Leveraging Supervised Learning and Automated Reasoning for Robot Sequential Decision-Making

Saeid Amiri[1], **Mohammad Shokrolah Shirazi**[2], **Shiqi Zhang**[1]
[1] Department of Computer Science, SUNY Binghamton
[2] Department of Electrical Engineering and Computer Science, Cleveland State University
samiri1@binghamton.edu; m.shokrolahshirazi@csuohio.edu; szhang@cs.binghamton.edu

Abstract

Sequential decision-making (SDM) plays a key role in intelligent robotics, and can be realized in very different ways, such as supervised learning, automated reasoning, and probabilistic planning. The three families of methods follow different assumptions and have different (dis)advantages. In this work, we aim at a robot SDM framework that exploits the complementary features of learning, reasoning, and planning. We use long short-term memory (LSTM) for passive state estimation with streaming sensor data, and commonsense reasoning and probabilistic planning (CORPP) for active information collection and task accomplishment. In experiments, a mobile robot is tasked with estimating human intentions using human motion trajectories, declarative contextual knowledge, and human-robot interaction (dialog-based and motion-based). Results suggest that our framework performs better than its "no learning" and "no reasoning" versions in a real-world office environment.

1 Introduction

Mobile robots have been able to operate in everyday environments over extended periods of time, and travel long distances that have been impossible before, while providing services, such as escorting, guidance, and delivery [Hawes et al., 2017, Veloso, 2018, Khandelwal et al., 2017]. Sequential decision-making (SDM) plays a key role toward robot long-term autonomy, because real-world domains are stochastic, and a robot must repeatedly estimate the current state and decide what to do next.

We develop a robot SDM framework in this work, where robots are able to simultaneously learn from past experiences, reason with declarative contextual knowledge, and plan to achieve long-term goals under uncertainty.

We apply our general-purpose framework to the problem of *human intention estimation* using a mobile robot, as shown in Figure 1. The robot can observe human motion trajectories using streaming sensor data, has contextual knowledge (e.g., visitors tend to need guidance help), and is equipped with dialog-based and motion-based interaction capabilities. The goal is to identify human intention (e.g., human intending to interact or not) as accurate and early as possible. Note that human intention may change over time, and the robot wants to change its estimation accordingly.

Figure 1: Robot estimating human intention, e.g., human intending to interact or not, by analyzing human trajectories, reasoning with contextual knowledge (such as location and time), and active human-robot interaction.

There are at least three artificial intelligence (AI) paradigms, namely supervised learning, automated reasoning, and probabilistic planning, that can be used for SDM, but none of them completely meets the requirements in the context of robotics. I) A robot can learn to make decisions from previous experiences using supervised learning, e.g., learning from pairs of human trajectory and intention to decide whether to offer help or not. However, supervised learning can be biased; and the robot cannot make use of contextual knowledge or be active in this process. II) A robot can reason with rule-based contextual knowledge for decision-making, e.g., people show up in open-house events need guidance help, and students leaving classrooms do not. However, such knowledge can hardly be comprehensive; and a robot cannot actively seek information to recover from inaccurate or incomplete knowledge, or leverage previous experiences that is widely available in the big-data age. III) A robot can plan actions for active information collection and goal achievement, e.g., using decision-theoretic frameworks such as Markov decision processes (MDPs) [Puterman, 2014] and partially observable MDPs (POMDPs) [Kaelbling et al., 1998]. However, (PO)MDPs are not good at incorporating declarative knowledge.

In this work, we develop a robot SDM framework that exploits the complementary features of learning, reasoning, and planning in AI. Specifically, we use *long short-term memory* (LSTM) [Hochreiter and Schmidhuber, 1997] to learn a classifier for passive perception using streaming sensor data, and use *commonsense reasoning and probabilistic*

planning (CORPP) [Zhang and Stone, 2015] for active perception and task completions using contextual knowledge and human-robot interaction. We experimentally evaluate our approach using the human intention estimation problem. Results suggest that, in comparison to no-reasoning and no-learning baselines, integrating LSTM and CORPP improves accuracy and efficiency.

2 Background

We briefly overview long short-term memory (LSTM) neural network [Hochreiter and Schmidhuber, 1997] for supervised learning, and commonsense reasoning and probabilistic planning (CORPP) [Zhang and Stone, 2015].

