the policy gradient method in RL to backpropagate the error information from the generator to the discriminator.

More central to the work in this thesis, the adversarial imitation learning setup was developed by Ho *et al*. [5] to improve the scalability and computational efficiency of inverse reinforcement learning. This framework reduces the problem of policy learning to that of matching probability distributions by systematically establishing an equivalence relationship between IOC, min-max optimisation and GANs. On similar lines, Finn *et al*. [10] establish a mathematical equivalence between GANs, IOC and Energy-Based Models (EBM) to improve the otherwise unstable training dynamics of GANs. Specifically, the authors propose to model the generator as a parameterised autoregressive model. Next, they evaluate the generator density function and use the density function in the GAN's training objective. In section 4.2 we systematically re-derive this relationship in a more direct way.

Similar to the work presented in this thesis, albeit for text-only dialogue, Li *et al*. [11] extend the framework proposed by Li *et al.* [22] to incorporate the benefits of EBMs and GAIL. The authors also propose architectural modifications in the generator and the discriminator to facilitate precise recovery of the reward function structure. Apart from the application to image-based dialogue, the key difference between their model and the one presented in this thesis is the reward propagation mechanism. The discriminator design by Li *et al*. [11] enables it to evaluate partially generated sequences; our method, on the other hand, assigns a reward to a complete response.

Wu *et al*. [12] adapt the adversarial learning setup by Li *et al*. [22] using the REINFORCE method [40] for image-based dialogue on the Visdial v0.9 dataset. They also propose a novel sequential co-attention encoder architecture which computes joint attention weights for the dialogue history, the image and the question.

Improving the response quality in both grounded and un-grounded dialogue systems is still an open problem. Several works in the literature try to overcome this problem by meticulously engineered algorithm design. Our work seeks to bypass this engineering effort by using an alternative training methodology of IOC. At the surface level, our work is similar to Wu *et al*. but; we propose an entirely different algorithmic framework and feature extraction pipeline. Concretely, in our case, the GAN optimises an energy-based

IOC objective as opposed to the REINFORCE objective. On the feature extraction front, we use feature-wise linear modulation as compared to the sequential co-attention.

Chapter 3

Background

This chapter intends to familiarise the reader with the background needed to understand the rest of the thesis.

3.1 Dialogue Systems

Dialogue systems use generative models for the problem of dialogue generation. Consider a dialogue as a sequence of K utterances $D = \{\mathcal{U}_1, \ldots, \mathcal{U}_k\}$ involving two interlocutors. Each utterance $\mathcal{U}_k$ contains a sequence of N_k tokens, *i.e.*, $\mathcal{U}_k = \{w_{k,1}, \ldots, w_{k,N_k}\}$ where, $w_{k,n} \in \mathbb{W}$ is a random variable representing a n^{th} token in the k^{th} utterance from a finite but large vocabulary $\mathbb{W}$. The tokens represent both speech acts like end of sequence (EOS) and words. The dialogue model is a generative model which represents parametric probability distribution over the set of all possible dialogues of arbitrary lengths. The probability of a dialogue D can be decomposed as:

$$P_\theta\{\mathcal{U}_1, \ldots, \mathcal{U}_k\} = \prod_{w=1}^{k} P_\theta(\mathcal{U}_k | \mathcal{U}_{<k}) \tag{3.1}$$

$$= \prod_{k=1}^{K} \prod_{n=1}^{N_k} P_\theta(w_{k,n} | w_{k,<n}, \mathcal{U}_{<k}) \tag{3.2}$$

where $\mathcal{U}_{<k} = \{\mathcal{U}_1, \ldots, \mathcal{U}_{k-1}\}$ and $w_{k,<n} = \{w_{k,1}, \ldots, w_{k,n-1}\}$.

Sampling from the model can be performed one word at a time form the conditional distribution $P_\theta(w_{k,n}|w_{k,<n},\mathcal{U}_{<k})$ conditioned on the previously sampled words.

Using the standard n-grams to compute the joint probabilities over the dialogues, *e.g.*, computing the probability tables for each token given the n preceding tokens, suffers from the curse of dimensionality and is intractable for any realistic vocabulary. To overcome this, Bengio *et al.* [14] propose a distributed vector representation of words called word embeddings, which parameterises $P_\theta(w_{k,n}|w_{k,<n},\mathcal{U}_{<k})$ as a smooth function using a neural network. The authors used a recurrent neural networks to model the sequence of tokens $\{w_1,\dots,w_N\}$ using the recurrence:

$$h_n = f(h_{n-1}, w_n) \tag{3.3}$$

where, $h_n \in \mathbb{R}^{d_h}$ is called the recurrent, or hidden state with a fixed dimension d_h. The language model encodes the historical context information in h which is used to predict the next token using the function:

$$P(w_{n+1} = v|w_{\leq n}) = \frac{\exp(g(h_n, v))}{\sum_{v'} \exp(g(h_n, v'))} \tag{3.4}$$

In a simple RNN based model the functions f and g are defined as:

$$f(h_{n-1}, w_n) = \tanh(\mathcal{H}h_{n-1} + I_{w_n}) \tag{3.5}$$

$$g(h_n, v) = (\mathcal{O})h_n \tag{3.6}$$

The matrix $I \in \mathbb{R}^{d_h \times \mathbb{W}}$ represents the input word embedding such that each column I_j corresponds to a token in the vocabulary. Given the large size of the vocabulary the matrix I is approximated using a low rank decomposition $I = I'' \cdot I'$ where $I'' \in \mathbb{R}^{d_h \times d_{I'}}$ and $I' \in \mathbb{R}^{d_{I'} \times |\mathbb{W}|}$ and $d_h > d_{I'}$. This embedding matrix I' is generally learnt separately from a larger text corpora. Similarly $\mathcal{O} \in \mathbb{R}^{d_h \times |\mathbb{W}|}$ represents the output word embedding. The matrix $\mathcal{O}$ contains projections of every possible next word into a distributed vector representation compared with the hidden state. $\mathcal{H}$ is the recurrent parameter linking h_{n-1} to h_n. All the parameters are learnt using maximum likelihood estimation

3.1.1 LSTM based Sequence to Sequence (seq2seq) Models

The traditional RNN based models have trouble learning long term dependencies [41]. Therefore, modern dialogue systems use long short term memory networks (LSTM) [42], or gated recurrent units (GRU) [41] based dialogue models.

Instead of using simple functions as eq. (3.5) an LSTM processes a sequence of inputs differently. At each time step, it associates an input gate $i_n \in \mathbb{R}^h$, a memory gate $f_n \in \mathbb{R}^h$ and an output $o_n \in \mathbb{R}^h$ gate to the input sequence $m_n \in \mathbb{R}^d$ and accordingly updates the cell state $c_n \in \mathbb{R}^h$ and the hidden state $h_n \in \mathbb{R}^h$.

$$f_n = \sigma_g\left(\theta_f \cdot m_n + \theta''_f \cdot h_{n-1} + b_f\right) \tag{3.7}$$

$$i_n = \sigma_g\left(\theta_i \cdot m_n + \theta''_i \cdot h_{n-1} + b_i\right) \tag{3.8}$$

$$o_n = \sigma_g\left(\theta_o \cdot m_n + \theta''_o \cdot h_{n-1} + b_o\right) \tag{3.9}$$

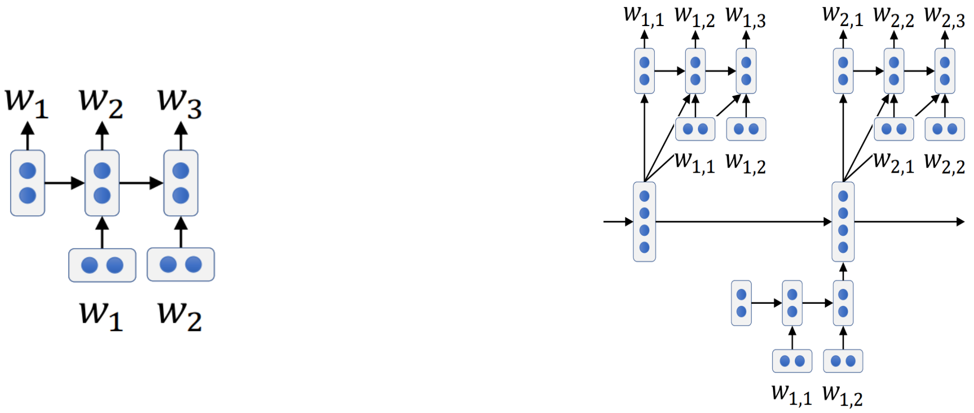
$$c'_n = \tanh\left(\theta_{c'} \cdot m_n + \theta''_{c'} \cdot h_{n-1} + b_{c'}\right) \tag{3.10}$$

$$c_n = f_n \odot c_{n-1} + i_n \odot c'_n \tag{3.11}$$

$$h_n = o_n \odot \tanh(c_n) \tag{3.12}$$

In Equations (3.7) to (3.11), each equation denotes specific operations performed by each gate. Also, $\theta_i, \theta_f, \theta_o, \theta_{c'} \in \mathbb{R}^{h\times d}$, $\theta''_i, \theta''_f, \theta''_o, \theta''_{c'} \in \mathbb{R}^{h\times h}$, $b_i, b_f, b_o, b_{c'} \in \mathbb{R}^h$, $\odot$ denotes the element-wise product and σ_g denotes the sigmoid activation function. In sequence to sequence generation tasks, each input sequence is paired with an output sequence. The output sequence is sampled from the probability distribution obtained by the same operation as in eq. (3.6). Commonly, the input and the output use different LSTMs with separate parameters to capture different compositional patterns. During decoding the algorithm terminates when an EOS token is predicted. At each time step, either a greedy approach or a beam search can be adopted for word prediction. Greedy search selects the token with the highest conditional probability, the embedding of which is then combined with the preceding output to predict the token at the next time step.

