GROUNDED LANGUAGE PROCESSING FOR ACTION UNDERSTANDING AND JUSTIFICATION

By

Shaohua Yang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science — Doctor of Philosophy

2019

ABSTRACT

GROUNDED LANGUAGE PROCESSING FOR ACTION UNDERSTANDING AND JUSTIFICATION

By

Shaohua Yang

Recent years have witnessed an increasing interest on cognitive robots entering into our life. In order to reason, collaborate and communicate with human in the shared physical world, the agents need to understand the meaning of human language, especially the actions, and connect them to the physical world. Furthermore, to make the communication more transparent and trustworthy, the agents should have human-like action justification ability to explain their decision-making behaviors. The goal of this dissertation is to develop approaches that learns to understand actions in the perceived world through language communication. Towards this goal, we study three related problems[1].

Semantic role labeling captures semantic roles (or participants) such as agent, patient and theme associated with verbs from text. While it provides important intermediate semantic representations for many traditional NLP tasks, it does not capture grounded semantics with which an artificial agent can reason, learn, and perform the actions. We utilize semantic role labeling to connect the visual semantics with linguistic semantics. On one hand, this structured semantic representation can help extend the traditional visual scene understanding instead of simply object recognition and relation detection, which is important for achieving human robot collaboration tasks. On the other hand, due to the shared common ground, not every language instruction is fully specified explicitly. We proposed to not only ground explicit semantic roles, but also implicit roles which is hidden

[1] This dissertation was supported by IIS-1617682 from the National Science Foundation and the DARPA XAI program under a subcontract from UCLA (N66001-17-2-4029)

during the communication. Our empirical results have shown that by incorporate the semantic information, we achieve better grounding performance, and also a better semantic representation of the visual world.

Another challenge for an agent is to explain to human why it recognizes what's going on as a certain action. With the recent advance of deep learning, A lot of works have shown to be very effective on action recognition. But most of them function like black-box models and have no interpretations of the decisions which are given. To enable collaboration and communication between humans and agents, we developed a generative conditional variational autoencoder (CVAE) approach which allows the agent to learn to acquire commonsense evidence for action justification. Our empirical results have shown that, compared to a typical attention-based model, CVAE has a significantly higher explanation ability in terms of identifying correct commonsense evidence to justify perceived actions. The experiment on communication grounding further shows that the commonsense evidence identified by CVAE can be communicated to humans to achieve a significantly higher common ground between humans and agents.

The third problem combines the action grounding with action justification in the context of visual commonsense reasoning. Humans have tremendous visual commonsense knowledge to answer the question and justify the rationale, but the agent does not. On one hand, this process requires the agent to jointly ground both the answers and rationales to the images. On the other hand, it also requires the agent to learn the relation between the answer and the rationale. We propose a deep factorized model to have a better understanding of the relations between the image, question, answer and rationale. Our empirical results have shown that the proposed model outperforms strong baselines in the overall performance. By explicitly modeling factors of language grounding and commonsense reasoning, the proposed model provides a better understanding of effects of these factors on grounded action justification.

ACKNOWLEDGMENTS

First and foremost, I am tremendously grateful for my advisor Dr. Joyce Y. Chai for her continuous support and guidance. She shared with me how to think critically, explore new problems, asking good questions and how to do good research. All of these experiences will have a great influence on my whole life. Besides, her great insights on the domain of human robot interaction and action understanding have always shed light on problems I have been working on. Without her continuous advice, inspiration and guidance for my PhD. study, this work would have been impossible.

I would also like to thank my dissertation committee members: Dr. Arun Ross, Dr. Xiaoming Liu and Dr. Taosheng Liu. I greatly appreciate their valuable feedback on every step of my PhD journey.

I'm very happy to have had the opportunity to collaborate with an amazing group of students and researchers: Dr. Changsong Liu, Dr. Lanbo She and Dr. Rui Fang provide great suggestions and directions when I start my research career as a PhD student. Thanks to Dr. Qiaozi Gao and Sari Saba-Sadiya for their great efforts and enlightening comments. I also appreciate my co-authors on various papers: Dr. Yu Cheng, Dr. Yunyi Jia, Dr. Ning Xi, Dr. Caiming Xiong, Dr. Songchun Zhu, Malcolm Doering, Nishant Shukla, Yunzhong He, Guangyue Xu and Dr. Lucy Vanderwende.

I would like to thank all my friends at MSU, who made my time at MSU enjoyable.

Finally, this thesis is dedicated to my parents, for all the years of your selfless love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Background

It has been a long dream to have intelligent agents which can help and collaborate with humans in everyday life such as house keeping, medical assistance and so on. Imagine such a scene: you can instruct a robot to finish some cooking task in the kitchen in natural language. When you say "take out the vegetable", the agent can quickly move to the location where the vegetable is and bring it back to you. In order for the agent to successfully complete this task, it needs to have advanced perception and reasoning abilities.

Compared with traditional robots which are hard-coded to finish specific tasks, this new generation of robots need to adapt to dynamic tasks and environments. That is to say, we cannot fixed the programmed procedures. It's impossible for us to enumerate all the possible tasks and environments to write programs for them. Instead we need to develop robots which can learn and generalize to a new environment and new tasks through language communications.

To achieve this goal, we need to develop intelligent agents which can connect natural language and the robot's perception. During this process, recognizing and understanding actions are of significant importance as verbs constitute a core part of natural language instructions. Although there has been a lot of research on object recognition and referential grounding, less work has been done to bridge the structured linguistic semantics and visual perception, which is essential for human robot interaction.

Machine learning approaches, especially deep learning approaches have achieved exciting performances on various tasks. However, they often function as a black box and are hard to interpret models' behaviors. To address this problem, recent years have witnessed an increasing interesting on explainable artificial intelligence (XAI). The goal is to make the black-box machine learning models more transparent and trustworthy. For example, in the medical diagnosis field, it's helpful for the agents to provide some evidence or justifications to the doctors regarding any predictions or recommendations made by the agent. Some recent work also tries to do post-hoc analysis to understand the behaviors of neural networks through feature visualization, attention visualization and so on. Different from post-hoc explanations, we propose to jointly model the the prediction and justification process as they are often coupled in the human cognitive process. In this way, the agent not only provides the predictions, but also their reasoning justifications. Such justification will allow human users to better interpret machines' behaviors and enable more mutual understanding and common ground.

In the following sections of this chapter, we first identify what are the research challenges. Then we describe our contributions towards the goal of grounded action understanding and justification.

1.2 Challenges

As discussed in the background section, it is important for robots not only to automatically connect the visual world to natural language, but also to infer justifications behind decision making. To address this issue, the following questions need to be addressed:

1. Natural language is represented as discrete and symbolic word sequences, but the robots' surrounding worlds are commonly continuous in nature such as image and videos. There

exists a big gap between representations of natural language and the visual world. How to connect the discrete language representation and the continuous visual representation so that AI agents can understand the meanings of language with respect to the physical world?

2. During human human conversation, some information is not be explicitly mentioned as it is commonly known and shared by each other. However, for the robot which is lack of commonsense knowledge, it's important to infer the implicit information to get a comprehensive understanding of human language. Is it possible to acquire the implicit knowledge and how to acquire it still remain open problems.

3. To allow agents to reason and justify recognized actions, the first key step is to understand how humans make justifications and what strategies humans apply to justify a perceived action. Exploring an effective action justification representation is of significant importance to further problem formulation.

4. The relations between the justifications and predictions are complex. On one hand, the justifications (or evidence) provide strong supports for the predicted action; on the other hand, the action prediction gives more contexts on what justifications to select. The physical world contains multiple actions where different actions may have different justifications, so jointly modeling the action prediction and justification is really important to tackle the diversities between the actions and justifications. Furthermore, the annotation process of action justifications is really time consuming and expensive, whether we can build models to alleviate human efforts to learn to predict and justify efficiently?

5. Finally, instead of understanding action prediction and justification purely based on textual relations, it's more meaningful to connect action prediction and action justification with the visual world. The agent not only needs to ground the target action, but also the action

justifications. How can we model the relations between the visual world, actions and action justifications to allow agents to get a comprehensive understanding of actions?

1.3 Contributions

Towards the goal of building agents that can ground action semantics and action justifications, the contributions of this dissertation are listed as follows:

1. To understand natural language, a variety of semantic representations are proposed and studied by linguists. One of the semantic representations is distributed semantics which represents each word as a continuous vector in a high dimension space, but it is limited in learning structured relations between different words or entities in the sentence. To explicitly model the relations, the frame-based verb semantics defines a frame of thematic roles (also referred to as semantic roles or verb arguments) for each verb to capture the semantics for different verbs [52]. For example, a verb can be characterized by *agent* (i.e., the animator of the action) and *patient* (i.e., the object on which the action is acted upon), and other roles such as *instrument*, *source*, *destination*, etc. Given a verb frame, the goal of Semantic Role Labeling is to identify linguistic entities from the text that serve different thematic roles [14, 28, 67, 111]. For instance, given the sentence *the woman takes out a cucumber from the refrigerator.*, *takes out* is the main verb (also called predicate); the noun phrase *the woman* is the agent of this action; *a cucumber* is the patient; and *the refrigerator* is the source. The frame-based verb semantics capture the verb meaning explicitly by connecting with other words/entities, which is easier for humans to interpret and has a more fine-grained structure. The linguists have developed several large frame-based knowledge bases including VerbNet [85], FrameNet [3], and Propbank [45] et al. However, based on pure symbolic

representations, It's difficult for the robot to connect texts to the situated world. For example, in the previous example, the robot needs to understand what the cucumber means, where is it, and what's the relation between the object cucumber and action takes out. To overcome this limitation, We developed an approach to jointly understand language and vision by incorporating linguistic semantic role information. To be specific, we propose a probabilistic graphical model to ground each semantic role to the possible trackings in the perceived world. In this way, we connect the low-level image pixels to high-level linguistic structure.

2. During human human conversation, it's not always the case that all content is explicitly stated as some experience or information is assumed known by each other. As in the previous example, the semantic role *destination* is not explicitly mentioned in the human instruction, but it is important to allow the agent to execute the action. Motivated by this phenomenon, we simultaneously grounds both explicit and implicit semantic roles. Although the destination is missing, its grounding closely depends on other roles' groundings including the action and patient and so on. Filling all the semantic slots of the physical actions is especially important for the robot to build the semantic map and conduct planing to connect with low-level actions.

3. Our empirical results on grounded semantic role labeling demonstrate a significant performance improvement compared with previous benchmark without modeling semantic context information and implicit semantic roles. Besides, we collect an additional layer of annotation on top of part of the TACOS dataset which captures the structure of actions informed by semantic roles from the video. The annotated data is publicly available for download [1]. It

[1] https://github.com/yangshao/gsrl

will provide a benchmark for future work on the grounded semantic role labeling task.

4. Before diving into the action justification problem, we need to first identify key structures of human justifications for action recognition. We conducted a human study to collect real human justifications. After some careful manual analysis, we identify several key dimensions of commonsense knowledge, from a human's perspective, to justify concrete actions in the physical environment. These dimensions provide an important basis to justify explanations that are aligned with human's commonsense knowledge about actions. More importantly, we can make use of these key dimensions to derive useful structured representations for action justification modeling.

5. We proposed an unsupervised conditional variational autoencoder (CVAE) based method to jointly learn to predict actions and select commonsense evidence as action justification. The CVAE model naturally models the generation process of both action prediction and commonsense evidence selection. Inferring commonsense evidence is equivalent to the posterior inference of the CVAE model: what are the possible support given the predicted action, which is flexible and powerful to incorporate actions as context. Inferring actions can be seen as the forward process by first selecting evidence, then using the evidence to do prediction. These two processes are jointly learned in our proposed framework. Furthermore we extend the unsupervised setting to the semi-supervised setting which adds supervisions to the latent commonsense evidence to verify whether it can improve both the action prediction and action justification performance. Our empirical experimental results show better performance on both the action prediction and justification. To test the communication grounding efficiency, we design human studies to show that our method can achieve higher communication grounding compared with strong baselines. The dataset will be made available to the

community, which will serve as a baseline for the future work on this topic.

6. Despite the success of the joint modeling of action prediction and justification, our previous work is limited in simple scenarios which only contain limited actions which do not have complex justifications. In addition, our previous work made a strong assumption that all justifications are pre-extracted from the image, which is not realistic. To overcome these limitations, we propose a joint factorized model by extending the traditional visual question answering task with extra justification inference. In this new setting, the question is designed for complex actions which may not limited to only one verb, but may contain multiple actions with different arguments. At the same time, the correct justification not only needs to explain the answer well, but also be grounded to the visual world. This requires the agent to understand the joint relation between the image, question, answer and justification. Compared with previous works based on a two step inference process, we factorize the complex interaction into small local interactions, our empirical experimental results demonstrate the effectiveness of the proposed factorized modeling. In addition, we carefully analyzed how different factors influence the final performance through detailed ablation studies.

1.4 Organization of Dissertation

The rest of chapters are organized as follows. Related works are introduced in Chapter 2. Then in Chapter 3, we detail how to formulate the grounded semantic role labeling problem as well as how we collect the dataset for this problem. Then we formally introduced the graphical model used to solve this problem. A series of experiments are conducted to demonstrate the effective of the proposed method. Chapter 4 introduces the joint action recognition and justification problem, as well as how different variants of methods can be used to alleviate the data annotation effort.

Detailed experimental results are shown to prove the effectiveness of the proposed method. In Chapter 5, we show how to use deep factorized model to solve the grounded action justification problem, Detailed ablation study results are shown to demonstrate how different factors help for the final joint grounded action recognition and justification problem. Finally we discuss some possible future directions in Chapter 6.

Chapter 2

Related Work

Learning to understand grounded action meanings is related to multiple research areas, from traditional linguistic studies on verb semantics, to grounded language learning, explainable artificial intelligence and commonsense reasoning. In this chapter, we discuss some related works in these areas.

2.1 Verb Semantics

How to represent the meaning of words, sentences or even documents has been a long studied problem in natural language processing [3, 45, 52, 72]. Actions, especially verbs, usually indicate some happening events in the physical world. Verbs are one of the most important components as they act as key parts to connect with other different components including nouns, adverbs and so on. According to the linguistic theory of verbs [35], the action verbs are mainly divided into two sub-categories: manner verbs and result verbs. The manner verbs are defined as "*verbs specify as part of their meaning a manner of carrying out an action*", and the result verbs are "*verbs specify the coming about of a result state*". Example manner verbs include *nibble, rub, laugh* and so on. Example result verbs contain *clean, fill, chop* and so on. Different kinds of verbs show different properties. For example, compared with the manner verbs, the result verbs have more obvious state of changes to indicate the possible action justifications. While for manner verbs, sub-actions are important indicators for the commonsense evidence.

In our work, we are interested in how to represent and understand action meanings.

2.1.1 Distributed Semantics

One of the established semantic categories is distributed semantics: The words/verbs are represented as continuous vectors in a high dimension space. The core idea behind the distributed semantics is that the word meaning is closely related with words occurring in similar linguistic contexts.

One of the earliest distributed representations is the bag-of-words representation. For a fixed vocabulary V which contains n unique words. Each word $w \in V$ can be represented as a vector v of which $v_i = 1$ and all other values are 0. i is the corresponding word index of w in the vocabulary V. The problem with the bag-of-words representation is that it's difficult to reflect the similarity between similar words.

To alleviate this limitation, some works [64] try to represent the words using lower dimension continuous vectors thus making similar words are close in the space. Some Other works use matrix factorization based methods for the distributed representation [33, 50]. Topic Model [6] is a specific generative bayesian method to model the relations between the words and documents based on latent topics.

With the recent advances of deep learning in natural language processing, many new algorithms are proposed to learn effective word representations. Word2vec [62] and Glove [69] are two of the most popular methods to utilize the large unlabeled text to learn distributed representations of words based on the distributed assumption. However, both of them are context independent meaning representations whose meanings do not depend on the specific context. Even more recently, ELMO [70] and Bert [18] are proposed to extend the context independent representations to context depended representations: For the same word, if they appears in different contexts, they may have different meanings. Nowadays, the distributed pre-trained word embeddings have been widely used for most natural language processing tasks. One of the main reasons is that the pre-trained word embeddings carry a lot of commonsense knowledge which can be beneficial to the target task which is beneficial to alleviate the large dataset requirement of deep learning.

2.1.2 Semantic Role Labeling

Although the distributed semantics is a good choice for exploring word similarities, it's hard to explore fine grained relations between words or entities.

To understand the semantics of actions/verbs, a more structured way is to identify the semantic relations between entities and the events they participate in. To be concrete, our goal is to identify *who did what to whom when and where* given the natural sentence. The linguists proposed frame-based verb semantic representation for this purpose: for each verb sense(one verb may contain more than one sense), a frame including different slots is defined as a template whose values are instantiated for a specific sentence describing the situation where the verb is happening. For example, the slots of the verb cut contains *agent*, *patient*, *source*, *instrument* and so on.

To facilitate the study of semantic role labeling, the researchers have developed semantic grammars and manually annotated linguistic resources including several large scale frame-based knowledge bases including VerbNet [85], FrameNet [3] and PropBank [45] and so on. These linguistic resources greatly accelerate a variety of statistical approaches towards semantic role labeling tasks [14, 67, 71, 111]. For example, the sentence: "*the man is cutting the vegetable with the knife*" contains the *predicate* cutting, the *patient* vegetable, and the *tool* knife. The semantic role labeling has been widely used on a lot of natural language processing applications including information extraction [22], question answering [86], summarization [42] et al.

2.2 Grounded Language Learning

Traditional semantic role labeling plays a very important role in many applications in natural language processing. However, it doesn't ground to the physical world, which makes the agents hardly understand the situation and then to perform the specified action.

Recent years have witnessed an increasing amount of works on muti-modal learning integrating language and vision including image annotation [40,73], image/video caption generation [19,21,48,66,96], video sentence alignment [59,65], scene generation [8], and multi-modal embedding incorporating language and vision [7,51].

One of the fundamental tasks in grounded language learning is to associate words with perceptual input. Words are discrete symbols and perceptions are usually represented by continuous sensory data. Therefore a common way of connecting them is to discretize the sensory feature space into categories that are associated with linguistic words such as *grounding color names* [27,36,61,76] and *grounding spatial terms* [29,75,87].

What is more relevant to our work is recent progress on grounded language understanding, which involves learning meanings of words through connections to machine perception [82] and grounding language expressions to the shared visual world. For example, to visual objects [55,56], to physical landmarks [90,91], and to perceived actions or activities [2,91].

Different approaches and emphases have also been explored for grounded language learning. For example, linear programming has been applied to mediate perceptual differences between humans and robots for referential grounding [55]. Approaches to semantic parsing have been applied to ground language to internal world representations [2, 9]. Logical Semantics with Perception (LSP) [47] was applied to ground natural language queries to visual referents through jointly parsing natural language (combinatory categorical grammar (CCG)) and visual attribute classification.

Graphical models have been applied to word grounding. For example, a generative model was applied to integrate And-Or-Graph representations of language and vision for joint parsing [93]. A Factorial Hidden Markov Model (FHMM) was applied to learn the meaning of nouns, verbs, prepositions, adjectives and adverbs from short video clips paired with sentences [104]. Discriminative models have also been applied to ground human commands or instructions to perceived visual entities, mostly for robotic applications [90, 91]. More recently, deep learning has been applied to ground phrases to image regions [39].

2.3 Explainable Artificial Intelligence

Advanced machine learning - especially deep learning approaches have proven the effectiveness in many applications such as image classification and machine translation. However, they do not provide meaningful and human interpretable explanations of model behaviors. This makes it difficult for artificial agents to collaborate with humans as it's crucial for humans to understand the agent's capabilities and limitations. To address this problem, there is a growing interest in Explainable artificial intelligence recently. For example, approaches are proposed to generate high precision rules to explain classifiers' decisions [79, 80]. Specifically for Convolutional Neural Networks (CNNs), recent work addresses its interpretability by mining semantic meanings of filters [108, 109] or by generating language explanations [32, 68]. Interpreting the neural models also helps to analyze the linguistic characteristics of the Alzheimer disease patients [38]. An increasing amount of works on the Visual Question Answering (VQA) task [1, 58] have also looked into more interpretable approaches by utilizing attention-based models [24] or reasoning based on explicit evidence [99].

Another trend is to understand physical actions by modeling physical attributes (including causal attributes) [23, 25, 26, 106]. Physical attributes and the related commonsense knowledge are important sources for explanation generation. Previous works try to acquire commonsense knowledge from image annotations [103] or learn commonsense knowledge from visual abstraction [95]. Recently related work also focus on commonsense knowledge related with human's mental states [74]. Different from above works, our work here focuses on learning to acquire commonsense evidence for action justification.

2.4 Visual Commonsense Reasoning

Deep learning based methods have achieved great performance on many vision tasks and applications, for some tasks it even surpasses the human-level performance. However, most of of these applications still capture superficial semantics such as recognizing the objects in an image, identify basic properties including colors, locations and so on. Humans have much stronger reasoning abilities for the situated visual situation. Most of this knowledge for the visual world are visual commonsense knowledge, from low-level spatial understanding to high level causal inference. In order to develop agents who behave like a human, the agents must have the ability to acquire such kind of visual commonsense and also can infer more complex scenarios based on the visual commonsense knowledge.

Some work [103] starts from extracting visual commonsense directly from image annotations. Specifically they are trying to mine object-object relations and further extend to entailment relations based on statistics from an annotated image corpus. However, this method requires a large scale of dense annotations of the images.

Other works try to extend the traditional recognition-based visual task to reasoning based tasks [1]. One of the most popular tasks is visual question answering(VQA): given an image and a question, the model needs to learn how to answer the visual question. However, the main limitation of the VQA task is that most of questions are not related with the visual commonsense, and are still related recognition-level semantics in most of current open source datasets.

Recently the Recognition to Cognition(R2C) work [105] collect a new large dataset focusing on the visual commonsense phenomenon in the movie domain. Their setting is similar to the VQA setting. The difference is that their questions require in-depth understanding of the visual semantics to answer, which they call the cognition semantics. For each sample, they also provide explanation choices to enable the model learns which explanation can be used to justify the answer. Compared with other explanation generation based tasks, making it a multiple choices problem can make the evaluation easier and more robust. And the R2C task can be seen as a combination of Visual Question Answering task and Visual Commonsense Reasoning task.

Chapter 3

Grounded Semantic Role Labeling

In the previous chapters, we briefly review the background of grounded language learning and some related works. This chapter [1] is organized as follows: first we introduce some backgrounds and motivate the new grounded semantic role labeling task. Then we show how we formulate the problem into the graphical model framework. Third we investigate a subset of the TACOS corpus and analyze the dataset statistics. Finally we conduct a set of experiments and discuss the results. Last we conclude the current work.

3.1 Introduction

Linguistic studies capture semantics of verbs by their frames of thematic roles (also referred to as semantic roles or verb arguments) [52]. For example, a verb can be characterized by *agent* (i.e., the animator of the action) and *patient* (i.e., the object on which the action is acted upon), and other roles such as *instrument*, *source*, *destination*, etc. Given a verb frame, the goal of Semantic Role Labeling (SRL) is to identify linguistic entities from the text that serve different thematic roles [14, 28, 67, 111]. For example, given the sentence *the woman takes out a cucumber from the refrigerator.*, *takes out* is the main verb (also called *predicate*); the noun phrase *the woman* is the *agent* of this action; *a cucumber* is the *patient*; and *the refrigerator* is the *source*.

[1] Grounded Semantic Role Labeling. Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, Joyce Y. Chai. Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), San Diego, CA, June 12-17, 2016.