2.1 LSTM

Recurrent neural networks (RNNs) are a kind of neural networks that use their internal state (memory) to process sequences of inputs. LSTM [Hochreiter and Schmidhuber, 1997] network, is a type of RNN that includes LSTM units. Each memory unit in the LSTM hidden layer has three gates for maintaining the unit state: input gate defines what information is added to the memory unit; output gate specifies what information is used as output; and forget gate defines what information can be removed. LSTMs use memory cells to resolve the problem of vanishing gradients, and is widely used in problems that require the use of long-term contextual information, e.g., speech recognition [Graves et al., 2013] and caption generation [Vinyals et al., 2015]. We use LSTM-based supervised learning for passive state estimation with streaming sensor data in this work.

2.2 CORPP

Commonsense reasoning and probabilistic planning (CORPP) is an algorithm that integrates automated reasoning and planning under uncertainty Zhang and Stone [2015]. The reasoning component represents and reasons with declarative contextual knowledge (both logical and probabilistic). The planning component's state space is used for computing an action policy that suggests actions toward achieving long-term goals. CORPP uses P-log [Baral et al., 2009, Balai and Gelfond, 2017] for knowledge representation and reasoning, and partially observable Markov decision processes (POMDPs) [Kaelbling et al., 1998] for probabilistic planning. In CORPP, the reasoning results are used to specify the state space and initial belief state for probabilistic planning.

P-log: Answer Set Prolog (ASP) is a logic programming paradigm that is strong in non-monotonic reasoning [Gelfond and Kahl, 2014, Lifschitz, 2016]. An ASP program includes a set of rules, each in the form of:

$$l_0 \leftarrow l_1,\ \cdots,\ l_n,\ \text{not}\ l_k,\ \cdots,\ \text{not}\ l_{n+k}$$

where l's are literals that represent whether a statement is true or not, and symbol *not* is called default negation. The right side of a rule is the *body*, and the left side is the *head*. A rule reads the head is true if the body is true.

P-log extends ASP by allowing probabilistic rules for quantitative reasoning. A P-log program consists of logical and probabilistic part. The logical part inherits the syntax and semantics of ASP. The probabilistic part contains *pr-atoms* in the form of:

$$pr_r(a(t) = y|B) = v$$

where $a(t)$ is a random variable, B is a set of literals and $v \in [0, 1]$. The pr-atom states that, if B holds and experiment r is fixed, the probability of $a(t) = y$ is v. Reasoning with a P-log program produces a set of possible worlds, and a distribution over the possible worlds.

There are representations for probabilistic inference that build on first-order logic (FOL), such as Probabilistic Soft Logic (PSL) [Bach et al., 2017] and Markov Logic Network (MLN) [Richardson and Domingos, 2006]. P-log directly takes probabilistic, declarative knowledge as the input, whereas PSL and MLN use data to learn weights of FOL rules. Informally, P-log is good at incorporating (declarative) human knowledge, and PSL and MLN are strong in learning from data for probabilistic inference.

POMDPs: Markov decision processes (MDPs) can be used for sequential decision-making under full observability. Partially observable MDPs (POMDPs) [Kaelbling et al., 1998] generalize MDPs by assuming partial observability of the current state. A POMDP model is represented as a tuple $(S, A, T, R, Z, O, \gamma)$ where S is the state-space, A is the action set, T is the state-transition function, R is the reward function, O is the observation function, Z is the observation set and γ is discount factor that determines the planning horizon.

An agent maintains a belief state distribution b with observations ($z \in Z$) using the Bayes update rule:

$$b'(s') = \frac{O(s', a, z) \sum_{s \in S} T(s, a, s') b(s)}{pr(z|a, b)}$$

where s is the state, a is the action, $pr(z|a, b)$ is a normalizer, and z is an observation. Solving a POMDP produces a policy that maps the current belief state distribution to an action toward maximizing long-term utilities.