Dialogue histories can often be long, and there is sometimes a need to exploit a longer-term context. Hierarchical models (HRED) address this limitation by capturing more extended contexts. Conceptually, hierarchical models use two RNNs, one for encoding the word level hierarchy and another for encoding the utterance level hierarchy. Figure 3.1 shows the difference between standard RNN architectures and hierarchical architectures. A special feature of this architectural modification is that it results in a direct temporal dependency between the hidden state of the current dialogue with the hidden state of the previous dialogue, widening the temporal span of information flow.



(a) Recurrent architecture models which use single or multiple layers of RNNs, GRUs, LSTMs, etc.

(b) Two-level hierarchy representative of HRED. Note: To simplify the notation, the figure represents utterances of length 3.

Fig. 3.1 Simple RNN and HRED models.

3.2 Generative Adversarial Networks

Generative Adversarial Network (GAN) [6, 7] is a generative model which proposes simultaneous training of two models: the generator (G) and the discriminator (D). The goal of the discriminator is to classify the inputs presented to it as synthetically generated or belonging to the true data distribution $p_{\text{data}}(\cdot)$. The goal of the generator is to fit the true data distribution by "convincing" the discriminator that it is generating samples from $p_{\text{data}}(\cdot)$.

Formally, the generator takes noise as input and outputs a sample $i \sim p_{\text{gen}}$, while the discriminator takes sample i and outputs the probability of the sample being from $p_{\text{data}}(\cdot)$. As the discriminator simultaneously learns to classify real and generated samples its loss is the average log probability it assigns to the correct classification from a mixture of real and generated samples.

$$\mathcal{L}_{D_{\theta_D}} = \underset{i \sim p_{\text{data}}}{\boldsymbol{E}}[-\log(D_{\theta_D}(i))] + \underset{i \sim p_{\text{gen}}}{\boldsymbol{E}}[-\log(1 - D_{\theta_D}(i))] \tag{3.13}$$

In the simplest model, the generator's loss is defined in terms of the discriminator's competence.

$$\mathcal{L}_{G_{\theta_G}} = \underset{i \sim p_{\text{gen}}}{\boldsymbol{E}}[-\log(D_{\theta_D}(i))] + \underset{i \sim p_{\text{gen}}}{\boldsymbol{E}}[-\log(1 - D_{\theta_D}(i))] \tag{3.14}$$

3.3 Energy Based Models

Energy based models [9] associate an energy value $E_\theta(i)$ to a sample i; specifically it uses the Boltzmann's distribution:

$$P_\theta(i) = \frac{\exp(E_\theta(i))}{Z} \tag{3.15}$$

The energy function parameters θ are often chosen to maximise the likelihood of the data; the main challenge in this setup however is the evaluation of the partition function Z. In most high dimensional problems, Z results in an intractable integral. A common approach to estimating Z requires sampling from the Boltzmann distribution within the inner loop of learning.

Approximate sampling methods like Markov chain Monte Carlo (MCMC) methods can be used to sample from $p_\theta(\cdot)$. But these methods also run into problems when working with multimodal distributions. The use of MCMC methods often results in arbitrary long compute cycles to produce diverse samples. Approximate inference methods can also be used during training, but this runs the risk of some modes being assigned low energies if the approximation is unable discover them.

3.4 Inverse Reinforcement Learning

3.4.1 RL-problem

Consider a standard MDP model : $(S,\ A,\ R,\ P,\ \alpha)$, where an agent takes an action $a \in A$ on encountering a state $s \in S$ and transitions to the next state $s' \in S$ according to a transition probability kernel $P_a^{ss'}$ which gives the probability of making the transition $s \rightarrow s'$ on executing an action a. In our model we assume that all transitions are deterministic, *i.e.*, $\underset{s,s' \in S}{\forall} \underset{a \in A}{\exists} P(s'|s,a) = 1$

An episode refers to a sequence of transitions $s_1 \rightarrow a_1 \rightarrow s_2 \rightarrow a_2 \rightarrow \cdots \rightarrow a_t \rightarrow s_T$ until the agent reaches terminal state s_T. At the terminal state s_T, the agent receives a reward r as per a reward function $R : S \times A \rightarrow \mathbb{R}$. The action a is chosen from the stochastic policy $\pi(a|s)$ which gives probability of taking action a in state s. The performance of the policy π is quantified as:

$$Q^{\pi}(s_t, a_t) = \underset{a_t \sim \pi(s_t)}{\boldsymbol{E}} \left[\sum_{t=0}^{N} \lambda^t r_t(s_t, a_t) | S_1 = s_1 \right] \tag{3.16}$$

The objective is to find a policy π that maximises the performance as in eq. (3.16).

3.4.2 Maximum Entropy Inverse Reinforcement learning

Given a set of demonstrated (expert) behaviour, *i.e.*, trajectories resulting from executing expert policy π_E, Inverse Reinforcement learning (IRL) [43] aims to find a reward function that can rationalise the given behaviour. In maximum entropy inverse reinforcement learning [8], the demonstrated behaviour $\mathcal{D}_{\text{demo}} = \{\varsigma_1, \cdots, \varsigma_n\}$ is assumed to be the result of an expert acting stochastically and near-optimally with respect to an unknown reward function. Trajectories with equivalent rewards have equal probability and trajectories are sampled from the distribution:

$$p(\varsigma|\theta) = \frac{1}{Z(\theta)}\exp(r_\theta(\varsigma_i)) = \frac{1}{Z(\theta)}\exp\left(\sum_{t=0}^{|\varsigma_i|-1} r_\theta(s_t, a_t)\right) \tag{3.17}$$

and

$$Z(\theta) = \int \exp(r_\theta(\varsigma))d\varsigma \tag{3.18}$$

Where $Z(\theta)$ is the partition function and r_θ is the reward function. MaxEnt-IRL maximises the likelihood of the demonstrated data D_{demo} under the maximum entropy (exponential family) distribution and the objective is given as :

$$\mathcal{L}(\theta) = -\underset{\varsigma\sim\pi_E}{\boldsymbol{E}} r_\theta(\varsigma) + \log(Z(\theta)) \tag{3.19}$$

However, it is difficult to apply vanilla MaxEnt-IRL to complex high dimensional settings since computing the partition function in eq. (3.18) is intractable. To overcome this drawback, Finn *et al.* [44] combine sample-based MaxEnt-IRL with forward reinforcement learning to estimate the partition function $Z(\theta)$.

Let $p_{\mathcal{D}_{\text{samp}}}$ denote the probability density function of the trajectories generated during training and let $\mathcal{D}_{\text{samp}} = \{\varsigma_1, \cdots, \varsigma_n\}$ denote the collection of all these trajectories. The sample-based MaxEnt-IRL uses the samples form $p_{\mathcal{D}_{\text{samp}}}$ to estimate the expectation given by eq. (3.18). Therefore the $Z(\theta)$ term in eq. (3.19) is replaced by the importance-sampling estimate in eq. (3.20).

$$\mathcal{L}_{\text{IRL}_r}(\theta) = -\underset{\varsigma_i\sim p_{\mathcal{D}_{\text{demo}}}}{\boldsymbol{E}} r_\theta(\varsigma_i) + \log\left(\underset{\varsigma_j\sim p_{\mathcal{D}_{\text{samp}}}}{\boldsymbol{E}}\left[\frac{\exp(r_\theta(\varsigma_j))}{p_{\mathcal{D}_{\text{samp}}}(\varsigma_j)}\right]\right) \tag{3.20}$$

The algorithm proposed by Finn *et al.* alternates between updating the reward function to maximise the likelihood of the demonstrated data and optimising the distribution $p_{\mathcal{D}_{\text{samp}}}$ to minimise the variance of the importance-sampling estimate. Note that the optimal importance sampling distribution for estimating the partition function in eq. (3.18) is given by $p_{\mathcal{D}_{\text{samp}}} = \exp(r_\theta(\varsigma))$. Hence, in the forward RL loop the sampling distribution

$p_{\mathcal{D}_{\text{samp}}}$ is updated to match $\frac{1}{Z}\exp(r_\theta(\varsigma))$ by minimising the KL-divergence between them or optimising the following objective function:

$$\mathcal{L}_{\text{IRL}_\pi}(p_{\mathcal{D}_{\text{samp}}}) = -\underset{\varsigma\sim p_{\mathcal{D}_{\text{samp}}}}{\boldsymbol{E}}[r_\theta(\varsigma)] + \underset{\varsigma\sim p_{\mathcal{D}_{\text{samp}}}}{\boldsymbol{E}}\left[\log[p_{\mathcal{D}_{\text{samp}}}(\varsigma)]\right] \tag{3.21}$$

3.4.3 Maximum Causal Entropy

Motivated by the task of prediction in sequential interactions, Ziebart *et al.* [45] propose to use maximum causal entropy to model the availability and influence of sequentially revealed side information. The causal entropy policy π is defined as:

$$H(\pi) = \underset{\pi}{\boldsymbol{E}}[-\log\pi(a|s)] \tag{3.22}$$

which measures the uncertainty presented in a policy π.

3.4.4 Generative Adversarial Imitation Learning:

To make the discovery of the reward function in real scenarios tractable, Ho *et al.* [5] use the paradigm of generative adversarial networks to cast the problem of inverse reinforcement learning into the min-max optimisation framework. This formulation bypasses the need for learning the reward function to recover the expert policy.

