Natural Language as a Scaffold for Visual Recognition

Mark Yatskar

A dissertation

submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Luke Zettlemoyer, Chair

Ali Farhadi

Yejin Choi

Program Authorized to Offer Degree:

Computer Science and Engineering

©Copyright 2017

Mark Yatskar

University of Washington

Abstract

Natural Language as a Scaffold for Visual Recognition

Mark Yatskar

Chair of the Supervisory Committee:

Professor Luke Zettlemoyer

Computer Science and Engineering

A goal of artificial intelligence is to create a system that can perceive and understand the visual world through images. Central to this goal is defining what exactly should be recognized, both in structure and coverage. Numerous competencies have been proposed, ranging from low level tasks such as edge detection to high level tasks, such as semantic segmentation. In each case, a specific set of visual targets is considered (e.g. particular objects or activities to be recognized) and it can be difficult to define a comprehensive set of everything that could be present in the images. In contrast to these efforts, we consider taking a broader view of visual recognition. We propose to use natural language as a guide for what people can perceive about the world from images and what ultimately machines should emulate.

We show it is possible to use unrestricted words and large natural language processing ontologies to define relatively complete sets of targets for visual recognition. We explore several core questions centered around this theme: (a) what kind of language can be used, (b) what it means to label everything and (c) can structure in language be used to define a recognition problem. We make progress in several directions, showing for example that highly ambiguous sentimental language can be used to formulate concrete targets and that linguistics feature norms can be used to densely annotate many complex aspects of images.

Finally, this thesis introduces situation recognition, a novel representation of events in images that relies on two natural language processing resources to achieve scale and expressivity. The formalism combines WordNet, an ontology of nouns, with FrameNet, an ontology of verbs and implicit argument types, and is supported by a newly collected large scale image resource imSitu. Situation recognition significantly improves over existing formulations for activities in images, allowing for higher coverage, increased richness of the representation, and more accurate models. We also identify new challenges with our proposal, such as rarely observed target outputs, and develop methods for addressing them.

Acknowledgement

This thesis absolutely could not exist without Luke Zettlemoyer and Ali Farhadi. Luke, you have been there from the first day supporting me through the academic, social and mental journey. More than anything, your trust and care have nursed me through more trying moments than I can remember. Ali, your technical wisdom and uniquely broad perspective on AI have been invaluable. Without you sparking the interest of AI2 into our research, much of this work could not have been done.

Thank you to all of the people I have collaborated with: Vicente Ordóñez, Yannis Konstas, Michel Galley, Lillian Lee, Cristian Danescu-Niculescu-Mizil, Bo Pang, Svitlana Volkova, Bill Dolan, Lucy Vanderwende, Asli Celikyilmaz, Srinivasan Iyer, Yejin Choi, Jieyu Zhao, Tianlu Wang, Kai-Wei Chang. I would like to especially thank Vicente, Yannis, Michel and Cristian, who through the process of working together mentored me and pushed me to explore new areas and ideas. Also, Lillian Lee, without whom I would have never thought to start, who has supported and advised me, and who has been a role model for me for many years. Also, many of these collaborations were born through internships at Microsoft Research or Allen Institute of Artificial Intelligence, and I appreciate the supportive space they have provided my research.

I would like to acknowledge my first home at UW, the LIL group. Yoav Artzi, Nicholas FitzGerald, Eunsol Choi you have all been amazing contributors to the work in this thesis through hallways discussions, philosophical arguments, random nuggets of advice, draft editing and role playing as the devil's advocate. Sameer Singh, Hoifun Poon, Tom Kwiatkowski and Mike Lewis, discussions with you have shaped many of my perspectives on AI. I would also like to thank many people broadly who have been associated with the UW NLP group: Chenhao Tan, Omer Levy, Tony Fader, Dan Garrette, Adrienne Wang, Chloe Kiddon, Alan Ritter, Wei Xu, Raphael Hoffmann, Gabriel Schubiner, Roy Schwartz, Kenton Lee, Luheng He, Victoria Lin, Maarten Sap, Mandar Joshi, Julian Michael, Minjoon Seo, Swabha Swayamdipta, Rowan Zellers, Maxwell

Forbes, Dallas Card, Sam Thomson, Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Noah Smith, Mari Ostendorf, Gina Levow, Emily Bender, Oren Etzoni, Dan Weld, Pedro Domingos, Jayant Krishnamurthy. You have formed the vibrant environment around NLP at CSE and together through reading groups, arguments, critical feedback, and random bouts of conference drinking we have grown our understanding. Increasingly I have become associated with Grail through Ali's group, and I would like to thank Joe Redmond, Hessam Bagherinezhad, Kiana Ehssani, and Daniel Gordon, Max Horton, without whom I couldn't begin to explore computer vision. Also, the computer vision team at AI2, Santosh Divvala, Ani Kembhavi, Eric Klove, Roozbeh Mottaghi, Mohammad Rastegari who have always been a wonderful sounding board for crazy ideas. Lastly, I would like to thank Elise DeGoede, Lindsay Michimoto, Chiemi Yamaoka-Vismale, and Aleesha Thurber who have been welcoming and caring guides at UW.

My time at UW would have been impossible without the close friendships that have supported me here. Paris Koutris and Ricardo Martin, you two have been my wards since the very first year. Together we've spent hours, hiking, strolling, climbing, watching basketball, throwing things, inventing games, and being true to heart. Thank you. Eunsol Choi, you have been my family here, and your candor, humor and energy have guided, supported and enriched my life. Your presence has been ever calming and focusing and without the countless little moments of levity you created at UW, the time would have felt endless. Thank you. Nicholas FitzGerald, Matt Kay, Yoav Atrzi, Natalie Zervou, Julija Lazutkaite, Qi Shan, Melanie Gens, Rob Gens, Daniel Perelman, Cynthia Matuszek, Sophie Ostlund, Abe Friesen, thank you for being the wonderful company, and Smith, for being our tireless host. Finally, I would like to thank my family, Yulia, Yuri, Maya, Roman Yatskar, and Tera Schoeneck who supported me, offering advice, curiosity, and the safety to explore.

DEDICATION

To Simion Yatskar

Contents

List of Figures

List of Tables

This thesis absolutely could not exist without Luke Zettlemoyer and Ali Farhadi. Luke, you have been there from the first day supporting me through the academic, social and mental journey. More than anything, your trust and care have nursed me through more trying moments than I can remember. Ali, your technical wisdom and uniquely broad perspective on AI have been invaluable. Without you sparking the interest of AI2 into our research, much of this work could not have been done.

Thank you to all of the people I have collaborated with: Vicente Ordóñez, Yannis Konstas, Michel Galley, Lillian Lee, Cristian Danescu-Niculescu-Mizil, Bo Pang, Svitlana Volkova, Bill Dolan, Lucy Vanderwende, Asli Celikyilmaz, Srinivasan Iyer, Yejin Choi, Jieyu Zhao, Tianlu Wang, Kai-Wei Chang. I would like to especially thank Vicente, Yannis, Michel and Cristian, who through the process of working together mentored me and pushed me to explore new areas and ideas. Also, Lillian Lee, without whom I would have never thought to start, who has supported and advised me, and who has been a role model for me for many years. Also, many of these collaborations were born through internships at Microsoft Research or Allen Institute of Artificial Intelligence, and I appreciate the supportive space they have provided my research.

I would like to acknowledge my first home at UW, the LIL group. Yoav Artzi, Nicholas FitzGerald, Eunsol Choi you have all been amazing contributors to the work in this thesis through hallways discussions, philosophical arguments, random nuggets of advice, draft editing and role playing as the devil's advocate. Sameer Singh, Hoifun Poon, Tom Kwiatkowski and Mike Lewis, discussions with you have shaped many of my perspectives on AI. I would also like to thank many people broadly who have been associated with the UW NLP group: Chenhao Tan, Omer Levy, Tony Fader, Dan Garrette, Adrienne Wang, Chloe Kiddon, Alan Ritter, Wei Xu, Raphael Hoffmann, Gabriel Schubiner, Roy Schwartz, Kenton Lee, Luheng He, Victoria Lin, Maarten Sap, Mandar Joshi, Julian Michael, Minjoon Seo, Swabha Swayamdipta, Rowan Zellers, Maxwell Forbes, Dallas Card, Sam Thomson, Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Noah Smith, Mari Ostendorf, Gina Levow, Emily Bender, Oren Etzoni, Dan Weld, Pedro Domingos, Jayant Krishnamurthy. You have formed the vibrant environment around NLP at CSE and together through reading groups, arguments, critical feedback, and random bouts of conference drinking we have grown our understanding. Increasingly I have become associated with Grail through Ali's group, and I would like to thank Joe Redmond, Hessam Bagherinezhad, Kiana Ehssani, and Daniel Gordon, Max Horton, without whom I couldn't begin to explore computer vision. Also, the computer vision team at AI2, Santosh Divvala, Ani Kembhavi, Eric

Klove, Roozbeh Mottaghi, Mohammad Rastegari who have always been a wonderful sounding board for crazy ideas. Lastly, I would like to thank Elise DeGoede, Lindsay Michimoto, Chiemi Yamaoka-Vismale, and Aleesha Thurber who have been welcoming and caring guides at UW.

My time at UW would have been impossible without the close friendships that have supported me here. Paris Koutris and Ricardo Martin, you two have been my wards since the very first year. Together we've spent hours, hiking, strolling, climbing, watching basketball, throwing things, inventing games, and being true to heart. Thank you. Eunsol Choi, you have been my family here, and your candor, humor and energy have guided, supported and enriched my life. Your presence has been ever calming and focusing and without the countless little moments of levity you created at UW, the time would have felt endless. Thank you. Nicholas FitzGerald, Matt Kay, Yoav Atrzi, Natalie Zervou, Julija Lazutkaite, Qi Shan, Melanie Gens, Rob Gens, Daniel Perelman, Cynthia Matuszek, Sophie Ostlund, Abe Friesen, thank you for being the wonderful company, and Smith, for being our tireless host. Finally, I would like to thank my family, Yulia, Yuri, Maya, Roman Yatskar, and Tera Schoeneck who supported me, offering advice, curiosity, and the safety to explore.

DEDICATION

To Simion Yatskar

Chapter 1

Introduction

A goal of artificial intelligence is to create a system that can perceive and understand the visual world through images. Central to this goal is defining what exactly should be recognized, both in structure and coverage. Numerous competencies have been proposed, ranging from low level tasks such as edge detection to high level tasks, such as semantic segmentation. In each case, a specific set of visual targets is considered (e.g. particular objects or activities to be recognized) and it can be difficult to define a comprehensive set of everything that could be present in the images. In contrast to these efforts, we consider taking a broader view of visual recognition. We propose to use natural language as a guide for what people can perceive about the world from images and what ultimately machines should emulate.

In this thesis we show it is possible to use unrestricted words and large natural language ontologies for defining what a machine understands about the visual world. While most computer vision tasks are can be viewed abstractly (i.e. classification: the problem of producing a label index given an image), language already plays a significant role in these problems. Labels are communicated to annotators using natural language names, i.e. ImageNet [30], where labels are WordNet [106] synsets, and some labels are even combinations of many words, forming phrases, i.e. labels from Stanford-40 activity recognition [151], such as `r`iding-a-bike or `r`iding-a-horse. Attributes are usually communicated using adjectives [85] (i.e. smooth) and actions are usually communicated using verbs [138], sometimes combined with nouns to disambiguate sense. In this work, we make this relationship more explicit and study (a) what kind of language can be used for defining a recognition problem, (b) what it means to label *everything* in an image using words and

(c) whether structure in language, such as frame semantic structure, can be used to define a recognition problem. Fundamentally, our proposal is to use unrestricted words and large natural language processing ontologies to define targets of visual recognition systems to substantially extend capabilities and coverage.