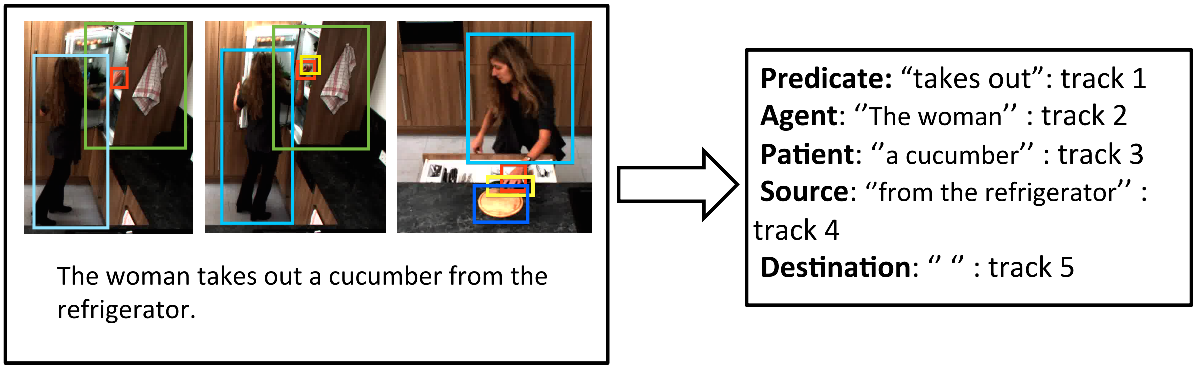

The woman takes out a cucumber from the refrigerator.
Predicate: "takes out": track 1
Agent: "The woman" : track 2
Patient: "a cucumber" : track 3
Source: "from the refrigerator" : track 4
Destination: " " : track 5

Figure 3.1: An example of grounded semantic role labeling for the sentence the woman takes out a cucumber from the refrigerator. The left hand side shows three frames of a video clip with the corresponding language description. The objects in the bounding boxes are tracked and each track has a unique identifier. The right hand side shows the grounding results where each role including the implicit role (destination) is grounded to a track id.

SRL captures important semantic representations for actions associated with verbs, which have shown beneficial for a variety of applications such as information extraction [22] and question answering [86]. However, the traditional SRL is not targeted to represent verb semantics that are grounded to the physical world so that artificial agents can truly understand the ongoing activities and (learn to) perform the specified actions. To address this issue, we propose a new task on grounded semantic role labeling.

Figure 3.1 shows an example of grounded SRL. The sentence *the woman takes out a cucumber from the refrigerator* describes an activity in a visual scene. The semantic role representation from linguistic processing (including implicit roles such as `destination`) is first extracted and then grounded to tracks of visual entities as shown in the video. For example, the verb phrase *take out* is grounded to a trajectory of the right hand. The role `agent` is grounded to the person who actually does the *take-out* action in the visual scene (*track 1*) ; the `patient` is grounded to the cucumber taken out (*track 3*); and the `source` is grounded to the refrigerator (*track 4*). The implicit role of `destination` (which is not explicitly mentioned in the language description) is grounded to the cutting board (*track 5*).

To tackle this problem, we have developed an approach to jointly process language and vision by incorporating semantic role information. In particular, we use a benchmark dataset (TACOS) which consists of parallel video and language descriptions in a complex cooking domain [77] in our investigation. We have further annotated several layers of information for developing and evaluating grounded semantic role labeling algorithms. Compared to previous works on language grounding [47,90,104], our work presents several contributions. First, beyond arguments explicitly mentioned in language descriptions, our work simultaneously grounds explicit and implicit roles with an attempt to better connect verb semantics with actions from the underlying physical world. By incorporating semantic role information, our approach has led to better grounding performance. Second, most previous works only focused on a small number of verbs with limited activities. We base our investigation on a wider range of verbs and in a much more complex domain where object recognition and tracking are notably more difficult. Third, our work results in additional layers of annotation to part of the TACOS dataset. This annotation captures the structure of actions informed by semantic roles from the video. The annotated data is available for download [2]. It will provide a benchmark for future work on grounded SRL.

3.2 Grounded Semantic Role Labeling

3.2.1 Problem Formulation

Given a sentence S and its corresponding video clip V, our goal is to ground explicit/implicit roles associated with a verb in S to video tracks in V. In this paper, we focus on the following set of semantic roles: {*predicate, patient, location, source, destination, tool*}. In the cooking domain, as actions always involve hands, the *predicate* is grounded to the hand pose represented by a trajectory

[2] https://github.com/yangshao/gsrl

of relevant hand(s). Normally *agent* would be grounded to the person who does the action. As there is only one person in the scene, we thus ignore the grounding of the *agent* in this work.

Video tracks capture tracks of objects (including hands) and locations. For example, in Figure 3.1, there are 5 tracks: human, hand, cucumber, refrigerator and cutting board. Regarding the representation of locations, instead of discretization of a whole image to many small regions(large search space), we create locations corresponding to five spatial relations (center, up, down, left, right) with respect to each object track, which means we have 5 times the number of locations compared with the number of objects. For instance, in Figure 3.1, the *source* is grounded to the center of the bounding boxes of the refrigerator track; and the *destination* is grounded to the center of the cutting board track. We use Conditional Random Field(CRF) to model this problem. An example CRF factor graph is shown in Figure 3.2. The CRF structure is created based on information extracted from language. More specifically, $s_1, ..., s_6$ refers to the observed text and its semantic role. Notice that s_6 is an implicit role as there is no text from the sentence describing *destination*. Also note that the whole prepositional phrase “from the drawer" is identified as the

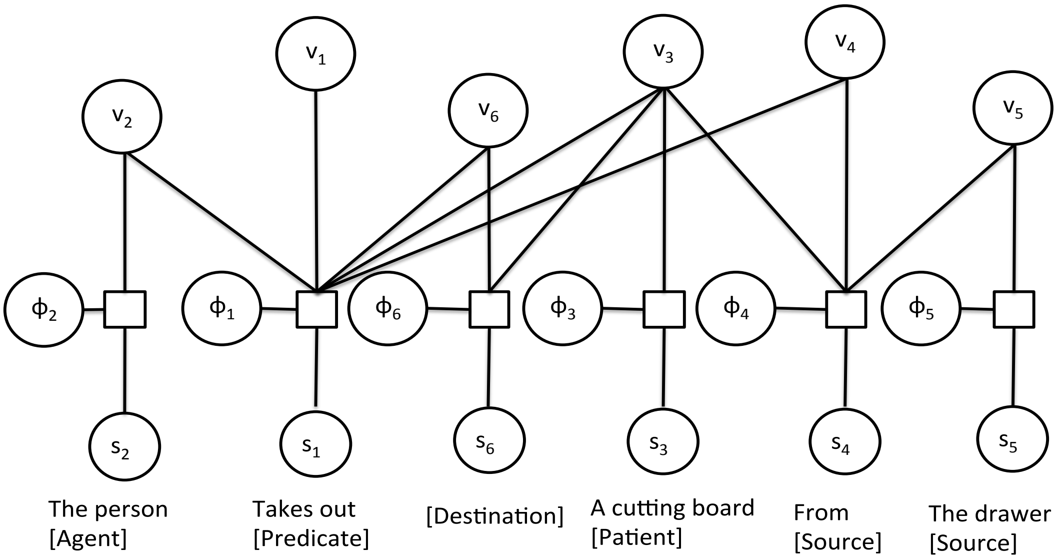


Figure 3.2: The CRF structure of sentence “the person takes out a cutting board from the drawer”. The text in the square bracket indicates the corresponding semantic role.

source rather than “the drawer" alone. This is because the prepositions play an important role in specifying location information. For example, “near the cutting boarding" is describing a location that is near to, but not exactly at the location of the cutting board. Here $v_1, ..., v_6$ are grounding random variables which take values from object tracks and locations in the video clip, and $\phi_1, ..., \phi_6$ are binary random variables which take values $\{0,1\}$. When ϕ_i equals to 1, it means v_i is the correct grounding of corresponding linguistic semantic role, otherwise it is not. The introduction of random variables ϕ_i follows previous work from Tellex and colleagues [90], which makes CRF learning more tractable.

3.3 CRF-based Learning and Inference

3.3.1 Conditional Random Field

Conditional Random Field is a kind of discriminative graphical model which models the conditional probability distribution $p(Y|X)$ in which X is the structured input and Y is structured output. The output space Y composes of an markov random field. Of all the CRF model variants, the linear chain CRF is the most commonly used which is first proposed for segmentating and labeling

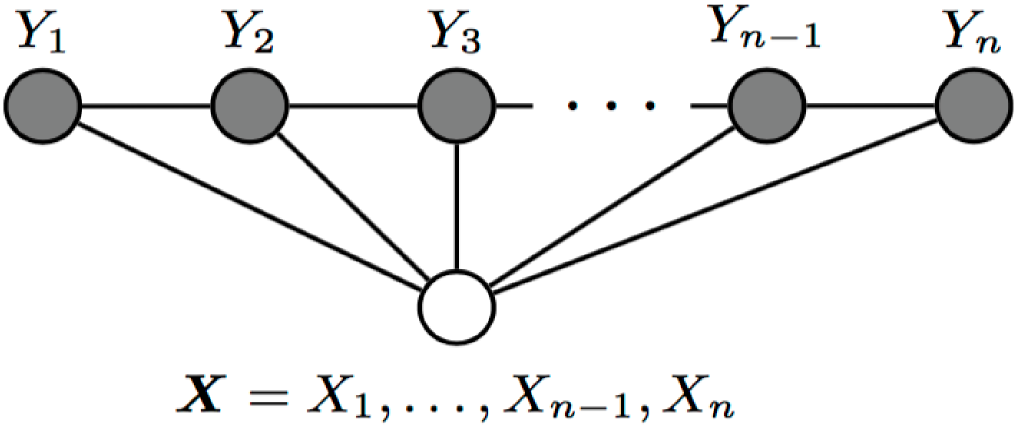


Figure 3.3: The linear conditional random field structure.

sequence data [49]. The conditional probability is represented as:

$$p(Y = y|X) = \frac{1}{Z(X)} \prod_{i} \Phi_i(X,y)$$

in which $\Phi_i(X,y)$ are the factors' potential functions and $Z(X)$ is the normalization constant. An example CRF for linear sequence tagging is shown in Figure 3.3.

3.3.2 CRF for Grounded Semantic Role Labeling

Different from the most common sequence labeling problem, grounded semantic role labeling is a more complex structure prediction task.

In the GSRL CRF model, we do not directly model the objective function as:

$$p(v_1,...,v_k|S,V)$$

where S refers to the sentence, V refers to the corresponding video clip and v_i refers to the grounding variable. Because the gradient based learning method needs the expectation of $v_1,...,v_k$, which is infeasible, we instead use the following objective function:

$$P(\phi|s_1,s_2,\ldots,s_k,v_1,v_2,\ldots,v_k,V)$$

where ϕ is a binary random vector $[\phi_1,...,\phi_k]$, indicating whether the grounding is correct. The objective function is factorized as follows:

$$\begin{aligned}
&P(\phi|s_1, s_2, \ldots, s_k, v_1, v_2, \ldots, v_k, V) \\
&= \frac{1}{Z}\prod_i \psi(\phi_i, s_i, v_i, V) \\
&= \prod_i \frac{1}{Z_i}\exp\{w^\top F(\phi_i, s_i, v_i, V)\} \\
&= \prod_i P(\phi_i|s_i, v_i, V)
\end{aligned}$$

where ψ is the potential function, w refers to parameters, $F(\phi_i, s_i, v_i, V)$ denotes a factor feature vector. Z and Z_i are normalization constant:

$$Z = \sum_\phi \prod_i \psi(\phi_i, s_i, v_i, V)$$

$$Z_i = \sum_{\phi_i} \psi(\phi_i, s_i, v_i, V)$$

We can see Z can be decomposed as the product of $Z_i s$ because each factor only relates to one ϕ_i. In this way, the objective function factorizes according to the structure of language with local normalization at each factor.

Gradient ascent with L2 regularization was used for parameter learning to maximize the objective function:

$$L = logP(P(\Phi|\lambda_1, \lambda_2, \ldots, \lambda_n, \gamma_1, \gamma_2, \ldots, \gamma_N, V)$$

Taking derivative of L we can get

$$\frac{\partial L}{\partial w} = \sum_i F(\phi_i, s_i, v_i, V) - \sum_i E_{P(\phi_i|s_i, v_i, V)} F(\phi_i, s_i, v_i, V)$$

where F refers to the feature function. During the training, we also use random grounding as negative samples for discriminative training. And the updating rule is

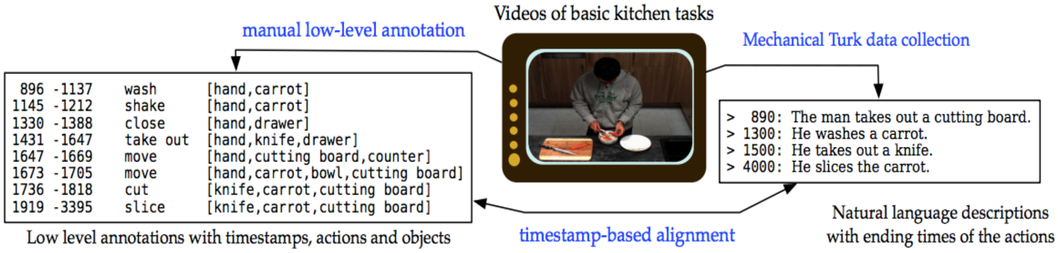
$$w_{t+1} = w_t + \eta \frac{\partial L}{\partial w},$$

where η is the step size. We can see the learning in tractable as ϕ is a binary random variable and calculate it's expectation is not that hard.

For the inference, The linear chain CRF have polynomial extract algorithm for decoding. But for the general graph structure, there is no efficient exact algorithm. The search space can be very large when the number of objects in the world increases. To address this problem we apply beam search to do the approximate inference. Specifically we select an easy to head inference order: we first ground roles including *patient*, *tool*, and then other roles including *location*, *source*, *destination* and *predicate*. We empirical try different beam search orders and find that this order achieves the best performance.

3.4 Dataset for Grounded Semantic Role Labeling

In this section, we first detail the statistics of the collected data for GSRL. Then introduce the approaches to process language and vision automatically.


Videos of basic kitchen tasks
manual low-level annotation
Mechanical Turk data collection
timestamp-based alignment
Low level annotations with timestamps, actions and objects
Natural language descriptions
with ending times of the actions

```
 896 -1137    wash      [hand,carrot]
1145 -1212    shake     [hand,carrot]
1330 -1388    close     [hand,drawer]
1431 -1647    take out  [hand,knife,drawer]
1647 -1669    move      [hand,cutting board,counter]
1673 -1705    move      [hand,carrot,bowl,cutting board]
1736 -1818    cut       [knife,carrot,cutting board]
1919 -3395    slice     [knife,carrot,cutting board]
```

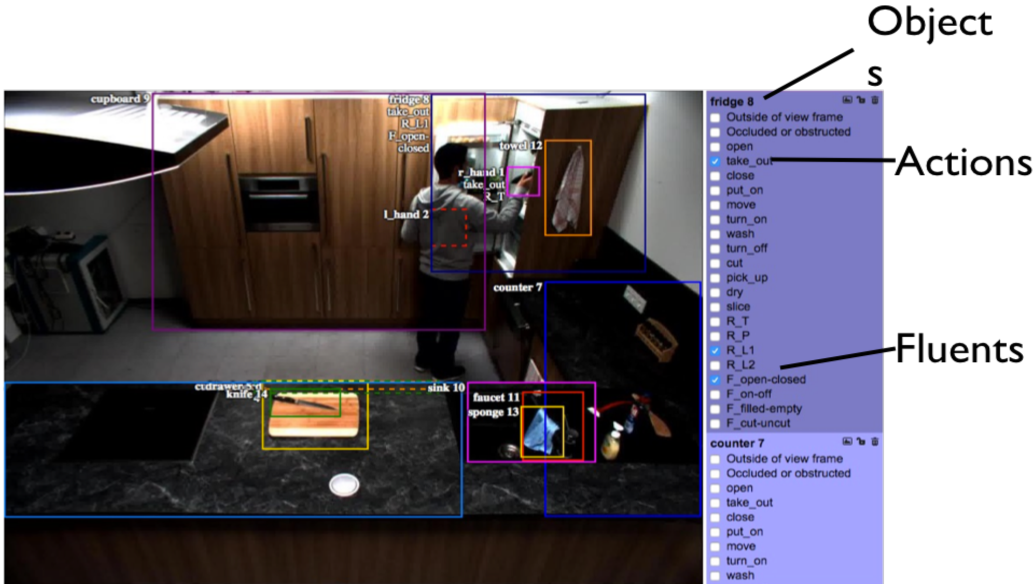
```
>  890: The man takes out a cutting board.
> 1300: He washes a carrot.
> 1500: He takes out a knife.
> 4000: He slices the carrot.
```

Figure 3.4: The TACOS dataset.

3.4.1 Dataset Collection

We conducted our investigation based on a subset of the TACOS corpus [77]. This dataset contains a set of video clips paired with natural language descriptions related to several cooking tasks. The natural language descriptions were collected through crowd-sourcing on top of the "MPII Cooking Composite Activities" video corpus [81]. The overview of the TACOS is shown in Figure 3.4. The middle part is the video clip corresponding to one specific task in the kitchen domain. On the left side, there show different manual low-level annotations with time segment, verbs and related objects. The right side shows the human language descriptions for specific time periods.

In this paper, we selected two tasks "cutting cucumber" and "cutting bread" as our experimental data. Each task has 5 videos showing how different people perform the same task. Each video is segmented to a sequence of video clips where each video clip comes with one or more language descriptions.



fridge 8
- ☐ Outside of view frame
- ☐ Occluded or obstructed
- ☐ open
- ☑ take_out
- ☐ close
- ☐ put_on
- ☐ move
- ☐ turn_on
- ☐ wash
- ☐ turn_off
- ☐ cut
- ☐ pick_up
- ☐ dry
- ☐ slice
- ☐ R_T
- ☐ R_P
- ☑ R_L1
- ☐ R_L2
- ☑ F_open-closed
- ☐ F_on-off
- ☐ F_filled-empty
- ☐ F_cut-uncut

counter 7
- ☐ Outside of view frame
- ☐ Occluded or obstructed
- ☐ open
- ☐ take_out
- ☐ close
- ☐ put_on
- ☐ move
- ☐ turn_on
- ☐ wash

Figure 3.5: The VATIC Annotation Interface for the TACOS dataset.

[2] For some verbs (e.g., *get*), there is a slight discrepancy between the sum of implicit/explicit roles across different categories. This is partly due to the fact that some verb occurrences take more than one objects as grounding to a role. It is also possibly due to missed/duplicated annotation for some categories.

The original TACOS dataset does not contain annotation for grounded semantic roles. To support our investigation and evaluation, we had made significant efforts adding the following annotations. For each video clip, we annotated the objects' bounding boxes, their tracks, and their labels (cucumber, cutting_board, etc.) using VATIC [97]. On average, each video clip is annotated with 15 tracks of objects. The annotation interface is shown in Figure 3.5.

For each sentence, we annotated the ground truth parsing structure and the semantic frame for each verb. The ground truth parsing structure is the representation of dependency parsing results. The semantic frame of a verb includes slots, fillers, and their groundings. For each semantic role (including both explicit roles and implicit roles) of a given verb, we also annotated the ground truth grounding in terms of the object tracks and locations. In total, our annotated dataset includes 976 pairs of video clips and corresponding sentences, 1094 verbs occurrences, and 3593 groundings of semantic roles. To check annotation agreement, 10% of the data was annotated by two annotators. The kappa statistics is 0.83 [13].

Table 3.1: Statistics for a set of verbs and their semantic roles in our annotated dataset. The entry indicates the number of explicit/implicit roles for each category. "–" denotes no such role is observed in the data.

Verb	Patient	Source	Destn	Location	Tool
take	251 / 0	102 / 149	2 / 248	–	–
put	94 / 0	–	75 / 19	–	–
get	247 / 0	62 / 190	0 / 239	–	–
cut	134 / 1	64 / 64	–	3 / 131	5 / 130
open	23 / 0	–	–	0 / 23	2 / 21
wash	93 / 0	–	–	26 / 58	2 / 82
slice	69 / 1	–	–	2 / 68	2 / 66
rinse	76 / 0	0 / 74	–	8 / 64	–
place	104 / 1	–	105 / 7	–	–
peel	29 / 0	–	–	1 / 27	2 / 27
remove	40 / 0	34 / 6	–	–	–

From this dataset, we selected 11 most frequent verbs (i.e., *get, take, wash, cut, rinse, slice, place, peel, put, remove, open*) in our current investigation for the following reasons. First, they are used more frequently so that we can have sufficient samples of each verb to learn the model. Second, they cover different types of actions: some are more related to the change of the state such as *take*, and some are more related to the process such as *wash*. As it turns out, these verbs also have different semantic role patterns as shown in Table 3.1. The *patient* roles of all these verbs are explicitly specified. This is not surprising as all these verbs are transitive verbs. There is a large variation for other roles. For example, for the verb *take*, the *destination* is rarely specified by linguistic expressions (i.e., only 2 instances), however it can be inferred from the video. For the verb *cut*, the *location* and the *tool* are also rarely specified by linguistic expressions. Nevertheless, these implicit roles contribute to the overall understanding of actions and should also be grounded too.

3.4.2 Automated Processing

To build the structure of the CRF as shown in Figure 3.2 and extract features for learning and inference, we have applied the following approaches to process language and vision.

Language Processing. Language processing consists of three steps to build a structure containing syntactic and semantic information. First, the Stanford Parser [60] is applied to create a dependency parsing tree for each sentence. Second, Senna [14] is applied to identify semantic role labels for the key verb in the sentence. The linguistic entities with semantic roles are matched against the dependency nodes in the tree and the corresponding semantic role labels are added to the tree. Third, for each verb, the PropBank [67] entries are searched to extract all relevant semantic roles. The implicit roles (i.e., not specified linguistically) are added as direct children of verb nodes in the tree. Through these three steps, the resulting tree from language processing has both explicit and implicit semantic roles. These trees are further transformed to the CRF structures based on a set of rules.

Vision Processing. A set of visual detectors are first trained for each type of objects. Here a random forest classifier is adopted. More specifically, we use 100 trees with HoG features [15] and color descriptors [94]. Both HoG and Color descriptors are used, because some objects are more structural such as knives and humans; While some are more textured such as towels. With the learned object detectors, given a candidate video clip, we run the detectors at each 10th frame (less than 0.5 seconds), and find the candidate windows for which the detector score corresponding to the object is larger than a threshold (set as 0.5). Then using the detected window as a starting point, we adopt tracking-by-detection [16] to go forward and backward to track this object and obtain the candidate track with this object label.

Feature Extraction. Features in the CRF model can be divided into the following three categories:

1. *Linguistic features* include word occurrence and semantic role information. They are extracted by language processing.

2. *Track label features* are the label information for tracks in the video. The labels come from human annotation or automated visual processing depending on different experimental settings (described in Section 3.5.1).

3. *Visual features* are a set of features involving geometric relations between tracks in the video. One important feature is the histogram comparison score. It measures the similarity between distance histograms. Specifically, histograms of distance values between the tracks of the *predicate* and other roles for each verb are first extracted from the training video clips. For an incoming distance histogram, we calculate its Chi-Square distances [107] from the

pre-extracted training histograms with the same verb and the same role. Its histogram comparison score is set to be the average of top 5 smallest Chi-Square distances. Other visual features include geometric information for single tracks and geometric relations between two tracks. For example, size, average speed, and moving direction are extracted for a single track. Average distance, size-ratio, and relative direction are extracted between two tracks. For features that are continuous, we discretized them into uniform bins.

To ground language into tracks from the video, instead of using track label features or visual features alone, we use a Cartesian product of these features with linguistic features. To learn the behavior of different semantic roles of different verbs, visual features are combined with the presence of both verbs and semantic roles through Cartesian product. To learn the correspondence between track labels and words, track label features are combined with the presence of words also through Cartesian product.