Mathematically, the authors establish an equivalence relation between the primitive MaxEnt-IRL objective of finding a reward function r from a set of functions R and then using this reward function r to recover the expert policy π_E using the standard RL objective as shown in eq. (3.23) to a regularised occupancy matching problem as in eq. (3.24):

$$\max_{r\in R}\left(\underbrace{\min_{\pi\in\Pi} -\lambda H(\pi) - \underset{\pi}{\boldsymbol{E}}[r(s_t,a_t)]}_{\text{Inner loop of solving the RL problem}}\right) + \underset{\pi_E}{\boldsymbol{E}}[r(s_t,a_t)] \tag{3.23}$$

$$\approx \min_{\pi\in\Pi} -\lambda H(\pi) + D_{JS}(\rho_\pi, \rho_{\pi_E}) \tag{3.24}$$

In eq. (3.24), ρ_π and ρ_{π_E} are the occupancy measures of the target and the expert policy. The occupancy measure can be interpreted as an unnormalised distribution of state-action pairs that an agent encounters when navigating the environment following a particular policy π. $D_{JS}(\rho_\pi, \rho_{\pi_E})$ is the Jensen-Shannon Divergence between the expert policy π_E and the target policy. The objective of the optimisation process is thus to search for a target policy which minimises the Jensen-Shannon Divergence in eq. (3.24). The minimisation of the Jensen-Shannon Divergence is approximated by the GAN training objective, *i.e.*, eq. (3.24) is solved by finding the saddle point (π, D) of the expression:

$$\mathcal{L} = \underset{\pi_G}{\boldsymbol{E}}[\log(D(s_t, a_t))] + \underset{\pi_E}{\boldsymbol{E}}[\log(1 - D(s_t, a_t))] - \lambda H(\pi_G) \tag{3.25}$$

It can be seen from eq. (3.25) that the algorithmic setup uses a discriminator (D) to approximate the reward function and the generator (G) to approximate the agent's policy whereas the maximum causal entropy (H) acts as the regulariser.

3.5 Feature-Wise Layer Modulation

Feature-wise Linear Modulation (FiLM) [46] was introduced in the context of image stylisation and extended and shown to be highly effective for multi-modal tasks such as visual question-answering [47–49]. FiLM layer applies a per-channel scaling and shifting to the convolutional feature maps. Such layers are parameter efficient (only two scalars per feature map) while still retaining high capacity, as they can scale up or down, zero-out, or negate whole feature maps. In vision-and-language tasks, another network, the so-called FiLM generator $\mathfrak{f}$, predicts these modulating parameters from the linguistic input e_l. More formally the FiLM layer computes a modulated feature map $\hat{F}$ as follows:

$$[\lambda, \beta] = \mathfrak{f}(e_l) \tag{3.26}$$

$$\hat{F} = \lambda \odot F + \beta \tag{3.27}$$

where λ, β are scaling and shifting parameters which which modulate the activations of the original feature map F. FiLM layers may be inserted throughout the hierarchy of a convolutional network, either pre-trained and fixed [34] or trained from scratch [48]. Prior FiLM-based models [47–49] have used a single-hop FiLM generator to predict the FiLM parameters in all layers, *e.g.*, an MLP which takes the language embedding e_l as input [47–49].

Chapter 4

Proposed Approach

4.1 Visual Dialogue as Markov Decision Process

Given a large but finite vocabulary $\mathbb{W}$ of discrete tokens, we consider dialogue as a sequence of t utterances, $\mathfrak{D} = \left\langle \{x, y\}_{(1:t)} | t \in \mathbb{N} \right\rangle$, where every t^{th} utterance of dependent variable or output y is a sequence of tokens with length t', such that $y_{t,t'} = \underset{t \in \mathbb{N}}{\forall} \left\langle y_{t,t'} | y_t \in \mathbb{W}^{t'} \right\rangle$. Similarly, t^{th} utterance of independent variable or input x_t is a vector containing $[\mathfrak{K}_t, y_{(t,<t')}]$, where $y_{(t,<t')} = [y_{(t,1)}, \ldots, y_{(t,t')}]$, is the partially generated sequence of t' tokens preceding t and $\mathfrak{K}_t$ is some transformation on a tuple of Image $I \in \mathbb{R}^{d_I}$, Caption: $\mathfrak{C} \in \mathbb{W}^{d_\mathfrak{C}}$, Question: $\underset{t \in \mathbb{N}}{\forall} q_t | q \in \mathbb{W}^{d_q}$, History of all the utterances preceding t: $\mathfrak{H}_t = \{x, y\}_{<t} | \mathfrak{H}_t \in \mathbb{W}^{d_q + t'}$, where $d_I, d_\mathfrak{C}$& d_q are dimensions of the image, caption and question vectors respectively.

Our objective is to learn a function $\mathfrak{z}$, which maps the input x_t to an output y_t.

$$\mathfrak{z} : x_t \to y_t \tag{4.1}$$

Specifically, the function $\mathfrak{z}$ is akin to a statistical language model that gives the probability of generating the next token given the previous tokens and context.

$$P(y_{t,t'}, x_t) = \prod_{t \in \mathbb{N}} \prod_{t' \in \mathbb{N}} P(y_{t,t'} | y_{(t,>t')}, x_t) \tag{4.2}$$

Instead of modelling every realisation of x_t as an independent and identically distributed random variable, we assume that $x_t \in X$ is the state of a discrete time MDP sampled from the conditional distribution $x_t \sim P\left(\cdot|x_{(t-1)}\right)$. It follows from subsection 3.4.1 that this assumption gives us the flexibility to formulate the dialogue generation problem as an MDP. However, x_t is a joint distribution of multi-dimensional random variables and cannot be readily used in the RL framework. Thus, x_t needs to be projected into a lower dimensional space such that it satisfies the Markov property and compactly represents the dialogue history.

To overcome this hurdle, we use seq2seq model for the composition $\mathfrak{z}$ as in eq. (4.3).

$$\mathfrak{z} = \mathfrak{d} \circ \mathfrak{e}(x_t) \tag{4.3}$$

$$\mathfrak{e} : x_t \rightarrow s_t \tag{4.4}$$

$$\mathfrak{d} : s_t \rightarrow y_{t'} \tag{4.5}$$

Here $\mathfrak{e}$ is a mapping from x_t to the state representation s_t and $\mathfrak{d}$ is a mapping from the state s_t to the answer y_t.

4.1.1 State Space model

4.1.1.1 State s_t

The state s_t is an abstraction obtained by mapping the input x_t containing sequences with variable dimensions and lengths to a fixed size vector representation of dimension d_h. The function $\mathfrak{e}$ generates s_t by recursively modelling the sequence x_t generated up to the t^{th} instance as:

$$s_t = h_t^{\mathfrak{e}} = f_{\mathfrak{e}}(h_{t-1}^{\mathfrak{e}}, x_t) \tag{4.6}$$

In eq. (4.6), $h^{\mathfrak{e}} \in \mathbb{R}^{d_{h^{\mathfrak{e}}}}$ is a real-valued representation that encodes historical information received until the t^{th} instance. The mapping $f_{\mathfrak{e}}$ can be as simple as a logistic sigmoid, a recurrent neural network or a combination of both as well other more sophisticated operations. Subsection 4.3.1 will delve into it's specifics.

4.1.1.2 Policy(π_θ)

The function $\mathfrak{d}$ takes the state vector s_t as the input and returns a parametric policy $\pi_\theta(\cdot|s)$ which is simply a discrete distribution over the tokens in the vocabulary $\mathbb{W}$ given the state s_t, characterised by the parameters $\theta_\mathfrak{d}$. In other words, the $\mathfrak{d}$ returns the conditional probability of generating token $y_{t,t'}$ given all previously generated tokens $y_{(t,<t')}$ and the state s_t. The function $f_\mathfrak{d}$ at time t, generates an order-sensitive compact summary of the historical information $h_t^\mathfrak{d} \in \mathbb{R}^{d_{h^\mathfrak{d}}}$. Similar to the $\mathfrak{e}$, the function $f_\mathfrak{d}$ can be any function or composition with varying degrees of sophistication.

$$h_t^\mathfrak{d} = f_\mathfrak{d}(h_{t-1}^\mathfrak{d}, s_t) \tag{4.7}$$

$$\pi_\theta(\cdot|s) = P(y_{(t,t')}|\, h_t^\mathfrak{d}) \tag{4.8}$$

4.1.1.3 Action a_t

Action a_t is a token $y_{t,t'}$ sampled from the policy π

$$a_t = y_{(t,t')} \sim \pi_\theta(\cdot|s_t) \tag{4.9}$$

4.2 Dialogue generation with adversarial imitation learning

We assume that humans optimise a utility function when forming responses in a dialogue. This utility function drives them to vary their responses to different dialogue contexts. The model presented in this thesis aims to use samples from human participants for the discovery of a dialogue policy which is optimal as per this implicit reward function. To this end, we propose a model which unifies GAIL and sample-based MaxEnt-IRL frameworks.

The other motivation for this unification is the stability of the GAN's training dynamics. GANs are notoriously hard to train as the simultaneous optimisation method used in their training guarantees only local asymptotic stability under very strong assumptions [50, 51]. As a result, most dialogue models trained using the adversarial setup get stuck in a local optimum which, results in learning of degenerate policies that either produce identical responses irrespective of the input or produce uninformative or nonsensical replies. In the parlance of generative modelling, this phenomenon is called mode-collapse. Many solutions have been proposed to stabilise GAN training, most of which involve modification of the optimisation objective. Finn *et al*. [10] posit an interesting proposition of combining EBMs, sample-based MaxEnt-IRL and GANs by modifying the discriminator's objective function. The authors suggest that the derivative of the MaxEnt-IRL in eq. (3.20) is mathematically equivalent to the derivative of the discriminator's objective in eq. (3.13) if the density function of the generated data p' is substituted in the discriminator's objective function. We re-derive Finn *et al*.'s claim for completeness.