Using language for the purpose of defining visual recognition systems has several advantages. First, natural language statements about images can be elicited from people at scale, largely without training, in both task oriented and open settings. This observation has led to the collection of large-scale datasets such as ImageNet [30] and the introduction of tasks such as image captioning [46] and visual question answer [3]. Grounded language is also naturally occurring, for example in captions associated with Flikr images [116], general images found on the web with associated alt text [33], or images described for the deaf [124]. Such data can serve as a significant source of weak supervision for visual recognition systems based on language. Furthermore, we can reuse significant research from natural language processing to understand the structure of statements. For example, parsers can be used to decompose multi-word or sentential statements into meaningful sub-units for recognition [83]. Natural language based visual recognition can be used to provide a shared set of conceptual units between a person and a computer. In a grounded settings, such as interpreting commands given by a human to a machine shared conceptual spaces can significantly simplify the job of the computer [140]. Finally, defining recognition in terms of language can yield significant benefits in terms of coverage visual concepts. Instead of defining particular concepts as targets of recognition, entire classes of natural language statements can used instead, even implicitly defining target behavior of recognition systems on unobserved concepts. This work significantly expands the connection between recognition and language both in terms of coverage of groundable concepts and methods for fusing the two.

1.1 Challenges

Several core challenges emerge when trying to frame visual recognition around natural language. Some of these issues follow directly from natural language processing, such as sparsity from heavy tailed distributions. On the other hand, several new challenges also arise.

Non-visual Language Not all words correspond to concepts that can easily ground to images. This problem comes up in several forms. First, while many concepts can be in general grounded, images can signifi-

cantly limit the number of unambiguous concepts. For example, an image of a person opening and a person closing a door can be indistinguishable without video context. Furthermore, some language that initially may seem highly subjective or ambiguous can have significant quantities of ground-able information, such as `b`usiness-oriented or `a`loof. This challenge can strongly impact the quality of collected data and cause unwarranted exclusion in efforts to affordably create datasets.

Lexical Ambiguity While defining visual recognition to encompass any possible word makes it easier to cover more concepts, words function to highlight aspects rather than as unambiguous names for something in the world. This challenge is typified by multiple ways to name a single object in image (for example, person, male, male child), and requires significant effort to address. Datasets must be collected with multiple different ground truth annotations to allow for robust evaluation. Furthermore, methods must be able to deal with potentially ambiguous data, or training with multiple equivalent, correction annotations.

Sparsity and Combinatorial Explosion Using language as labels for visual recognition introduces sparsity into the structure of labels. Many labels occur infrequently and these tend to be the challenging cases. Furthermore, while most existing methods for computer vision focus on generalizing across image variation, recognition structured around language requires systems produce novel combinations of labels.

1.2 Approaches

In our efforts to explore how language can be used to structure visual recognition, we have tackled several aspects of the problem. Initial work was small in scope, focusing on questions such as : what language is groundable? what does it mean to annotate everything with language? Later work focused on using semantic role labeling resources such as FrameNet to formulate a large scale, high coverage event recognition formalism for images.

Sentimental Language for Visual Tasks Language can describe varied aspects of our visual world, including not only what is literally there but also the social, cultural, and emotional sentiment it invokes. Recently, there has been a growing effort to study *literal* language that describes directly observable properties, such as object color, shape, or category [44; 109; 104]. Here, we add a focus on *sentimental* visual

language, which compactly describes more subjective properties such as if a person looks determined, if a resume looks professional, or if a restaurant looks romantic. Such models enable many new applications, such as text editors that automatically select properties including font, color, or text alignment to best match high level descriptions such as "professional" or "artistic."

We study visual language, both literal and sentimental, that describes the overall appearance and style of virtual characters. We use literal language as feature norms, a tool used for studying semantic information in cognitive science [105]. Literal words, such "black" or "hat," are annotated for objects to indicate how people perceive visual properties. Such feature norms provide our gold-standard visual detectors, and allow us to focus on learning to model sentimental language, such as "youthful" or "goth."

Description with Dense Language Annotation In an effort to approximate relatively complete visual recognition, we collected manually labeled representations of objects, parts, attributes and activities for a benchmark caption generation dataset that includes images paired with human authored descriptions [122]. For example, such labels include object boundaries and descriptive text, here including the facts that the children are "riding" and "walking" and that they are "young." Our goal is to be as exhaustive as possible, giving equal treatment to all objects. Labels gathered in this way are a type of feature norms [105], which have been used in the cognitive science literature to approximate human perception and were recently used as a visual proxy in distributional semantics [134]. We present the first effort, that we are aware of, for using feature norms to study image description generation.

In this work, we instead study generation with more complete visual support, as provided by human annotations, allowing us to develop more comprehensive models than previously considered. Such models have the dual benefit of (1) providing new insights into how to construct more human-like sentences and (2) allowing us to perform experiments that systematically study the contribution of different visual cues in generation, suggesting which automatic detectors would be most beneficial for generation.

Common Sense using WordNet How can we discover that bowls can hold broccoli, that if a knife touches a cake then a person is probably cutting cake, or that cutlery can be on dining tables? We propose to leverage the effort of computer vision researchers in creating large scale datasets for object detection and use these resources instead to extract symbolic representations of visual common sense. The knowledge we compile

is physical, not commonly covered in text and more exhaustive than what people can usually produce. We show that such derived visual knowledge can be placed in the context of large language resources such as WordNet to extend the amount of inferred knowledge by several orders of magnitude.

Our focus is particularly on visual common sense, defined as the information about spatial and functional properties of entities in the world. We propose to extract three types of knowledge from the Microsoft Common Objects in Context dataset [95] (MS-COCO), consisting of 300,000 images, covering 80 objects, with object segments and natural language captions. First, we find spatial relations, e.g. *holds*(bed, dog), from outlines of co-occurring objects. Next, we construct entailment rules like *holds*(bed, dog) $\Rightarrow$ *laying-on*(dog, bed) by associating spatial relations with text in captions. Finally, we uncover general facts such as *holds*(furniture, domestic animal), applicable to object types not present in MS-COCO by using WordNet [106] and a novel submodular k-coverage formulation.

Situation Recognition When we look at an image, we instantly and effortlessly recognize not only what is happening (e.g., clipping) but who and what is involved (e.g., a person, shears, a sheep, wool) and how these entities relate to each other, the *roles* that they play (e.g., the person does the clipping, the shears are the clipping tool, and the wool is being clipped from the sheep). In this paper, we argue for explicitly encoding such semantic roles, a key missing ingredient in current paradigms of recognition, in image understanding. We introduce *situation recognition*, a problem that involves predicting activities along with actors, objects, substances, and locations and how these pieces fit together (semantic roles). Situation recognition generalizes activity recognition and human-object interaction. In essence, we are building representations that support the understanding not just of "What is happening?" but also "Who is doing it?" (the `agent` role), "What are they doing it to?" (`patient`), "What are they doing it with?" (`tool`), "Where did it start?" (`source`), and so on, as appropriate for each activity.

It is difficult to know a priori what roles entities can play in each activity. However, we can draw inspiration from the way verbs are used in the English language by building on FrameNet [53], a linguist-authored verb lexicon. FrameNet pairs every verb with a *frame*, which specifies a set of *semantic roles*. Semantic roles categorize how objects can participate in the activity described by a verb. Such frames have been used to build semantic parsers that match verbs to their arguments in English sentences, for example see [26]. However, here we instead use them to define the space of possible situations, much like how

WordNet [106] was used to define ImageNet [126] object classes. For each frame, the verb defines an activity label, and the semantic roles specify how WordNet entities participate in the activity. For example, Figure 6.1 shows situations where the FrameNet verb `spraying` has a semantic role `tool` that is filled with WordNet synsets such as `spray can` or `hose`.

To demonstrate the generality of the situation recognition task, we introduce *imSitu*, a collection of over 125,000 images depicting 200,000 distinct situations. Each situation includes one of 500 possible activities and values for up to 6 activity-specific roles (3.5 on average and 1,700 unique roles in total with 190 types). The images were gathered from Google image search with query expansion techniques and labeled with complete situations on Amazon Mechanical Turk. The annotators specified one of 80,000 possible WordNet sysnets for each role, providing over 11,000 unique values for this image collection. In addition to being large scale, this data is also high quality.

Sparsity in Situation Recognition Situation recognition can be challenging because many activities, such as `carrying`, have very open ended semantic roles, such as `item`, the thing being carried: nearly any object can be carried and the training data will never contain all possibilities. This is a prototypical instance of semantic sparsity: rare outputs constitute a large portion of required predictions (35% in the imSitu dataset [152]), and current state-of-the-art performance for situation recognition drops significantly when even one participating object has few samples for it's role. We propose to address this challenge in two ways by (1) building models that more effectively share examples of objects between different roles and (2) semantically augmenting our training set to fill in rarely represented noun-role combinations.

We introduce a new compositional Conditional Random Field formulation (CRF) to reduce the effects of semantic sparsity by encouraging sharing between nouns in different roles. Like previous work [152], we use a deep neural network to directly predict factors in the CRF. In such models, required factors for the CRF are predicted using a global image representation through a linear regression unique to each factor. In contrast, we propose a novel tensor composition function that uses low dimensional representations of nouns and roles, and shares weights across all roles and nouns to score combinations. Our model is compositional, independent representations of nouns and roles are combined to predict factors, and allows for a globally shared representation of nouns across the entire CRF.

This model is trained with a new form of semantic data augmentation, to provide extra training samples

for rarely observed noun-role combinations. We show that it is possible to generate short search queries that correspond to partial situations (i.e. “man carrying baby” or “carrying on back”) which can be used for web image retrieval. Such noisy data can then be incorporated in pre-training by optimizing marginal likelihood, effectively performing a soft clustering of values for unlabeled aspects of situations. This data also supports self training where model predictions are used to prune the set of images before training the final predictor.

1.3 Contributions

The contributions of this thesis can be summarized as:

- A proposal for grounding sentimental language
- A proposal for dense labeling of images using natural language statements
- Methods for extracting visual common sense
- Situation recognition, a novel formalism for event recognition in images
- ImSitu, the highest coverage, currently available resource for events in images
- Models for addressing sparsity in structured prediction of situations

1.4 Thesis Outline

The rest of the thesis will be organized with smaller explorations first, followed by two larger sections about situation recognition. Chapter 3 discusses relating literal descriptions to sentimental ones in the context of avatar generation. Chapter 4 considers a proposal for labeling everything in an image and introduces a caption generation system based on the labeling formalism. Chapter 5 introduces methods for extracting visual common sense using a object detection dataset and Wordnet. Finally, Chapters 6 and Chapter 7 deal with situation recognition, the first introducing data and baseline models, and the second identifying sparsity as a central challenge and proposing methods for addressing it.

Chapter 2

Related Work

2.1 Overview

Our work has examined the limits of what language is groundable. While previous explorations considered retrieving images [79] or generating graphics scenes from literal descriptions [24], we were the first consider high level, sentimental intents for finding objects in our avatar work. Furthermore, we have proposed a representation of an image entirely driven by language labels. Such a representation resembles structures in other grounded contexts, such as databases [159; 157; 84; 8], sports commentary [18], or navigation [5; 15; 140], but focuses on semantic elements important for image captioning. Our dense image annotation work was the first to try to exhaustively model all semantic elements found in captions, such as groups, properties, events, and parts.