3 Framework

We develop a robot SDM framework that tightly couples the LSTM-based supervised learning, and CORPP-based reasoning and planning. Streaming sensor data, e.g., from RGB-D sensors, is fed into a LSTM-based classifier. The classifier's output is provided to the reasoner. The reasoner reasons with declarative contextual knowledge, the classifier's output, and the classifier's accuracies. The reasoner produces a prior belief distribution over all possible states for the probabilistic planner. The planner suggests actions to enable the robot to actively interact with people, and determines when and what (estimated intention) to report. Figure 2 is an overview of our robot SDM framework.

In the following subsections, we explain in detail how each of these components are develped for the problem of human intention estimation.

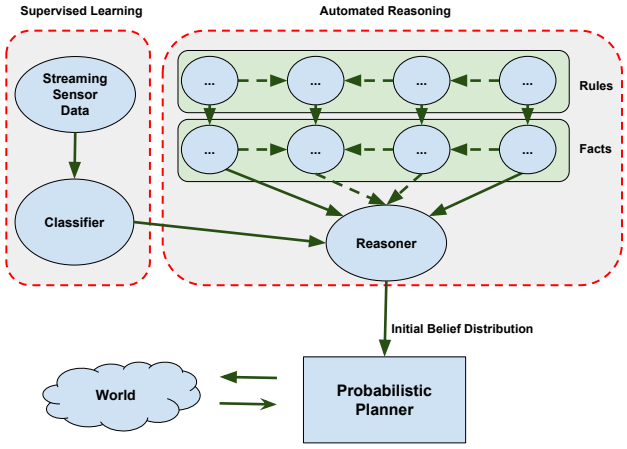

Figure 2: An overview of our robot SDM framework that integrates supervised learning, automated reasoning, and probabilistic planning.

3.1 Passive perception with streaming sensor data

We train a classifier to make passive intention estimations using a dataset from the literature [Kato et al., 2015]. Since the human trajectories are in the form of time sequence data, we use LSTM to train a classifier for estimating human intentions based on human motion trajectory. Next we present some implementation details of our LSTM network.

Features of the input vectors include the x and y components of human motion trajectories. The input vector length is 60 including 30 pairs of x and y values. Our LSTM hidden layer includes 50 memory units. In order to output binary classification results, we use a dense layer with sigmoid activation function in the output layer.

We use Adam [Kingma and Ba, 2014], a first-order gradient method, for optimization. The loss function was calculated using binary cross entropy. For regularization, we use a dropout value of 0.2. The memory units and the hidden states of the LSTM are initialized to zero. The epoch size (number of passes over the entire dataset) is 300. The batch size is 30. The data was split into sets for training (70%) and testing (30%).

3.2 Reasoning with contextual knowledge

Domain knowledge provided by an expert human can help the robot make better estimations. For instance, in the early mornings of work days, people are less likely to be interested in interacting with the robot, in comparison to the university open-house days. The main purpose of the reasoning component is to incorporate such contextual knowledge into passive state estimation with sensor readings.

Our reasoning program contains random variables $\{location,\ time,\ \cdots,\ intention\}$, where the range of each variable is defined as below:

$$Location : \{classroom, library\}$$
$$Time : \{morning, afternoon, evening\}$$
$$Identity : \{student, professor, visitor\}$$
$$Intention : \{interested, notinterested\}$$

We further include probabilistic rules into the reasoning component. For instance, the following two rules state that the probability of a visitor showing up in the afternoon is 0.7, and the probability of a professor showing up in the library (instead of other places) is 0.1, respectively.

$$pr(time = afternoon|identity = visitor) = 0.7.$$
$$pr(location = library|identity = professor) = 0.1.$$

It should be noted that *time* and *location* are facts that are fully observable to the robot, whereas human *identity* is a hidden variable that must be estimated using observations. It is necessary to introduce *identity*, because there is no direct causal relation between time (and location) and human intention. Instead, time and location probabilistically determine human identity, which then probabilistically determines human intention, as shown in Figure 3.

For instance, when time is *afternoon*, location is *library*, and the LSTM-based classifier outputs *positive* (meaning human motion trajectory suggests the human is interested in interaction), our probabilistic reasoner infers the following distribution over possible identities.

$$[student = 0.16, visitor = 0.48, professor = 0.36]$$

where $pr(identity = student)$ is small because 1) from human knowledge, it is unlikely that students are interested in interacting with the robot to get guidance service, and 2) the LSTM-based classifier suggests the human is interested.