4.2.1 Reducing the variance in Importance Sampling

The importance sampling estimate in eq. (3.20) can have very high variance if the sampling distribution $p_{\mathcal{D}_{\text{samp}}}$ does not have trajectories with high rewards. To reduce the variance we define a mixture distribution η which can sample high reward trajectories from the data distribution $p_{\mathcal{D}_{\text{demo}}}$.

Definition 4.2.1 (Mixture distribution)**.** η is a mixture distribution which uniformly samples trajectories from $p_{\mathcal{D}_{\text{demo}}}$ and $p_{\mathcal{D}_{\text{samp}}}$. The probability density function of η is therefore defined as:

$$\eta(\varsigma) = \frac{1}{2} p_{\mathcal{D}_{\text{demo}}}(\varsigma) + \frac{1}{2} p_{\mathcal{D}_{\text{samp}}}(\varsigma) \tag{4.10}$$

As per definition 4.2.1 we modify the importance sampling correction ratio in eq. (3.20) using a rough estimate of the demonstration distribution $p'_{\mathcal{D}_{\text{demo}}}$ sampled from η.

$$\mathcal{L}(\theta) = -\underset{\varsigma_i \sim p_{\mathcal{D}_{\text{demo}}}}{\boldsymbol{E}} r_\theta(\varsigma_i) + \log\left(\underset{\varsigma_j \sim \eta}{\boldsymbol{E}}\left[\frac{\exp(r_\theta(\varsigma_j))}{\frac{1}{2}p'_{\mathcal{D}_{\text{demo}}}(\varsigma_j) + \frac{1}{2}p_{\mathcal{D}_{\text{samp}}}(\varsigma_j)}\right]\right) \tag{4.11}$$

$$\text{Let } p'_{\mathcal{D}_{\text{demo}}}(\varsigma) = \frac{1}{Z}\exp(r_\theta(\varsigma))$$

Let us use the shorthand η' for the approximate mixture density given by:

$$\eta'(\varsigma) = \frac{1}{2Z}\exp(r_\theta(\varsigma)) + \frac{1}{2}p_{\mathcal{D}_{\text{samp}}}(\varsigma) \tag{4.12}$$

$$\therefore \mathcal{L}(\theta) = -\underset{\varsigma_i \sim p_{\mathcal{D}_{\text{demo}}}}{\boldsymbol{E}} r_\theta(\varsigma_i) + \log\left(\underset{\varsigma_j \sim \eta}{\boldsymbol{E}}\left[\frac{\exp(r_\theta(\varsigma_j))}{\eta'(\varsigma_j)}\right]\right) \tag{4.13}$$

Lemma 4.2.1. Given a set of probability distributions $P = \{p_1, \ldots, p_i\}$ and mixture distribution $f(x) = \sum_i w_i p_i(x)$ where $\underset{i \in \mathbb{N}}{\forall} w_i \in (0,1]$, the expected value of a random variable x sampled from f is given by: $\underset{x\sim f}{\boldsymbol{E}}[x] = \sum_i \left(w_i \underset{x \sim p_i}{\boldsymbol{E}}[x]\right)$

Proof. By definition of an expectation of operator:

$$\begin{aligned} \underset{x\sim f}{\boldsymbol{E}} &= \sum_j x_j f(x_j) \\ &= \sum_j x_j \sum_i w_i p_i(x_j) && \text{by definition} \\ &= \sum_i \left(w_i \underset{x\sim p_i}{\boldsymbol{E}}[x]\right) \end{aligned}$$

□

4.2.2 Modelling discriminator using Boltzmann's function

In GAN [6] the optimal discriminator is given by:

$$D^*(\varsigma) = \frac{p_{\text{data}}(\varsigma)}{p_{\text{data}}(\varsigma) + p_{\text{gen}}(\varsigma)} \tag{4.14}$$

We modify the GAN's optimal discriminator given in eq. (4.14) such that it can estimate the density of generated samples. Let $p'_{\text{data}}(\varsigma)$ denote the rough estimate of samples from $p_{\text{gen}}(\varsigma)$. Next we use the Boltzmann distribution in eq. (3.15) to model $p'_{\text{data}}(\varsigma)$.

Definition 4.2.2 (Boltzmann Discriminator)**.** The Boltzmann Discriminator D_θ is defined as:

$$D_{\theta_D}(\varsigma) = \frac{\frac{1}{Z}\exp(r_{\theta_D}(\varsigma))}{\frac{1}{Z}\exp(r_{\theta_D}(\varsigma)) + p_{\text{gen}}(\varsigma)} \tag{4.15}$$

$\therefore$ As per Definition 4.2.2, eq. (3.13) changes to

$$\begin{aligned}
\mathcal{L}_{D_{\theta_D}} &= \mathop{\boldsymbol{E}}_{\varsigma\sim p_{\text{data}}}\left[-\log\left(\frac{\frac{1}{Z}\exp(r_{\theta_D}(\varsigma))}{\frac{1}{Z}\exp(r_{\theta_D}(\varsigma)) + p_{\text{gen}}(\varsigma)}\right)\right] \\
&\mathop{\boldsymbol{E}}_{\varsigma\sim p_{\text{gen}}}\left[-\log\left(\frac{p_{\text{gen}}(\varsigma)}{\frac{1}{Z}\exp(r_{\theta_D}(\varsigma)) + p_{\text{gen}}(\varsigma)}\right)\right] && (4.16) \\
\text{Let } \mu &= \frac{1}{Z}\exp(r_{\theta_D}(\varsigma)) + p_{\text{gen}}(\varsigma) && (4.17) \\
\therefore \mathcal{L}_{D_{\theta_D}} &= \mathop{\boldsymbol{E}}_{\varsigma\sim p_{\text{data}}}\left[-\log\left(\frac{\frac{1}{Z}\exp(r_{\theta_D}(\varsigma))}{\mu}\right)\right] + \mathop{\boldsymbol{E}}_{\varsigma\sim p_{\text{gen}}}\left[-\log\left(\frac{p_{\text{gen}}(\varsigma)}{\mu}\right)\right] && (4.18) \\
&= \log(Z) - \mathop{\boldsymbol{E}}_{\varsigma\sim p_{\text{data}}}\left[r_{\theta_D}(\varsigma)\right] + \mathop{\boldsymbol{E}}_{\varsigma\sim p_{\text{data}}}[\log\mu(\varsigma)] - \mathop{\boldsymbol{E}}_{\varsigma\sim p_{\text{gen}}}\left[\log p_{\text{gen}}(\varsigma)\right] \\
&+ \mathop{\boldsymbol{E}}_{\varsigma\sim p_{\text{gen}}}[\log(\mu(\varsigma))] && (4.19) \\
&= \log(Z) - \mathop{\boldsymbol{E}}_{\varsigma\sim p_{\text{data}}}\left[r_{\theta_D}(\varsigma)\right] - \mathop{\boldsymbol{E}}_{\varsigma\sim p_{\text{gen}}}\left[\log p_{\text{gen}}(\varsigma)\right] + 2\mathop{\boldsymbol{E}}_{\varsigma\sim\mu}[\log(\mu(\varsigma))] \quad \ldots\text{from lemma 4.2.1} \\
& && (4.20)
\end{aligned}$$

Lemma 4.2.2. Given a discriminator as in definition 4.2.2, if $\eta' = \mu$ then the optimal partition function Z for eq. (4.16) is an importance sampling estimator for the MaxEnt-IRL objective in eq. (4.13).

Proof. Expanding eq. (4.16) we get:

$$\nabla_Z \mathcal{L}_{D_{\theta_D}} = 0 \tag{4.21}$$

$$\therefore \frac{1}{Z} = \frac{1}{Z^2} \mathop{\boldsymbol{E}}_{\varsigma\sim\mu}\left[\frac{\exp(r_{\theta_D}(\varsigma))}{\mu(\varsigma)}\right] \tag{4.22}$$

$$\therefore Z = \mathop{\boldsymbol{E}}_{\varsigma\sim\mu}\left[\frac{\exp(r_{\theta_D}(\varsigma))}{\mu(\varsigma)}\right] \tag{4.23}$$

□

Proposition 4.2.3. Given a partition function $Z = \mathop{\boldsymbol{E}}_{\varsigma\sim\eta}\left[\frac{\exp(r_{\theta_D}(\varsigma))}{\mu(\varsigma)}\right]$ and $\eta' = \mu$, the gradient of $\mathcal{L}_{D_{\theta_D}}$ and $\mathcal{L}_{\mathrm{IRL}_r}$ are analytically equivalent *i.e.*, $\nabla_{\theta_D}\mathcal{L}_{D_{\theta_D}} = \nabla_\theta \mathcal{L}(\theta)$

Proof. In eq. (4.13) the approximate mixture distribution η' is used for importance-sampling correction and is constant.

$$\therefore \nabla_\theta \mathcal{L}_{\mathrm{IRL}_r}(\theta) = -\mathop{\boldsymbol{E}}_{\varsigma\sim p_{\text{data}}}[\nabla_\theta r_\theta(\varsigma)] + \mathop{\boldsymbol{E}}_{\varsigma\sim\eta'}\left[\left(\frac{-\exp(r_\theta(\varsigma))\nabla_\theta r_\theta(\varsigma)}{\eta'(\varsigma)}\right)\left(\frac{\eta'(\varsigma)}{\exp(r_\theta(\varsigma))}\right)\right] \tag{4.24}$$

$$= -\mathop{\boldsymbol{E}}_{\varsigma\sim p_{\text{data}}}[\nabla_\theta r_\theta(\varsigma)] + \mathop{\boldsymbol{E}}_{\varsigma\sim\eta'}\left[\frac{-\frac{1}{Z}\exp(r_\theta(\varsigma))\nabla_\theta r_\theta(\varsigma)}{\eta'(\varsigma)}\right] \tag{4.25}$$

$$\nabla_{\theta_D}\mathcal{L}_{D_{\theta_D}} = -\mathop{\boldsymbol{E}}_{\varsigma\sim p_{\text{data}}}\left[\nabla_{\theta_D} r_{\theta_D}(\varsigma)\right] + \mathop{\boldsymbol{E}}_{\varsigma\sim\mu}\left[\frac{\frac{1}{Z}\exp(r_{\theta_D}(\varsigma))\nabla_{\theta_D} r_{\theta_D}(\varsigma)}{\mu(\varsigma)}\right] \tag{4.26}$$

$$\therefore \nabla_{\theta_D}\mathcal{L}_{D_{\theta_D}} = \nabla_\theta \mathcal{L}_{\mathrm{IRL}_r}(\theta) \quad \because \eta' = \mu \tag{4.27}$$

□

Corollary 4.2.4. The generator's objective function $\mathcal{L}_{G_{\theta_G}}$ is equivalent to MaxEnt-IRL's objective $\mathcal{L}_{\mathrm{IRL}_\pi}$.