Our work builds on an increasing trend to use language as a basis for forming recognition problems. Most simply, all common visual recognition datasets use label spaces defined by language but few works exploit this fact. For example, ImageNet [126] uses WordNet [106] to define a recognition problem but the underlying challenge [127] ignores the relationship of the categories to WordNet. Several works have used the hierarchical structure of the WordNet labels [31; 29; 115], but no work has shown the utility of explicitly reasoning about the labels as used in language beyond zero-shot learning settings [90; 56; 99; 87] work work embedding analogy tasks [134]. Many problems have been proposed that explicitly tie language and vision, such as captions [46; 116; 95], visual question answer [4; 123; 154; 60; 41], or referring expression [74] but

none of these proposals demonstrate improvement on existing recognition tasks. Our situation recognition work contributes a missing piece in all of these works: semantics of events . We use FrameNet [53] to formulate a language based representation of events in and provide the first large scale example of a language representation improving a core problem, such as activity recognition.

Each of the following sections provides more detailed related work for each of our contributions.

2.2 Sentimental Language

To the best of our knowledge, our focus on learning to understand visual sentiment descriptions is novel. However, visual sentiment has been studied from other perspectives. Jrgensen [72] provides examples which show that visual descriptions communicate social status and story information in addition to literal object and properties. Tousch et al. [142] draw the distinction between "of-ness" (objective and concrete) and "about-ness" (subjective and abstract) in image retrieval, and observe that many image queries are abstract (for example, images about freedom). Finally, in descriptions of people undergoing emotional distress, Fussell and Moss [58] show that literal descriptions co-occur frequently with sentimental ones.

There has been significant work on more literal aspects of grounded language understanding, both visual and non-visual. The WordsEye project [24] generates 3D scenes from literal paragraph-length descriptions. Generating literal textual descriptions of visual scenes has also been studied, including both captions [82; 147; 52] and descriptions [47]. Furthermore, Chen and Dolan [16] collected literal descriptions of videos with the goal of learning paraphrases while Zitnick and Parikh [163] describe a corpus of descriptions for clip art that supports the discovery of semantic elements of visual scenes.

There has also been significant recent work on automatically recovering visual attributes, both absolute [44] and relative [79], a challenge that we avoid having to solve with our use of feature norms [105].

Grounded language understanding has also received significant attention, where the goal is to learn to understand situated non-visual language use. For example, there has been work on learning to execute instructions [11; 15; 5], provide sports commentary [17], understand high level strategy guides to improve game play [12; 35], and understand referring expression [104].

Finally, our work is similar in spirit to sentiment analysis [118], emotion detection from images and speech [158], and metaphor understanding [131; 132]. However, we focus on more general visual context.

2.3 Dense Labeling

A number of approaches have been proposed for constructing sentences from images, including copying captions from other images [47; 114], using text surrounding an image in a news article [51], filling visual sentence templates [82; 147; 37], and stitching together existing sentence descriptions [66; 83]. However, due to the lack of reliable detectors, especially for activities, many previous systems have a small vocabulary and must generate many words, including verbs, with no direct visual support.

The Midge algorithm [108] is most closely related to our approach, and will provide a baseline in our experiments. Midge is syntax-driven but again uses a small vocabulary without direct visual support for every word. It outputs a large set of sentences to describe all triplets of recognized objects in the scene, but does not include a content selection model to select the best sentence. In the evaluation, we extend Midge with content and sentence selection rules so that we can use it as a baseline.

The visual facts we annotate are motivated by research in machine vision. Attributes have been shown to be a good intermediate representation for categorization [45]. Activity recognition in still images is also emerging [92; 148; 130], although significantly less studied than object recognition. Also, parts have been widely used in object recognition [50]. Yet, no work tests the contribution of these labels for sentence generation.

There is also a significant amount of work on other grounded language problems, where related models were developed. Visual referring expression generation systems [80; 110; 54] aim to identify specific objects, a sub-problem we deal with when describing images more generally. Other research generates descriptions in simulated worlds and, like this work, uses feature rich models [2], or syntactic structures like PCFGs [17; 77] but does not combine the two. Finally, Zitnick and Parikh [163] study sentences describing clipart scenes. They present a number of factors influencing overall descriptive quality, several of which we use in sentence generation for the first time.

2.4 Common Sense

Common sense knowledge has been predominately created directly from human input or extracted from text [91; 98; 14]. In contrast, our work is focused on visual common sense extracted from images annotated with

regions and descriptions.

There has also been recent interest in the vision community to build databases of visual common sense knowledge. Efforts have focused on a small set of relations, such as *similar to* or *part of* [20]. Webly supervised techniques [33; 20] have also been used to test whether a particular object-relation-object triplet occurs in images [128]. In contrast, we use seven spatial relations and allow natural language relations that represent a larger array of higher level semantics. We also leverage existing efforts on annotating large scale image datasets instead of relying on the noisy outputs of a computer vision system.

On a technical level, our methods for extracting common sense facts from images rely on Pointwise-Mutual Information (PMI), analogous to other rule extraction systems based on text [94; 129]. We view objects as an analogy for words, images as documents, and object-object configurations as typed bigrams. Our methods for generalizing relations are inspired by work that tries to predict a class label for an image given a hierarchy of concepts [29; 113; 115]. Yet our work is the first to deal with visual relations between pairs of concepts in the hierarchy by using a sub-modular formulation that maximizes the amount of coverage of subordinate categories while avoiding contradictions with an initial set of discovered common-sense assertions.

2.5 Situation Recognition

Activity recognition in still images has been widely studied [65], and it is generally accepted that objects and scenes are important for recognition [93]. These intuitions are often built directly into datasets by framing activity recognition as a discrete classification problem, with a small set of multi-word category labels that combine a verb with a scene or object [28; 42; 68; 138; 149; 151]. Although recent work has scaled the number of classes [88], they are still hand selected and it can be difficult to know what should be included in the set. For example, while "cutting-vegetables" is a category in Stanford-40, many others possibilities, like "cutting-grass" or the more generic "cutting," are missing (similar examples can be found in all current activity recognition datasets). In contrast, our task formulation uses linguistic resources to define a very large and significantly more comprehensive space of possible situations.

Many methods have been proposed for modeling semantic context in activity recognition [34]. Our approach is most closely related to work that models object co-occurrence [120] and uses graphical models

to combine many sources of contextual information [59; 49]. Actions have been a particularly fruitful source of context [103], especially when combined with pose to create human-object interactions [100; 150]. However, we present the first approach to define how multiple objects participate in a single activity, allowing us to systematically recover activity-specific facts such as "Who is doing it?" (the `agent` role), "What are they doing it to?" (`patient`), etc.

There is also significant related work in the intersection of language and vision. WordNet [106] is used to define ImageNet [30] classes, much like how we use FrameNet [53] to define our situation space. Recent work has also explored other areas of cross pollination, including video recognition [64], cross modal mappings [135; 87; 57], coreference [40; 76], and affordances [162]. In particular, sentence generation is closely related and has received significant attention [153; 73; 22; 43; 145; 102; 70; 116; 97]. Our situations are inspired by semantic role labeling models [26; 75], which are designed to provide a type of shallow semantics for verbs; in essence, our frames correspond to simple declarative sentences. However, we sidestep the evaluation challenges that come with generating sentences [144; 39], while also providing visual evidence for verbs that should aid captioning. At least partially motivated by the same concerns, there are recent efforts to formulate Visual Question Answering (VQA) tasks [4; 123; 154; 60; 41], where the system must answer questions like "What is the person using to cut the grass?" In a pilot study on a VQA dataset [4], we found that up to 20% of questions ask about a semantic role, suggesting that situation recognition could be beneficial.

Finally, situation recognition is related to two parallel efforts to define visual semantic role labeling tasks. Both provide instance-level information with bounding regions for objects [69; 125]. We instead focus on classification, annotate an order of magnitude more images and are the first to consider more than two semantic roles.

2.6 Sparsity in Situation Recognition

Learning to cope with semantic sparsity is closely related to zero-shot or k-shot learning. Attribute-based learning [85; 86; 48], cross-modal transfer [135; 90; 57; 87] and using text priors [99; 64] have all been proposed but they study classification or other simplified settings. For the structured case, image captioning models [153; 73; 22; 43; 102; 70; 116; 97] have been observed to suffer from a lack of diversity and

generalization [145]. Recent efforts to gain insight on such issues extract subject-verb-object (SVO) triplets from captions and count prediction failures on rare tuples [6]. Our use of imSitu to study semantic sparsity circumvents the need for intermediate processing of captions and generalizes to verbs with more than two arguments.

Compositional models have been explored in a number of applications in natural language processing, such as sentiment analysis [137], dependency parsing [89], text similarity [7], and visual question answering [1] as effective tools for combining natural language elements for prediction. Recently, bilinear pooling [96] and compact bilinear pooling [61] have been proposed as second-order feature representations for tasks such as fine grained recognition and visual question answer. We build on such methods, using low dimensional embeddings of semantic units and expressive outer product computations.

Using the web as a resource for image understanding has been studied through NEIL [21], a system which continuously queries for concepts discovered in text, and Levan [33], which can create detectors from user specified queries. Web supervision has also been explored for pretraining convolutional neural networks [19] or for fine-grained bird classification [19] and common sense reasoning [128]. Yet we are the first to explore the connection between semantic sparsity and language for automatically generating queries for semantic web augmentation and we are able to show improvement on a large scale, fully supervised structured prediction task.

Chapter 3

Sentimental Language

Language can describe our visual world at many levels, including not only what is literally there but also the sentiment that it invokes. In this paper, we study visual language, both literal and sentimental, that describes the overall appearance and style of virtual characters. Sentimental properties, including labels such as "youthful" or "country western," must be inferred from descriptions of the more literal properties, such as facial features and clothing selection. We present a new dataset, collected to describe Xbox avatars, as well as models for learning the relationships between these avatars and their literal and sentimental descriptions. In a series of experiments, we demonstrate that such learned models can be used for a range of tasks, including predicting sentimental words and using them to rank and build avatars. Together, these results demonstrate that sentimental language provides a concise (though noisy) means of specifying low-level visual properties.

3.1 Introduction

Language can describe varied aspects of our visual world, including not only what is literally there but also the social, cultural, and emotional sentiment it invokes. Recently, there has been a growing effort to study *literal* language that describes directly observable properties, such as object color, shape, or category [44; 109; 104]. Here, we add a focus on *sentimental* visual language, which compactly describes more subjective properties such as if a person looks determined, if a resume looks professional, or if a restaurant looks romantic. Such models enable many new applications, such as text editors that automatically select

(A) This is a *light tan young* man with *short* and *trim* haircut. His hair is *light brown* and *grows closely to this face*. He has *straight* eyebrows and *large brown* eyes. He has a *neat* and *trim* appearance.

(B) *State of mind:* angry, upset, determined. *Likes:* country western, rodeo. *Occupation:* cowboy, wrangler, horse trainer. *Overall:* youthful, cowboy.

Figure 3.1: (A) Literal avatar descriptions and (B) sentimental descriptions of four avatar properties, including possible occupations and interests.

properties including font, color, or text alignment to best match high level descriptions such as "professional" or "artistic."

In this paper, we study visual language, both literal and sentimental, that describes the overall appearance and style of virtual characters, like those in Figure 3.1. We use literal language as feature norms, a tool used for studying semantic information in cognitive science [105]. Literal words, such "black" or "hat," are annotated for objects to indicate how people perceive visual properties. Such feature norms provide our gold-standard visual detectors, and allow us to focus on learning to model sentimental language, such as "youthful" or "goth."

We introduce a new corpus of descriptions of Xbox avatars created by actual gamers. Each avatar is specified by 19 attributes, including clothing and body type, allowing for more than 10^{20} possibilities. Using Amazon Mechanical Turk,[1] we collected literal and sentimental descriptions of complete avatars and many of their component parts, such as the cowboy hat in Figure 3.1(B). In all, there are over 100K descriptions. To demonstrate potential for learning, we also report an A/B test which shows that native

[1] www.mturk.com

speakers can use sentimental descriptions to distinguish the labeled avatars from random distractors. This new data will enable study of the relationships between the co-occurring literal and sentimental text in a rich visual setting.[2]

We describe models for three tasks: (i) classifying when words match avatars, (ii) ranking avatars given a description, and (iii) constructing avatars to match a description. Each model includes literal part descriptions as feature norms, enabling us to learn which literal and sentinel word pairs best predict complete avatars.