According to the above distribution (over identities), our reasoner will generate the posterior distribution over human intentions. Finally, this binary distribution (over whether human being interested in interaction or not) is provided to the POMDP-based planner as informative priors.

3.3 Active Perception via POMDP-based HRI

Robots can take actions to reach out to people and actively gather information. We use POMDPs to build probabilistic controllers.

- $S : S_i \times S_l \cup \{term\}$ is the factored state space. S_i includes two states representing human being interested to interact with the robot or not. S_l includes two states representing whether the robot has greeted the human or not. $term$ is the terminal state.
- $A : A_a \cup A_r$ is the set of actions. A_a are perception actions including turning to the human, greeting, and slightly move toward the human. A_r includes two actions for reporting the human being interested in interaction or not.
- $Z : Z_v \cup Z_p \cup \{not\ applicable\}$ is the observation set where Z_v contains human verbal feedback and Z_p is the set of human physical reactions, such as turning toward the robot.

The transition and reward functions are defined accordingly. Reporting actions deterministically lead to the $term$ state. Reporting human intention yields a big bonus or a big penalty, depending on the report being correct or not. Each perception action $a \in A_a$ has a small cost that is in the form of a small negative reward. We use the discount factor

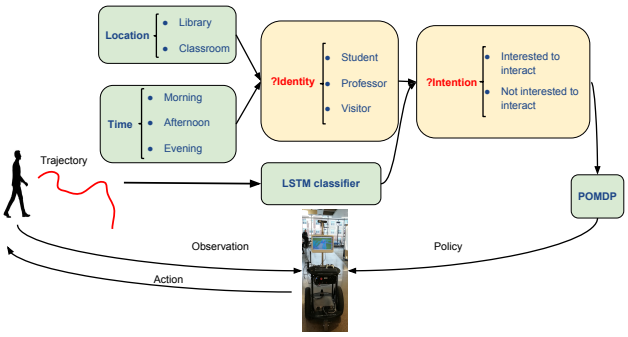

Figure 3: The schematic representation of human intention estimation problem based on the motion trajectory.

$\gamma = 0.99$ to give the robot a relatively long planning horizon. Using an off-the-shelf solver [Kurniawati et al., 2008], the robot can generate a behavioral policy that maps its belief state to an action toward efficiently and accurately estimating human intentions.

Recap: The robot's LSTM-based classifier estimates human intention based on the human trajectories. Part of the structured knowledge in the reasoning component is shown in Figure 3. Facts of time and location are fully observable, and can be used for reasoning about human identity, and intention. Finally, the POMDP-based planner maintains an initial belief distribution over all possible states, where the reasoning results serve as informative prior beliefs.

The reasoning and planning components of CORPP are constructed using human knowledge, and do not involve learning. The reasoning component aims at correcting and unbiasing outputs of the LSTM-based classifiers, and the planning component is for active perception. Possible improvements are discussed in Section 6.

4 Experiments

We have conducted experiments in simulation to evaluate the necessity of integrating supervised learning, automated reasoning, and probabilistic planning. Baseline methods include the no-learning, and no-reasoning versions of the developed framework.

4.1 Human motion trajectory dataset

We use a publicly available dataset [Kato et al., 2015] to train the LSTM-based classifier. Each sample in the dataset includes a human motion trajectory in 2D space, and a label of whether the human eventually interacts with the robot or not. There are totally 2286 instances in the dataset, where 63 are positive instances (2.7%), and 2223 are negative (97.3%). Each trajectory includes a sequence of data fields with the sampling rate of 33 milliseconds. Each data field is in the form of a vector: $(x_i,\ y_i,\ z_i,\ v_i,\ \theta_m,\ \theta_h)$. Index i denotes the timestep. x_i and y_i are the coordinates in millimeter. z_i is human height. v_i is human linear velocity in mm/s. θ_m is the motion angle in $radius$. θ_h is the face orientation in $radius$. They used multiple 3D range sensors mounted on ceilings to track human motion trajectories and collect the dataset [Brscic et al., 2013].

While all data fields are potentially useful for training the classifiers, we only use the features of x and y coordinates because of the limitations of our robot's perception capabilities.