Proof.

$$\mathcal{L}_{G_{\theta_G}} = \mathop{\boldsymbol{E}}_{\varsigma \sim p_{\mathcal{D}_{\text{samp}}}} [\log(1 - D(\varsigma)) - \log(D(\varsigma))] \tag{4.28}$$

$$= \mathop{\boldsymbol{E}}_{\varsigma \sim p_{\mathcal{D}_{\text{samp}}}} \left[-r_{\theta_G}(\varsigma) + \log(Z) + \log(p_{\mathcal{D}_{\text{samp}}}(\varsigma)) \right] \tag{4.29}$$

$$= \mathcal{L}_{\text{IRL}_\pi} + \log(Z) \tag{4.30}$$

□

4.2.3 Energy based GAIL

As per proposition 4.2.3 the discriminator's objective (loss) function in our model is given by:

$$\mathcal{L}_{D_{\theta_D}}(\theta_D) = -\mathop{\boldsymbol{E}}_{\varsigma_i \sim \pi_E} \left[r_{\theta_D}(\varsigma_i) \right] + \log \mathop{\boldsymbol{E}}_{\varsigma_j \sim \pi_{\theta_G}} \left(\frac{\exp(r_{\theta_D}(\varsigma_j))}{p_{\mathcal{D}_{\text{samp}}}(\varsigma_j)} \right) \tag{4.31}$$

Where π_E denotes the policy of demonstrated trajectories and π_{θ_G} is the agent's policy. The term $p_{\mathcal{D}_{\text{samp}}}$ denotes the background distribution from which the dialogues ς_j were sampled. In our setting, $p_{\mathcal{D}_{\text{samp}}}$ is the distribution of dialogues generated with the current dialogue policy π_{θ_G}. We use $\mathcal{D}_{demo}$ and $\mathcal{D}_{samp}$ to represent the set of dialogues generated with policy π_E and π_{θ_G} respectively. Subsequently, the gradient of the discriminator/reward function is given by:

$$\nabla_{\theta_D} \mathcal{L}_{D_{\theta_D}} = -\mathop{\boldsymbol{E}}_{\varsigma_i \in p_{\mathcal{D}_{\text{demo}}}} \left[\nabla_{\theta_D} r_{\theta_D}(\varsigma_i) \right] + \frac{1}{\sum_j \left(\frac{\exp(r_{\theta_D}(\varsigma_i))}{p_{\mathcal{D}_{\text{samp}}}(\varsigma_j)} \right)} \sum_{\varsigma_j \in D_{samp}} \left(\frac{\exp(r_\theta(\varsigma_i))}{p_{\mathcal{D}_{\text{samp}}}(\varsigma_j)} \right) \nabla_{\theta_D} r_{\theta_D}(\varsigma_j) \tag{4.32}$$

With the reward function as eq. (4.32), the expert policy can be found by solving a common reinforcement learning problem:

$$\pi_{\theta_G} \in \mathop{\arg\min}_{\pi \in \Pi} \left(-\lambda H(\pi) - \mathop{\boldsymbol{E}}_{\varsigma \sim p_{\mathcal{D}_{\text{samp}}}} \left[r_{\theta_D}(\varsigma) \right] \right) \tag{4.33}$$

where ς represents the sampled dialogue and $H(\pi)$ is the causal entropy regularisation term; $r_{\theta_D}(\varsigma)$ is the reward that can be accessed from the reward model. The goal of the generator is to generate dialogues that can achieve higher rewards from the reward model. The found policy therefore maximises the expected cumulative reward while maintaining high entropy. The derivative can be inferred as follows:

$$\nabla_{\theta_G}\mathcal{L}_{G_{\theta_G}} = -\underset{\varsigma\sim\pi_{\theta_G}}{\boldsymbol{E}} \nabla_{\theta_G}\left[\log(\pi_{\theta_G}(\varsigma))\right]\left[r_{\theta_D}(\varsigma) - \lambda\log(\pi_{\theta_G}(\varsigma))\right] \tag{4.34}$$

We unify all these modifications in the algorithm used to train the dialogue model.

Algorithm

Algorithm 1: GAIL Algorithm

```
Input: Expert Trajectories $\varsigma_E \sim \pi_E$, initial policy and discriminator parameters
  $\theta_{G_0}, \theta_{D_0}$ for $i = 0, 1, 2...$ do
    Sample trajectories $\varsigma_i \sim \pi_{\theta_{G_i}}$;
    Update the discriminator parameters from $\theta_{D_i}$ to $\theta_{D_{i+1}}$ with the gradient
      from eq. (4.32) ;
    Take a policy step from $\theta_{G_i}$ to $\theta_{G_{i+1}}$ using the TRPO [52] rule, i.e., take a
      KL-constrained natural gradient step with eq. (4.34);
end
```

4.3 Architecture Design

In the GAIL setup, the generator models the policy, and the discriminator approximates the reward function. The seq2seq model used in the generator is responsible for combining signals from image and text modalities, generate the state representation and use this representation for approximating the policy function. The discriminator identifies the source of the trajectory presented to it.

4.3.1 Generator Architecture

With the seq2seq model as its main component, we use several specialised operations in the generator. The encoder generates a feature representation $\mathfrak{K}_t$, and the decoder, on the other hand carries out two functions: (1) First, it generates the state vector s_t using $\mathfrak{K}_t$ and previously generated tokens $y_{t,<t'}$., and (2) It learns the policy π for generating the response y_t.

4.3.1.1 Encoder Architecture

A dialogue corpus with human-human conversations has many instances of pronouns referring to an object(entity) in an image. In linguistics parlance, these are called co-reference(s), and the task of finding the relations between co-references and referred entities is called co-reference resolution. Contextual resolution of co-references is necessary to avoid repetition of co-referred entities and maintain coherency in generated responses. The attention mechanism [53] has empirically demonstrated success in performing implicit co-reference resolution by identifying the portions of the dialogue history that can help in answering the question.

Given that the dialogues in our setting are grounded on images we hypothesise that integrating information from both image and text modalities for accurate co-reference resolution can be beneficial in learning a meaningful representation.

To this end, we design an encoder architecture with an attention mechanism that uses the question and the dialogue history to identify the portion of image features useful for producing an answer.

Specifically, we use a hierarchical pipeline to learn the dialogue features, *i.e.*, we first identify the image (c_{iq}) and history-question features (c_{hq}) which are related to the current question (ϕ_q). Next we use the history-question features (c_{hq}) to identify the corresponding image features (c_{ihq}). Finally, we fuse all these features to recursively generate a context representation $\mathfrak{K}_t$.

$$\lambda_{hq} = \text{softmax}(\theta_{\lambda_{hq}}^{\top} \cdot \phi_q) \tag{4.35}$$

$$\beta_{hq} = \text{softmax}(\theta_{\beta_{hq}}^{\top} \cdot \phi_q) \tag{4.36}$$

$$\lambda_{iq} = \text{softmax}(\theta_{\lambda_{iq}}^{\top} \cdot \phi_q) \tag{4.37}$$

$$\beta_{iq} = \text{softmax}(\theta_{\beta_{iq}}^{\top} \cdot \phi_q) \tag{4.38}$$

$$c_{iq} = \lambda_{iq} \odot \phi_i + \beta_{iq}(\phi_i) \tag{4.39}$$

$$c_{hq} = \lambda_{hq} \odot \phi_h + \beta_{hq}(\phi_h) \tag{4.40}$$

$$\lambda_{ihq} = \text{softmax}(\theta_{\lambda_{ihq}}^{\top} \cdot c_{hq}) \tag{4.41}$$

$$\beta_{ihq} = \text{softmax}(\theta_{\beta_{ihq}}^{\top} \cdot c_{hq}) \tag{4.42}$$

$$c_{ihq} = \lambda_{ihq} \odot \phi_i + \beta_{ihq}(\phi_i) \tag{4.43}$$

$$g = \tanh(\theta_s^{\top} \cdot [\phi_q; c_{ihq}; c_{iq}]) \tag{4.44}$$

$$\mathfrak{K}_t = h_t^{\mathfrak{e}'} = \varphi_{\mathfrak{e}'}(h_{(t-1)}^{\mathfrak{e}'}, g) \tag{4.45}$$

Equations (4.35) to (4.45) represent the sequence of operations performed to generate the context vector. All the operations are on spatial image features (ϕ_i), encoded question (ϕ_q), and encoded history (ϕ_h) such that $\{\phi_i, \phi_h, \phi_q\} \in \mathbb{R}^d$, where d is some fixed dimension and θ are tunable parameters. The image feature vector(ϕ_i) is encoded by a VGG-CNN and the question(ϕ_q) and history features(ϕ_h) are individually encoded by an RNN at every time step as the dialogue evolves.