Experiments demonstrate the potential for jointly modeling literal and sentimental visual descriptions on our new dataset. The approach outperforms several baselines and learns varied relationships between the sentimental and literal descriptions. For example, in one experiment "nerdy student" is predictive of an avatar with features indicating its shirt is "plaid" and glasses are "large" and faces that are not "bearded." We also show that individual sentimental words can be predicted but that multiple avatars can match a single sentimental description. Finally, we use our model to build complete avatars and show that we can accurately predict the sentimental terms annotators ascribe to them.

3.2 Related Work

3.3 Data Collection

We gathered a large number of natural language descriptions from Mechanical Turk (MTurk). They include: (1) literal descriptions of specific facial features, clothing or accessories and (2) high level subjective descriptions of human-generated avatars.[3]

Literal Descriptions We showed annotators a single image of clothing, a facial feature or an accessory and asked them to produce short descriptions. Figure 3.2 shows the distribution over object types. We restricted descriptions to be between 3 and 15 words. In all, we collected 33.2K descriptions and had on average 7 words per descriptions.

[2]Data available at http://homes.cs.washington.edu/~my89/avatar.

[3](2) also has phrases describing emotional reactions. We also collected (3) multilingual literal, (4) relative literal and (5) comprehensive full-body descriptions. We do not use this data, but it will be included in the public release.

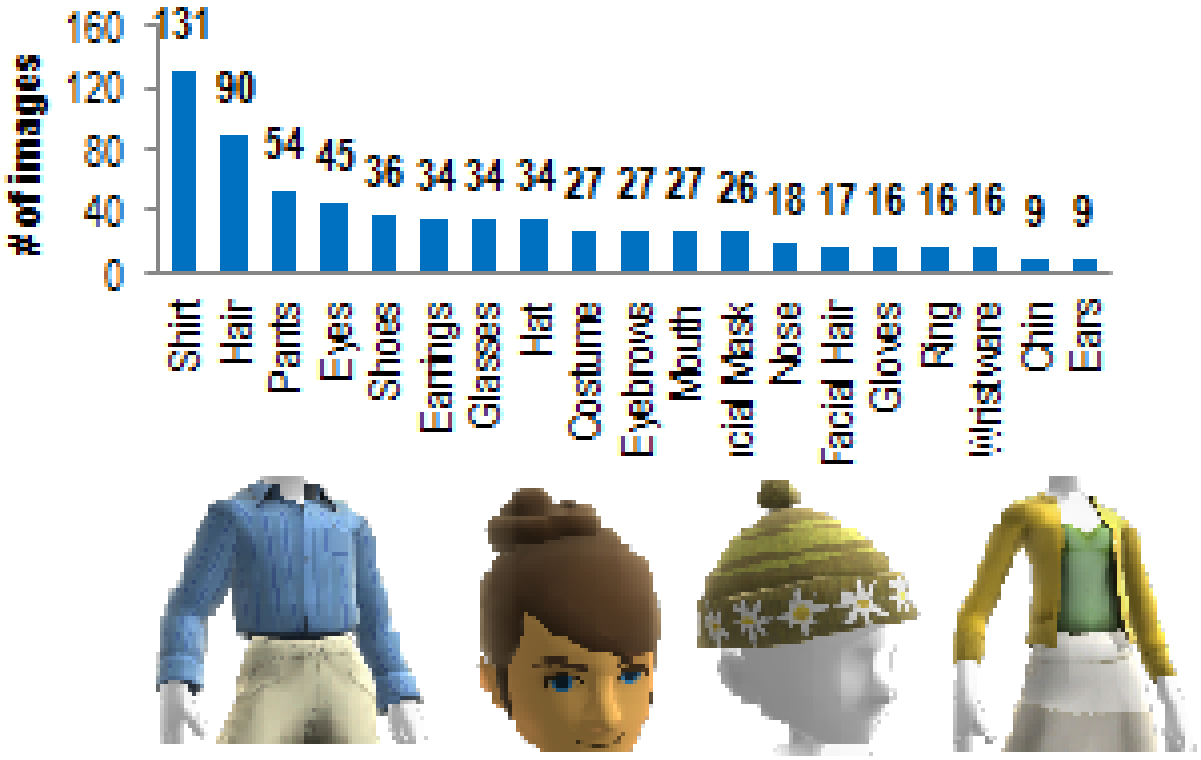


Figure 3.2: The number of assets per category and example images from the *hair, shirt* and *hat* categories.

Sentimental Descriptions We also collected 1913 gamer-created avatars from the web. The avatars were filtered to contain only items from the set of 665 for which we gathered literal descriptions. The gender distribution is 95% male.

To gather high level sentimental descriptions, annotators were presented with an image of an avatar and asked to list phrases in response to the follow different aspects:

- State of mind of the avatar.
- Things the avatar might care about.
- What the avatar might do for a living.
- Overall appearance of the avatar.

6144 unique vocabulary items occurred in these descriptions, but only 1179 occurred more than 10 times. Figure 3.1 (B) shows an avatar and its corresponding sentimental descriptions.

Quality Control All annotations in our dataset are produced by non-expert annotators. We relied on manual spot checks to limit poor annotations. Over time, we developed a trusted crowd of annotators who

produced only high quality annotations during the earliest stage of data collection.

3.4 Feasibility

Our hypothesis is that sentimental language does not uniquely identify an avatar, but instead summarizes or otherwise describes its overall look. In general, there is a trade off between concise and precise descriptions. For example, given a single word you might be able to generally describe the overall look of an avatar, but a long, detailed, literal description would be required to completely specify their appearance.

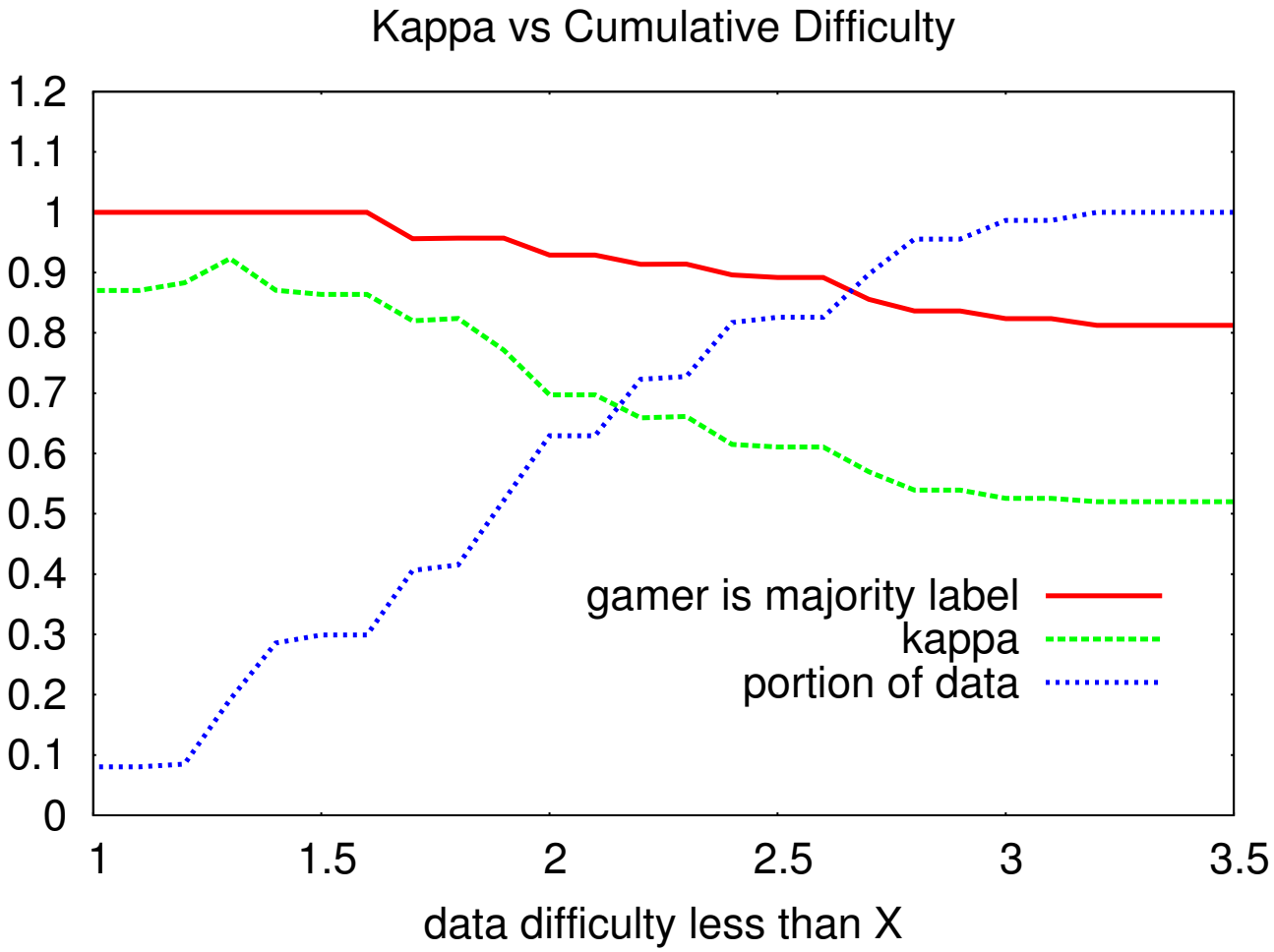


Figure 3.3: Judged task difficulty versus agreement, gamer avatar preference, and percentage of data covered. The difficulty axis is cumulative.

To demonstrate that the sentimental descriptions we collected are precise enough to be predictive of appearance, we conducted an experiment that prompts people to judge when avatars match descriptions. We

State of mind: playful, happy; Likes: sex Occupation: hobo Overall: dumb

State of mind: content, humble, satisfied, peaceful, relaxed, calm. Likes: fashion, friends, money, cars, music, education. Occupation: teacher, singer, actor, performer, dancer, computer engineer. Overall: nerdy, cool, smart, comfy, easygoing, reserved

Figure 3.4: Avatars rated as difficult.

created an A/B test where we show English speakers two avatars and one sentimental description. They were asked to select which avatar is better matched by the description and how difficult they felt, on a scale from 1 to 4, it was to judge. For 100 randomly selected descriptions, we asked 5 raters to compare the gamer avatars to randomly generated ones (where each asset is selected independently according to a uniform distribution). Figure 3.3 shows a plot of Kappa and the percent of the time a majority of the raters selected the gamer avatar. The easiest 20% of the data pairs had the strongest agreement, with kappa=.92, and two thirds of the data has kappa = .70. While agreement falls off to .52 for the full data set, the gamer avatar remains the majority judgment 81% of the time.

The fact that random avatars are sometimes preferred indicates that it can be difficult to judge sentimental descriptions. Consider the avatars in Figure 3.4. Neither conforms to a clear sentimental description based on the questions we asked. The right one is described with conflicting words and the words describing the left one are very general (like “dumb”). This corresponds to our intuition that while many avatars can be succinctly summarized with our questions, some would be more easily described using literal language.

3.5 Tasks and Evaluation

We formulate three tasks to study the feasibility of learning the relationship between sentimental and literal descriptions. In this section, we first define the space of possible avatars, followed by the tasks.