To train the model, we randomly selected 75% of annotated 976 pairs of video clips and corresponding sentences as training set. The remaining 25% were used as the testing set.

3.5 Evaluation of Grounded Semantic Role Labeling

3.5.1 Experimental Setup

Comparison. To evaluate the performance of our approach, we compare it with two approaches.

- **Baseline**: To identify the grounding for each semantic role, the first baseline chooses the most possible track based on the object type conditional distribution given the verb and semantic role. If an object type corresponds to multiple tracks in the video, e.g., multiple drawers or knives, we then randomly select one of the tracks as grounding. We ran this baseline method five times and reported the average performance.

- **Tellex (2011)**: The second approach we compared with is based on an implementation [90]. The difference is that they don't explicitly model fine-grained semantic role information. For a better comparison, we map the grounding results from this approach to different explicit semantic roles according to the SRL annotation of the sentence. Note that this approach is not able to ground implicit roles.

More specifically, we compare these two approaches with two variations of our system:

- **$\mathbf{GSRL_{wo\_V}}$**: The CRF model using linguistic features and track label features (described in Section 3.4.2).
- **GSRL**: The full CRF model using linguistic features, track label features, and visual features(described in Section 3.4.2).

Configurations. Both automated language processing and vision processing are error-prone. To further understand the limitations of grounded SRL, we compare performance under different configurations along the two dimensions: (1) the CRF structure is built upon annotated ground-truth language parsing versus automated language parsing; (2) object tracking and labeling is based on annotation versus automated processing. These lead to four different experimental configurations.

Evaluation Metrics. For experiments that are based on annotated object tracks, we can simply use the traditional *accuracy* that directly measures the percentage of grounded tracks that are correct. However, for experiments using automated tracking, evaluation can be difficult as tracking itself poses significant challenges. The grounding results (to tracks) cannot be directly compared with the annotated ground-truth tracks. To address this problem, we have defined a new metric called *approximate accuracy*. This metric is motivated by previous computer vision work that evaluates

Table 3.2: Evaluation results based on annotated language parsing.

Accuracy On the Gold Recognition/Tracking Setting														
Methods	Predicate	Patient		Source		Destination		Location		Tool		Explicit All	Implicit All	All
		explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit			
Baseline	0.856	0.372	NA	0.225	0.314	0.311	0.569	NA	0.910	NA	0.853	0.556	0.620	0.583
Tellex(2011)	0.865	0.745	–	0.306	–	0.763	–	NA	–	NA	–	0.722	–	–
GSRL$_{\mathbf{wo\_V}}$	0.854	0.794^{*}_{+}	NA	0.375^{*}	0.392^{*}_{+}	0.658^{*}	0.615^{*}_{+}	NA	0.920_{+}	NA	0.793_{+}	0.768^{*}_{+}	0.648^{*}_{+}	0.717^{*}
GSRL	**0.878**$^{*}_{+}$	**0.839**$^{*}_{+}$	NA	**0.556**$^{*}_{+}$	**0.684**$^{*}_{+}$	**0.789***	**0.641**$^{*}_{+}$	NA	**0.930**$_{+}$	NA	**0.897**$^{*}_{+}$	**0.825**$^{*}_{+}$	**0.768**$^{*}_{+}$	**0.8***
Approximated Accuracy On the Automated Recognition/Tracking Setting														
Methods	Predicate	Patient		Source		Destination		Location		Tool		Explicit All	Implicit All	All
		explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit			
Baseline	0.529	0.206	NA	0.169	0.119	0.236	0.566	NA	**0.476**	NA	0.6	0.352	0.393	0.369
Tellex(2011)	**0.607**	0.233	–	0.154	–	0.333	–	NA	–	NA	–	0.359	–	–
GSRL$_{\mathbf{wo\_V}}$	0.582^{*}	0.244^{*}	NA	**0.262**$^{*}_{+}$	**0.126**$_{+}$	**0.485**$^{*}_{+}$	**0.613**$^{*}_{+}$	NA	0.467_{+}	NA	**0.714**$^{*}_{+}$	**0.410**$^{*}_{+}$	**0.425**$^{*}_{+}$	**0.417***
GSRL	0.548	**0.263***	NA	**0.262**$^{*}_{+}$	0.086_{+}	0.394^{*}	0.514_{+}	NA	0.456_{+}	NA	0.688^{*}_{+}	0.399^{*}_{+}	0.381_{+}	0.391^{*}
Upper_Bound	0.920	0.309	NA	0.277	0.252	0.636	0.829	NA	0.511	NA	0.818	0.577	0.573	0.575

Table 3.3: Evaluation results based on automated language parsing.

Accuracy On the Gold Recognition/Tracking Setting														
Methods	Predicate	Patient		Source		Destination		Location		Tool		Explicit All	Implicit All	All
		explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit			
Baseline	0.881	0.318	NA	0.203	0.316	0.235	0.607	NA	0.877	NA	**0.895**	0.539	0.595	0.563
Tellex(2011)	**0.903**	0.746	–	0.156	–	0.353	–	NA	–	NA	–	0.680	–	–
GSRL$_{\mathbf{wo\_V}}$	0.873	0.813^{*}_{+}	NA	0.328^{*}_{+}	0.360_{+}	**0.412***	0.648^{*}_{+}	NA	0.877_{+}	NA	0.818_{+}	0.769^{*}_{+}	0.611_{+}	0.7^{*}
GSRL	0.873	**0.875**$^{*}_{+}$	NA	**0.453**$^{*}_{+}$	**0.667**$^{*}_{+}$	**0.412***	**0.667**$^{*}_{+}$	NA	**0.891**$_{+}$	NA	0.891_{+}	**0.823**$^{*}_{+}$	**0.741**$^{*}_{+}$	**0.787***
Approximated Accuracy On the Automated Recognition/Tracking Setting														
Methods	Predicate	Patient		Source		Destination		Location		Tool		Explicit All	Implicit All	All
		explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit			
Baseline	0.543	0.174	NA	0.121	0.113	0.093	0.594	NA	**0.612**	NA	0.567	0.327	0.405	0.362
Tellex(2011)	0.598	0.218	–	0.086	–	0.00	–	NA	–	NA	–	0.322	–	–
GSRL$_{\mathbf{wo\_V}}$	**0.618***	**0.243***	NA	**0.190**$^{*}_{+}$	**0.120**$_{+}$	**0.133**$_{+}$	**0.641**$^{*}_{+}$	NA	0.585_{+}	NA	**0.723**$^{*}_{+}$	**0.401**$^{*}_{+}$	**0.434**$^{*}_{+}$	**0.415***
GSRL	0.493	**0.243***	NA	**0.190**$^{*}_{+}$	0.063_{+}	**0.133**$_{+}$	0.612_{+}	NA	0.554_{+}	NA	0.617_{+}	0.367^{*}_{+}	0.386_{+}	0.375
Upper_Bound	0.908	0.277	NA	0.259	0.254	0.4	0.854	NA	0.631	NA	0.830	0.543	0.585	0.561

tracking performance [4]. Suppose the ground truth grounding for a role is track *gt* and the predicted grounding is track *pt*. The two tracks *gt* and *pt* are often not the same (although may have some overlaps). Suppose the number of frames in the video clip is *k*. For each frame, we calculate the distance between the centroids of these two tracks. If their distance is below a predefined threshold, we consider the two tracks overlapping in this frame. We consider the grounding is correct if the ratio of the overlapping frames between *gt* and *pt* exceeds 50%. As can be seen, this is a lenient and an approximate measure of accuracy.

3.5.2 Results

The results based on the ground-truth language parsing are shown in Table 3.2, and the results based on automated language parsing are shown in Table 3.3. For results based on annotated object tracking, the performance is reported in *accuracy* and for results based on automated object tracking, the performance is reported in *approximate accuracy*. When the number of testing samples is less than 15, we do not show the result as it tends to be unreliable (shown as *NA*). Tellex (2011) does not address implicit roles (shown as "–"). The best performance score is shown in bold. We also conducted a two-tailed bootstrap significance testing [20]. The score with a "*" indicates it is statistically significant ($p < 0.05$) compared to the baseline approach. The score with a "+" indicates it is statistically significant ($p < 0.05$) compared to the approach [90].

For experiments based on automated object tracking, we also calculated an *upper_bound* to assess the best possible performance which can be achieved by a perfect grounding algorithm given the current vision processing results. This *upper_bound* is calculated based on grounding each role to the track which is closest to the ground-truth annotated track. For the experiments based on annotated tracking, the *upper_bound* would be 100%. This measure provides some understandings about how good the grounding approach is given the limitation of vision processing. Notice that the grounding results in the gold/automatic language processing setting are not directly comparable as the automatic SRL can misidentify frame elements.

3.5.3 Discussion

As shown in Table 3.2 and Table 3.3, our approach consistently outperforms the baseline (for both explicit and implicit roles) and the Tellex (2011) approach. Under the configuration of gold recognition/tracking, the incorporation of visual features further improves the performance. However,

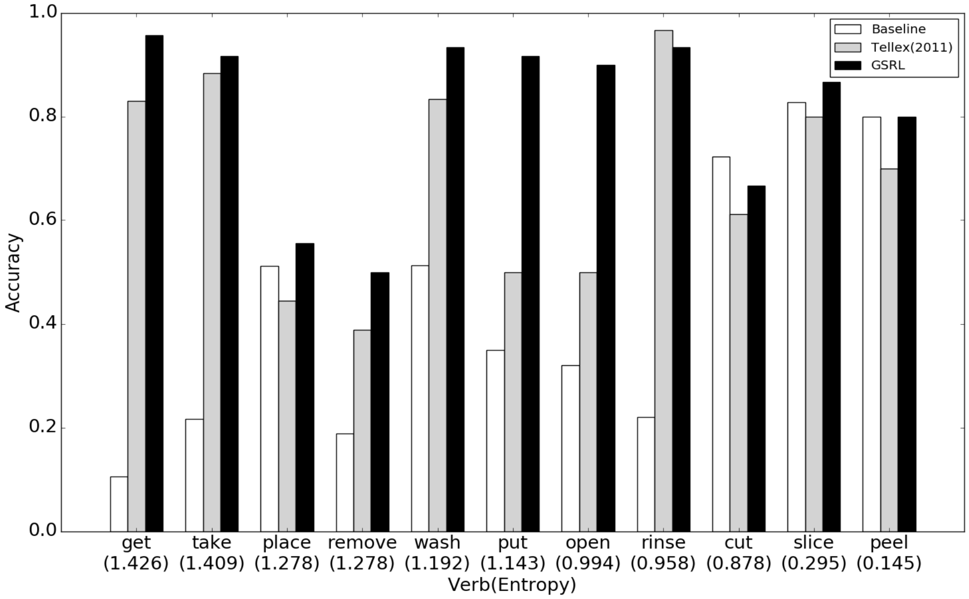


Figure 3.6: The relation between the accuracy and the entropy of each verb's patient from the gold language, gold visual recognition/tracking setting. The entropy for the patient role of each verb is shown below the verb.

this performance gain is not observed when automated object tracking and labeling is used. One possible explanation is that as we only had limited data, we did not use separate data to train models for object recognition/tracking. So the GSRL model was trained with gold recognition/tracking data and tested with automated recognition/tracking data.

By comparing our method with Tellex (2011), we can see that by incorporating fine grained semantic role information, our approach achieves better performance on almost all the explicit roles (except for the *patient* role under the automated tracking condition).

The results have also shown that some roles are easier to ground than others in this domain. For example, the *predicate* role is grounded to the hand tracks (either left hand or right hand), there are not many variations such that the simple baseline can achieve pretty high performance, especially when annotated tracking is used. The same situation happens to the *location* role as most of the locations happen near the *sink* when the verb is *wash*, and near the *cutting board* for verbs like *cut*, etc. However, for the *patient* role, there is a large difference between our approach and baseline approaches as there is a larger variation of different types of objects that can participate in the role for a given verb.

For experiments with automated tracking, the *upper_bound* for each role also varies. Some roles (e.g., *patient*) have a pretty low upper bound. The accuracy from our full GSRL model is already quite close to the upper bound. For other roles such as *predicate* and *destination*, there is a larger gap between the current performance and the upper bound. This difference reflects the model's capability in grounding different roles.

Figure 3.6 shows a close-up look at the grounding performance to the *patient* role for each verb under the gold parsing and gold tracking configuration. The reason we only show the results of *patient* role here is every verb has this role to be grounded. For each verb, we also calculated its entropy based on the distribution of different types of objects that can serve as the *patient* role in the training data. The entropy is shown at the bottom of the Figure. For verbs such as *take* and *put*, our full GSRL model leads to much better performance compared to the baseline. As the baseline approach relies on the entropy of the potential grounding for a role, we further measured the improvement of the performance and calculated the correlation between the improvement and the entropy of each verb. The result shows that Pearson coefficient between the entropy and the improvement of GSRL over the baseline is 0.614. This indicates the improvement from GSRL is positively correlated with the entropy value associated with a role, implying the GSRL model can deal with more uncertain situations. For the verb *cut*, The GSRL model performs slightly worse than the baseline. One explanation is that the possible objects that can participate as a patient for *cut* are relatively constrained where simple features might be sufficient. A large number of features may introduce noise, and thus jeopardizing the performance.

We further compare the performance of our full GRSL model with Tellex (2011) (also shown in Figure 3) on the *patient* role of different verbs. Our approach outperforms Tellex (2011) on most of the verbs, especially *put* and *open*. A close look at the results have shown that in those cases, the *patient* roles are often specified by pronouns. Therefore, the track label features and linguistic features are not very helpful, and the correct grounding mainly depends on visual features. Our full GSRL model can better capture the geometry relations between different semantic roles by incorporating fine-grained role information.

3.6 Conclusion

This chapter investigates a new problem on grounded semantic role labeling. Besides semantic roles explicitly mentioned in language descriptions, our approach also grounds implicit roles which are not explicitly specified. As implicit roles also capture important participants related to an action (e.g., tools used in the action), our approach provides a more complete representation of action semantics which can be used by artificial agents for further reasoning and planning towards the physical world. Our empirical results on a complex cooking domain have shown that, by incorporating semantic role information with visual features, our approach can achieve better performance compared to baseline approaches. Our results have also shown that grounded semantic role labeling is a challenging problem which often depends on the quality of automated visual processing (e.g., object tracking and recognition).

There are several directions for future improvement. First, the current alignment between a video clip and a sentence is generated by some heuristics which are error-prone. One way to address this is to treat alignment and grounding as a joint problem. Second, our current visual features have not shown effective especially when they are extracted based on automatic visual processing. This is partly due to the complexity of the scene from the TACOS dataset and the lack of depth information. Recent advances in object tracking algorithms [63, 101] together with 3D sensing can be explored in the future to improve visual processing. Moreover, linguistic studies have shown that action verbs such as *cut* and *slice* often denote some change of state as a result of the action [34, 35]. The change of state can be perceived from the physical world. Thus another direction is to systematically study causality of verbs. Causality models for verbs can potentially provide top-down information to guide intermediate representations for visual processing and improve grounded language understanding.

The capability of grounding semantic roles to the physical world has many important implications. It will support the development of intelligent agents which can reason and act upon the shared physical world. For example, unlike traditional action recognition in computer vision [98], grounded SRL will provide deeper understanding of the activities which involve participants in the actions guided by linguistic knowledge. For agents that can act upon the physical world such as robots, grounded SRL will allow the agents to acquire the grounded structure of human commands and thus perform the requested actions through planning (e.g., to follow the command "put the cup on the table"). Grounded SRL will also contribute to robot action learning where humans can teach the robot new actions (e.g., simple cooking tasks) through both task demonstration and language instruction.

Chapter 4

Commonsense Action Explanation in Human-Agent Communication

In the previous chapter, we conduct a comprehensive study on the grounded semantic role labeling task in order to understand the verb semantics in the physical situated world. In this chapter [1], we start looking into a more interesting problem: the grounded semantic role labeling task targets to answer questions like "what's the relation between the action and the object/location in the physical world?". But for the commonsense reasoning and human interpretation, we are more interested in asking the question: why do you think this action happens, which is a more challenging task compared with grounded semantic role labeling.

In this chapter, we will first conduct a study on human justifications. Then give a formal formulation for commonsense action justification. We also detail the process of data crowd sourcing . Empirical experiments are conducted to prove the effectiveness of the proposed method. Finally we propose a novel human study to verify the communication grounding efficiency compared with a variety of different methods.

[1] Commonsense Justification for Action Explanation. Shaohua Yang, Qiaozi Gao, Sari Saba-Sadiya, Joyce Y. Chai. Conference on Empirical Methods in Natural Language Processing (EMNLP), Brussels, Belgium, November 1-4, 2018.

4.1 Introduction

When collaborating with artificial agents, it's important for humans to understand agents' abilities and limitations (e.g., understand why a decision is made by the agent) so that humans can be more cooperative in joint tasks (e.g., decide when to trust the agent's prediction). To address this issue, recent years have seen an increasing effort on Explainable AI (XAI) which attempts to develop explainable models that can explain the agent's decision making while maintaining a high-level of performance.

There are two types of explanation: *introspective explanation* which addresses decision making process and *justification explanation* which gathers evidence to support a certain decision [5, 68]. In this paper we focus on justification explanation - identifying *commonsense evidence* particularly for action justification. Although one of the end goals of this investigation is to support perception, our current focus is on higher level commonsense reasoning for action explanation. Therefore this work is based on a symbolic representation of the world without concerning vision processing. Specifically our task is framed as: given many symbolic descriptions of the physical world (e.g., object relations and attributes as a result of vision or other processing), how to identify a small set of descriptions which can justify an action in line with humans' commonsense knowledge? The lack of commonsense knowledge is a major bottleneck in artificial agents which jeopardizes the common ground between humans and agents for successful communication. If artificial agents ever become partners with humans in joint tasks, the ability to learn and acquire commonsense evidence for action justification is fundamental. This paper intends to address this important yet less studied problem.

As a first step in our investigation, we initiated a human study to identify key dimensions of commonsense reasoning, from the human's point of view, that justify an action. We then developed an explainable model based on the generative conditional variational autoencoder (CVAE) that models perceived attributes/relations as latent variables to learn the association between commonsense evidence and actions. Our empirical results on a subset of the Visual Genome data [46] show that, compared to a typical attention-based model, CVAE has a significantly higher explanation ability in terms of identifying correct commonsense evidence to justify the recognized action. When adding the supervision of commonsense evidence during training, both the explainability and the performance (i.e., action prediction) are further improved. In addition, we evaluated the role of commonsense evidence in communication grounding between humans and agents. Our experimental results show that the commonsense evidence generated by CVAE leads to a significantly higher common ground of actions.

The contributions of this chapter are three folds. First we identified several key dimensions of commonsense knowledge, from a human's perspective, to justify concrete actions in the physical environment. These dimensions provide a basis to justification explanation that is aligned with human's commonsense knowledge about the action. Second we proposed a method using CVAE to jointly learn to predict actions and select commonsense evidence as action justification. CVAE naturally models the generation process of both actions and commonsense evidence. Inferring commonsense evidence is equivalent to the posterior inference of the CVAE model, which is flexible and powerful to incorporate actions as context. Our experimental results have shown a higher explainability of CVAE in action justification without sacrificing performance. Finally our dataset of commonsense evidence for action explanation, together with our proposed methods, will be made available to the community. It will serve as a baseline for the future work on this topic.

4.2 A Study On Justification Explanation

While there is a rich literature on explanations in Psychology, Philosophy, and Linguistics [17, 57, 92], the kind of concrete physical actions we are interested in are rarely studied by previous work. To address this issue, our work began with a small human study that would enable us to identify a low level and quantitatively useful taxonomy of commonsense in explaining actions that can be perceived from the physical world.

We created a set of 12 short video clips (each about 14 seconds) from the Microsoft Research Video to Text dataset [100]. For each video clip, we asked human subjects to explain why they think a certain action is happening in the video. A total of about 170 responses from 67 participants were collected [2] After a careful examination, we came up with the following dimensions which capture commonsense explanation for actions.

- **Transitive-relations**: This kind of explanation does not directly focus on the structural relations between an action and its participants, but rather transits to the relation between the participant and something else (potentially related). For example, use *a woman wears an apron* to justify the *cook* action. In the collected responses, 64% of them used transitive relations. (Most subject responses contain multiple categories of explanation.)
- **Sub-actions**: Almost 75% of the responses used the existence of sub-actions as evidence (for example, the action is cook because there are sub-actions of cutting and heating meat).
- **Spatial-relations**: Around 15% of the responses used spatial relations involving the participants of the action, for example, *the knife is on the cutting board*, and *the water is in the bottle*, etc.
- **Effect-state**: Over 28% of the responses cited a change in the state of an object, in other

[2] The full survey along with other collected data will be released.

words the effect state, as evidence, such as *cucumber in small pieces* as the evidence for *chop*.

- **Associated-attributes**: Other attributes associated with the participants of the action, but not the effect state of the participants (20%). While these attributes are not directly related to the action, they are linked to the action by association. For example, *banana is sliced* is used as evidence to justify *blend*.
- **Other**: Participants have also cited other commonsense such as the “definition" of the action (5%), or the manner associated with different sub actions(12%).

Except for the category `Other` which cannot be perceived from an image, the other five categories can be potentially perceived and used as commonsense evidence to justify a perceived action. These five categories of commonsense explanations are used in our computational models described next.

4.3 Method

We formulate our task as the following: given a set of relations **R** and a set of attributes **E**, the goal is to jointly select evidence **z** and predict target action $\mathbf{a} \in \mathbf{A}$ where **A** is the vocabulary of actions. **R** is represented as $\{r_1, r_2, ..., r_m\}$ where each r_i is a tuple (r_i^p, r_i^s, r_i^o) corresponding to **predicate**, **subject** and **object**. **E** refers to $\{e_1, e_2, ..., e_n\}$ where each e_i is a tuple (e_i^o, e_i^p) corresponding to the object and attribute. We introduce **z** as a discrete vector $(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{m+n})$ where $\mathbf{z}_i \in \{0, 1\}$ represents the hidden explainable variable. **z** is interpreted as an evidence selector: $\mathbf{z}_i = 1$ means the corresponding relation/attribute justifying the target action **a**. Given all these definitions, our target is to learn the probability $\mathbf{p}(\mathbf{a}, \mathbf{z} | \mathbf{R}, \mathbf{E})$.

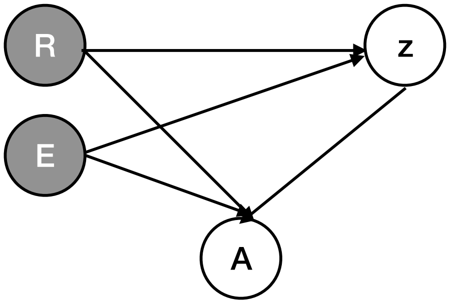


Figure 4.1: Grphaical Model representation of the Conditional Variational Auto-encoder.

4.3.1 Conditional Variational Autoencoder

The varational autoencoder(VAE) [44] is proposed as a generative model to combine the power of both directed continuous or discrete graphical model and neural network with latent variables. The VAE models the generative process of random variable $\mathbf{x}$ as following: first the latent variable $\mathbf{z}$ is generated from a prior probability distribution $\mathbf{p}(\mathbf{z})$, then a data sample $\mathbf{x}$ is generated from a conditional probability distribution $\mathbf{p}(\mathbf{x}|\mathbf{z})$. The CVAE [110] is a natural extension of VAE: Both the prior distribution and conditional distribution now are conditioned on an additional context $\mathbf{c}$: $\mathbf{p}(\mathbf{z}|\mathbf{c})$ and $\mathbf{p}(\mathbf{z}|\mathbf{x},\mathbf{c})$.