Table 1: Experiment results using five different approaches.

Method	Accuracy	Precision	Recall	F1 score
Learning	0.61	0.56	0.30	0.39
Reasoning	0.60	0.54	0.62	0.58
Learning + Reasoning	0.58	0.51	0.72	0.60
CORRP	0.79	0.67	0.94	0.78
Learning + CORPP (ours)	0.83	0.74	0.86	0.80

4.2 Simulation

We did pairwise comparisons of the following methods, where each data point corresponds to 500 trials. **Learning:** the robot only uses its LSTM-based classifier to passively estimate human intention. **Reasoning:** the robot uses only contextual knowledge to reason about human intention. **Learning+reasoning:** the robot reasons about the classifier's output and contextual knowledge. **Reasoning+planning (CORPP):** the robot uses contextual knowledge to compute priors for POMDP-based planning. **Learning+CORPP (ours):** the framework developed in this work.

In each simulation trial, we first randomly generate a sample of human identity. We then sample time, and location according to our prior knowledge (in the form of distributions) of how likely people show up in different times and location. After that, we sample human intention according to time, location, and identity. Finally, we sample a trajectory from the dataset of human trajectories according to human intention.

Table 1 shows the results from simulation experiments, where the developed framework that includes learning, reasoning, and planning produces the best overall performance in F1 score — the F1 score is a harmonic average of the precision and recall. Another observation is that our approach that combines supervised learning and CORPP requires lower costs (13.1) in human-robot interaction, i.e., dialog-based and motion-based, in comparison to CORPP (21.6). This suggests that our approach enables the robot to take less perception actions (c.f., CORPP), while producing higher F1 scores in human intention estimation.

5 Related Work

This work is related to existing research that incorporates knowledge representation and reasoning (KRR) into sequential decision-making (SDM) in stochastic worlds. SDM can be realized via either probabilistic planning (e.g., MDPs and POMDPs [Kaelbling et al., 1998]) or reinforcement learning (RL) [Sutton et al., 1998].

When world model is unavailable, one can use RL algorithms to learn an action policy. Declarative action knowledge has been used to help an agent to select only the reasonable actions in RL exploration [Leonetti et al., 2016].

Researchers have developed an algorithm called PEORL that integrates hierarchical RL with task planning [Yang et al., 2018]. In that work, RL (low-level) helps learn action costs for the task planner, and the task planner guides RL (high-level) to accomplish complex tasks. These works cannot learn complex representations from previous annotated decision-making experiences.

In case of world model being available, probabilistic planning methods can be used for computing action policies. Contextual knowledge and logical reasoning have been used to help better estimate the current world state in probabilistic planning [Zhang et al., 2015]. Hybrid reasoning (both logical and probabilistic) was used to guide probabilistic planning by calculating a probability for each possible state, producing an algorithm called CORPP [Zhang and Stone, 2015]. We use CORPP in this work. More recently, hybrid reasoning has been used to reason about world dynamics [Zhang et al., 2017], enabling planners to generate behaviors that are adaptive to dynamic world dynamics. Sridharan et al. [2015] developed a refinement-based architecture, where declarative knowledge is used for task planning and reasoning tasks, such as diagnosis and history explanation, and high-level deterministic plans are implemented via probabilistic planners. Very recently, researchers have used human-provided declarative information to improve robot probabilistic planning [Chitnis et al., 2018]. Learning was not involved in these works.

To the best of our knowledge, this is the first work that simultaneously supports supervised learning for passive perception, automated reasoning with contextual knowledge, and active information gathering via probabilistic planning.

6 Conclusions and Future Work

In this work, we develop a robot sequential decision-making framework that integrates supervised learning for passive state estimation, automated reasoning for incorporating declarative contextual knowledge, and probabilistic planning for active perception and task completions. The developed framework has been applied to a human intention estimation problem using a mobile robot. Results suggest that the integration of supervised deep learning, logical-probabilistic reasoning, and probabilistic planning enables simultaneous passive and active state estimation, producing the best performance in estimating human intentions.