The modulation of feature vectors is carried out using a specialised procedure called feature-wise linear transformation (FiLM). The modulating feature vector(s) λ, β are generated using exponentially normalised linear transformation(s) (eqs. (4.35) to (4.38), eqs. (4.41) and (4.42)). These features are then used to perform per-feature affine transformation(s) on the image and dialogue history feature vectors eqs. (4.39), (4.40) and (4.43).

The question and hierarchical image features are then concatenated and projected to a lower-dimensional space via a non-linear transformation, eq. (4.44), where $\theta_s \in \mathbb{R}^{3d \times d}$ and $[;]$ represents the concatenation operation.

Finally, we obtain the context vector by recursively modelling the resulting feature vector using an LSTM($\varphi_{\mathfrak{e}}$), eq. (4.45). We name our encoder as Hierarchical Image History Encoder (HIH).

4.3.1.2 Decoder

The decoder uses two LSTM cells. The first LSTM $\varphi_{\mathfrak{e}}$ recursively processes the partially generated sequence $y_{t,<t'}$ and context vector $\mathfrak{K}_t$ to generate the state s_t of the dialogue. The next LSTM $\varphi_{\mathfrak{d}}$ acts on the state vector s_t to form an intermediate representation $h_t^{\mathfrak{d}}$ of dimension $d_{h_t^{\mathfrak{d}}}$. Finally we obtain the policy π by exponentially normalising a non-linear transformation of $h_t^{\mathfrak{d}}$.

$$s_t = h_t^{\mathfrak{e}} = \varphi_{\mathfrak{e}}(h_{(t-1)}^{\mathfrak{e}}, [\mathfrak{K}_t, y_{(t,<t')}]) \tag{4.46}$$

$$h_t^{\mathfrak{d}} = \varphi_{\mathfrak{d}}(h_{(t-1)}^{\mathfrak{d}}, s_t) \tag{4.47}$$

$$\pi_{\theta_G} = \text{softmax}(\tanh(\theta_{\mathfrak{d}}^{\top} \cdot h_t^{\mathfrak{d}})) \tag{4.48}$$

Equations (4.46) to (4.48) show the decoder operations, where $h_t^{\mathfrak{d}} \in \mathbb{R}^{d_{h_t^{\mathfrak{d}}}}$ and $\theta_{\mathfrak{d}} \in \mathbb{R}^{d_{h_t^{\mathfrak{d}}} \times \mathbb{W}}$.

4.3.2 Discriminator Design

As mentioned previously, the discriminator identifies the source of the trajectory and approximates the reward signal needed for the optimisation process. With the same rationale as in the generator, to correctly determine the source of the response (generated by the agent or real) we use feature-wise modulation to capture the interplay of visual features and dialogue semantics.

$$\begin{aligned}
y_t &\sim \pi_{\theta_G}(\cdot|s_t) && (4.49)\\
\phi_{qa} &= \varphi_q(h_{t-1}^{\varphi_{qa}}, [q_t; y_t]) && (4.50)\\
\hat{\mathfrak{H}}_t &= [q_{1:t-1}; y_{1:t-1}] && (4.51)\\
\phi_{\hat{h}} &= \varphi_{\mathfrak{H}}(h_{t-1}^{\varphi_{\hat{\mathfrak{H}}}}, \hat{\mathfrak{H}}_t) && (4.52)\\
\lambda_{hqa} &= \text{softmax}(\theta_{\lambda_{hqa}}^{\top} \cdot \phi_{qa}) && (4.53)\\
\beta_{hqa} &= \text{softmax}(\theta_{\beta_{hqa}}^{\top} \cdot \phi_{qa}) && (4.54)\\
c_{hqa} &= \lambda_{hqa} \odot \phi_{\hat{h}} + \beta_{hqa} && (4.55)\\
\lambda_{ic} &= \text{softmax}(\theta_{\lambda_{ic}}^{\top} \cdot c_{hqa}) && (4.56)\\
\beta_{ic} &= \text{softmax}(\theta_{\beta_{ic}}^{\top} \cdot c_{hqa}) && (4.57)\\
c_D &= \lambda_{ic} \odot \phi_i + \beta_{ic} && (4.58)\\
r_{\theta_D}(\varsigma_t) &= \sigma(\theta_D^{\top} \cdot c_D) && (4.59)
\end{aligned}$$

Equations (4.49) to (4.59) describe the discriminator architecture in detail. At the end of every episode after the t' $^{\text{th}}$ instance *i.e.*, after generating a complete answer y_t, we concatenate it to the question vector q_t. We also keep a track of the dialogue history, which is nothing but a stack of question answer pairs from the previous $t-1$ rounds. The image features, dialogue history and the current state action-pair, form the trajectory ς_t. The information in ς_t is encoded into a context-vector and further classified using a binary classifier. Thus the discriminator's output is the reward $r(\varsigma_t)$ the agent gets at the end of trajectory.

Chapter 5

Experiments, Results and Discussion

5.1 Experimental Setup

To study the effectiveness of our approach, we evaluate the effect of both the learning framework and encoder architecture in isolation. First, the HIH encoder is evaluated using both RL and supervised learning. Next, to assess the contribution of visual features in policy learning; we modify the HIH encoder to exclude image features. We call this new encoder Hierarchical History (HH) and test it with both RL and supervised learning.

The models are trained on both VisDial v0.9 and v1.0 containing a training set of 82k and 123k images and a validation set of 40k and 2k images. The training set is split into sets of 60k and 100k for training and 22k and 23k for validation respectively The test-set consisting of additional 8k images provided as a part of v1.0 is used reserved for testing.

Captions/Questions/Answers longer than 24/16/8 words are truncated, and a vocabulary of 8964 words is built using words that occur at least five times in the training set.

5.1.1 Training details

In all our experiments, we use Deep-LSTMs with two layers, each with a hidden layer dimension of 1024 and 512. We use the output of the last pooling layer (512x7x7) of

VGG-19 to get the image features. We use the Adam optimiser [39] with a base learning rate of 4e-4. We use Beam Search to sample answers from the vocabulary and set the beam size to 8. We determine the optimal value for λ to be 0.01 using hyper-parameter optimisation.

All the models are pre-trained with supervised learning using half of the training data. We make an assumption here that the maximum likelihood objective captures some part of the RL objective. This setup reflects naturally occurring practical scenarios for the algorithm where a large amount of labelled data can be used to boot-strap a policy if the maximum likelihood and RL objective are at least partially aligned. We train the RL algorithm for 400 epochs.

5.1.2 Evaluation Protocol:

Following the evaluation protocol setup of Das *et al*. [36] we use a retrieval setting to evaluate the responses at each round in the dialogue. Specifically, every question in VisDial is coupled with a list of 100 candidate answer options, which the models are asked to sort for evaluation purposes. As we are evaluating models that generate answers, we use the model's likelihood scores to rank the candidate answers. Another way of interpreting this setup is that we evaluate the likelihood of the candidate answers being generated by the learnt policy. The Evaluation metrics are as follows:

1. Rank(M): This metric gives the rank of the human response(lower is better).

2. recall@k: Existence of human response in the top k responses (higher is better).

3. Mean Reciprocal Rank(MRR): this metric gives the mean reciprocal rank of the human response(higher is better).

5.2 Results

We compare the performance of our methods with two state of the art models: (1) Memory Network Encoder-Decoder (MN) architecture by Das *et al*. [36], (2) History Conditioned Image Attentive Encoder (HCIAE) by Lu *et al*. [37]. The nomenclature *model type-learning method* is used to represent models used in the experiments. Performance on Visdial v0.9 is shown in table 5.1 and that on Visdial v1.0 is shown in table 5.2. Table 5.3 shows the percentage improvement of our models over the baselines.

Model	MRR	R@1	R@5	R@10	Mean
MN-QIH-G [36][1]	0.5259	42.29	62.85	68.88	17.06
HCIEA-DIS [37]	0.5467	44.35	65.28	71.55	14.23
MN-RL	**0.5486**	**44.12**	**65.59**	**72.32**	**13.68**
HCIEA-RL	**0.5470**	**44.78**	**65.91**	**72.18**	**14.18**
HIH-MLE	**0.5471**	**44.70**	**65.39**	**73.48**	**14.06**
HIH-RL	**0.5569**	**47.20**	**66.75**	**71.21**	**13.71**
HH-RL	**0.5463**	**46.91**	**66.26**	**71.42**	**13.98**
HH-MLE	**0.5459**	**46.23**	**65.12**	**69.32**	**14.57**

Table 5.1 Performance of methods on VisDial 0.9 measures by Mean Reciprocal Rank(MRR), recall@k and mean rank. Higher is better for MRR and recall@k, lower is better for mean rank.

Model	MRR	R@1	R@5	R@10	Mean
MN-QIH-G [36]	0.5912	47.13	65.17	72.01	11.47
HCIEA-DIS [37]	0.6216	46.56	66.75	71.68	10.40
MN-RL	**0.6137**	**47.21**	**65.28**	**72.32**	**10.39**
HCIEA-RL	**0.6291**	**48.77**	**69.93**	**73.77**	**10.58**
HIH-MLE	**0.6301**	**47.55**	**72.03**	**72.89**	**10.35**
HIH-RL	**0.6341**	**48.84**	**71.38**	**73.00**	**09.98**
HH-RL	**0.6319**	**47.92**	**70.78**	**73.15**	**10.07**
HH-MLE	**0.6297**	**48.10**	**68.44**	**72.29**	**10.13**

Table 5.2 Performance of methods on VisDial 1.0(latest release) measures by Mean Reciprocal Rank(MRR), recall@k and mean rank. Higher is better for MRR and recall@k, lower is better for mean rank.