For our task, we decompose the inference problem $\mathbf{p}(\mathbf{a},\mathbf{z}|\mathbf{R},\mathbf{E})$ into two smaller problems. The first sub-problem is to infer $\mathbf{p}(\mathbf{a}|\mathbf{R},\mathbf{E})$, which we call performer. The second problem is to infer $\mathbf{p}(\mathbf{z}|\mathbf{a},\mathbf{R},\mathbf{E})$ which we call explainer. These two problems are closely coupled, hence we will model them jointly. The probability distribution $\mathbf{p}(\mathbf{a}|\mathbf{R},\mathbf{E})$ can be written as :

$$\mathbf{p}(\mathbf{a}|\mathbf{R},\mathbf{E}) = \sum_{\mathbf{z}} \mathbf{p}_\theta(\mathbf{a}|\mathbf{z},\mathbf{R},\mathbf{E})\mathbf{p}(\mathbf{z}|\mathbf{R},\mathbf{E})$$

The corresponding graphical representation is shown in Figure 4.1

Directly optimizing this conditional probability is not feasible. Usually the Evidence Lower

Bound (ELBO) [88] is optimized, which can be derived as following:

$$
\begin{aligned}
&\mathrm{ELBO}(\mathbf{a},\mathbf{R},\mathbf{E};\theta,\phi) \\
&= -\mathrm{KL}(\mathbf{q}_\phi(\mathbf{z}|\mathbf{a},\mathbf{R},\mathbf{E})||\mathbf{p}_\theta(\mathbf{z}|\mathbf{R},\mathbf{E})) \\
&+ \mathbf{E}_{\mathbf{q}_\phi(\mathbf{z}|\mathbf{a},\mathbf{R},\mathbf{E})}[\log p_\theta(\mathbf{a}|\mathbf{z},\mathbf{R},\mathbf{E})] \\
&\leq \log \mathbf{p}(\mathbf{a}|\mathbf{R},\mathbf{E})
\end{aligned}
\tag{4.1}
$$

For the first KL divergence term, we are minimizing the distance between the posterior distribution and the prior distribution. For the second term, we are maximizing the expectation of the target action based on the posterior latent distribution.

In most previous work using VAE, there is no explicit meaning for the hidden representation $\mathbf{z}$, thus it's hard for humans to interpret. For example, $\mathbf{z}$ is simply assumed as a Gaussian distribution or a categorical distribution. In order to have a more explicit representation for the purpose of explanation, our latent discrete variable $\mathbf{z}$ is used to indicate whether the corresponding relation or attribute can be used for explanation. In the above ELBO equation, $\mathbf{p}(\mathbf{a}|\mathbf{R},\mathbf{E})$ is the performer and $\mathbf{q}_\phi(\mathbf{z}|\mathbf{a},\mathbf{R},\mathbf{E})$ is the explainer. Thus we can learn the performer and explainer jointly.

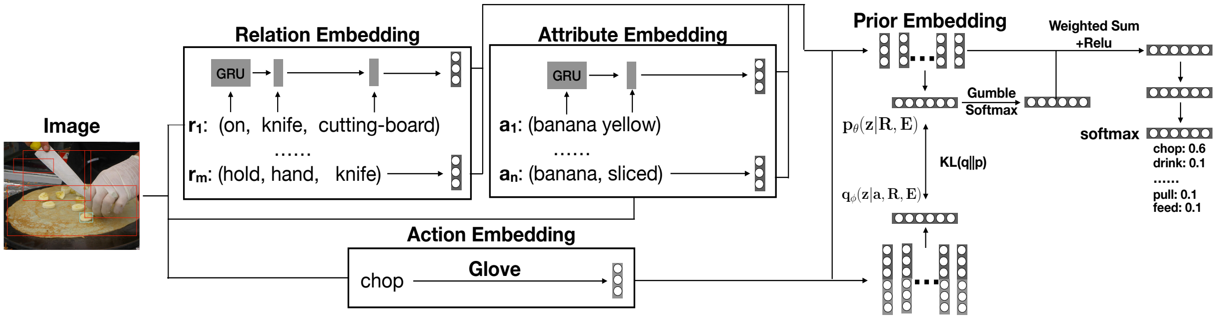


Figure 4.2: System Architecture for the CVAE model.

The whole system architecture is shown in Figure 4.2. From an image, we first extract candidate relation set **R** and attribute set **E** from human image descriptions or trained visual detectors. Every relation **r** and attribute **e** are embedded using a Gated Recurrent Neural Network [11].

$$\mathbf{r}^{emb} = \mathrm{GRU}([r^{p}, r^{s}, r^{o}])$$

$$\mathbf{e}^{emb} = \mathrm{GRU}([e^{o}, e^{p}])$$

The action **a** is represented by a Glove embedding [69], followed by another non-linear layer:

$$\mathbf{a}^{emb} = \mathrm{ReLU}(\mathbf{W}_i \mathbf{a}^{glove} + \mathbf{b}_i)$$

where $\mathbf{a}^{glove} \in \mathbb{R}^{k}$ is the pre-trained glove embedding. Then the latent variable **z** can be calculated as:

$$\mathbf{q}_{\phi}(\mathbf{z}|\mathbf{a}, \mathbf{R}, \mathbf{E}) = \mathrm{softmax}(\mathbf{W}_{\mathbf{z}}[\mathbf{U}; \mathbf{a}^{emb}] + \mathbf{b}_{\mathbf{z}})$$

where $\mathbf{U} = [\mathbf{r}_1^{emb}, ..., \mathbf{r}_m^{emb}, \mathbf{e}_1^{emb}, ..., \mathbf{e}_n^{emb}]$ and $[\mathbf{U}, \mathbf{a}^{emb}]$ means the concatenation of **U** and $\mathbf{a}^{emb}$. and $W_{\mathbf{z}} \in \mathbb{R}^{2\times 2k}$ as we assume each $\mathbf{z}_i$ belongs to one of the two classes $\{0, 1\}$. The prior distribution can be calculated as:

$$\mathbf{p}_{\theta}(\mathbf{z}|\mathbf{R}, \mathbf{E}) = \mathrm{softmax}(\mathbf{W}_{\mathbf{z}}^{'} \mathbf{U} + \mathbf{b}_{\mathbf{z}}^{'})$$

The main idea is that in order to sample one discrete $\mathbf{z}_d$ from the softmax distribution, it's equivalent to get the sample from

$$z = \mathrm{one\_hot}(\arg\max_{i}(\log(\pi_i) + g_i))$$

where π_i is the i-th logit for the softmax distribution **p**(z) and g_i is a sample drawn from Gumbel(0,1). The $\arg\max$ is further approximated as a continuous, differentiable function:

$$z_i = \frac{\exp(((\log(\pi_i) + g_i)/\tau)}{\sum_{j=1}^{2} \exp(((\log(\pi_j) + g_j)/\tau)}$$

τ is the temperature to control the accuracy of this approximate, the smaller the τ, the closer between the approximated distribution with the true distribution. We denote

$$\hat{z} = \text{gumble\_softmax}(z)$$

as the discrete approximate of z. Here $\hat{z} \in \mathbb{R}^{m+n,1}$

The KL divergence between the prior random variable $\mathbf{z}_{prior}$ from $\mathbf{p}_\theta(\mathbf{z}|\mathbf{R},\mathbf{E})$ and the posterior random variable $\mathbf{z}_{posterior}$ from $\mathbf{q}_\phi(\mathbf{z}|\mathbf{a},\mathbf{R},\mathbf{E})$ is:

$$\text{KL}(\mathbf{z}_{prior}, \mathbf{z}_{posterior}) = -p_i \log \frac{p_i}{p_i^{'}} - (1-p_i) \log \frac{1-p_i}{1-p_i^{'}}$$

here $\mathbf{z}_{prior} \sim \text{Bern}(p_i)$, $\mathbf{z}_{posterior} \sim \text{Bern}\left(p_i^{'}\right)$.

Another challenge is that $\mathbf{z}$ is a discrete variable which blocks the gradient and makes the end-to-end training infeasible. Gumbel-Softmax [37] is a re-parameterization trick to deal with the discrete variables in the neural network. We use this trick to sample discrete $\mathbf{z}$. Then we do a weighted sum pooling between discretized $\mathbf{z}$ and $\mathbf{U}$:

$$\mathbf{h}_z = \text{ReLU}(\sum_i \mathbf{z}_i * \mathbf{U}_i)$$

$$\mathbf{h} = \text{ReLU}(\mathbf{W}_h \mathbf{h}_z + \mathbf{b}_h)$$

$$\mathbf{p}_\theta(\mathbf{a}|\mathbf{z},\mathbf{R},\mathbf{E}) = \text{softmax}(\mathbf{W}\mathbf{h} + \mathbf{b})$$

During training, we also add a sparsity regularization on the latent variable $\mathbf{z}$ besides the `ELBO`. So our final training objective is

$$\begin{aligned}\mathscr{L}_{CVAE} = &-\mathrm{ELBO}(\mathbf{a},\mathbf{R},\mathbf{E};\theta,\phi)\\ &+\beta\,\mathrm{KL}(\mathbf{q}_\phi(\mathbf{z}|\mathbf{a},\mathbf{R},\mathbf{E})||\mathrm{Bern}(0))\end{aligned} \tag{4.2}$$

During testing, we have two objectives. First we want to infer the target action $\mathbf{a}$, which can be computed through sampling:

$$\begin{aligned}\mathbf{p}(\mathbf{a}|\mathbf{R},\mathbf{E}) &= \sum_{\mathbf{z}}\mathbf{p}_\theta(\mathbf{z}|\mathbf{R},\mathbf{E})\mathbf{p}_\theta(\mathbf{a}|\mathbf{z},\mathbf{R},\mathbf{E})\\ &\approx \frac{1}{S}\sum_{s=1}^{S}\mathbf{p}_\theta(\mathbf{a}|\mathbf{z}_s,\mathbf{R},\mathbf{E})\end{aligned} \tag{4.3}$$

where $\mathbf{z}_s \sim \mathbf{p}(\mathbf{z}|\mathbf{R},\mathbf{E})$ and S is the number of samples. After obtaining the predicted action $\hat{\mathbf{a}}$, the posterior explanation is inferred as $\mathbf{q}_\phi(\mathbf{z}|\hat{\mathbf{a}},\mathbf{R},\mathbf{E})$.

4.3.2 Conditional Variational Autoencoder with Supervision (`CVAE+SV`)

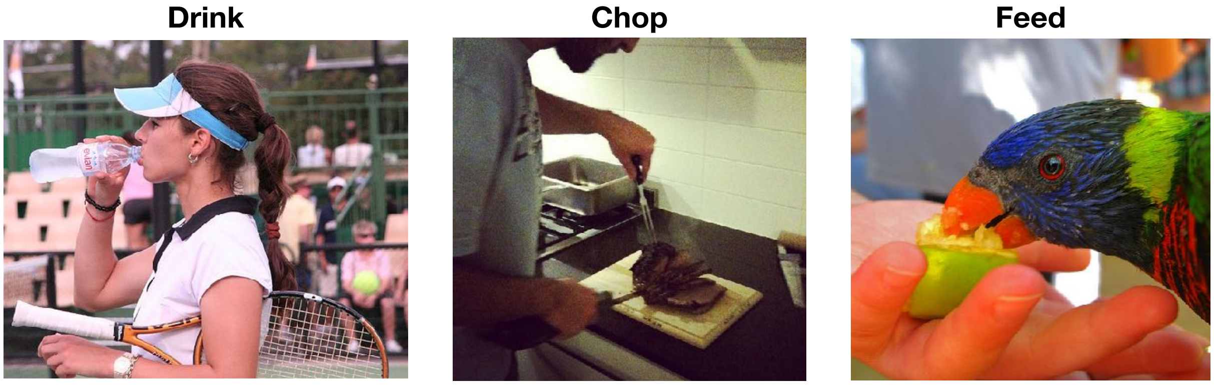
In this setting, we assume we have the supervision for the discrete latent variable z. which is more like a multi-task setting. We optimize both the action prediction loss and the evidence selection loss. The final loss function is defined as:

$$\mathscr{L}_{SV} = \lambda\mathscr{L}_{CVAE} + (1-\lambda)\mathscr{L}_{evidence}$$

where

$$\mathscr{L}_{evidence} = -\sum_k(\mathbf{z}_k\log\mathbf{p}(\hat{\mathbf{z}}_k) + (1-\mathbf{z}_k)\log(1-\mathbf{p}(\hat{\mathbf{z}}_k)))$$

in which $\mathbf{z}_k \in \{0,1\}$ is the ground truth label, $\hat{\mathbf{z}}_k$ is the predicted label and λ is a hyper-parameter.



Drink	Chop	Feed
(hold, hand, bottle)	**(carve, knife, meat)**	**(eat, bird, fruit)**
(near, bottle, mouth)	**(use, man, knife)**	**(on, fruit, hand)**
(in, water, bottle)	(on, fork, meat)	**(on, bird, hand)**
(hold, woman, racket)	(under, stove, pan)	(on, neck, bird)
(racket, orange)	**(meat, sliced)**	(apple, green)
(shirt, white)	(fork, long)	(beak, orange)

Figure 4.3: Example Crowdsourcing Annotations in which bold relations/attributes are annotated as gold.

4.4 Data Collection

To evaluate our method, we created a dataset based on the Visual Genome (VG) data [46]. Each image in the VG dataset is annotated with bounding boxes, relations and attributes describing the bounding boxes. The available annotations provided an ideal setup which allowed us to focus on studying commonsense explanation.

Table 4.1: Statistics for the average relations/attributes Mean and Std for each verb in the dataset.

	feed	pull	ride	drink	chop	brush	fry	bake	blend	eat
Ave_Rel	15.49 ± 7.55	14.62 ± 9.36	12.42 ± 7.18	15.16 ± 9.89	12.00 ± 7.22	15.40 ± 8.93	14.02 ± 7.02	13.31 ± 7.27	14.37 ± 6.37	15.08 ± 6.87
Ave_Gold_Rel	2.79 ± 1.28	1.86 ± 0.84	1.69 ± 0.83	2.41 ± 1.14	2.41 ± 1.66	2.26 ± 1.08	2.72 ± 2.06	2.25 ± 1.69	2.56 ± 1.84	2.52 ± 1.08
Ave_Att	12.48 ± 7.11	13.60 ± 7.52	12.20 ± 7.13	10.86 ± 6.52	15.09 ± 6.82	12.31 ± 8.91	15.31 ± 7.16	13.44 ± 6.84	15.22 ± 7.18	11.98 ± 6.50
Ave_Gold_Att	0.26 ± 0.48	0.20 ± 0.45	0.13 ± 0.40	0.30 ± 0.56	1.60 ± 1.33	0.22 ± 0.49	0.91 ± 1.26	0.93 ± 1.06	0.15 ± 0.40	0.41 ± 0.70

More specifically, we selected ten frequently occurred actions: *feed, pull, ride, drink, chop, brush, fry, bake, blend, eat* and manually identified a set of images depicting these actions. This leads to a dataset of 853 images, where each image comes with a ground-truth action and annotated bounding boxes as well as corresponding relations and attributes. We then showed each image to the crowd (through Amazon Mechanical Turk) and instructed the turkers to choose justifying relations and attributes from a list. Each image was annotated by three turkers. The relations or attributes that were selected by two or more turkers are considered *gold* that can be used to explain or justify the perceived action.

Table 4.1 shows some basic statistics for each action. The number of average relations and attributes in each image for different actions varies slightly. However, only a small percentage of them are considered gold. What's interesting is that the percentage of attributes considered gold is significantly less than the percentage of the relations. The sparsity of gold relations/attributes shows that it's a challenging task to learn an explainer for a target action. Some example image annotations are shown in Figure 4.3. For each image, we show a subset of relations/attributes and gold commonsense features are marked bold. We further categorize gold relations and attributes into different commonsense categories as discussed in Section 4.2. As shown in Table 4.2, the ratios of transitive relations are similar across different actions. The ratios of spatial relations and sub_actions vary for different verbs. For instance, *ride, bake, blend* tend to be explained by spatial relations more often than sub-actions. In terms of attributes, *feed, pull, ride* cannot be explained by effect states while *chop* is mainly explained by the effect state of its direct object.

Table 4.2: Statistics for the categories of annotated relations/attributes for each verb.

	feed	**pull**	**ride**	**drink**	**chop**	**brush**	**fry**	**bake**	**blend**	**eat**
Rel_Transitive	0.10	0.14	0.15	0.11	0.11	0.13	0.12	0.18	0.15	0.09
Rel_Sub_Action	0.45	0.46	0.13	0.32	0.29	0.39	0.17	0.11	0.09	0.43
Rel_Spatial	0.45	0.40	0.72	0.57	0.60	0.48	0.71	0.71	0.76	0.48
Att_Effect	0.0	0.0	0.0	0.14	0.82	0.05	0.53	0.34	0.22	0.27
Att_Associated	1.0	1.0	1.0	0.86	0.18	0.95	0.47	0.66	0.78	0.73

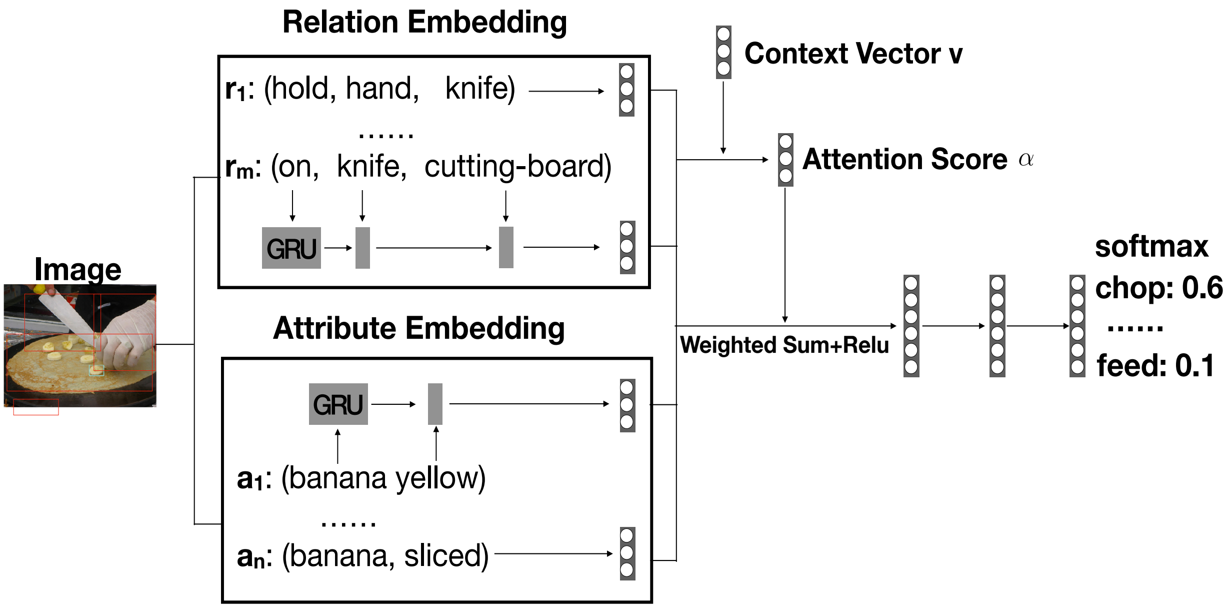


Figure 4.4: The system architecture for attention-based method.

4.5 Evaluation on Action Explanation

In this section, we first evaluate the performance of action prediction and explainer. We then compare two naive incremental learning strategies to show how they will influence the final model's performance. We also show that when we have limited data annotation, a semi-supervised method can help to improve the performance.

To evaluate our model, we randomly split our dataset (853 images) into 60% for training, 20% for validation, and 20% for test. For all the models we use the Adam optimizer with a starting learning rate 1e-4. All other hyperparameters are tuned on the validation set.

Table 4.3: Action Prediction Accuracy and Evidence Selection MAP.

	Action Accuracy	**Evidence MAP**
Random	0.1	0.251
Attention	0.789	0.442
CVAE	0.835	0.572
CVAE+SV	0.871	0.690
Upper Bound	0.918	1.0

4.5.1 Action Prediction and Explanation

For all experiments in this paper, we use the annotated relations/attributes from the original Visual Genome data. As the state-of-the-art recall@50 on the relation detection with a limited vocabulary is only around 20% [53], using annotated relations and attributes allows us to focus on the study of commonsense evidence and its role in justification and communication grounding.

We use Mean Average Precision (MAP) metric for evidence evaluation as we want to rank good evidence higher than others.

Methods for Comparison

We compare the following methods:

(1) CVAE. The conditional variational autoencoder model presented in Section 4.3.1.

(2) CVAE+SV. The CVAE model with supervision as presented in Section 4.3.2.

(3) Upper Bound. We also calculate the upper bound of the CVAE model using the human annotated gold evidence.

(4) Attention. We use an attention-based model as one of the baseline methods. It is similar with the model presented by [102]. The architecture is shown in Figure 4.4. The attention is calculated as:

$$\alpha_i = \frac{\exp(\mathbf{u}_i^T \mathbf{v})}{\sum_j \exp(\mathbf{u}_i^T \mathbf{v})}$$

where $\mathbf{v}$ is the context parameter, and $\mathbf{u}_i$ is the GRU embedding of the corresponding relation/attribute.

Table 4.4: Results from the human study on communication grounding.

	Easy				**Hard**			
	Attenton	**CVAE**	**CVAE+SV**	**Gold**	**Attention**	**CVAE**	**CVAE+SV**	**Gold**
M+H+	0.665	0.776	0.818	0.888	0.576	0.718	0.788	0.841
M+H-	0.124	0.059	0.047	0.024	0.212	0.118	0.076	0.071
M-H+	0.165	0.129	0.129	0.064	0.135	0.076	0.076	0.041
M-H-	0.046	0.035	0.006	0.024	0.077	0.088	0.059	0.047

(5) `Random`. A baseline method that randomly ranks all actions and evidence.

Evaluation Results

The results are shown in Table 4.3. Since the `Upper Bound` method directly uses the human annotated gold evidence, its MAP for selecting evidence is always 1.0.

The `CVAE` model outperforms the attention-based model in both action prediction and evidence selection tasks. This indicates that the `CVAE` model can incorporate a better guidance for evidence selection during the training process. Furthermore, after adding the evidence supervision, the `CVAE+SV` model gives even better performance in both action prediction and evident selection. We notice that for the `CVAE+SV` model, its action prediction accuracy is approaching the upper bound 91.8%, however the evidence selection MAP is still far from the upper bound even with supervision.

4.5.2 Incremental Study

Human learn new knowledge through interactions incrementally. But it's very challenging for machine to learn in an incremental way. In this section, we are interested to explore how the *CVAE* and *CVAE+SV* models work under a simulated incremental setting. Specifically, we assume the training data are received in a sequential order instead of all at once.