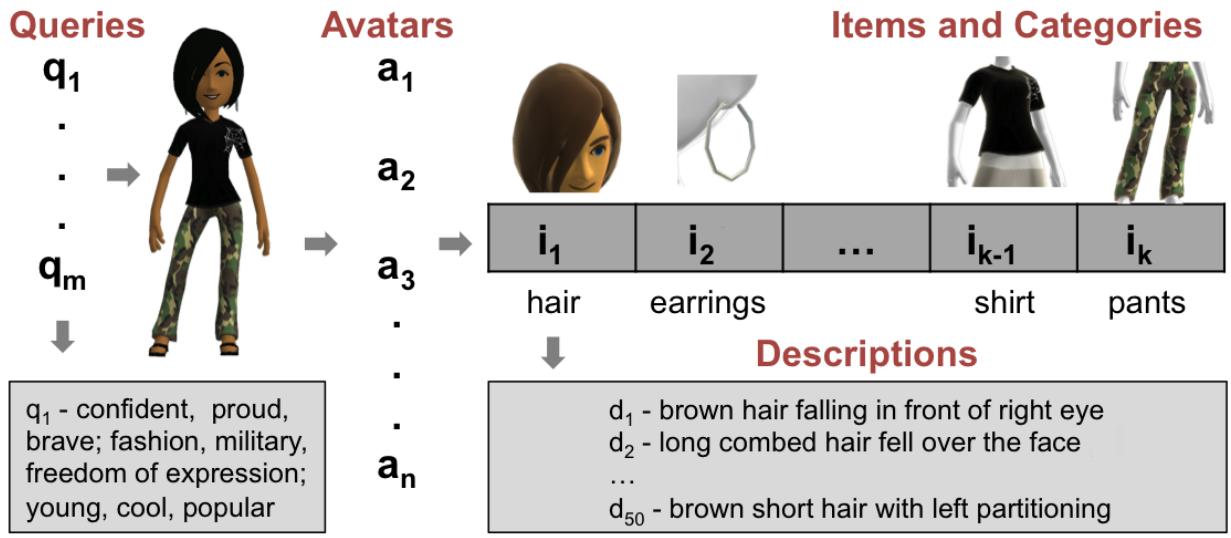


Figure 3.5: Avatars, queries, items, literal descriptions.

Avatars Figure 3.5 summarizes the notation we will develop to describe the data. An avatar is defined by a 19 dimensional vector $\vec{a}$ where each position is an index into a list of possible items $\vec{i}$. Each dimension represents a position on the avatar, for example, *hat* or *nose*. Each possible item is called an asset and is associated with a set of positions it can fill. Most assets take up exactly one position, while there are a few cases where assets take multiple positions.[4] An avatar $\vec{a}$ is valid if all of its mandatory positions are filled, and no two assets conflict on a position. Mandatory positions include hair, eyes, ears, eyebrows, nose, mouth, chin, shirt, pants, and shoes. All other positions are optional. We refer to this set of valid $\vec{a}$ as A. Practically speaking, if an avatar is not valid, it cannot be reliably rendered graphically.

Each item i is associated with the literal descriptions $\vec{d_i} \in D$ where D is the set of literal descriptions. Furthermore, every avatar $\vec{a}$ is associated a list of sentimental query words $\vec{q}$, describing subjective aspects

[4]For example, long sleeve shirts cover up watches, so they take up both shirt and wristwear positions. Costumes tend to span many more positions, for example there a suit that takes up shirt, pants, wristwear and shoes positions.

of an avatar.[5]

Sentimental Word Prediction We first study individual words. The word prediction task is to decide whether a given avatar can be described with a particular sentimental word q^*. We evaluate performance with F-score.

Avatar Ranking We also consider an avatar retrieval task, where the goal is to rank the set of avatars in our data, $\cup_{j=1...n}\vec{a_j}$, according to which one best matches a sentimental description, $\vec{q_i}$. As an automated evaluation, we report the average percentile position assigned to the true $\vec{a_i}$ for each example. However, in general, many different avatars can match each $\vec{q_i}$, an interesting phenomena we will further study with human evaluation.

Avatar Generation Finally, we consider the problem of generating novel, previously unseen avatars, by selecting a set of items that best embody some sentimental description. As with ranking, we aim to construct the avatar $\vec{a_i}$ that matches each sentimental description $\vec{q_i}$. We evaluate by considering the item overlap between $\vec{a_i}$ and the output avatar $\vec{a^*}$, discounting for empty positions:[6]

$$f = \frac{\sum_{j=1}^{|\vec{a^*}|} I(\vec{a_j^*} = \vec{a_{ij}})}{\max(numparts(\vec{a^*}), numparts(\vec{a_i}))}, \tag{3.1}$$

where $numparts$ returns the number of non-empty avatar positions. The score is a conservative measure because some items are significantly more visually salient than others. For instance, shirts and pants occupy a large portion of the physical realization of the avatar, while rings are small and virtually unnoticeable. We additionally perform a human evaluation in Section 7.5 to better understand these challenges.

3.6 Methods

We present two different models: one that considers words in isolation and another that jointly models the query words. This section defines the models and how we learn them.

[5] We do not distinguish which prompt (e.g., "state of mind" or "occupation") a word in $\vec{q}$ came from, although the vocabularies are relatively disjoint.

[6] Optional items are infrequently used. Therefore not predicting them at all offers a strong baseline. Yet doing this demonstrates nothing about an algorithm's ability to predict items which contribute to the sentimental qualities of an avatar.

3.6.1 Independent Sentimental Word Model

The independent word model (S-Independent) assumes that each word independently describes the avatar. We construct a separate linear model for each word in the vocabulary.

To train these model, we transform the data to form a binary classification problem for each word, where the positive data includes all avatars the word was seen with, $(q, \vec{a_i}, 1)$ for all i and $q \in \vec{q_i}$, and the rest are negative, $(q, \vec{a_i}, 0)$ for all i and $q \notin \vec{q_i}$.

We use the following features:

- an indicator feature for the cross product of a sentiment query word q, a literal description word $w \in D$, and the avatar position index j (for example, $q =$ "angry" with $w =$ "pointy" and $j =$ eyebrows):

$$I(q \in \vec{q_i}, w \in \vec{d_{a_{ij}}}, j)$$

- a bias feature for keeping a position empty:

$$I(q \in \vec{q_i}, a_{ij} = empty, j)$$

These features will allow the model to capture correlations between our feature norms which provide descriptions of visual attributes, like black, and sentimental words, like gothic.

S-Independent is used for both word prediction and ranking. For prediction, we train a linear model using averaged binary perceptron. For ranking, we try to rank all positive instances above negative instances. We use an averaged structured perceptron to train the ranker [23]. To rank with respect to an entire query $\vec{q_i}$, we sum the scores of each word $q \in \vec{q_i}$.

3.6.2 Joint Sentimental Model

The second approach (S-Joint) jointly models the query words to learn the relationships between literal and sentimental words with score s:

$$s(\vec{a}|\vec{q}, D) = \sum_{i=1}^{|\vec{a}|} \sum_{j=1}^{|\vec{q}|} \theta^T f(\vec{a_i}, \vec{q_j}, \vec{d_{a_i}})$$

Where every word in the query has a separate factor and every position is treated independently subject to the constraint that $\vec{a}$ is valid. The feature function f uses the same features as the word independent model above.

This model is used for ranking and generation. For ranking, we try to rank the avatar a_i for query q_i above all other avatars in the candidate set. For generation, we try to score a_i above all other valid avatars given the query q_i. In both cases, we train with averaged structured perceptron [23] on the original data, containing query, avatar pairs $(\vec{q_i}, \vec{a_i})$.

3.7 Experimental Setup

Random Baseline For the ranking and avatar generation tasks, we report random baselines. For ranking, we randomly order the avatars. In the generation case, we select an item randomly for every position. This baseline does not generate optional assets because they are rare in the real data.

Sentimental-Literal Overlap (SL-Overlap) We also report a baseline that measures the overlap between words in the sentiment query $\vec{q_i}$ and words in the literal asset descriptions D. In generation, for each position in the avatar, $\vec{a_i}$, SL-Overlap selects the item whose literal description has the most words in common with $\vec{q_i}$. If no item had overlap with the query, we backoff to a random choice. In the case of ranking, it orders avatars by the sum over every position of the number of words in common between the literal description and the query, $\vec{q_i}$. This baseline tests the degree to which literal and sentimental descriptions overlap lexically.

Feature Generation For all models that use lexical features, we limited the number of words. 6144 unique vocabulary items occur in the query set, and 3524 in the literal description set. There are over 400 million entries in the full set of features that include the cross product of these sets with all possible avatar positions, as described in Section 3.6. Since this would present a challenge for learning, we prune in two ways. We stem all words with a Porter stemmer. We also filter out all features which do not occur at least 10 times in our training set. The final model has approximately 700k features.

Word	F-Score	Precision	Recall	N
happi	0.84	0.89	0.78	149
student	0.78	0.82	0.74	129
friend	0.76	0.84	0.70	153
music	0.74	0.89	0.63	148
confid	0.74	0.82	0.76	157
sport	0.69	0.62	0.76	76
casual	0.63	0.6	0.67	84
youth	0.6	0.57	0.64	88
waitress	0.59	0.42	1	5
smart	0.57	0.54	0.6	88
fashion	0.54	0.54	0.54	70
monei	0.54	0.52	0.56	76
cool	0.54	0.52	0.56	84
relax	0.53	0.52	0.56	90
game	0.51	0.44	0.62	61
musician	0.51	0.44	0.61	66
parti	0.51	0.43	0.62	58
content	0.5	0.47	0.53	75
friendli	0.49	0.42	0.6	56
smooth	0.49	0.4	0.63	57

Table 3.1: Top 20 words (stemmed) for classification. N is the number of occurrences in the test set.

3.8 Results

We present results for the tasks described in Section 3.5 with the appropriate models from Section 3.6.

3.8.1 Word Prediction Results

The goal of our first experiment is to study when individual sentiment words can be accurately predicted. We computed sentimental word classification accuracy for 1179 word classes with 10 or more mentions. Table 3.1 shows the top 20 words ordered by F-score.[7] Many common words can be predicted with relatively high accuracy. Words with strong individual cues like happy (a smiling mouth), and confidence (wide eyes) and nerdi (particular glasses) can be predicted well.

The average F-score among all words was .085. 33.2% of words have an F-score of zero. These zeros include words like: unusual, bland, sarcastic, trust, prepared, limber, healthy and poetry. Some of these words indicate broad classes of avatars (e.g., unusual avatars) and others indicate subtle modifications to looks that without other words are not specific (e.g., a prepared surfer vs. a prepared business man). Fur-

[7] Accuracy numbers are inappropriate in this case because the number of negative instances, in most cases, is far larger than the number of positive ones.

pensive,confrontational; music,socializing; musician,bar tending,club owner; smart,cool.

Figure 3.6: A sentimental description paired with the highest ranked avatars found by S-Joint.

Algorithm	Percentile Rank
S-joint	77.3
S-independant	73.5
SL-overlap	60.4
Random	48.8

Table 3.2: Automatic evaluation of ranking. The average percentile that a test avatar was ranked given its sentimental description.

thermore, evaluation was done assuming that when a word is not mentioned, it is should be predicted as negative. This fails to account for the fact that people do not mention everything that's true, but instead make choices about what to mention based on the most relevant qualities. Despite these difficulties, the classification performance shows that we can accurately capture usage patterns for many words.

3.8.2 Ranking Results

Ranking allows us to test the hypothesis that multiple avatars are valid for a high level description. Furthermore, we consider the differences between S-Joint and S-Independent, showing that jointly modelings all words improves ranking performance.

Automatic Evaluation The results are shown in Table 3.2. Both S-Independent and S-Joint outperform the SL-overlap baseline. SL-Overlap's poor performance can be attributed to low direct overlap between sentimental words and literal words. S-Joint also outperforms the S-Independent.