In the future, we plan to implement this framework on a mobile robot, where in particular, we will evaluate the real-time performance of our system. There is the potential of applying learning algorithms into our reasoning and planning components. For instance, model-based reinforcement learning algorithms can be used to learn world dynamics to update parameters of our planning component [Lu et al., 2018], and data mining algorithms [Han et al., 2011] can be used to learn probabilistic reasoning rules that can be formalized using P-log.

References

Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *The Journal of Machine Learning Research*, 18(1):3846–3912, 2017.

Evgenii Balai and Michael Gelfond. Refining and generalizing p-log–preliminary report. In *Proceedings of the 10th Workshop on Answer Set Programming and Other Computing Paradigms*, 2017.

Chitta Baral, Michael Gelfond, and Nelson Rushton. Probabilistic reasoning with answer sets. *Theory and Practice of Logic Programming*, 9(1):57–144, 2009.

Drazen Brscic, Takayuki Kanda, Tetsushi Ikeda, and Takahiro Miyashita. Person tracking in large public spaces using 3-d range sensors. *IEEE Transactions on Human-Machine Systems*, 43(6):522–534, 2013.

Rohan Chitnis, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Integrating human-provided information into belief state representation using dynamic factorization. *arXiv preprint arXiv:1803.00119*, 2018.

Michael Gelfond and Yulia Kahl. *Knowledge representation, reasoning, and the design of intelligent agents: The answer-set programming approach*. Cambridge University Press, 2014.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

Nick Hawes, Christopher Burbridge, Ferdian Jovan, Lars Kunze, Bruno Lacerda, Lenka Mudrova, Jay Young, Jeremy Wyatt, et al. The strands project: Long-term autonomy in everyday environments. *IEEE Robotics & Automation Magazine*, 24(3):146–156, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Yusuke Kato, Takayuki Kanda, and Hiroshi Ishiguro. May i help you?: Design of human-like polite approaching behavior. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 35–42. ACM, 2015.

Piyush Khandelwal, Shiqi Zhang, Jivko Sinapov, Matteo Leonetti, Jesse Thomason, Fangkai Yang, et al. BWIBots: A platform for bridging the gap between AI and human–robot interaction research. *The International Journal of Robotics Research*, 36(5-7):635–659, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Hanna Kurniawati, David Hsu, and Wee Sun Lee. Sarsop: Efficient point-based pomdp planning by approximating

optimally reachable belief spaces. In *Robotics: Science and Systems*, volume 2008. Zurich, Switzerland., 2008.

Matteo Leonetti, Luca Iocchi, and Peter Stone. A synthesis of automated planning and reinforcement learning for efficient, robust decision-making. *Artificial Intelligence*, 241:103–130, 2016.

Vladimir Lifschitz. Answer sets and the language of answer set programming. *AI Magazine*, 37(3):7–12, 2016.

Keting Lu, Shiqi Zhang, Peter Stone, and Xiaoping Chen. Robot representing and reasoning with knowledge from reinforcement learning. *arXiv preprint arXiv:1809.11074*, 2018.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1-2):107–136, 2006.

Mohan Sridharan, Michael Gelfond, Shiqi Zhang, and Jeremy Wyatt. A refinement-based architecture for knowledge representation and reasoning in robotics. *arXiv preprint arXiv:1508.03891*, 2015.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*. 1998.

Manuela M Veloso. The increasingly fascinating opportunity for human-robot-ai interaction: The cobot mobile service robots. *ACM Transactions on Human-Robot Interaction (THRI)*, 7(1):5, 2018.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

Fangkai Yang, Daoming Lyu, Bo Liu, and Steven Gustafson. Peorl: Integrating symbolic planning and hierarchical reinforcement learning for robust decision-making. *arXiv preprint arXiv:1804.07779*, 2018.

Shiqi Zhang and Peter Stone. Corpp: Commonsense reasoning and probabilistic planning, as applied to dialog with a mobile robot. In *AAAI*, pages 1394–1400, 2015.

Shiqi Zhang, Mohan Sridharan, and Jeremy L Wyatt. Mixed logical inference and probabilistic planning for robots in unreliable worlds. *IEEE Transactions on Robotics*, 31(3): 699–713, 2015.

Shiqi Zhang, Piyush Khandelwal, and Peter Stone. Dynamically constructed (po) mdps for adaptive robot planning. In *AAAI*, pages 3855–3863, 2017.