We evaluate two simple incremental strategies. The naive incremental strategy is to retrain the model using all available training data when new pieces of data come. The local incremental

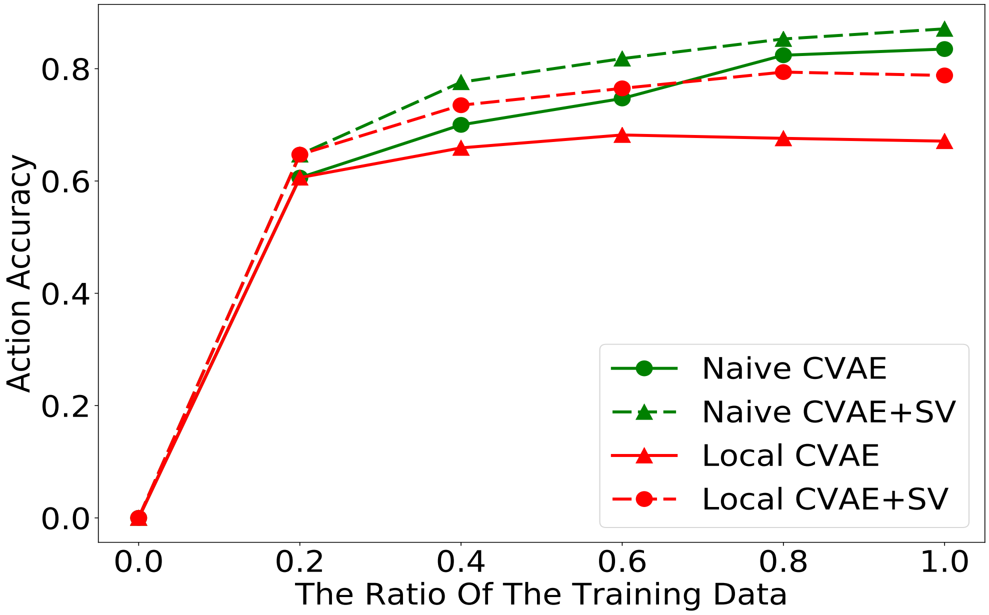


Figure 4.5: Action prediction accuracy in the incremental study.

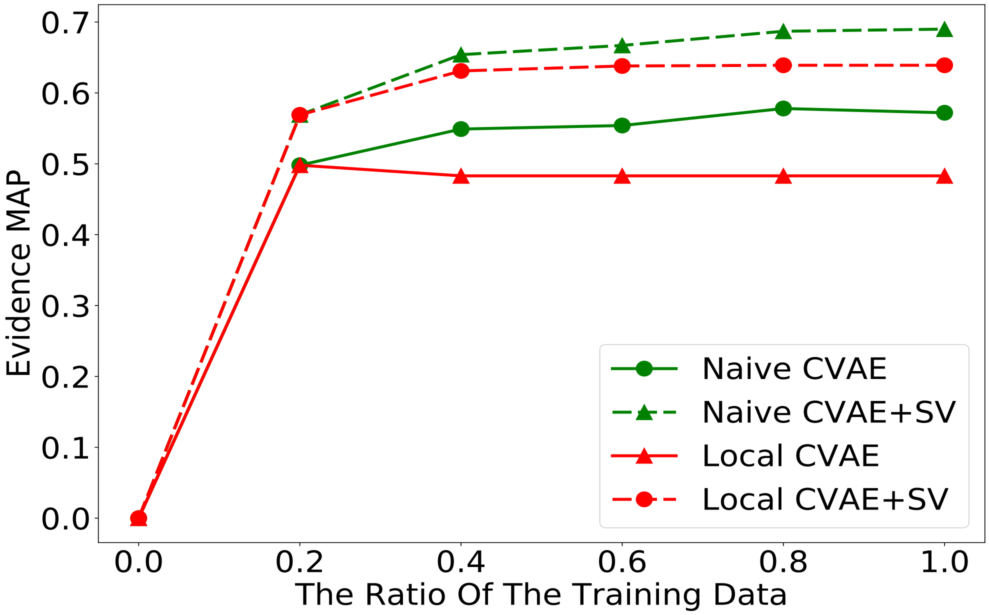


Figure 4.6: Evidence selection MAP in the incremental study.

strategy is to finetune the model only using the newly arrived data, with parameters initialized as the previous best model. Usually the local incremental strategy has shorter training time compared with the naive incremental strategy, since the local strategy has less training samples at each time. We use the same training/validation/test split as in Section 4.5.1.

Figure 4.5 and Figure 4.6 show the results of the incremental study. Overall, the naive incremental strategy outperforms the local incremental strategy.

The local incremental strategy performs worse than the naive incremental strategy with the increase of the ratio of training data. When adding the supervision, the local *CVAE* with supervison model performs better, But still worse than the naive incremental strategy.

4.5.3 Semi-Supervised Learning

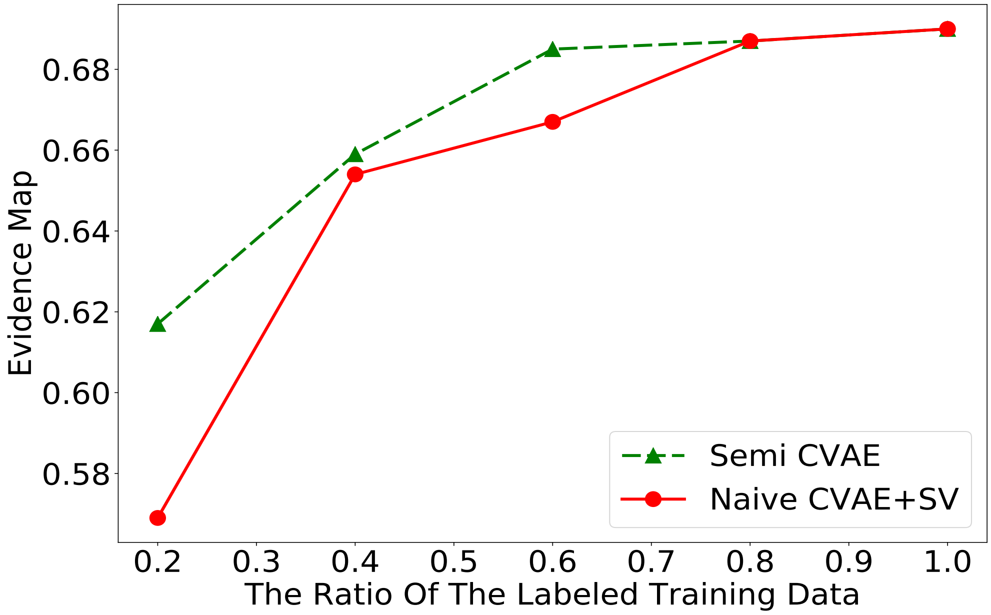


Figure 4.7: Evidence selection MAP for semi-supervised learning.

Although we have shown that it improves the model performance when we add supervision on the latent variable **z**, collecting this label information through human annotation is usually time consuming and expensive. In this section, we explore how semi-supervised learning can help to alleviate this difficulty.

As a generative model, VAE has shown its advantage on semi-supervised learning [43]. In fact our task is simpler as our latent variable $\mathbf{z}$ is also the target label $\mathbf{y}$. Following the method in [43], our semi-supervised learning loss function is defined as:

$$\mathscr{L} = \sum_{(\mathbf{a},\mathbf{R},\mathbf{E},\mathbf{z})\sim\mathbf{p}_l} \mathscr{L}_{SV} + \sum_{(\mathbf{a},\mathbf{R},\mathbf{E})\sim\mathbf{p}_u} \mathscr{L}_{CVAE}$$

where $\mathscr{L}_{SV}$ is defined in section 4.3.2 and $\mathscr{L}_{CVAE}$ is detailed in section 4.3.1. In other words, the data sample with evidence label is fed to $\mathscr{L}_{SV}$, otherwise is fed to $\mathscr{L}_{CVAE}$.

The results are shown in Figure 4.7 where the x-axis shows the ratio of labeled examples. The incremental Naive CVAE+SV model only uses the labeled evidence examples while the Semi CVAE model also uses unlabeled evidence examples. The Figure shows that the Semi CVAE model outperforms the Naive CVAE+SV model. This indicates that the semi-supervised method can improve the evidence selection by making use of unlabeled examples.

4.5.4 Visual Simulator

In the previous experiment, we assume no visual preprocessing and directly use textual input provided by visual genome.

Detecting relations from image is a very hard problem: even the close state-of-the-art recall@50 is only around 20% with a small and limited objects and relation predicate vocabulary, which is far from practical usage for our case. First Our data contains a very large vocabulary(more than 1000 objects and predicates in total) and the common way to select most frequrent occuring vocabulary can hardly cover the human annotated gold evidence due to the sparsity. Second The

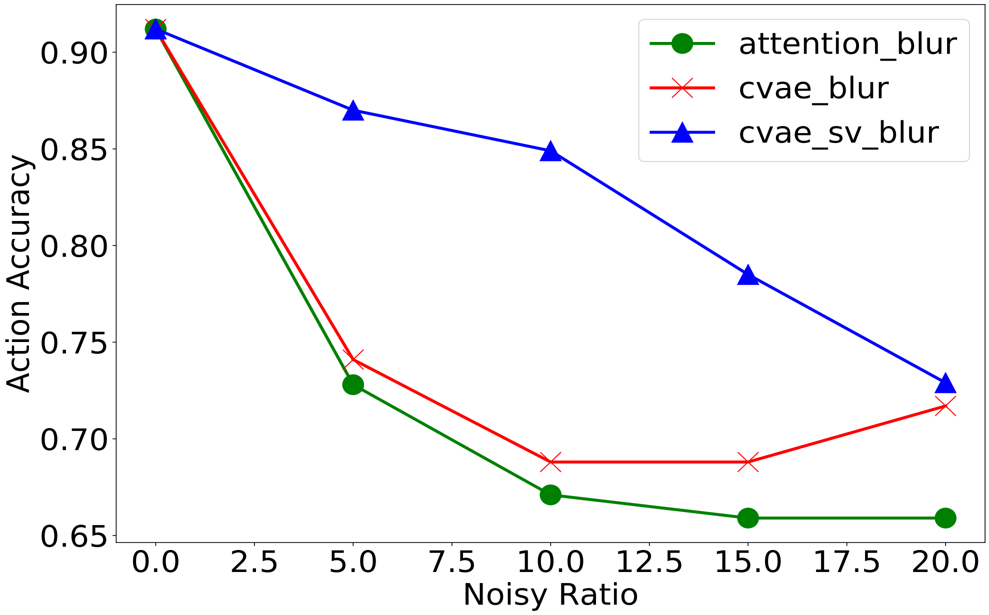


Figure 4.8: Action Accuracy For the Visual Simulator

label set also contains overlapping labels which are very similar. Third the visual genome data has the problem of missing annotations: an image contains a lot of relations while the textual annotation only covers a small ratio of them.

So We use faster rcnn [78] to detect the object bounding boxes and object classes. Then we use the structural ranking based method [54] to classify relation predicates and object attributes.

We use all data to build the visual model. Then for our task, we blurred all image with kernel size 5, and using the top predicted relations/attributes as noise mixed with human annotated gold evidence as input. We define the **Noisy Ratio** as the ratio between noisy evidence and gold evidence to estimate the robustness of the model. The reason behind doing so is that we find even we train and test on the same dataset, the recall for the gold evidence is not high due to the existence of a lot of ordinary relatios and attributes such as “in”, “on”, “color” and so on.

In Figure 4.8 and Figure 4.9, we compared different methods’ performance under different noisy ratio conditions. Each value is computed by taking the average of 3 runs. The CVAE model’s performance is slighly better the attention based model. But the gap is not large. Through man-

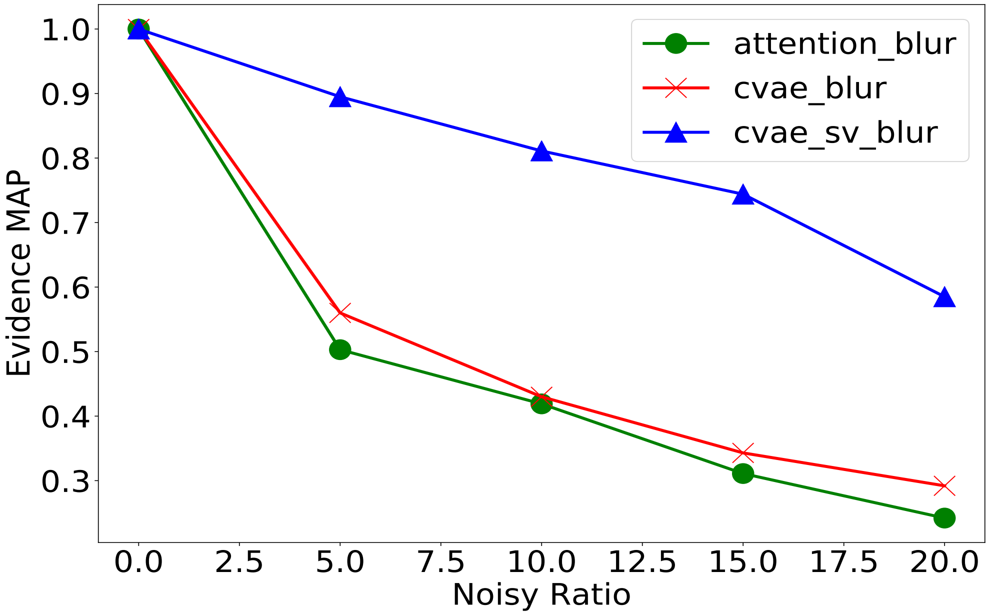


Figure 4.9: Evidence MAP For the Visual Simulator

nually checking the predicted noisy evidence, we find they have a different distribution compared with the noisy evidence extracted from image descriptions and not human-like due to the limitation of 2D image spatial bias(most top relation predicates include “in” and “on”.). How to improve the performance of the visual simulator to generate human-like evidence candidates is still an open problem we will explore in the future work.

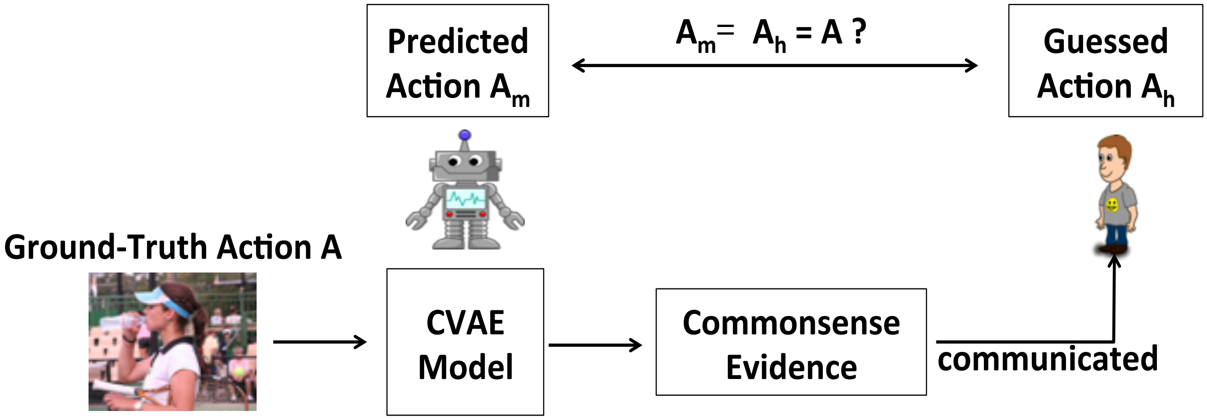


Figure 4.10: The experimental setup for the human subject study examining the role of commonsense justification towards common ground.

4.6 Commonsense Justification towards Common Ground

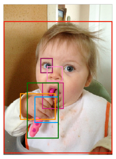
	Attention	CVAE	CVAE+SV	Gold
Gold Action: Bake	A_m: Eat A_h: Bake	A_m: Bake A_h: Bake	A_m: Bake A_h: Bake	A_m: Bake A_h: Bake
	• The bread is next to the bread. • The bread is on the rack. • The bread is on the pan. • The man has keys. • The man has the band.	• The bread is on the rack. • The bread is on the pan. • The bread is on the tray. • The bread is next to the bread. • The bread is baked.	• The bread is baked. • The bread is next to the bread. • The person is pushing the tray. • The bread is on the pan. • The bread is on the rack.	• The bread is on the tray. • The person is pushing the tray. • The bread is baked.
Gold Action: Brush	A_m: Brush A_h: Skin	A_m: Brush A_h: Brush	A_m: Brush A_h: Brush	A_m: Brush A_h: Brush
	• The baby has a mouth. • The baby has a hand. • The baby has eyeballs. • The baby has fingers. • The baby has a nose.	• The hand holds the toothbrush. • The toothbrush is in the mouth. • The baby has a mouth. • The baby has fingers. • The baby has a nose.	• The hand holds the toothbrush. • The toothbrush is in the mouth. • The baby has eyeballs. • The baby has a mouth. • The baby has a hand.	• The toothbrush is in the mouth. • The hand holds the toothbrush.

Figure 4.11: Examples of the communication grounding study based on different models.

n human-agent communication, the success of communication is largely dependent on common ground which captures shared knowledge, beliefs, or past experience [12]. As commonsense evidence what humans use to justify actions, To validate this hypothesis, we conducted a human-subject experiment to examine the role of commonsense justification in facilitating common ground.

4.6.1 Experiment Setup

Figure 4.10 shows the setup of our experiment. The agent is provided with an image and applies various models (e.g., `CVAE`) to jointly predict the action and identify commonsense evidence. The human is provided with a list of six action choices and does not have access to the image. The agent communicates to the human only the identified commonsense evidence and the human makes a guess on the action from the candidate list purely based on the communicated evidence. The idea is that, if the human and the agent share the same beliefs about evidence to justify an action, then the action guessed by the human should be the same as the action predicted by the agent.

Generating Distracting Verbs. For each image, the human is provided with a list of six action/verb candidates. To generate this list, we mix four distracting verbs with the ground-truth action verb plus a default `Other`. Most of the distracting verbs come from the concrete action verbs made available by [26]. We first manually filtered out the verbs which have the same meaning with the ground-truth verb. We then selected two groups of distracting verbs: an *easy* group (where the distracting verbs have larger distance from the ground-truth verb in the embedding space, with an average similarity of 0.284) and a *hard* group (more close to the ground-truth verbs with an average similarity of 0.479). The temperature based softmax distribution [10] was used to sample the easy and the hard distracting verbs based on the pre-trained GloVe [69] embedding cosine similarity. The selected confusion verbs are list in table 4.5.

Process. A total of 170 images were used in this experiment, and 24 workers from AMT participated in our study. For each image, we applied three different models: `Attention` baseline, `CVAE`, and `CVAE+SV` to generate the commonsense evidence. An upper bound based on gold commonsense evidence was also measured. Note that, the agent has no knowledge of the human's action choices when generating the commonsense evidence. Theory of mind is an important aspect in human-agent communication. Incorporating human's action choices in justifying action is an interesting however a different problem which requires different solutions. In this paper, we only focus on the situation where the mind of the human is opaque to the agent.

For each model and each image under the easy or hard configurations, the top five predicted commonsense evidence (associated with the predicted action) were shown to a worker. The the worker was requested to select the most probable action from the distracting list only based on these five pieces of evidence. We randomly assigned three workers to each image. The majority of three selections was considered as the final answer. If all three selections disagreed, one worker's

Table 4.5: Target Actions and Simple/Hard Confusion Actions.

Target Action	Easy Confusion Actions	Hard Confusion Actions
Bake	Grate, Knot, Burn, Frame, Other	Batter, Boil, Fry, Fold, Other
Blend	Bolt, Seperate, Bind, Wrap, Other	Weave, Chop, Twist, Squash, Other
Brush	Split, Block, Catch, Roll, Other	Frame, Stain, Skin, Scrape, Other
Chop	Crack, Wipe, Kick, Tear, Other	Grate, Burn, Bake, Scrape, Other
Drink	Nail, Smoke, Crush, Shoot, Other	Eat, Ride, Smoke, Bake, Other
Eat	Crush, Catch, Light, Ride, Other	Drink, Feed, Bite, Get, Other
Feed	Squash, Loose, Kick, Build, Other	Eat, Get, Assemble, Skin, Other
Fry	Scrape, Insert, Knock, Frame, Other	Boil, Bake, Batter, Scrape, Other
Pull	Twist, Feed, Drink, Coil, Other	Put, Drop, Lift, Pack, Other
Ride	Park, Pack, Crack, Label, Other	Get, Open, Throw, Sail, Other

Table 4.6: Results from the human subject study on common ground.

	Attenton	**CVAE**	**CVAE+SV**	**Gold**
Easy	0.665	0.776	0.818	0.888
Hard	0.576	0.718	0.788	0.841

choice was randomly selected as the final answer.

Metrics for Common Ground. We use the agreement between the action guessed by the human and the action predicted by the agent to measure how well the selected commonsense evidence serves to bring the human and the agent to a common ground of perceived actions. More formally, as shown in Figure 4.10, given an image, suppose its ground-truth action is A, the action predicted by the agent/machine is A_m, and the action guessed by the human is A_h, the *Common Ground* is defined as: $A_m = A_h = A$. Here we also enforce that the predicted action should be the same as the ground-truth action. The percentage of trials based on different models that have led to a common ground is measured and compared.

4.6.2 Experimental Results

Table 4.6 shows the comparison results among various models and the upper bound where the gold commonsense evidence provided to the human. It's not surprising that performance on common ground is worse in the *hard* configuration as the distracting verbs are more similar to the target action. The CVAE-based method is better than the attention-based method in facilitating common ground.

Figure 4.11 shows two examples of the top five predicted evidence under different models. For each model, it also shows the agent predicted action (A_m) and the human guessed action (A_h). In both examples, all models were able to establish a common ground except for the attention-based model. The evidence selected by the CVAE+SV model is clearly more accurate than the CVAE model and is more close to the ground-truth evidence. The second example shows that although the attention-based model predicts a correct target action, it fails to convey correct commonsense evidence to establish a common ground with the human.

4.7 Conclusion

This chapter describes an approach to action justification using commonsense evidence. As demonstrated in our experiments, commonsense evidence is selected to align with humans' justification of an action and is therefore critical in establishing a common ground between humans and agents. As a first step in our investigation, this work is based on annotated symbolic descriptions from perception. This assumption allows us to focus on higher level commonsense reasoning and supports a better understanding of the role of commonsense evidence in explanation and communication grounding. Our future work will extend the model and findings from this work to vision processing that will not only identify commonsense evidence but also explain where and how in the perceived environment the evidence is gathered.

Chapter 5

Grounded Action Justification

In the previous chapter, we conduct some pilot studies on relations between physical actions and the latent structured commonsense justifications. However, previous experiments are based on pure textual inputs (we assume the relations and attributes of the image are given), which is not realistic in the real world. Furthermore, the empirical results are conducted on a limited action vocabulary, how to efficiently deal with diverse actions in the real world is still an open and challenging problem. In this chapter, we try to addresses the problem of learning to justify perceived actions through natural language rationales. We propose a deep factorized network which jointly models the relations between the shared environment, perceived actions, and action justifications. Our empirical results have shown that the proposed model outperforms strong baselines in the overall performance. By explicitly modeling factors of language grounding and commonsense reasoning, the proposed model provides a better understanding of effects of these factors on grounded action justification.

5.1 Introduction

To boost the research of multi-modal visual understanding, researchers propose different language and vision tasks as test beds for different methods. One of the most popular tasks is visual question answering: given an query and image, how to select the correct answer from a list of candidate answers? However, for commonly used visual question answering datasets, most questions are targeted for objects or properties, which do not consider higher level cognitive reasoning ability. Motivated by this, the Recognition to Cognition [105]) dataset is collected and a new visual commonsense reasoning (VCR) task is proposed: given an image and a question, the goal is to select the correct answer and the corresponding rationale from a list of candidates.

The VCR task is closely related with our previous work, but they are different on following perspectives:

- Our previous work assumes that the given inputs are textual representations, and we directly model the linguistic texts without grounding action justifications.
- Our previous work performs experiments on a limited action vocabulary, while the VCR dataset are collected from complex movie scenes which contains complex and diverse actions. Besides, our previous work represent actions as single verbs instead action phrases or sentences. But in the VCR dataset, the gold answers could be long sentence descriptions. The VCR task is more challenging compared to our previous task in Chapter 4.
- Our previous work assumes structured but simplified justifications containing independent relations and attributes, while in the VCR task, the rationales are complex sentences including multiple verbs, objects, attributes and their interactions.