Inspection of the parameters shows that S-Joint does better than S-Independent in modeling words that

Model	Overlap
Random	0.041
SL-Overlap	0.049
S-Joint	0.126

Table 3.3: Automatic generation evaluation results. The item overlap metric is defined in Section 3.5.

only relate to a subset of body positions. For example, in one case we found that for the word "puzzled" nearly 50% of the weights were on features that related to eyebrows and eyes. This type of specialization was far more pronounced for S-Joint. The joint nature of the learning allows the features for individual words to specialize for specific positions. In contrast, S-Independent must independently predict all parts for every word.

Human Evaluation We report human relevancy judgments for the top-5 returned results from S-Joint. On average, 56.2% were marked to be relevant. This shows that S-Joint is performing better than automatic numbers would indicate, confirming our intuition that there is a one-to-many relationship between a sentimental description and avatars. Sentimental descriptions, while having significant signal, are not exact. These results also indicate that relying on automatic measures of accuracy that assume a single reference avatar underestimates performance. Figure 3.6 shows the top ranked results returned by S-Joint for a sentimental description where the model performs well.

3.8.3 Generation Results

Finally we evaluate three models for avatar generation: Random, SL-Overlap and S-Joint using automatic measures and human evaluation.

Automatic Evaluation Table 3.3 presents results for automatic evaluation. The Random baseline performs badly, on average assigning items correctly to less than 1 position in the generated avatar. The SL-Overlap baseline improves, but still performs quite poorly. The S-Joint model performs significantly better, correctly guessing 2-3 items for each output avatar. However, as we will see in the manual evaluation, many of the non-matching parts it produces are still a good fit for the query.

	Kappa	Majority	Random	Sys.
SL-Overlap	0.20	0.25	0.34	0.32
S-Joint	0.52	0.90	0.07	0.81
Gamer	0.52	0.81	0.08	0.77

Table 3.4: Human evaluation of automatically generated avatars. Majority represents the percentage of time the system output is preferred by a majority of raters. Random and System (Sys.) indicate the percentage of time each was preferred.

Sentiment	positive features	negative features
happi	mouth:thick, mouth:smilei, mouth:make, mouth:open	mouth:tight, mouth:emotionless, mouth:brownish, mouth:attract
gothic	shoes:brown, shirt:black, pants:hot, shirt:band	shirt:half, shirt:tight, pants:sexi, hair:brownish
retro	eyebrows:men, eyebrows:large, hair:round, pants:light	eyebrows:beauti, pants:side; eyebrows:trim, pants:cut
beach	pants:yello, pants:half, nose:narrow, pants:white	shirt:brown, shirt:side; shoes:long, pants:jean

Table 3.5: Most positive and negative features for a word stem. A feature is [position]:[literal word].

Human Evaluation As before, there are many reasonable avatars that could match as well as the reference avatars. Therefore, we also evaluated generation with A/B tests, much like in Section 3.4. Annotators were asked to judge which of two avatars better matched a sentimental description. They could rate System A or System B as better, or report that they were equal or that neither matched the description. We consider two comparisons: SL-Overlap vs. Random and S-Joint vs Random. Five annotators performed each condition, rating 100 examples with randomly ordered avatars.

We report the results for human evaluation including kappa, majority judgments, and a distribution over judgments in Table 3.4. The SL-Overlap baseline is indistinguishable from a random avatar. This contrasts with the ranking case, where this simple baseline showed improvement, indicating that generation is a much harder problem. Furthermore, agreement is low; people felt the need to make a choice but were not consistent.

We also see in Table 3.4 that people prefer the S-Joint model outputs to random avatars as often as they prefer gamer to random. While this does not necessarily imply that S-Joint creates gamer-quality avatars, it indicates substantial progress by learning a mapping between literal and sentimental words.

Qualitative Results Table 6 presents the highest and lowest weighted features for different sentimental query words. Figure 3.8 shows four descriptions that were assigned high quality avatars.

In general, many of the weaker avatars had aspects of the descriptions but lacked such distinctive overall looks. This was especially true when the descriptions contained seemingly contradictory information. For

Ambition; business, fashion, success; salesman; smooth, professional.

Capable, confident, firm; heavy metal, extreme sports, motorcycles; engineer, mechanic, machinist; aggressive, strong, protective.

Figure 3.7: Avatars automatically generated with the S-Joint model.

example, one avatar was described as being both nerdy and popular. We generated a look that had aspects of both of these descriptions, including a head that contained both conservative elements (like glasses) and less conservative elements (like crazy hair and earrings). However, the combination would not be described as nerdy or popular, because of difficult to predict global interactions between the co-occurring words and items. This is an important area for future work.

Stressed, bored, discontent; emo music; works at a record store; goth, dark, drab.

Happy, content, confident, home, career, family, secretary,student, classy,clean,casual

Figure 3.8: Avatars automatically generated with the S-Joint model.

weight	position	literal stem	figurative stem
1.23	mouth	grin	happi
1.19	mouth	youth	happi
1.17	mouth	friendli	happi
1.13	mouth	wide	happi
1.12	mouth	semicircl	happi
-1.32	mouth	anxieti	happi
-1.11	mouth	tight	happi
-1.03	mouth	lack	happi
-1.02	mouth	upset	happi
-0.98	mouth	angri	happi
1.05	shirt	joi	bank
0.97	costume	formal	bank
0.96	costume	fit	bank
0.91	costume	super	bank
0.90	costume	overcoat	bank
-1.13	eyes	brow	bank
-0.74	eyes	alert	bank
-0.71	shirt	symbol	bank
-0.68	eyes	natur	bank
-0.67	hair	smooth	bank

Table 3.6: positive and negative indicators for query words

Chapter 4

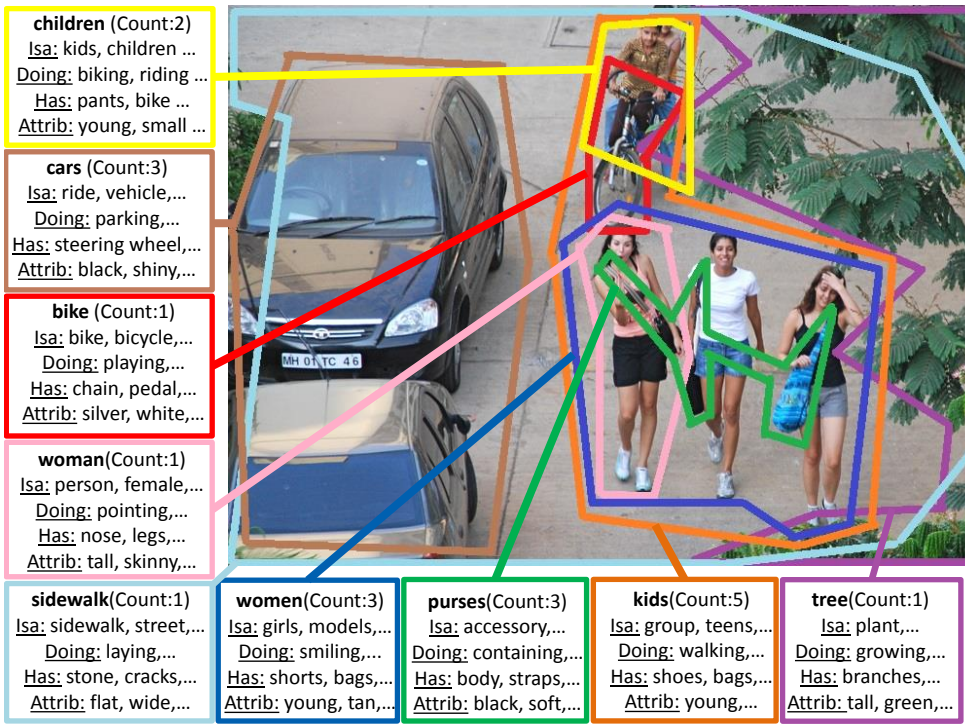
Dense Labeling

This paper, for the first time, studies generation of descriptive sentences from densely annotated images. Previous work studied generation from automatically detected visual information but produced a limited class of sentences, hindered by currently unreliable recognition of activities and attributes. Instead, we collect human annotations of objects, parts, attributes and activities in images. These annotations allow us to build a significantly more comprehensive model of language generation and also allow us to study what visual information is required to generate human-like descriptions. Experiments demonstrate high quality output, relative to two baselines, and that activity annotations and relative spatial location of objects contribute the most to producing high quality sentences.

4.1 Introduction

Image descriptions compactly summarize complex visual scenes. For example, consider the descriptions of the image in Figure 4.1, which vary in content but focus on the women and what they are doing. Automatically generating such descriptions is challenging: a full system must understand the image, select the relevant visual content to present, and construct complete sentences. Existing systems aim to address all of these challenges but use visual detectors for only a small vocabulary of words, typically nouns, associated with objects that can be reliably found.[1] Such systems are blind to much of the visual content needed to

[1]While object recognition is improving (ImageNet accuracy is over 90% for 1000 classes) progress in activity recognition has been slower; the state of the art is below 50% mean average precision for 40 activity classes [148].



Five young people on the street, two sharing a bicycle.
Several young people are walking near parked vehicles.
Three girls with large handbags walking down the sidewalk.
Three women walk down a city street, as seen from above.
Three young woman walking down a sidewalk looking up.

Figure 4.1: An annotated image with human generated sentence descriptions. Each bounding polygon encompasses one or more objects and is associated with a count and text labels.This image has 9 high level objects annotated with over 250 textual labels.

generate complete, human-like sentences.

In this paper, we instead study generation with more complete visual support, as provided by human annotations, allowing us to develop more comprehensive models than previously considered. Such models have the dual benefit of (1) providing new insights into how to construct more human-like sentences and (2) allowing us to perform experiments that systematically study the contribution of different visual cues in generation, suggesting which automatic detectors would be most beneficial for generation.

In an effort to approximate relatively complete visual recognition, we collected manually labeled representations of objects, parts, attributes and activities for a benchmark caption generation dataset that includes images paired with human authored descriptions [122].[2] As seen in Figure 4.1, the labels include object boundaries and descriptive text, here including the facts that the children are "riding" and "walking" and that they are "young." Our goal is to be as exhaustive as possible, giving equal treatment to all objects. For

[2]Our annotations will be made publicly available.

example, the annotations in Figure 4.1 contain enough information to generate the first three sentences and most of the content in the remaining two. Labels gathered in this way are a type of feature norms [105], which have been used in the cognitive science literature to approximate human perception and were recently used as a visual proxy in distributional semantics [134]. We present the first effort, that we are aware of, for using feature norms to study image description generation.

Such rich data allows us to develop significantly more comprehensive generation models. We divide generation into choices about which visual content to select and how to realize a sentence that describes that content. Our approach is grammar-based, feature-rich, and jointly models both decisions. The content selection model includes latent variables that align phrases to visual objects and features that, for example, measure how visual salience and spatial relationships influence which objects are mentioned. The realization approach considers a number of cues, including language model scores, word specificity, and relative spatial information (for example to produce the best spatial prepositions), when producing the final sentence. When used with a reranking model, including global cues such as sentence length, this approach provides a full generation system.

Our experiments demonstrate high quality visual content selection, within 90% of human performance on unigram BLEU, and improved complete sentence generation, nearly halving the difference from human performance to two baselines on 4-gram BLEU. In ablations, we measure the importance of different modeling considerations and visual cues, showing that annotation of activities and relative bounding box information between objects are crucial contributors to generating human-like description.

4.2 Dataset

We collected a dataset of richly annotated images to approximate gold standard visual recognition. In collecting the data, we sought a visual annotation with sufficient coverage to support the generation of as many of the words in the original image descriptions as possible. We also aimed to make it as visually exhaustive as possible—giving equal treatment to all visible objects. This ensures less bias from annotators' perception about which objects are important, since one of the problems we would like to solve is content selection. This dataset will be made available for future experiments.

We built on the dataset from [122] which contained 8,000 Flickr images and associated descriptions

gathered using Amazon Mechanical Turk (MTurk). Restricting ourselves to Creative Commons images, we sampled 500 images for annotation.