From the results in table 5.3, we observe that on Visdial v0.9 HIH-RL model improves the MRR by 3.84% and the Rank (M) by 12.37%. On Visdial v1.0 same models improve

[1] This was our initial baseline for evaluating the results of GAIL architecture.

Model	% improvements in metrics over baselines on V0.9				
	% improvements in metrics over baselines on V1.0				
MN-RL	2.29	1.87	2.40	3.00	12.56
	1.21	0.78	-1.03	0.66	4.98
HCIEA-RL	2.00	3.37	2.88	-0.05	9.36
	3.74	4.11	6.02	2.68	3.25
HIH-MLE	2.01	3.19	2.07	0.38	10.13
	3.90	1.50	9.20	1.46	5.35
HIH-RL	3.84	8.96	4.20	-0.01	12.37
	4.57	4.26	8.21	1.60	8.73
HH-RL	1.86	8.29	3.43	1.72	10.65
	4.21	2.29	7.31	1.82	7.91
HH-MLE	1.79	6.72	1.65	-1.28	6.87
	3.84	2.68	3.76	0.62	7.36

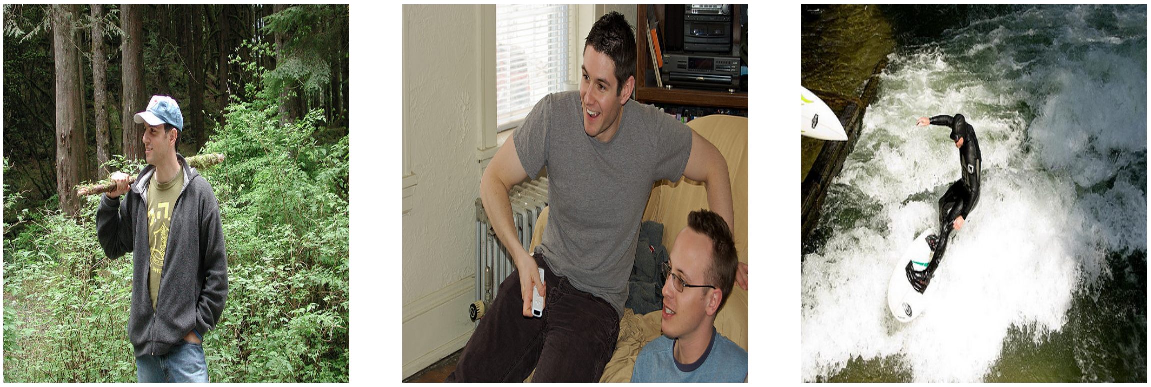
Table 5.3 % improvement over the mean baseline scores Visdial v0.9 and v1.0

MRR by 4.21% and M by 8.73%. Additionally, the maximum likelihood trained HIH-MLE model improves the MRR and M by 2.01% and 10.13% on v0.9 and 3.90%, 5.34% on v1.0 versions of Visdial, respectively. The baseline methods trained using our method also improve their performance over the supervised learning counterparts. Using RL for training all the models results in a mean improvement of 2.53% on MRR and 10.5% on M for Visdial v0.9 and 3.49% on MRR and 6% on M for v1.0.

Surprisingly we also observe nominal improvements using the HH-RL and HH-MLE models.

5.3 Discussion

From the dialogues in tables 5.4 to 5.6 on images in fig. 5.1, it can be seen that the models trained using our method can be less prone to producing responses like *I don't know*. Intuitively the min-max optimisation setup pitches the generator and discriminator against each other. The discriminator's job is to identify the source of the generated response (human or artificial). The generator's job is to produce human-like responses. As humans are least likely to generate bland or deflective responses in most scenarios,



(a) Caption: A man standing with a stick on his shoulder near the woods.

(b) Caption: 2 men sit on a couch while 1 plays a game.

(c) Caption: A man in a black wetsuit rides on a white surfboard.

Fig. 5.1 Random images sampled from Visdial v1.0 test set.

Q1: How old does the man look?	Q2: what color is the stick?	Q3: Is the man wearing a jacket or coat?
MN-QIH-G:don't know.	MN-QIH-G:black	MN-QIH-G:yes.
HCIEA-DIS:20	HCIEA-DIS:Yellow	HCIEA-DIS:Can't tell
MN-RL: Looks 30	MN-RL: Black	MN-RL: Yes
HCIEA-RL:Jacket	HCIEA-RL:Yes	HCIEA-RL:No
HIH-MLE: Almost 30	HIH-MLE: There is no stick	HIH-MLE: Jacket
HIH-RL: He seems to be in his 30's	HIH-RL: Brown	HIH-RL: The man is wearing a jacket
HH-RL:In his 20's	HH-RL:brown	HH-RL:He has a jacket
HH-MLE:30	HH-MLE:brown	HH-MLE:jacket

Table 5.4 Dialogues for fig. 5.1a

the optimisation process discourages the generator from learning a policy that is prone to producing such responses. Moreover, we can establish a direct analogy between the improvement in the quality of the generated samples due to the evaluation of generator density and the ability of inverse reinforcement learning to discover novel behaviour. This is because both the modified optimisation objective and increased capacity due to over-parameterisation improves the ability of the generator to "fill in" the modes of the data distribution. On the contrary, maximum likelihood training is ineffective for learning multimodal distributions as it results in the generator putting most of the prob-

Q1: How old do the men look?	Q2: Do they look like brothers	Q3: What kind of clothing are they wearing?
MN-QIH-G:30.	MN-QIH-G:yes	MN-QIH-G:can't tell.
HCIEA-DIS:no	HCIEA-DIS:Yellow	HCIEA-DIS:no
MN-RL: Yes they do	MN-RL: Hair are black	MN-RL: Casual clothing
HCIEA-RL:30	HCIEA-RL:no	HCIEA-RL:jeans
HIH-MLE: old	HIH-MLE: yes	HIH-MLE: shirt and pant
HIH-RL: I would say in their late 20's	HIH-RL: I don't think so	HIH-RL: They are wearing t-shirts
HH-RL:In their 30's	HH-RL:Yes they are brothers	HH-RL:tshirt and jeans
HH-MLE:30	HH-MLE:No	HH-MLE:t shirts and jeans

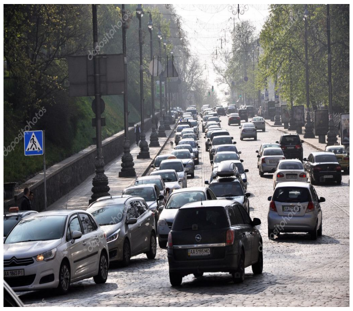
Table 5.5 Dialogues for fig. 5.1b

Q1: Is the picture in color?	Q2: Is it daytime?	Q3: Can you see the sun?
MN-QIH-G:yes	MN-QIH-G:yes	MN-QIH-G:can't tell
HCIEA-DIS:yes	HCIEA-DIS:yes	HCIEA-DIS:no
MN-RL: yes	MN-RL: yes	MN-RL: no
HCIEA-RL:yes	HCIEA-RL:yes it is day time	HCIEA-RL:no I can't
HIH-MLE: yes	HIH-MLE: don't know	HIH-MLE: no
HIH-RL: Yes	HIH-RL: Yes	HIH-RL: No I cannot see it.
HH-RL:yes it is	HH-RL:no	HH-RL:no
HH-MLE:no	HH-MLE:no	HH-MLE:no

Table 5.6 Dialogues for fig. 5.1c

Fig. 5.2 A pathological example showing that models ignore visual information.

(a) Caption: A busy street filled with lots of traffic



(b) Dialogue for fig. 5.2a

Q1: How old is the lady?	Q2: Is it?
MN-QIH-G:30.	MN-QIH-G:30.
HCIEA-DIS:20	HCIEA-DIS: Yes
MN-RL: She is 30.	MN-RL: Yes
HCIEA-RL:20	HCIEA-RL:Yes
HIH-MLE: 20	HIH-MLE: Maybe
HIH-RL: I think she is 30.	HIH-RL: Sure
HH-RL:30 I think	HH-RL:Yes
HH-MLE:25	HH-MLE:Yes

ability mass in spaces that have negligible density under the original data distribution which results in the generation of irrelevant responses.

Another important aspect highlighted by the results is the high sample complexity of RL methods. The percentage improvement by IRL methods is higher in VisDial v1.0 as it has more training samples. This finding reaffirms the need for a larger training corpus to validate the effectiveness of RL based methods.

We now turn our attention to the observation where models without image features result in a competitive performance. Notice that this artefact has been studied at length by Massiceti *et al*. [54]. In our case, we also find that plausible dialogue policies can be learnt by exploiting the co-relations between the QA pairs in the data. However, this also means that the image and the caption are unaccounted for when generating responses at test time. We demonstrate this by a pathological example in fig. 5.2 where the agent can answer questions about the objects which are not present in the image. This observation calls into the question the ability of the current methods to learn representations useful for downstream tasks like dialogue. Consequently, in addition to overcoming the lack of well-defined evaluation metrics as argued by Massiceti *et al*., and Anand *et al*., [55] we need to better our understanding of the feature learning and combination methods for problems like image-based dialogue systems.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this thesis, we verify that with enough training data, inverse reinforcement learning in conjunction with min-max optimisation can effectively help in mitigating the issue of deflective and/or bland responses in image-based dialogue systems.