Mathematically, the visual commonsense task is defined as:

$$A, R = \arg\max_{A,R} p(A, R | Q, I)$$

where I is the given image, Q is the question, A is one of the answer choices and R is one of the rationale choices. In the previous work [105], this joint process is modeled as a two-step process

$$p(A,R|Q,I) \propto p(A|Q,I)p(R|Q,I,A)$$

where the first step is to select the best answer choice based on the question and image, and the second step is to select the most probable rationale according to the image, question and inferred answer. However, it's counter-intuitive for humans to solve a similar question using two-step process. When humans answer visual questions, the answering process and the rationale reasoning process interacts and often happens simultaneously. So we propose a joint learning and inference method to solve the problem $p(A,R|Q,I)$.

Specifically, to learn to justify perceived actions through natural language rationales, we formulate the problem as: given an image (*I*) and a question (*Q*) about the activity from the image (e.g., "what is person 1 doing?"), the goal is to identify an answer (*A*) from a list of potential answers (e.g., "person 1 is drinking water") and a rationale *R* that supports the answer from a set of rationale candidates (e.g., "person 1 is holding a cup") at the same time. Our solution to the problem is to jointly infer (A,R) that maximizes $P(A,R|Q,I)$ as rationales and predictions often support each other in the decision making process. We do not consider rationales as post-hoc justifications as in the the original VCR task. This is a key difference between our setup and the VCR setup. Because of this difference, the original VCR dataset (i.e., *R2C* dataset) cannot be directly applied here. Actually, we augment a portion of the origial *R2C* dataset for our investigation of the joint problem(details described in Section 5.2).

As there are intrinsic relations between the image, question, answers, and rationales, we develop a factorized deep neural network which explicitly models these relations to capture the following intuitions: **`(A,Q,I)`**: a good answer has to be grounded to the image content given the question.

(R,Q,I): a good rationale has to be grounded to the image content given the question.

(A,R): a good rationale to an answer should follow general commonsense knowledge.

The first two factors (i.e., (A,Q,I) and (R,Q,I)) concern about the ability to ground language to perception (i.e., together they are referred to as the *language grounding* factor) and the third factor addresses the ability of reasoning based on commonsense knowledge that may hold between answers and rationales (i.e., the *commonsense reasoning* factor).

The contributions of this work are that, instead of treating rationales as post-hoc justifications, we propose a model that jointly infers actions and rationales. By decomposing complex relations between images, questions, action answers, and action rationales, the proposed model not only outperforms strong baselines, but more importantly promotes a better understanding of the role of language grounding and commonsense reasoning on grounded action justification.

5.2 R2C Dataset Augmentation

The original R2C dataset was collected through amazon mechanical turk. Each turker is given an image with detections and contextual descriptions, and is requested to generate 1 to 3 questions with answers and rationales. The images are extracted from movie clips. To generate the negative candidates for the answers and rationales. The authors proposed a so-called adversarial matching method in two steps.

- For the first step, The candidate negative answers are selected from all the candidate answers which are the top most similar answers compared with the query.

- For the second step, the candidate rationales are selected from all the candidate rationales which are the top most similar rationales with the query and the ground truth answer.

After these two steps, one sample is composed of one query, one image, four candidate answer choices which contain the ground truth answer and four candidate rationale choices which contain the ground truth rationales. From the previous two step descriptions of dataset generation process, a problem when we jointly infer the answer and rationale is the answer bias problem. As currently all four rationale candidates are related with the correct answer, which leads to the leaking of answer information.

To fix the answer bias, one possible solution is to introduce negative rationales not only similar to the correct answer, but also similar to the negative answers. Fortunately, in the original R2C dataset, To avoid the linguistic bias, each answer candidate(including the wrong answer) has at least one time acting as the correct answer appearing in other samples. We can borrow rationale candidates for each negative answer from where it appears as the correct answer, which means that now each sample is composed of 16 rationale candidates(of which 12 are augmented rationales). In this way, we can eliminate the answer bias by enforcing the model not only learn to reason about the textual relations between the answer and the rationale, but also ground the justifications.

The last problem remaining with this kind of dataset augmentation is that how to fill in the object tagging for new augmented rationales. In the original collected dataset, the turkers directly use bounding box tagging numbers instead of the objects' names in the rationales. For example, one possible rationale could be "**[2] is a professional musician in an orchestra.**" Here [2] is the tagging number for one of the bounding boxes in the image. When we augment new rationales, we need to re-map the old tagging numbers to the new tagging numbers. We use a heuristic based method to do the re-mapping.

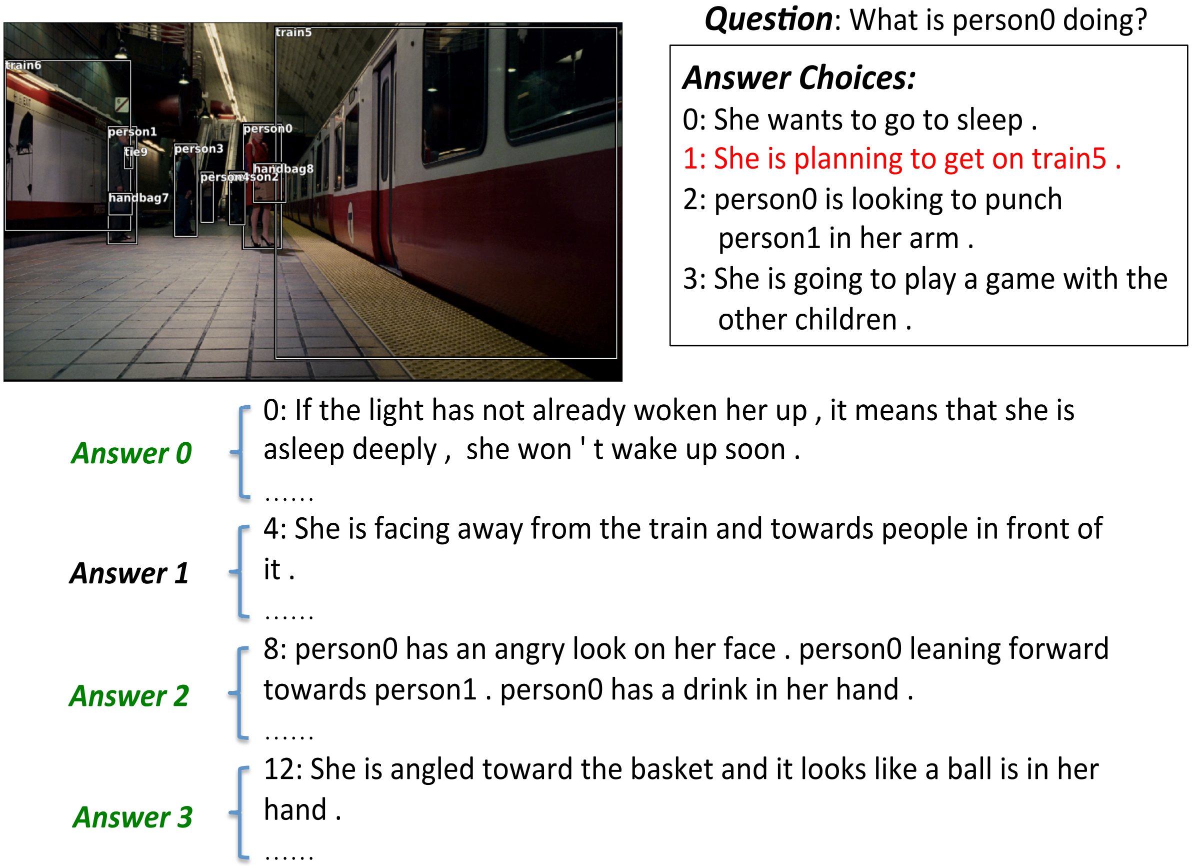
- As the same answer appears in two samples in which sample s_i is the correct one and sample s_j is the wrong one, we build object tagging mappings between these two samples for objects appearing in the same answer. Following this mapping, we map the sample s_j's taggings to sample s_j's taggings for all sample s_j's borrowed rationales.

- For tagging numbers not appearing in the first step's mapping, we first try to match with tagging numbers in sample s_i which have the same tagging labels as in the original sample s_j. If there exist multiple tagging numbers for the same label, we randomly assign one of the tagging numbers to build the mapping.

After these steps, we build an unbiased augmented dataset for the joint VCR task. Here we show an concrete example to illustrate the augmented sample in Figure 5.1. The upper left corner shows the image with auto-detected bounding boxes. The upper right corner shows the question and four answer choices. The red colored answer is the ground truth answer. Below the images we show all the candidate rationales. Each answer corresponds to 4 candidate rationales(for simplicity, here we only show one specific rationale). The rationales for answer 1 come from the original r2c dataset, and all the other rationales come from the data augmentation process, which are colored green.

5.3 Deep Factorized Joint Modeling for Grounded Question Answering and Explanation.

In this section, we will detailed our proposed method. First we will briefly review and discuss the behind model intuitions. Then we will mathematically detailed the model architecture and learning process.



Question: What is person0 doing?

Answer Choices:

0: She wants to go to sleep .

1: She is planning to get on train5 .

2: person0 is looking to punch person1 in her arm .

3: She is going to play a game with the other children .

Answer 0

0: If the light has not already woken her up , it means that she is asleep deeply , she won ' t wake up soon .

……

Answer 1

4: She is facing away from the train and towards people in front of it .

……

Answer 2

8: person0 has an angry look on her face . person0 leaning forward towards person1 . person0 has a drink in her hand .

……

Answer 3

12: She is angled toward the basket and it looks like a ball is in her hand .

……

Figure 5.1: An example augmented R2C sample.

5.3.1 Motivation

Humans tend to answer the visual questions and generate explanations for the answer within a joint process instead of the post-hoc two step process. Motivated by this intuition, we propose a joint algorithm to predict the answer and the rationale simultaneously. The benefits from this joint process are: first it is more consistent with the human thinking process, humans do not separate the visual question answering process and the explanation process into two processes. Second the joint modeling process can better incorporate interactions between the language and vision, which helps alleviate the error propagation problem in the two step process.

5.3.2 Deep Factorized Modeling

According to the theory of un-directed graphical model, we can write the probability distribution $p(A,R|Q,I)$ as:

$$p(A,R|Q,I) = \frac{1}{Z} exp(\Phi(A,R,Q,I))$$

where $\Phi(A,R,Q,I)$ is the potential function of the factor and Z is the normalization constant which is

$$Z = \sum_{A,R} exp(\Phi(A,R,Q,I)$$

The meaning of the random variables A,R,Q,I are consistent with previous definitions.

Then the key problem turns out to be **how do we factorize the whole big factor** $\Phi(A,R,Q,I)$. Before answering this question, let's first think: what are the good answers and rationales. Good principles include:

- A good answer must be grounded to the image well given the question.
- A good rationale must be grounded to the image well given the question.
- A good answer and a good rationale must match well.

Based on these three assumptions, we factorize the big factor into 3 smaller factors, shown in Figure 5.2 Formally, the factor model can be written as

$$\Phi(A,R,Q,I) = \Phi(A,Q,I) + \Phi(R,Q,I) + \Phi(A,R))$$

The factor $\Phi(A,Q,I)$ is kind of similar with the VQA task which captures the grounded semantics of the answer given the image and the question. The factor $\Phi(R,Q,I)$ captures the grounded semantics of the rationale given the image and the question. The last factor $\Phi(A,R)$ captures

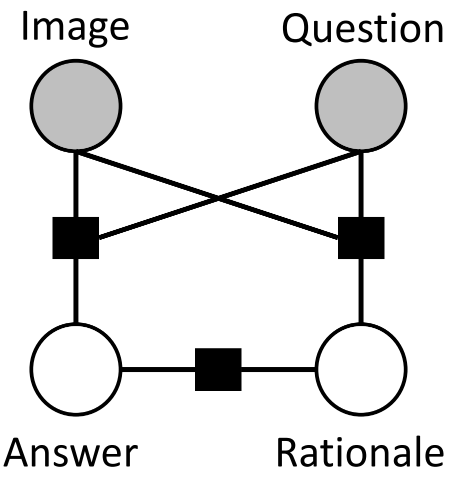


Figure 5.2: A Graphical model representation of the joint VCR task.

the correlations between the answer and the rationale. By jointly optimizing these three factors, we hope to infer the answers and rationales which are both grounded in the image and also have matching causal semantics well.

For each factor, we use an neural network to approximate the potential function. In the following parts, which is shown in Figure 5.3. We will detail each factor's neural architecture.

5.3.2.1 Visual Question Answering

In this subsection, we build an neural network to predict the factor potential $\Phi(A,Q,I)$. The whole architecture is shown in the left part of Figure 5.3.

Input Representation: In this paper, we use Bert [18] for both the question, answer and rationale representation. Let's denote the question Q as $q_1,q_2,...,q_{l_q}$ where each q_i is the word and l_q is the length of the question. Similarly we let the answer A as $a_1,a_2,...,a_{l_a}$ where each a_i is the word and l_a is the length of the answer. so using Bert, we can transform the question as a matrix $Q_I \in R^{l_q\times 768}$, and the answer as a matrix $A_I \in R^{l_a\times 768}$. For the image I, we use a pre-trained resnet [31] to extract visual features $I_h \in R^{c\times w\times h}$, and we reshape this 3 dimensional tensor as 2d

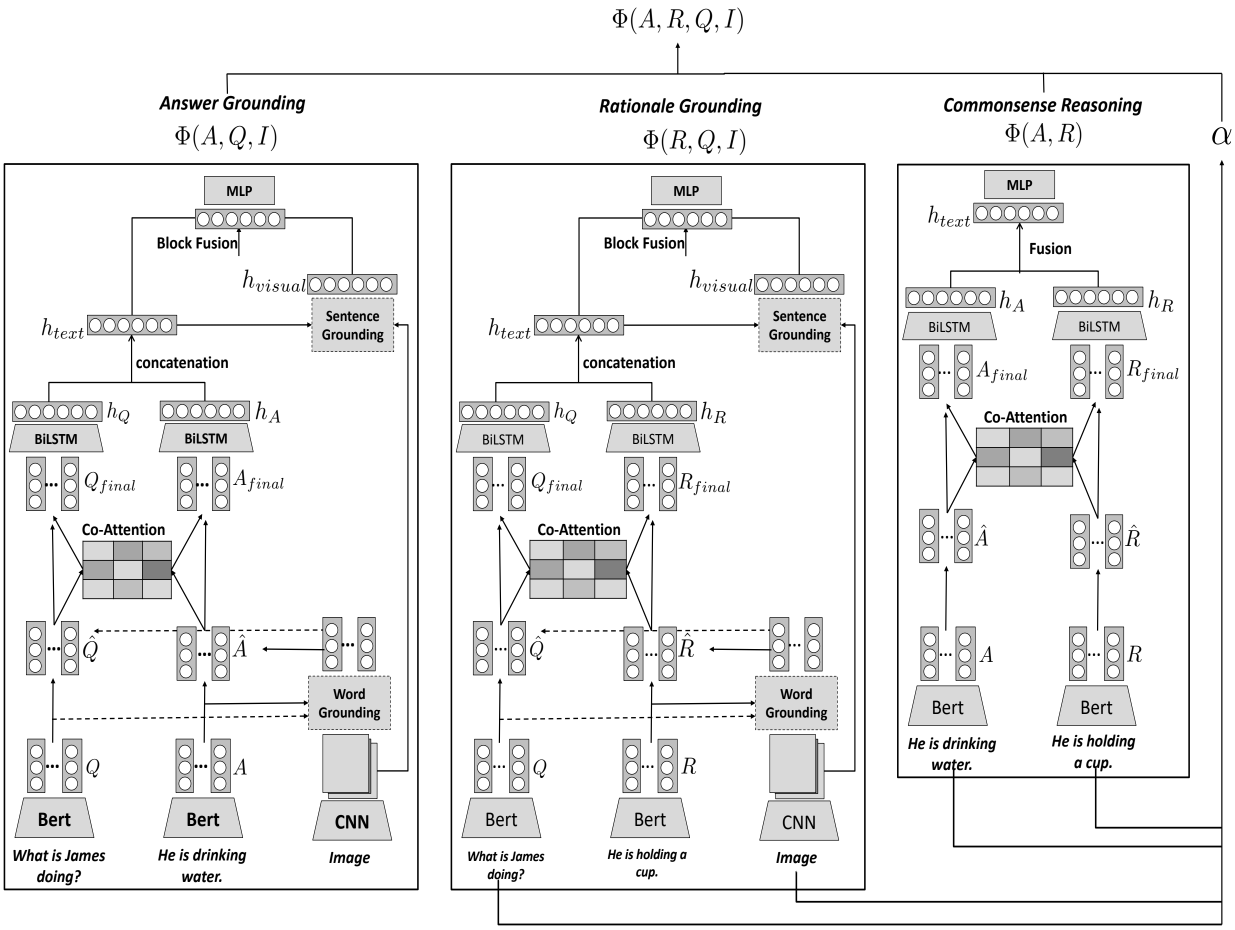


Figure 5.3: The neural network architecture for deep factorized network.

matrix $I_h \in R^{w*h,c}$.

Word-Level Grounding: The first step is to get word level grounding. the goal is to enrich each word with a corresponding visual representation. For example, for some objects or nouns presented in Figure 5.1, we hope the model can automatically learn where should be focused on. To achieve this, we apply a bi-linear attention mechanism between the sentence's words Bert representation and the image visual representation

$$s_{ij} = Q_I^i M_1 I_h^j$$

$$a_{ij} = \frac{s_{ij}}{\sum_j s_{ij}}$$

where Q_I^i is the Bert representation of the i_{th} word in the question, and I_h^j is the approximate j_{th} regional representation of the image. The attention vector a is gotten by the softmax of the scores s_{ij}. The final grounded visual representations for words is calculated as:

$$v_i = \sum_j a_{ij} I_h^j$$

In our augmented R2C dataset, some words are represented by bounding box tagging number as we denoted before. We directly using the corresponding bounding box visual feature extract from Mask-RCNN [30]. For words without bounding box tagging, we use attended visual representation. we directly concatenate the visual representation v_i with the original Bert representation Q_I^i to form a new sentence representation for the question and the answer:

$$\hat{Q}_I = [Q_I^1, v_1; ..., Q_I^{l_q}, v_{l_q}]$$

$$\hat{A}_I = [A_I^1, v_1^*; ..., A_I^{l_a}, v_{l_a}^*]$$

where v_i^* is the corresponding visual representations for words in the answer.

Contextualized word representation: To get a better meaning representation of words, we use co-attention to get contextualized representations for the question and answers. Following the previous step, we get grounded question representation $\hat{Q}_I \in R^{l_q \times d}$ and grounded answer representation $\hat{A}_I \in R^{l_a \times d}$. First we calculate a correlation score matrix:

$$S = \hat{Q}_I M_2 \hat{A}_I^{\,T}$$

The scores for each row represent how each word in the question attends the words in the answer. The scores for each column represent how each word in the answer attends the words in the question. Following the previous attention mechanism, we get answer-guided question representation $\hat{Q}_I^* \in R^{l_q \times d}$ and question-guided answer representation $\hat{A}_I^* \in R^{l_a \times d}$. The final contextualized question and answer representation are $Q = [\hat{Q}_I, \hat{Q}_I^*]$ and $A = [\hat{A}_I, \hat{A}_I^*]$.

To get a summary of textual information for both the question and answer, we run bi-directional LSTM on both the question and answer.

$$h_f^q = LSTM_{forward}(Q)$$

$$h_b^q = LSTM_{backward}(Q)$$

$$h^q = [h_f^q, h_b^q]$$

$$h_f^a = LSTM_{forward}(A)$$

$$h_b^a = LSTM_{backward}(A)$$

$$h^a = [h_f^a, h_b^a]$$

we concatenate h_q and h_a to get the final textual representation

$$h = [h_q, h_a]$$

Sentence-Level Grounding: We use the textual summary vector to re-attend the visual representation to get a better visual representation. Specifically, we use the summary vector h to attend the visual representation I_h. By similarly using the attention mechanism for word level grounding, we get a visual representation h_{visual}.

To get the final factor potential prediction, we first concatenate h and h_{visual} to get h_{final}. Then apply an fully connected linear layer:

$$f = Wh_{final} + b$$

5.3.2.2 Visual Question Rationale(VQR)

The goal of the module is to get the grounding score of the rationales given the image and the question, So we use a similar architecture with the module VQA. The detailed architecture is shown in the Figure 5.3. The difference is to replace the answers in the VQA module with rationales. Actually these two modules also share all the parameters used in the architecture, which helps to decrease the risk of overfitting.

5.3.2.3 Causal Matching Between the Answer and the Rationale

The last module for our work is the answer rationale matching. The essence of this module is trying to learn the causal relation between sentences. In the right part of the Figure 5.3, we show how we do the causal match between the answer and the rationale. The procedure follows a simplified process of VQA/VQR without visual information. The first step is to use Bert to embed both the answer and rationale, then we build contextualized answer representation and rationale representation using co-attention mechanism. We use a shared Long-Short Term Memory Unit to encode these two sequential representations. Finally we concatenate these two embeddings and use a final linear layer to get the causal factor potential.

5.3.3 Training and Inference

During the training, we use maximum likelihood as the training criteria.

$$
\begin{aligned}
L &= logP(A,R|Q,I) \\
&= log\frac{exp(\Phi(A,R,Q,I))}{\sum_{A,R} exp(\Phi(A,R,Q,I))} \\
&= \Phi(A,Q,I) + \Phi(R,Q,I) + \Phi(A,R) - log\sum_{A,R} exp(\Phi(A,R,Q,I))
\end{aligned}
\tag{5.1}
$$

All the parameters are optimized end-to-end jointly.

During the inference, we are trying to seek the answer A and rationale R where

$$A,R = \arg\max_{A,R}(\Phi(A,Q,I) + \Phi(R,Q,I) + \Phi(A,R))$$

Adaptive Factor Weighting(AFW): In previous loss function, we assume all the 3 factors are equally important, however, it's more intuitive that we use adaptive weights for different factors based on the whole context. So we instead decompose the large factor as:

$$\Phi(A,R,Q,I) = \alpha_1\Phi(A,Q,I) + \alpha_2\Phi(R,Q,I) + \alpha_3\Phi(A,R)$$

where the vector α is learned as model parameters:

$$s_\alpha = g(A,R,Q,I)$$

$$\alpha = softmax(s_\alpha)$$

The function g is a multi-layer network to be learned based on the concatenation of the representation of A, R, Q and I. Correspondingly during the inference, the best answer and rationale is inferred by

$$A,R = \arg\max_{A,R}(\alpha_1\Phi(A,Q,I) + \alpha_2\Phi(R,Q,I) + \alpha_3\Phi(A,R))$$

5.4 Experiments and Results

In this section, we first will show some basic statistics about the augmented statistics. Then we will show some experimental results compared with different methods. Third we will explore different ablations studies to better understand how important different modules play. Finally we will show some qualitative analysis on the attention mechanism we used in different modules.

5.4.1 Dataset Statistics

In this section, we will show some basic statistics for the training and testing.

As we stated before, After augmented the original R2C dataset, for each sample, we have 1 image, 1 visual question, 4 answer choices and 16 rationale choices. The full dataset is divided into 10 folds, where different folds contain unique image set. and we randomly choose 8 folds as training, 1 fold as validation and the left fold as testing. As the original dataset contains all kinds of question including activity, explanation, temporal and so on. In this work, we are focusing on actions, so we use a simple rule to filter the action related samples in the original dataset. Specifically we extract samples whose question conains one of these following keywords: **doing, looking, event, playing, preparing**.

In table 5.1, we show the basic statistics of of the train/validation/test dataset including the number of samples, the number of total tokens, and the number of unique images.

In Figure 5.4, Figure 5.5 and Figure 5.6, we show the histogram distribution of questions,

Table 5.1: Basic statistics for the augmented R2C dataset.

	Train	Validation	Test
No. of Samples	33687	4811	4816
No. of Tokens	10.8m	1.55m	1.57m
No. of Images	27193	3878	3869

Table 5.2: Validation and Test Accuracy for all the models.