We collected annotations of images in three stages using MTurk, and assigned each annotation task to 3-5 workers to improve quality through redundancy [13]. Below we describe the process for annotating a single image.

Stage 1: We prompted five turkers to list *all* objects in an image, ignoring objects that are parts of larger objects (e.g., the arms of a person), which we collected later in Stage 3. This list also included groups, such as crowds of people.

Stage 2: For each unique object label from Stage 1, we asked two turkers to draw a polygon around the object identified.[3] In cases where the object is a group, we also asked for the number of objects present (1-6 or many). Finally, we created a list of all references to the object from the first stage, which we call the *Object* facet.

Stage 3: For each object or group, we prompted three turkers to provide descriptive phrases of:

- *Doing* – actions the object participates in, e.g. "jumping."
- *Parts* – physical parts e.g. "legs", or other items in the possession of the object e.g. "shirt."
- *Attributes* – adjectives describing the object, e.g. "red."
- *Isa* – alternative names for a object e.g. "boy", "rider."

Figure 4.1 shows more examples for objects in a labeled image.[4] We refer to all of these annotations, including the merged *Object* labels, as facets. These labels provide feature norms [105], which have recently used as a visual proxy in distributional semantics [134; 133] but have not been previous studied for generation. This annotation of 500 images (2500 sentences) yielded over 4000 object instances and 100,000 textual descriptions.

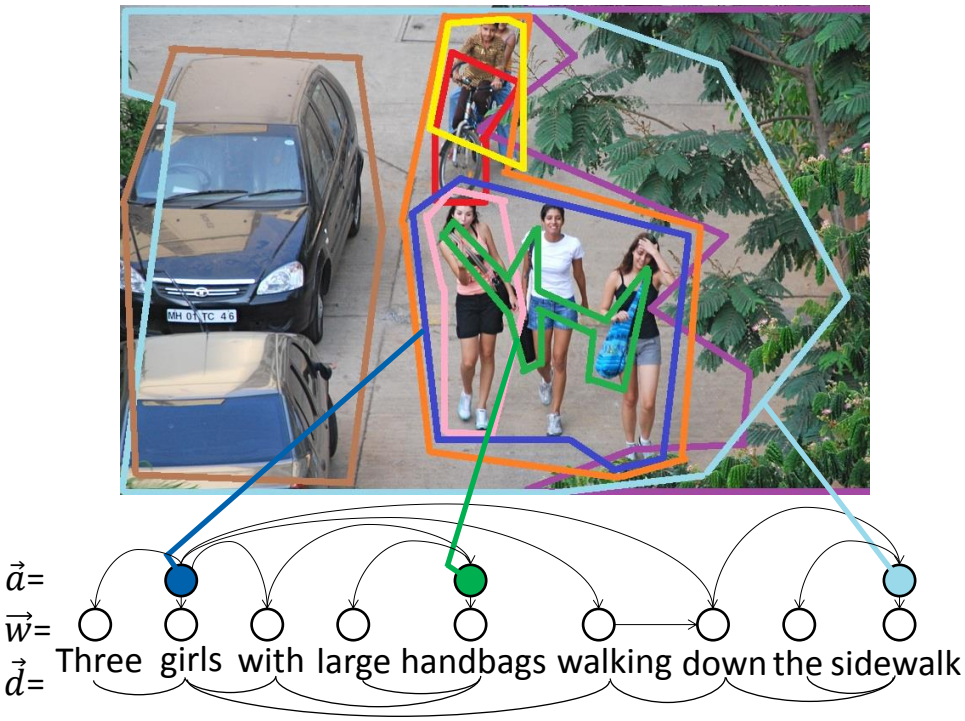


Figure 4.2: One path through the generative model and the Bayesian network it induces. The first row of colored circles are alignment variables to objects in the image. The second row is words, generated conditioned on alignments.

4.3 Approach

Given such rich annotations, we can now develop significantly more comprehensive generation models. In this section, we present an approach that first uses a generative model and then a reranker. The generative model defines a distribution over content selection and content realization choices, using diverse cues from the image annotations. The reranker trades off our generative model score, language model score (to encourage fluency), and length of the sentence to produce the final output.

Generative Model We want to generate a sentence $\vec{w} = \langle w_1 \dots w_n \rangle$ where each word $w_i \in V$ comes from a fixed vocabulary V. The vocabulary V includes all 2700 words used in descriptive sentences in the training set.[5]

The model conditions on an annotated image I that contains a set of objects O, where each object $o \in O$

[3]We modified LabelMe [141].

[4]In the experiments, Parts and Isa facets do not improve performance, so we do not use them in the final model. Isa is redundant with the Object facet, as can be seen in Figure 4.1. Also parts like clothing, were often annotated as separate objects.

[5]We do not generate from image facets directly, because only 20% of the sentences in our data can be produced from these words alone. Instead, we develop features which consider the similarity between labels in the image and words in the vocabulary.

1. for a main clause (d,e are optional), select:
   (a) subject a_s alignment from $p_a(a)$.
   (b) subject word w_s from $p_n(w \mid a_s, \vec{d}_c)$
   (c) verb word w_v from $p_v(w \mid a_s, \vec{d}_c)$
   (d) object alignment a_o from $p_a(a' \mid a_s, w_v, \vec{d}_c)$
   (e) object word w_o from $p_n(w \mid a_o, \vec{d}_c)$
   (f) end with p_{stop} or go to (2) with (w_s, a_s)
   (g) end with p_{stop} or go to (2) with (w_v, a_s)
   (h) end with p_{stop} or go to (2) with (w_o, a_o)
2. for a (word, alignment) (w', a) (a,b are optional):
   (a) if w' not verb: modify w' with noun, select:
      i. modifier word w_n from $p_n(w \mid a, \vec{d}_c)$.
      ii. end with p_{stop} or go to (2) with (a_m, w_n)
   (b) modify w' with preposition, select:
      i. preposition word w_p
      if w' not a verb: from $p_p(w \mid a, \vec{d}_c)$
      else: from $p_p(w \mid a, w_v, \vec{d}_c)$
      ii. object alignment a_p from $p_a(a' \mid a, w_p, \vec{d}_c)$
      iii. object word w_n from $p_n(w \mid a_p, \vec{d}_c)$.
      iv. end with p_{stop} or go to (2) with (a_p, w_n)

Figure 4.3: Generative process for producing words $\vec{w}$, alignments $\vec{a}$ and dependencies $\vec{d}$. Each distribution is conditioned on the partially complete path through generative process $\vec{d}_c$ to establish sentence context. The notation p_{stop} is short hand for $p_{stop}(STOP|\vec{w}, \vec{d}_c)$ the stopping distribution.

has a bounding polygon and a number of facets containing string labels. To model the naming of specific objects, words w_i can be associated with alignment variables a_i that range over O. One such variable is introduced for each head noun in the sentence. Figure 4.2 shows alignment variable settings with colors that match objects in the image I. Finally, as a byproduct of the hierarchal generative process, we construct an undirected dependency tree $\vec{d}$ over the words in $\vec{w}$.

The complete generative model defines the probability $p(\vec{w}, \vec{a}, \vec{d} \mid I)$ of a sentence $\vec{w}$, word alignments $\vec{a}$, and undirected dependency tree $\vec{d}$, given the annotated input image I. The overall process unfolds recursively, as seen in Figure 4.3. The main clause is produced by first selecting the subject alignment a_s followed by the subject word w_s. It then chooses the verb and optionally the object alignment a_o and word

w_o. The process then continues recursively, modifying the subject, verb, and object of the sentence with noun and prepositional modifiers. The recursion begins at Step 2 in Figure 4.3. Given a parent word w and that word's relevant alignment variable a, the model creates attachments where w is the grammatical head of subsequently produced words. Choices about whether to create noun modifiers or prepositional modifiers are made in steps (a) and (b). The process chooses values for the alignment variables and then chooses content words, adding connective prepositions in the case of prepositional modifiers. It then chooses to end or submits new word-alignment pairs to be recursively modified.

Each line defines a decision that must be made according to a local probability distribution. For example, Step 1.a defines the probability of aligning a subject word to various objects in the image. The distributions are maximum entropy models, similar to previous work [2], using features described in the next section. The induced undirected dependency tree $\vec{d}$ has an edge between each word and the previously generated word (or the input word w in Steps 2.a.i and 2.a.ii, when no previous word is available). Figure 4.2 shows a possible output from the process, along with the Bayesian network that encodes what each decision was conditioned on during generation.

Learning We learn the model from data $\{(\vec{w}_i, \vec{d}_i, I_i) \mid i = 1 \ldots m\}$ containing sentences $\vec{w}_i$, dependency trees $\vec{d}_i$, computed with the Stanford parser [27], and images I_i. The dependency trees define the path that was taken through the generative process in Figure 4.3 and are used to create a Bayesian network for every sentence, like in Figure 4.2. However, object alignments $\vec{a}_i$ are latent during learning and we must marginalize over them.

$$\mathcal{L}(\theta) = \sum_i \log \sum_{\vec{a}} p(\vec{a}, \vec{w}_i, \vec{d}_i \mid I_i; \theta) - r|\theta|^2$$

where θ is the set of parameters and r is the regularization coefficient. In essence, we maximize the likelihood of every sentence's observed Bayesian network, while marginalizing over content selection variables we did not observe.

Because the model only includes pairwise dependencies between the hidden alignment variables $\vec{a}$, the inference problem is quadratic in the number of objects and non-convex because $\vec{a}$ is unobserved. We optimize this objective directly with L-BFGS, using the junction-tree algorithm to compute the sum and the

gradient.[6]

Inference To describe an image, we need to maximize over word, alignment, and the dependency parse variables:

$$\arg\max_{\vec{w},\vec{a},\vec{d}} p(\vec{w}, \vec{a}, \vec{d} \mid I)$$

This computation is intractable because we need to consider all possible sentences, so we use beam search for strings up to a fixed length.

Reranking Generating directly from the process in Figure 4.3 results in sentences that may be short and repetitive because the model score is a product of locally normalized distributions. The reranker takes as input a candidate list c, for an image I, as decoded from the generative model. The candidate list includes the top-k scoring hypotheses for each sentence length up to a fixed maximum. A linear scoring function is used for reranking optimized with MERT [112] to maximize BLEU-2.

4.4 Features

We construct indicator features to capture variation in usage in different parts of the sentence, types of objects that are mentioned, visual salience, and semantic and visual coordination between objects. The features are included in the maximum entropy models used to parameterize the distributions described in Figure 4.3. To limit over-fitting we avoid using lexical features whenever possible, instead relying on WordNet synsets [107].

Features in the generative model use tests for local properties, such as the identity of a synset of a word in WordNet, conjoined with an identifier that indicates context in the generative process.[7] Generative model features indicate (1) visual and semantic information about objects in distributions over alignments (content selection) and (2) preferences for referring to objects in distributions over words (content realization). Features in the reranking model indicate global properties of candidate sentences. Exact formulas for computing the features are in the supplementary material.

[6] To compute the gradient, we differentiate the recurrence in the junction-tree algorithm by applying the product rule.

[7] For example, in Figure 4.2 the context for the word "sidewalk" would be "word,syntactic-object,verb,preposition" indicating it is a word, in the syntactic object of a preposition, which was attached to a verb modifying prepositional phrase.

Visual features, such as an object's position in the image, are used for content selection. Pairwise visual information between two objects, for example the bounding box overlap between objects or the relative position of the two objects, is included in distributions where selection of an alignment variable conditions on previously generated alignments. For verbs (Step 1.d in Figure 4.3) and prepositions (Step 2.b.ii), these features are conjoined with the stem of the connective.