However, qualitative analysis also reveals that contemporary methods are biased towards ignoring the visual information and exploit surface level co-relations in the text. These biases go unnoticed due to the retrieval based evaluation metrics and cast doubt on the deployability of these methods for real-world use cases.

We, therefore, need to take a step back and rigorously analyse the current methods in simpler scenarios to assess their capabilities thoroughly.

6.2 Future Work

As mentioned in section 2.3 the sub-field of image-based dialogue is still in its infancy, and there is a wide scope for building on the work presented in this thesis. A lack of well-defined evaluation metrics is a problem widely acknowledged by the dialogue community. As image-based dialogue systems are grounded in the visual domain, a conversation on an image at any given time requires the agent to be proficient in various types of visual reasoning *e.g.*, a question about the colour of an object requires the agent to reason about the differences in colour. In contrast, the position of an object in an image requires the spatial reasoning; alternatively, questions like *Is the person in the room happy?* requires sophisticated situational awareness). Therefore, conducting large scale human evaluations can help ameliorate the evaluation problem. Categorising questions asked by test subjects into broader aspects of visual reasoning and evaluating the agent's performance individually on every category can help researchers understand what kind of questions are easy or hard for the agent to answer. It can help identify the categories in which techniques like spatial attention and FiLM are not effective, thereby highlighting specific areas of improvements in these methods or motivate the development of new techniques.

Alternatively, another interesting avenue is to improve the current representation learning methods by borrowing techniques from the sensor fusion literature. Sensor fusion literature has systematically studied techniques which combine data derived from disparate sources to generate representations with less uncertainty as compared to the source. Rigorously understood techniques like Fischer's method [56] and multi-sensor integration [57] can be used for algorithmic and architectural improvements in models used for fusing information from image and text modalities. These improvements can direct the agents to fulfil their actual utility of conveying information about visual stimuli.

References

[1] J. Gao, M. Galley, and L. Li, “Neural Approaches to Conversational AI,” *Foundations and Trends in Information Retrieval*, vol. 13, no. 2-3, pp. 127–298, 2019.

[2] M. Ghazvininejad, C. Brockett, M. Chang, B. Dolan, J. Gao, W. Yih, and M. Galley, “A Knowledge-Grounded Neural Conversation Model,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 5110–5117, 2018.

[3] H. B. Deng and M. Krstić, “Stochastic nonlinear stabilization—I: a backstepping design,” in *Systems and Control Letters*, 1997.

[4] R. E. Kalman, “When Is a Linear Control System Optimal?,” *Journal of Basic Engineering*, vol. 86, pp. 51–60, 03 1964.

[5] J. Ho and S. Ermon, “Generative Adversarial Imitation Learning,” in *NIPS*, 2016.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, “Generative Adversarial Nets,” in *NIPS*, 2014.

[7] J. Schmidhuber, “Learning Factorial Codes by Predictability Minimization,” *Neural Computation*, vol. 4, pp. 863–879, 1992.

[8] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, “Maximum Entropy Inverse Reinforcement Learning,” in *AAAI*, 2008.

[9] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, “Equation of state calculations by fast computing machines,” in *The Journal of Chemical Physics*, 1953.

[10] C. Finn, P. F. Christiano, P. Abbeel, and S. Levine, “A Connection between Generative Adversarial Networks, Inverse Reinforcement Learning, and Energy-Based Models,” *ArXiv*, vol. abs/1611.03852, 2016.

[11] Z. Li, J. Kiseleva, and M. de Rijke, "Dialogue Generation: From Imitation Learning to Inverse Reinforcement Learning," in *AAAI*, 2018.

[12] Q. Wu, P. Wang, C. Shen, I. D. Reid, and A. van den Hengel, "Are You Talking to Me? Reasoned Visual Dialog Generation Through Adversarial Learning," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6106–6115, 2017.

[13] I. G. Goodfellow, Y. Bengio, and A. C. Courville, "Deep Learning," *Nature*, vol. 521, pp. 436–444, 2015.

[14] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A Neural Probabilistic Language Model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, 2000.

[15] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *NIPS*, 2014.

[16] A. Ritter, C. Cherry, and W. B. Dolan, "Data-Driven Response Generation in Social Media," in *EMNLP*, 2011.

[17] L. Shang, Z. Lu, and H. Li, "Neural Responding Machine for Short-Text Conversation," in *ACL*, 2015.

[18] O. Vinyals and Q. V. Le, "A Neural Conversational Model," *ArXiv*, vol. abs/1506.05869, 2015.

[19] I. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models," in *AAAI*, 2015.

[20] I. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. C. Courville, and Y. Bengio, "A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues," in *AAAI*, 2016.

[21] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction," *IEEE Transactions on Neural Networks*, vol. 16, pp. 285–286, 1988.

[22] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao, "Deep Reinforcement Learning for Dialogue Generation," in *EMNLP*, 2016.

[23] N. Asghar, P. Poupart, X. Jiang, and H. Li, "Deep active learning for dialogue generation," in **SEM*, 2016.

[24] P. Su, M. Gasic, N. Mrksic, L. M. Rojas-Barahona, S. Ultes, D. Vandyke, T. Wen, and S. J. Young, "Continuously Learning Neural Dialogue Management," *CoRR*, vol. abs/1606.02689, 2016.

[25] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *ECCV*, 2014.

[26] M. Malinowski and M. Fritz, "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input," in *Advances in Neural Information Processing Systems 27* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), pp. 1682–1690, Curran Associates, Inc., 2014.

[27] M. Ren, R. Kiros, and R. S. Zemel, "Exploring Models and Data for Image Question Answering," in *NIPS*, 2015.

[28] H. Gao, 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 *NIPS*, 2015.

[29] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, "VQA: Visual Question Answering," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015.

[30] Y. Zhu, O. Groth, M. S. Bernstein, and L. Fei-Fei, "Visual7W: Grounded Question Answering in Images," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4995–5004, 2015.

[31] R. Krishna, Y. Zhu, O. Groth, J. M. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," *International Journal of Computer Vision*, vol. 123, pp. 32–73, 2016.

[32] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep Modular Co-Attention Networks for Visual Question Answering," in *CVPR*, 2019.

[33] K. Kafle and C. Kanan, "An Analysis of Visual Question Answering Algorithms," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1983–1991, 2017.

[34] H. de Vries, F. Strub, A. P. S. Chandar, O. Pietquin, H. Larochelle, and A. C. Courville, "GuessWhat?! Visual Object Discovery through Multi-modal Dialogue," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4466–4475, 2016.

[35] N. Mostafazadeh, C. Brockett, W. B. Dolan, M. Galley, J. Gao, G. P. Spithourakis, and L. Vanderwende, “Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation,” in *IJCNLP*, 2017.

[36] A. Das, 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*, 2017.

[37] J. Lu, A. Kannan, , J. Yang, D. Parikh, and D. Batra, “Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model,” *NIPS*, 2017.

[38] L. Yu, W. Zhang, J. Wang, and Y. Yu, “Seqgan: Sequence generative adversarial nets with policy gradient.,” in *AAAI*, pp. 2852–2858, 2017.

[39] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[40] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Machine Learning*, vol. 8, pp. 229–256, May 1992.

[41] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, “Gated Feedback Recurrent Neural Networks,” in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 2067–2075, 2015.

[42] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[43] A. Y. Ng and S. J. Russell, “Algorithms for Inverse Reinforcement Learning,” in *ICML*, 2000.

[44] C. Finn, S. Levine, and P. Abbeel, “Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization,” in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML’16, pp. 49–58, JMLR.org, 2016.

[45] B. D. Ziebart, J. A. Bagnell, and A. K. Dey, “Modeling Interaction via the Principle of Maximum Causal Entropy,” in *ICML*, 2010.

[46] V. Dumoulin, J. Shlens, and M. Kudlur, “A Learned Representation For Artistic Style,” in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[47] H. de Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville, "Modulating early visual processing by language," in *NIPS*, 2017.

[48] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "FiLM: Visual Reasoning with a General Conditioning Layer," in *AAAI*, 2017.

[49] J.-B. Delbrouck and S. Dupont, "Modulating and attending the source image during encoding improves Multimodal Translation," *Visually-Grounded Interaction and Language Workshop (NIPS) (2017)*, vol. abs/1712.03449, 2017.

[50] V. Nagarajan and J. Z. Kolter, "Gradient descent GAN optimization is locally stable," in *NIPS*, 2017.

[51] L. M. Mescheder, A. Geiger, and S. Nowozin, "Which Training Methods for GANs do actually Converge?," in *ICML*, 2018.

[52] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust Region Policy Optimisation," *CoRR*, vol. abs/1502.05477, 2015.

[53] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. C. Courville, and Y. Bengio, "An Actor-Critic Algorithm for Sequence Prediction," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[54] N. S. P. H. T. Daniela Massiceti, Puneet K. Dokania, "Visual dialogue without vision or dialogue," in *Critiquing and Correcting Trends in Machine Learning, NeurIPS 2018*, dec 2018.

[55] K. K. H. L. A. C. Ankesh Anand, Eugene Belilovsky, "Blindfold baselines for Embodied QA," in *Visually Grounded Interaction and Language (ViGIL) 2.0, NeurIPS 2018*, dec 2018.

[56] S. R. A. Fischer, *Statistical Method for Research Works*. Edinburgh, Oliver and Boyd, 1938.

[57] H. F. Durrant-Whyte, "Sensor Models and Multisensor Integration," *Autonomous Robot Vehicles*, vol. 1, pp. 73–89, 1990.