Method\Accuracy	**Validation**	**Test**
R2C	0.348	0.344
R2C_beam	0.364	0.364
VAE	0.248	0.245
MLP	0.327	0.342
Factorized Model(Our Work)	**0.385**	**0.391**

answers and rationales on the train/validation/testing dataset. From these 3 Figures, we can see that the question/answer/rationale has a similar distribtuion across the train, validation and test dataset. When comparing the histrogram in the same dataset, we can clearly see that the rationales' average length is longer than the answers' average length.

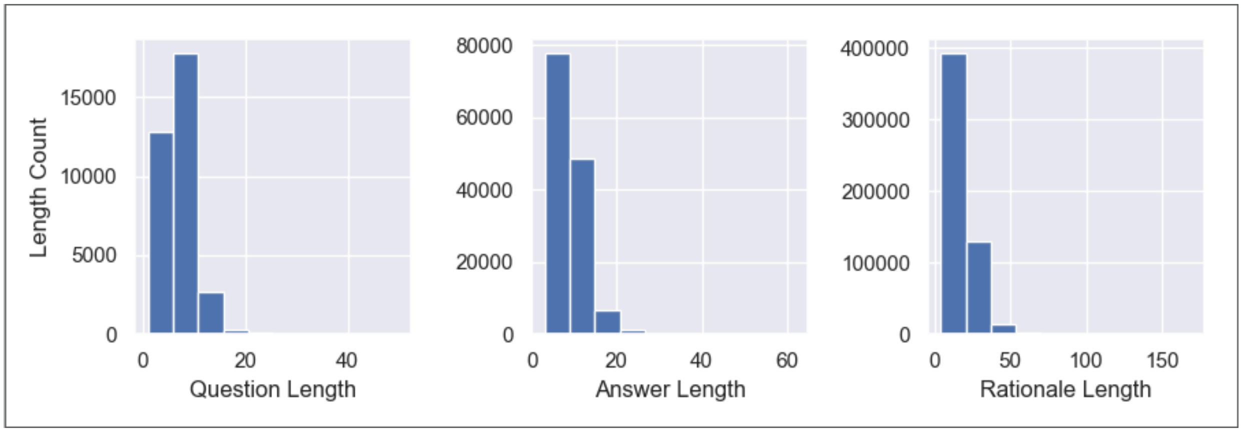


Figure 5.4: The length distribution of the sentences in the training dataset.

5.4.2 Results

To evaluate the effectiveness of methods, we compare our joint model with following baseslines:

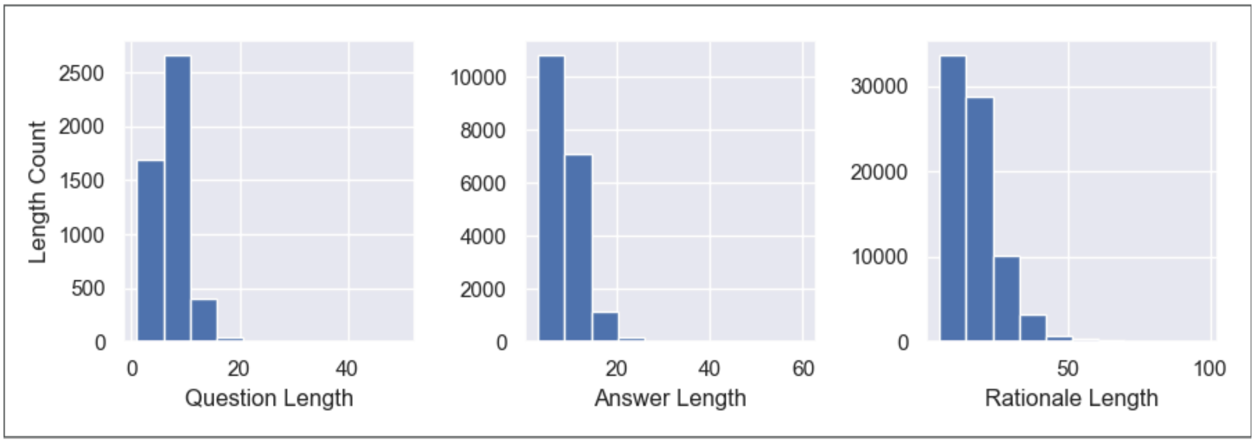


Figure 5.5: The length distribution of the sentences in the validation dataset.

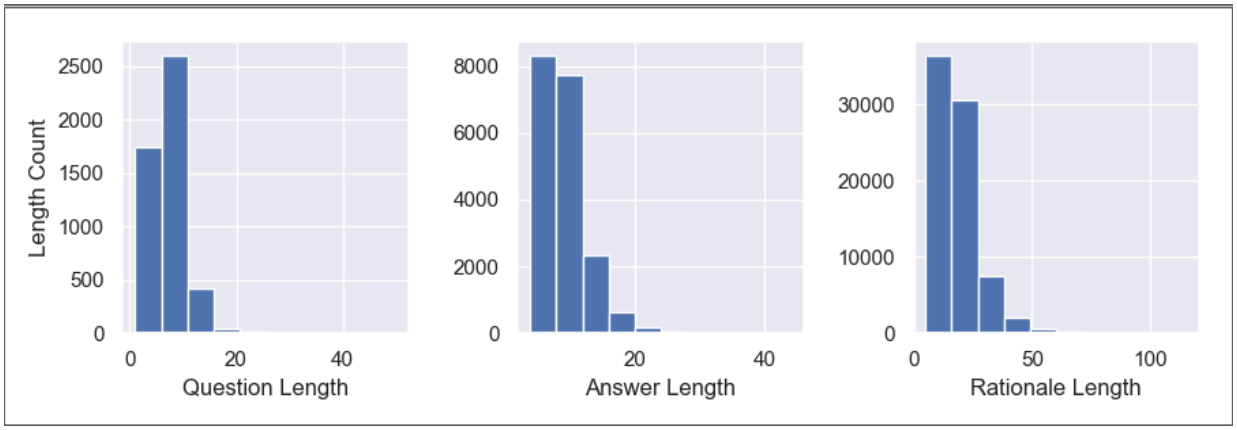


Figure 5.6: The length distribution of the sentences in the testing dataset.

- **R2C**: the original methods proposed by the work [105], which utilize a two step strategy: first predict the action, then predict the rationale.

- **R2C_beam**: an extension of the **R2C** method. when inferring the rationale, using the beam search to search the best combination of the prediction of the answer and rationale.

- **VAE:** We also extend the previous VAE based method to incorporate the question and visual information.

- **MLP:** As an important common baseline for the VQA task, Here we also build an MLP baseline for the joint R2C ask.

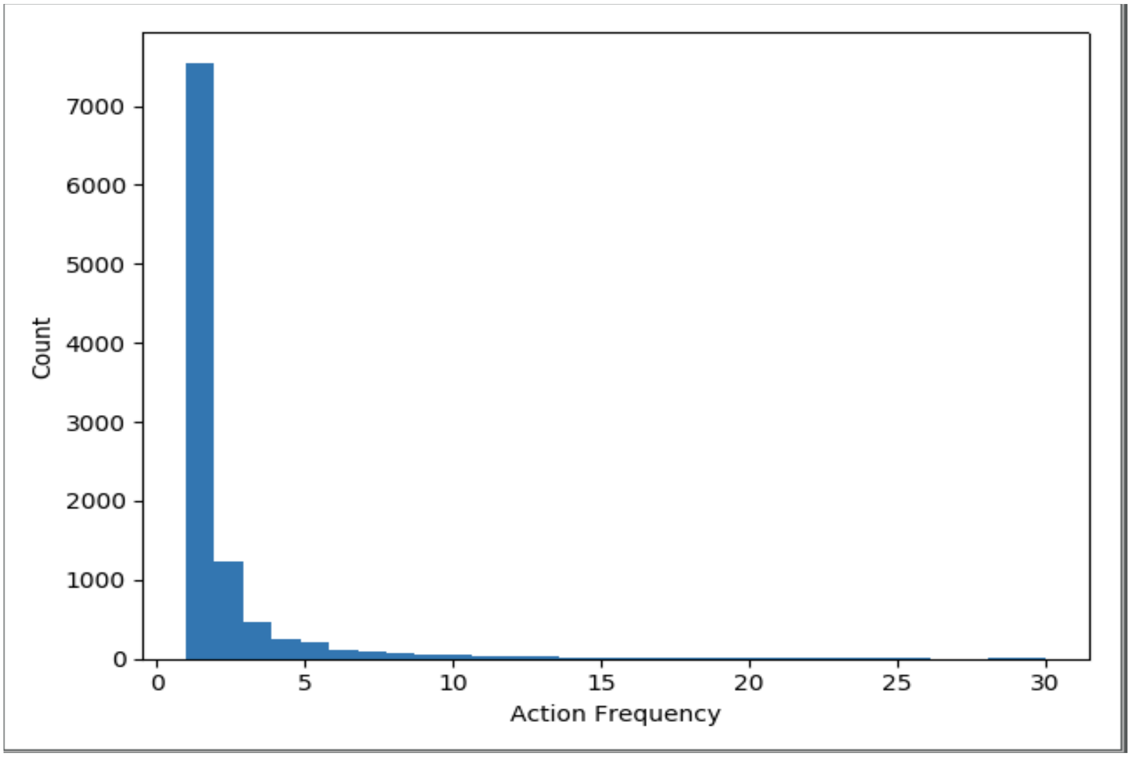


Figure 5.7: The Histogram of action frequency in the training dataset.

As we use Bert as linguistic representation, the first step in our training process is to finetune the Bert model on our dataset. Specifically we generate positive and negative pairs of question answer, answer rationale and question rationale pairs. Then we treat the Bert finetune as a two-way binary classification problem. Following the setup of the original R2C work, we use adam optimzer with learning rate 2×10^{-5}.

For all models' training process, we use adam optimizer with a learning rate of 2×10^{-4} and weight decay of 10^{-4}. The gradients are clipped to have L2 norm of at most 1.0. We set the hidden layer dimension as 512.

The table 5.2 shows different methods' performance on the validation and test dataset. First we can see that our factorized model performs best compared with other methods. Second we can see that the best model besides our work is the **R2C_beam** method although as a two-step method. Third we notice that although both the **VAE** and **MLP** are also joint model, they do not perform well compared with the R2C method. This shows that carefully modelling the interaction between different parts or modules are also very important. For **VAE** based method, we can see that it performs much worse than other methods. One Possible reason is that, to make the whole VAE system end-to-end differentiable, the effective approximate gumbeling sampling process for the rationale given the answer is more challenging. By comparing the **R2C** method and **R2C_beam**, we can see beam search can effectively mitigate the propagated errors during the independent inference process of **R2C**.

5.4.3 Ablation Study

To have a better understanding of how different factors performs in our model, we conducted some ablation studies how important different factors are both quantitatively and qualitatively.

First we want to test how important for each factors we designed for the Joint R2C problem. We are interested in the following settings for the Factorized Model(FM):

- **FM-AFW**: the factorized model without adaptive factor weighting. The goal of this study is to test how important the adaptive factor weighting plays in the system.
- **FM-QA**: the factorized model without the visual question answer factor. The goal is to test how important the VQA factor plays.
- **FM-QR**: the factorized model without the visual question rationale factor. The goal is to test how important the VQA factor plays.
- **FM-AR**: the factorized model without the answer rationale causal matching factor. The goal is to test how important the AR factor plays.
- **FM-IM**: The factorized model without the visual information. In this setting, we mask all

Table 5.3: Ablation Study Results.

Method\Accuracy	**Validation**	**Test**
FM	**0.385**	**0.391**
FM-AFW	0.371	0.382
FM-QA	0.333	0.346
FM-QR	0.309	0.312
FM-AR	0.194	0.209
FM-IM	0.265	0.271

the visual context, and try to test how the pure textual based system performs in the joint R2C task.

The results for the ablation studies are shown in table 5.3. From this table, we can see that different factors have different influence on the model's performance. We can see that when add the adaptive factor weighting, we actually improves the model's performance both on validation and test. Besides, among the three factors **QA**, **QR** and **AR**, the factor **AR** plays the most important role, as we can see that the performance dropped the most when removing the factor **AR**. Besides, the rank of these 3 factors are **AR**, **QR** and **QA** with decreasing importance. Besides, we can also see that even if we completely remove the visual image, the validation and testing performance decreased, but is even higher than the setting removing the factor **AR**.

Next we do some qualitative analysis to understand what exactly does model learn. Because we are mainly interested in the actions in this work. The first step we do is to show some distribution of the actions in our training dataset. To extract the actions in the ground truth answers in the training samples, we conduct dependency parsing on the ground truth answer, then select the possible combinations of **verb** or **verb+noun** pattern. The histogram of the action frequency is shown in Figure 5.7. The x-axis shows the frequency of actions, while the y-axis shows the histogram count. We can see a severe long train distribution where most of the action frequency are within 5. This also verifies that why introducing Bert helps a lot to improve the final performance in the R2C

Table 5.4: Effect of factors on different actions.

	No.	DFN	-AR	-IM
Look	516	0.389	**0.219**	0.264
Get	180	0.461	**0.267**	0.3
Watch	152	0.428	**0.184**	0.316
Talk	137	0.365	**0.182**	0.248
Dance	107	0.411	0.299	**0.215**
Wait	89	0.404	**0.146**	0.213
Drink	89	0.337	0.269	**0.213**
Eat	89	0.472	**0.213**	0.326

work [105] as the Bert is pre-trained on a large dataset using language modelling, thus the learned Bert representation can lead to better generalization performance.

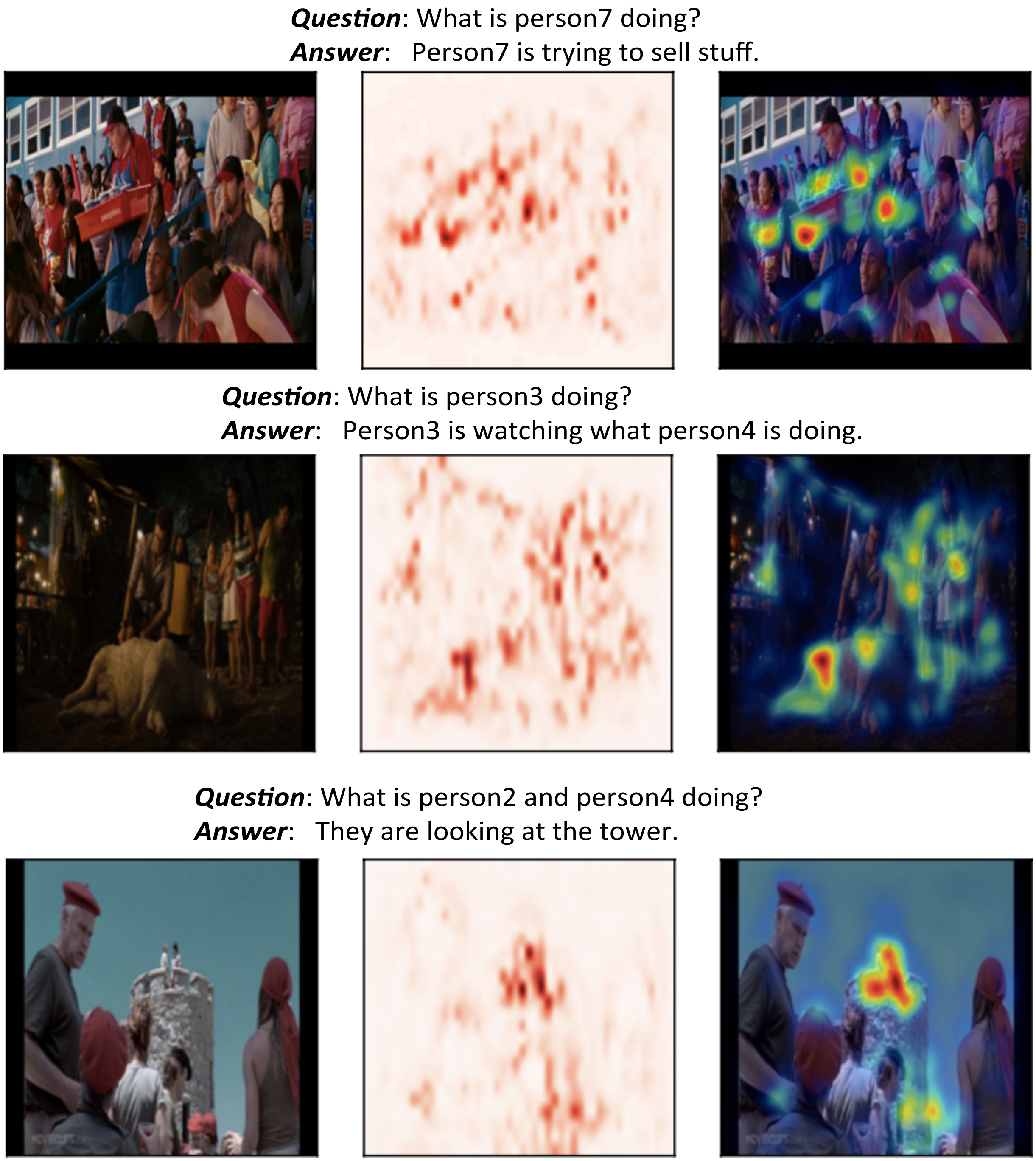
To have a more detailed understanding of how different factors influence different actions. We show the accuracy for specific actions under different ablation study setting. Specifically, For each one of the top eight frequent actions, we calculate its accuracy under different ablation studies. The results are shown in table 5.4. Different factors may have different influence for different actions. For the majority of actions except for the actions *dance* and *drink*, the reasoning factor plays a more important role compared to the language grounding factor. A possible explanation is that language grounding is an extremely challenging task and the learned model for language grounding is still quite limited. On the other hand, the reasoning model does a better job in capturing commonsense relations between an answer and a rationale. We think the use of Bert contributes to this advantage as Bert has shown superb performance on many commonsense reasoning tasks [41, 83, 84, 89].

Visual Attention Analysis. Here we will show some visualizations of the visual attention learned in the VQA and VQR factor. Some example visualizations are shown in Figure 5.8. Here we show the sentence-level answer attention on the image. We can see that in the first Figure, the model's attention mainly focus on the surroundings of the object **stuff** in the answer. In the second Figure, The main attention is on the man's hand and the sheep, but there are also some noisy attentions on other peoples. In the last Figure, The answer's visual attention is correctly put on the tower which the people are looking at. From these examples, we can see that the model actually learns to where to focus and extract useful visual feature representation to learn to answer the visual question. In Figure 5.9, Similarly given a pair of question and rationale, we show the corresponding attention map the model learns. For example, in the third example, the attention is mainly put on the map and arm to indicate this is a *hold* action.

Causal Attention Analysis. We also visualize the attention mapping between the answer and the rationale. This kind of correlation actually indicates what we called causal relation. In our model, we model this kind of relations as un-directed relation. Here we visualize the attention from an answer to a rationale, which means we normalize the correlation matrix learned along the direction of rationals. Thus we can see what are import words used for each word in the answer. The example visualizations are shown in Figure 5.10. In the first example, the word **performing** mainly correlates with words **higher pedestal** and words **playing music**. In the second example, almost all words in the answer correlate with words **someone** and **hear**. Two characteristics we observe here are: first we notice that the correspondence of the same people's names are not necessarily aligned in this process, second the answer words tend to attend much fewer words in the rationale sentence. One possible explanation is that the fine-tuned Bert representation is a contextualized representation, so even for same words, when they are in different context, their vector representation may not be very similar.

5.4.4 Error Analysis

From the previous results, we can see that there still exists a big gap between our results and human level performance. To have a better understanding of where the errors come from, we analyze the prediction errors from following two aspects. First we categorize errors in different ways:



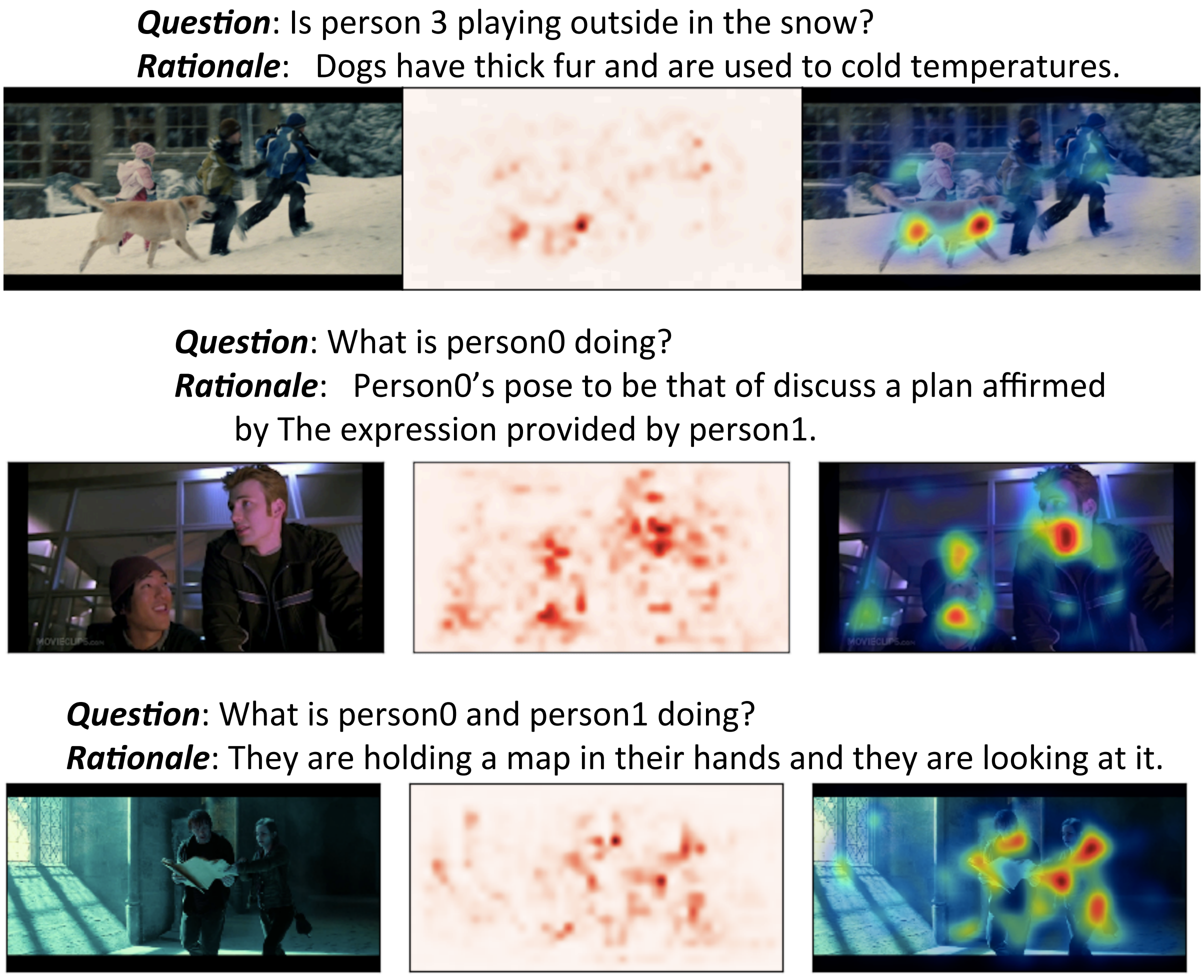
Question: What is person7 doing?
Answer: Person7 is trying to sell stuff.

Question: What is person3 doing?
Answer: Person3 is watching what person4 is doing.

Question: What is person2 and person4 doing?
Answer: They are looking at the tower.

Figure 5.8: Attention visualization for the VQA factor.

- A-R+: the predicted answer is wrong, but the predicted rationale is correct.
- A+R-: the predicted answer is correct, but the predicted rationale is wrong.
- A-R-: Both the predicted answer and the predicted rationale are wrong, and they also doesn't

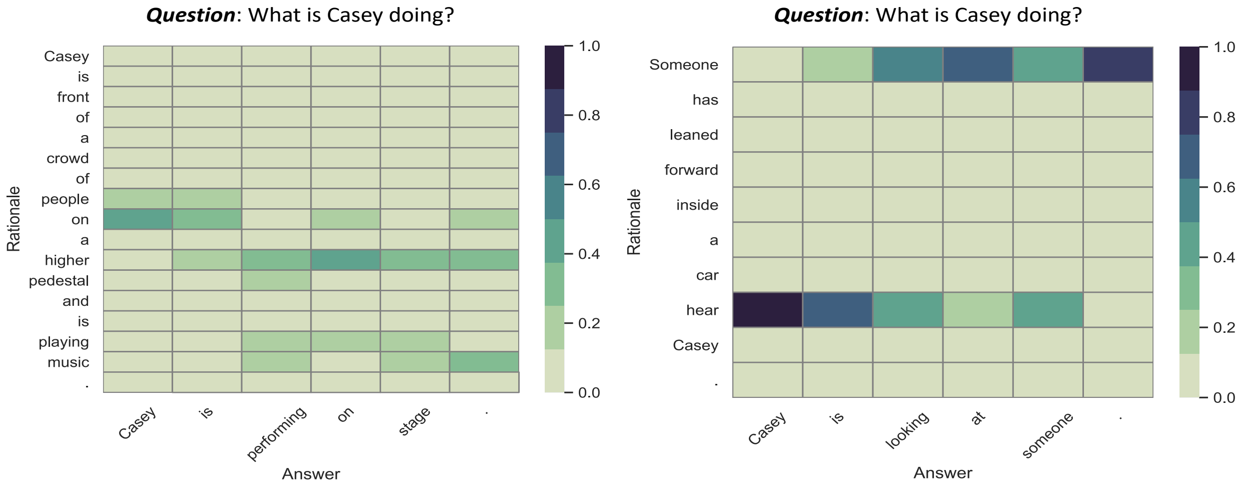


Question: Is person 3 playing outside in the snow?
Rationale: Dogs have thick fur and are used to cold temperatures.

Question: What is person0 doing?
Rationale: Person0's pose to be that of discuss a plan affirmed by The expression provided by person1.

Question: What is person0 and person1 doing?
Rationale: They are holding a map in their hands and they are looking at it.

Figure 5.9: Attention visualization for the VQR factor.



Question: What is Casey doing?

Question: What is Casey doing?

Figure 5.10: Attention visualization for the AR factor.

match. Noted that we augment each original answer with 4 rationales of which one of them is the correct rationale for the original answer in another sample.

- A-R-*: Both the predicted answer and the predicted rationale are wrong, but they are matched.

The ratios of different error types are shown in Figure 5.11. We can see that the most frequent error is R-A+, which means the answer prediction is correct, while the rationale prediction is wrong. The least occurred error is A-R+, which means the answer prediction is wrong and rationale is correct. Combining these information we learned that inferring answers seems easier than inferring rationales.

We also analyze the error from the factors perspective. For each sample, the gold answer rationale pair is (a, r), the predicted answer rationale pair is (a_p, r_p). For each pair we have 3 scores s_{vqa}, s_{vqr}, s_{ar}. Ideally the scores for the gold pair should be larger than the predicted pair. For each score, we calculate whether the score ordering is within our anticipation or not. We estimate them as these 3 ratios:

- AR_MM: the order of the AR factor scores is mis-matched.
- QR_MM: the order of the QR factor scores is mis-matched.
- QA_MM: the order of the QA factor scores is mis-matched.

The ratios of these 3 types of mis-matches are shown in Figure 5.12. We can see that the most frequent mis-match types is AR_MM, which means the most challenging part of this task is how to learn the commonsense matching factor for the answer and the rationale. This Figures also shows that the least amount of mis-match is for the factor QA, which coincides with the previous error types analysis: the error for the answer is lower than the error for the rationale.

Table 5.5: Error rates among samples with different lengths of ground truth rationales.

Length	**[0,10)**	**[10,20)**	**[20,30)**	**[30,40)**	**[40, $+\infty$)**
Error rate	0.758	0.631	0.530	0.470	0.455

A quick error analysis has shown that the model performance seems to correlate with the length of the ground truth rationales (as shown in Table 5.5). Although the ground truth length distribution is nearly uniform in the training dataset, our results show that the shorter the length of a rationale is, the less likely the model would pick it up. This shows that providing more contextual information in the rationales may help improve model performance.

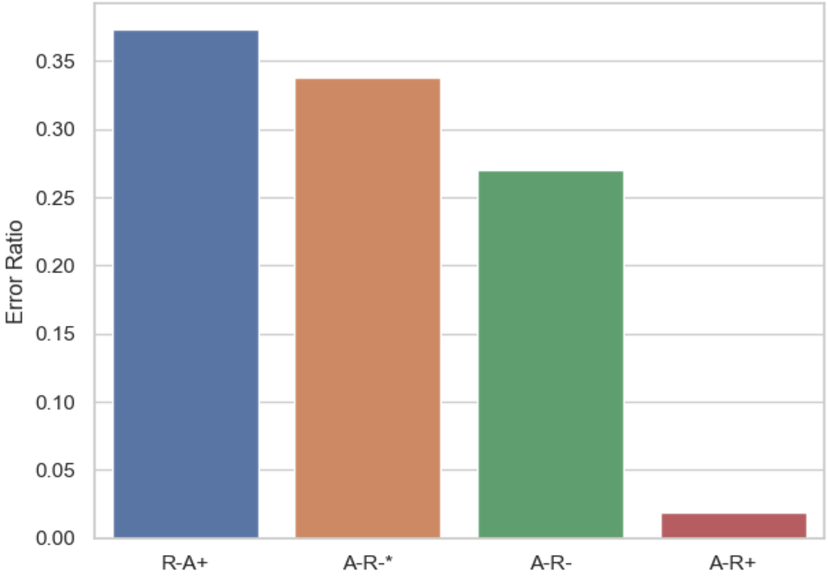


Figure 5.11: The ratio of different error types.

5.5 Conclusion

In this work, we propose a joint learning task setting for the visual question answer and rationale task. Compared with the previous R2C framework, we argue it's better to learn these two things jointly instead following a two step process. To adapt to the joint learning task setting, we augment

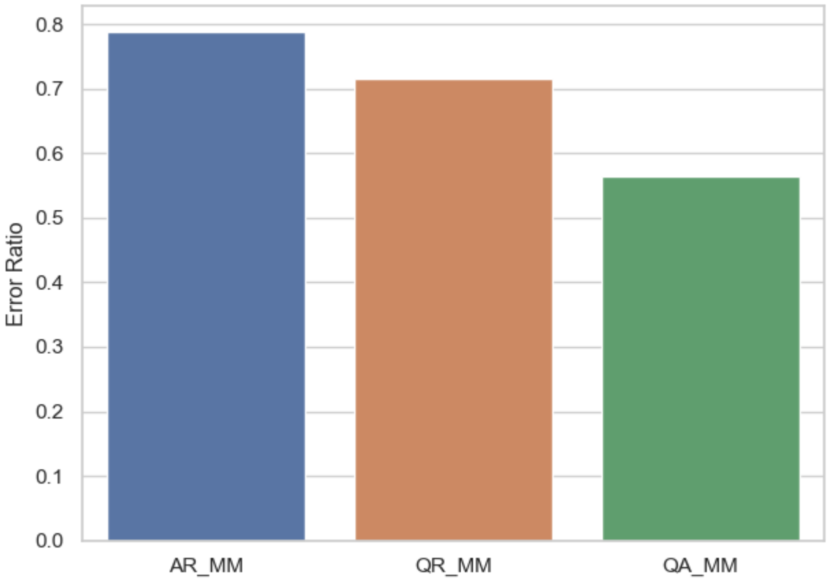


Figure 5.12: The inverse ratio of different factors.

the original R2C dataset with more negative rationales to eliminate the dataset answer bias. Based on the new dataset, we propose a pairwise deep factorization model to model the joint probability of the answer and the rationale given the image and the question. In essence, this joint task is trying to jointly solve visual question answering and natural language causal inference together. Finally we conduct comprehensive experiments and ablations to verify the effectiveness of our proposed method. However, the current best performance is still far from human performance. We showed the extreme action sparsity in the dataset analysis. In the future work, how to effective mitigate this kind of sparsity and introduce more external knowledge to the learning process will be an interesting direction worth exploring.

Chapter 6

Conclusion and Future Work

6.1 Conclusions

The process of human learning to understand the world is not isolated by using only linguistic descriptions or visual signals. Actually we learn through a combination of different sources of signals from our surrounding world, which is also called multi-modal learning. Imagine when you are young, how do you learn the action of "*pick up*"? It often comes with a combination of the visual demonstration and language instructions. Humans have strong abilities to synthesize information and learn from different sources including text, vision and speech et al. In order to build agents which can really understand human utterances, execute instructions and even generate explanations or rationales, the agents need to understand grounded meanings of text by connecting to the physical world to get a comprehensive understanding.

To achieve this goal, we investigate the problem of grounded language learning of physical actions through connecting low level visual semantics with high level linguistic semantics. At the same time, we also explore how to connect grounded rationales with action predictions.

- In Chapter 3, we propose a new task: grounded semantic role labeling to bridge the gap between the high level linguistic structured semantic information and the low level visual information including object trackings and attributes et al. From the linguistics perspective, the semantics of actions/verbs are represented as frames with slots and values which characterize key properties describing the action. These properties are called semantic roles including patient, location tool and so on. For each semantic role, we ground it to visual elements in the physical world: the objects in the video clips for our study. Besides explicit semantic roles mentioned in the linguistic descriptions, we also ground the implicit semantic roles which happens in the world but are not explicitly mentioned in the description because they are also very important for the agent to learn to understand and interpret the action. As shown in our experiments, the agent can have a better grounded understanding of the environment with the incorporation of the semantic information.

- Chapter 4 discusses an approach that attempts to infer commonsense justifications for physical actions in human-agent communication. To understand the action meaning, not only recognizing what is happening is important, but also providing commonsense evidence to support the decision the agent made. On one hand, it helps improves the trust between the human and the agent. On the other hand, it also helps improve the communication grounding during communication. We propose a generative modeling framework to jointly infer the action and the corresponding commonsense justification. We decompose explanations into relations and attributes, then model the evidence selection problem as a latent variable inference problem. Our empirical evaluations show that this joint inference model achieves better performance compared to previous competitive methods. Furthermore, as our latent variables are interpretable, we add the supervision to the latent variable and show that it actually improves both the evidence selection and action prediction performance. Lastly, we design a human study to verify that our propose joint model helps improve communication grounding between humans and agents.

- In Chapter 5, we focus on the problem of grounded action justification. We propose to solve the joint visual commonsense reasoning task: given an image and a question, the goal is to select the answer and provide the corresponding rationale. A new factorized neural model is developed to better understand relations between the image, question, answer and rationales. Specifically, we decompose the problem into three small factors including Image-Question-Answer factor, Image-Question-Rationale factor and Image-Answer-Rationale factor. Experimental results show that the factorized model achieves better accuracy. Besides, the comprehensive ablation study results show that different factors are essential for the final performance, and the Image-Answer-Rationale factor plays the most significant role for the final performance.

6.2 Future Directions

This dissertation explores different approaches for grounded action understanding and justification through language communication. To extend these approaches to a variety of real world applications, possible future works are described as follows:

- **Data efficient Learning for grounded semantic role labeling.** Currently the supervised graphical model based method for grounded semantic role labeling requires a lot of labeled data which are time-consuming and expensive. One important future research direction is to explore semi-supervised methods to alleviate the burden of data labeling. For example, how to effectively incorporate the unlabeled data to improve the generalization ability.

- **Deep Learning for grounded semantic role labeling.** Recently the deep learning based methods show great potentials on solving multi-modal related problems for several reasons: first they can learn better visual representations and contextual depended word representations. Second the deep architecture can better captures the interaction of information flows across different modals. One possible research direction to further improve grounded semantic role labeling by modeling both the visual context and linguistic semantic roles as graphs and use graph based neural network to predict target groundings.

- **Incorporating commonsense knowledge for grounded action justification.** Our current system's performance is still far from human performance on commonsense justification. We find that incorporating pre-trained word embeddings such as Bert is greatly helpful for improving the accuracy. The pre-trained Bert embeddings capture commonsense knowledge from large scale external datasets. Then one natural follow up question would be: whether we can effectively incorporate some action knowledge base into our model to improve the model performance. For example, how to inject ConceptNet or VerbNet knowledge to the deep neural network?

- **Grounded Justification Generation.** Currently we mainly frame the grounded action justification problem as a ranking or classification task. Compared with language generation tasks, this choice makes the evaluation easier and reasonable. But language generation is a more realistic setting as it doesn't need us to provide specific candidates. So a possible extended new task could be: given an image, a question and several answer candidates, how can we select the correct answer and generate natural language rationales justifying our prediction?

As more and more multi-media applications start to enter our daily lives such as movies with captions, image with descriptions, it will be more and more important to build intelligent agents to understand the world through multi-modal learning. Despite the efforts we have made in this dissertation, a lot of important and interesting problems still remain open. We believe that future research on this topic is of great values to make fundamental advances in AI.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[2] Y. Artzi and L. Zettlemoyer. Weakly supervised learning of semantic parsers for mapping instructions to actions. *TACL*, 1:49–62, 2013.

[3] C. F. Baker, C. J. Fillmore, and J. B. Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.

[4] F. Bashir and F. Porikli. Performance evaluation of object detection and tracking systems. In *Proceedings 9th IEEE International Workshop on PETS*, pages 7–14, 2006.

[5] O. Biran and K. McKeown. Human-centric justification of machine learning predictions. *IJCAI, Melbourne, Australia*, 2017.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[7] E. Bruni, N.-K. Tran, and M. Baroni. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49:1–47, 2014.

[8] A. Chang, W. Monroe, M. Savva, C. Potts, and C. D. Manning. Text to 3d scene generation with rich lexical grounding. *arXiv preprint arXiv:1505.06289*, 2015.

[9] D. L. Chen and R. J. Mooney. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*, pages 128–135. ACM, 2008.

[10] J. Chorowski and N. Jaitly. Towards better decoding and language model integration in sequence to sequence models. *arXiv preprint arXiv:1612.02695*, 2016.

[11] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[12] H. H. Clark. Using language. 1996. *Cambridge University Press: Cambridge*, 952:274–296, 1996.

[13] J. Cohen et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[14] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

[15] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[16] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer. Adaptive color attributes for real-time visual tracking. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1090–1097. IEEE, 2014.

[17] D. Dennett. *The intentional Stance*. MIT Press, 1987.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[19] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell. Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809*, 2015.

[20] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[21] D. Elliott and A. de Vries. Describing images using inferred visual dependency representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 42–52, Beijing, China, July 2015. Association for Computational Linguistics.

[22] B. Emanuele, G. Castellucci, D. Croce, and R. Basili. Textual inference and meaning representation in human robot interaction. In *Joint Symposium on Semantic Processing.*, page 65, 2013.

[23] M. Forbes and Y. Choi. Verb physics: Relative physical knowledge of actions and objects. *arXiv preprint arXiv:1706.03799*, 2017.

[24] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[25] Q. Gao, M. Doering, S. Yang, and J. Y. Chai. Physical causality of action verbs in grounded language understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 1814–1824, 2016.

[26] Q. Gao, S. Yang, J. Chai, and L. Vanderwende. What action causes this? towards naive

physical action-effect prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[27] P. Gärdenfors. *Conceptual spaces: The geometry of thought.* MIT press, 2004.

[28] D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288, 2002.

[29] S. Guadarrama, L. Riano, D. Golland, D. Go, Y. Jia, D. Klein, P. Abbeel, T. Darrell, et al. Grounding spatial relations for human-robot interaction. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 1640–1647. IEEE, 2013.

[30] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[31] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[32] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016.

[33] T. Hofmann. Probabilistic latent semantic indexing. In *ACM SIGIR Forum*, volume 51, pages 211–218. ACM, 2017.

[34] M. R. Hovav and B. Levin. Reflections on manner/result complementarity. *Lecture notes*, 2008.

[35] M. R. Hovav and B. Levin. Reflections on Manner / Result Complementarity. *Lexical Semantics, Syntax, and Event Structure*, pages 21–38, 2010.

[36] G. Jäger. Natural color categories are convex sets. In *Logic, language and meaning*, pages 11–20. Springer, 2010.

[37] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[38] S. Karlekar, T. Niu, and M. Bansal. Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models. *arXiv preprint arXiv:1804.06440*, 2018.

[39] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. June 2015.

[40] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects

in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics.

[41] N. S. Keskar, B. McCann, C. Xiong, and R. Socher. Unifying question answering and text classification via span extraction. *arXiv preprint arXiv:1904.09286*, 2019.

[42] A. Khan, N. Salim, and Y. J. Kumar. A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing*, 30:737–747, 2015.

[43] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.

[44] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[45] P. Kingsbury and M. Palmer. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*, 2002.

[46] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[47] J. Krishnamurthy and T. Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1:193–206, 2013.

[48] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Generalizing image captions for image-text parallel corpus. In *ACL (2)*, pages 790–796. Citeseer, 2013.

[49] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[50] T. K. Landauer. *Latent semantic analysis*. Wiley Online Library, 2006.

[51] A. Lazaridou, N. T. Pham, and M. Baroni. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*, 2015.

[52] B. Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.

[53] K. Liang, Y. Guo, H. Chang, and X. Chen. Visual relationship detection with deep structural

ranking. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.

[54] K. Liang, Y. Guo, H. Chang, and X. Chen. Visual relationship detection with deep structural ranking. 2018.

[55] C. Liu and J. Y. Chai. Learning to mediate perceptual differences in situated human-robot dialogue. In *The Twenty-Ninth Conference on Artificial Intelligence (AAAI-15)*. to appear, 2015.

[56] C. Liu, R. Fang, and J. Chai. Towards mediating shared perceptual basis in situated dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 140–149, Seoul, South Korea, 2012.

[57] T. Lombrozo. Explanation and abductive inference. *Oxford handbook of thinking and reasoning*, pages 260–276, 2012.

[58] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[59] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabinovich, and K. Murphy. What's cookin'? interpreting cooking videos using text, speech and vision. *arXiv preprint arXiv:1503.01558*, 2015.

[60] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[61] B. McMahan and M. Stone. A bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115, 2015.

[62] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[63] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(1):58–72, 2014.

[64] F. Morin and Y. Bengio. Hierarchical probabilistic neural network language model. In *Aistats*, volume 5, pages 246–252. Citeseer, 2005.

[65] I. Naim, Y. C. Song, Q. Liu, L. Huang, H. Kautz, J. Luo, and D. Gildea. Discriminative unsupervised alignment of natural language instructions with corresponding video segments.

In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 164–174, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[66] L. G. M. Ortiz, C. Wolff, and M. Lapata. Learning to interpret and describe abstract scenes. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1505–1515.

[67] M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106, 2005.

[68] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. *arXiv preprint arXiv:1802.08129*, 2018.

[69] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[70] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[71] S. S. Pradhan, W. Ward, K. Hacioglu, J. H. Martin, and D. Jurafsky. Shallow semantic parsing using support vector machines. In *HLT-NAACL*, pages 233–240, 2004.

[72] J. Pustejovsky. The syntax of event structure. *Cognition*, 41(1-3):47–81, 1991.

[73] V. Ramanathan, P. Liang, and L. Fei-Fei. Video event understanding using natural language descriptions. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 905–912. IEEE, 2013.

[74] H. Rashkin, A. Bosselut, M. Sap, K. Knight, and Y. Choi. Modeling naive psychology of characters in simple commonsense stories. *arXiv preprint arXiv:1805.06533*, 2018.

[75] T. Regier and L. A. Carlson. Grounding spatial language in perception: an empirical and computational investigation. *Journal of experimental psychology: General*, 130(2):273, 2001.

[76] T. Regier, P. Kay, and R. S. Cook. Focal colors are universal after all. *Proceedings of the National Academy of Sciences of the United States of America*, 102(23):8386–8391, 2005.

[77] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1:25–36, 2013.

[78] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[79] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining (KDD)*, 2016.

[80] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[81] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele. Script data for attribute-based recognition of composite activities. In *Computer Vision–ECCV 2012*, pages 144–157. Springer, 2012.

[82] D. Roy. Grounding words in perception and action: computational insights. *TRENDS in Cognitive Sciences*, 9(8):389–396, 2005.

[83] Y.-P. Ruan, X. Zhu, Z.-H. Ling, Z. Shi, Q. Liu, and S. Wei. Exploring unsupervised pretraining and sentence structure modelling for winograd schema challenge. *arXiv preprint arXiv:1904.09705*, 2019.

[84] M. Sap, H. Rashkin, D. Chen, R. LeBras, and Y. Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

[85] K. K. Schuler. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania, 2005.

[86] D. Shen and M. Lapata. Using semantic roles to improve question answering. In *EMNLP-CoNLL*, pages 12–21, 2007.

[87] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock. Spatial language for human-robot dialogs. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(2):154–167, 2004.

[88] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015.

[89] S. Storks, Q. Gao, and J. Y. Chai. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, 2019.

[90] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. J. Teller, and N. Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, 2011.

[91] S. Tellex, P. Thaker, J. Joseph, and N. Roy. Learning perceptually grounded word meanings from unaligned parallel data. *Machine Learning*, 94(2):151–167, 2014.

[92] Thagard. Probabilistic networks and explanatory coherence. *Cognitive Science Quarterly*, 2000.

[93] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu. Joint video and text parsing for understanding events and answering queries. *MultiMedia, IEEE*, 21(2):42–70, 2014.

[94] J. Van De Weijer and C. Schmid. Coloring local feature extraction. In *Computer Vision–ECCV 2006*, pages 334–348. Springer, 2006.

[95] R. Vedantam, X. Lin, T. Batra, C. Lawrence Zitnick, and D. Parikh. Learning common sense through visual abstraction. In *Proceedings of the IEEE international conference on computer vision*, pages 2542–2550, 2015.

[96] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1494–1504, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[97] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, 101(1):184–204, 2013.

[98] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

[99] P. Wang, Q. Wu, C. Shen, and A. van den Hengel. The vqa-machine: Learning how to use existing vision algorithms to answer new questions. In *Proc. CVPR*, 2017.

[100] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, June 2016.

[101] Y. Yang, C. Fermuller, and Y. Aloimonos. Detection of manipulation action consequences (mac). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2563–2570, 2013.

[102] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.

[103] M. Yatskar, V. Ordonez, and A. Farhadi. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of NAACL-HLT*, pages 193–198, 2016.

[104] H. Yu and J. M. Siskind. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 53–63, 2013.

[105] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. From recognition to cognition: Visual commonsense reasoning. *arXiv preprint arXiv:1811.10830*, 2018.

[106] R. Zellers and Y. Choi. Zero-shot activity recognition with verb attribute induction. *arXiv preprint arXiv:1707.09468*, 2017.

[107] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, 73(2):213–238, 2007.

[108] Q. Zhang, R. Cao, F. Shi, Y. N. Wu, and S.-C. Zhu. Interpreting cnn knowledge via an explanatory graph. *arXiv preprint arXiv:1708.01785*, 2017.

[109] Q. Zhang, Y. N. Wu, and S.-C. Zhu. Interpretable convolutional neural networks. *arXiv preprint arXiv:1710.00935*, 2017.

[110] T. Zhao, R. Zhao, and M. Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*, 2017.

[111] J. Zhou and W. Xu. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1127–1137, Beijing, China, July 2015. Association for Computational Linguistics.