Semantic types of objects are also used in content selection. We define semantic types by finding synsets of labels in objects that correspond to high level types, [8] a list motivated by the animacy hiearchy [155]. Type features indicate the type of the object referred to by an alignment variable as well as the cross product of types when an alignment variable is on conditioning side of a distribution (e.g. Step 1.d). Like above, in the presence of a connective word, these features are conjoined with the stem of the connective.

Content realization features help select words when conditioning on already chosen alignments (e.g. Step 1.b). These features include the identity of the WordNet synset corresponding to a word, the word's depth in the synset hierarchy, the language model score for adding that word[9] and whether the word matches any labels in facets corresponding to the object referenced by an alignment variable.

Reranking features are primarily used to overcome issues of repetition and length endemic to using generative distributions, more commonly used for alignment, to create sentences. We use only four features: length, the number of repetitions, generative model score, and language model score.

4.5 Experimental Setup

Data We used 80% of the data for training (2000 sentences, 400 images), 10% for development, and 10% for testing (250 sentences, 50 images).

Parameters We tuned the regularization parameter on the held out data and chose $r = 8$. The reranker candidate list included the top 500 sentences for each sentence length up to 15 and weights were optimized with ZMERT [156].

[8] For example, human, animal, artifact (a human created object), natural body (trees, water, ect), or natural artifact (stick, leaf, rock).

[9] We use tri-grams with Kneser-Ney smoothing over the 1 million caption data set [114].

Metrics Our evaluation is based on BLEU-n [119], which considers all n-grams up to length n. To assess human performance using BLEU, we score each of the five references against the four other ones and finally average the five BLEU scores. In order to make these results comparable to BLEU scores for our model and baselines, we perform the same five-fold averaging when computing BLEU for each system.

We also compute accuracy for different syntactic positions in the sentence. We look at a number of categories: the main clause's components (S,V,O), prepositional phrase components, the preposition (Po) and their objects (Pp) and noun modifying words (N), including determiners. Phrases match if they have an exact string match and share context identifiers as defined in the features sections.

Human Evaluation Annotators rated sentences output by our full model against either human or a baseline system generated descriptions. Three criteria were evaluated: grammaticality, which sentence is more complete and well formed; truthfulness, which sentence is more accurately capturing something true in the image; and salience, which sentence is capturing important things in the image while still being concise. Two annotators annotated all test pairs for all criteria for a given pair of systems. Six annotators were used (none authors) and agreement was high (cohen's kappa = 0.963, 0.823 and 0.703 for grammar, truth and salience).

Machine Translation Baseline The first baseline is designed to see if it is possible to generate good sentences from the facet string labels alone, with no visual information. We use an extension of phrase-based machine translation techniques [111]. We created a virtual bitext by pairing each image description (the target sentence) with a sequence[10] of visual identifiers (the source "sentence") listing strings from the facet labels. We use a standard features and a few task specific features, and optimized BLEU-4 using MERT [112]. Details of our adaptations are in the supplementary material.

Midge Baseline As described in related work, the Midge system attempts to create a set of sentences to describe everything in an input image. These sentences must all be true, but do not have to select the same content that a person would. It can be adapted to our task by adding object selection and sentence ranking rules. Details of how we adapted Midge are in the supplementary material.

[10] We defined a consistent ordering of visual identifiers and set the distortion limit of the phrase-based decoder to infinity.

	BL-1	BL-2	BL-3	BL-4
Human	61.0	42.0	27.8	18.3
Full Model	**57.1**	**35.7**	**18.3**	**9.5**
MT Baseline	39.8	23.6	13.2	6.1
Midge Baseline	43.5	20.2	9.4	0.0

Table 4.1: Results for the test set for the BLEU1-4 metrics.

Grammar	Full	Other	Equal
Full vs **Human**	7.65	19.4	72.94
Full vs MT	6.47	5.29	88.23
Full vs Midge	40.59	15.88	43.53
Truth	Full	Other	Equal
Full vs **Human**	0.59	67.65	31.76
Full vs MT	30.0	10.59	59.41
Full vs Midge	51.76	27.71	23.53
Salience	Full	Other	Equal
Full vs **Human**	8.82	88.24	2.94
Full vs MT	51.76	16.47	31.77
Full vs Midge	71.18	14.71	14.12

Table 4.2: Human evaluation of our Full-Model in heads up tests against Human authored sentences and baseline systems, the machine translation baseline (MT) and the Midge inspired baseline. **Bold** indicates the better system. Other is not the Full system.

4.6 Results

We report experiments for our generation pipeline and ablations that remove data and features.

Overall Performance Table 6.3 shows the results on the test set. The full model consistently achieves the highest BLEU scores. Overall, these numbers suggest strong content selection by getting high recall for individual words (BLEU-1), but fall further behind human performance as the length of the n-gram grows (BLEU-2 through BLEU-4). These number match our perception that the model is learning to produce high quality sentences, but does not always describe all of the important aspects of the scene or use exactly the expected wording. Table 4.4 presents example output, which we will discuss in more detail shortly.

Human Evaluation Table 4.2 presents the results of a human evaluation. The full model outperforms all baselines on every measure, but is not always competitive with human descriptions. It performs the best on grammaticality, where it is judged to be as grammatical as humans. However, surprisingly, in many cases it is also often judged equal to the other baselines. Examination of baseline output reveals that the MT baseline often generates short sentences, having little chance of being judged ungrammatical. Furthermore,

Model	BL-1	BL-2	BL-3	BL-4	BP	SVO	S	V	O	PP	Pp	Po	N
Human	64.7	46.0	31.5	20.1	-	-	-	-	-	-	-	-	-
Full-Model	**59.0**	36.9	**19.3**	**10.5**	96.4	15.8	**64.9**	**40.4**	36.8	5.7	50.0	20.7	69.1
– doing	51.1	32.6	16.9	9.2	94.0	10.5	63.2	15.8	10.5	8.0	45.5	**21.6**	69.7
– count	55.4	33.5	16.0	8.5	96.0	17.5	59.6	35.1	15.4	**8.5**	**53.7**	19.5	66.7
– properties	57.8	**37.2**	18.8	10.0	94.2	**19.3**	61.4	36.8	36.8	8.0	47.1	20.7	**73.5**
– visual	56.7	35.1	18.9	9.4	**97.3**	**19.3**	**64.9**	36.8	**50.0**	5.1	41.8	15.3	71.6
– pairwise	56.9	35.5	16.5	8.2	96.1	12.3	**64.9**	**40.4**	45.5	7.1	42.4	21.2	70.9

Table 4.3: Ablation results on development data using BLEU1-4 and reporting match accuracy for sentence structures.

the Midge baseline, like our system, is a syntax-based system and therefore often produces grammatical sentences. Although our system performs well with respect to the baselines on truthfulness, often the system constructs sentences with incorrect prepositions, an issue that could be improved with better estimates of 3-d position in the image. On truthfulness, the MT baseline is comparable to our system, often being judged equal, because its output is short. Our system's strength is salience, a factor the baselines do not model.

Data Ablation Table 4.3 shows annotation ablation experiments on the development set, where we remove different classes of data labels to measure the performance that can be achieved with less visual information. In all cases, the overall behavior of the system varies, as it tries to learn to compensate for the missing information.

Ablating actions is by far the most detrimental. Overall BLEU score suffers and prediction accuracy of the verb (V) degrades significantly causing cascading errors that affect the object of the verb (O). Removing count information affects noun attachment (N) performance. Images where determiner use is important or where groups of objects are best identified by the number (for example, three dogs) are difficult to describe naturally. Finally, we see a tradeoff when removing properties. There is an increase in noun modifier accuracy (N) but a decrease in content selection quality (BL-1), showing recall has gone down. In essence, the approach just learns to stop trying to generate adjectives and other modifiers that would rely on the missing properties. The difference in BLEU score with the Full-Model is small, even without these modifiers, because there often still exists a a short output with high accuracy.

Feature Ablation The bottom two rows in Table 4.3 show ablations of the visual and pairwise features, measuring the contribution of the visual information provided by the bounding box annotations. The ablated visual information includes bounding-box positions and relative pairwise visual information. The pairwise

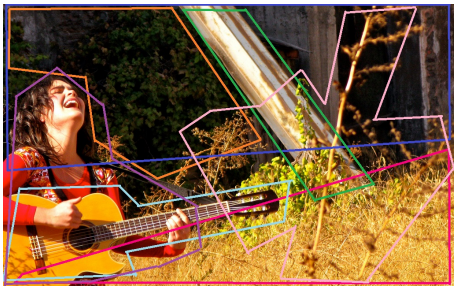
ablation removes the ability to model any interactions between objects, for example, relative bounding box or pairwise object type information.

Overall, prepositional phrase accuracy is most affected. Ablating visual features significantly impacts accuracy of prepositional phrases (Pp and Po), affecting the use of preposition words the most, and lowering fluency (BL-4). Precision in the object of the verb (O) rises; the model makes 50% fewer predictions in that position than the Full-Model because it lacks features to coordinate subject and object of the verb. Ablating pairwise features has similar results. While the model corrects errors in the object of the preposition (Po) with the addition of some visual features, fluency is still worse than the Full-Model, as reflected by BL-4.

Qualitative Results Table 4.4 has examples of good and bad system output. The first two images are good examples, including both system output (**S**) and a human reference (**R**). The second two contain lower quality outputs. Overall, the model captures common ways to refer to people and scenes. However, it does better for images with fewer sentient objects because content selection is less ambiguous.

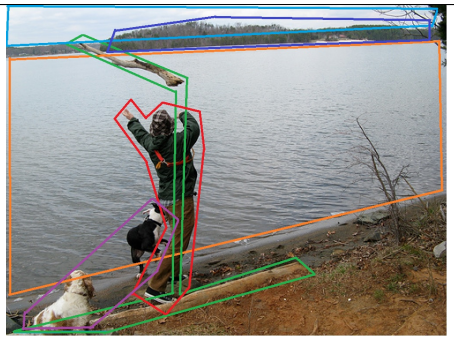
Our system does well at finding important objects. For example, in the first good image, we mention the guitar instead of the house, both of which are prominent and have high overlap with the woman. In the second case, we identify that both dogs and humans tend to be important actors in scenes but poorly identify their relationship.

The bad examples show difficult scenes. In the first description the broad context is not identified, instead focusing on the bench (highlighted in red). The second example identifies a weakness in our annotation: it encodes contradictory groupings of the people. The groupings covers all of the children, including the boy running, and many subsets of the people near the grass. This causes significant ambiguity and our methods cannot differentiate them, incorrectly choosing to mention just the children and picking an inappropriate verb (one participant in the group is not sitting). Improved annotation of groups would enable the study of generation for more complex scenes, such as these.



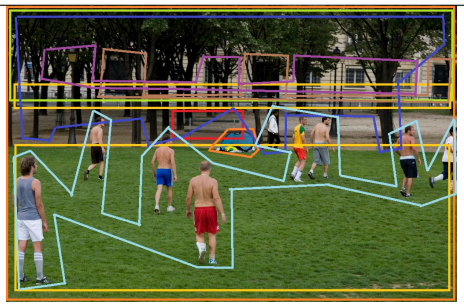
S: A girl playing a guitar in the grass

R: A woman with a nylon stringed guitar is playing in a field



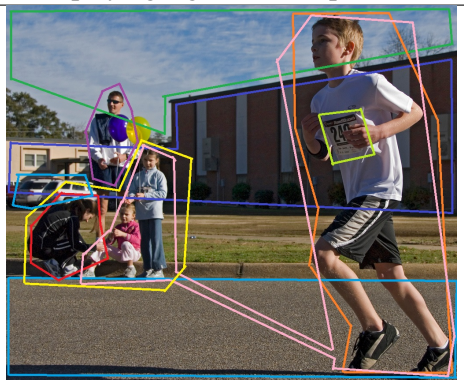
S: A man playing with two dogs in the water

R: A man is throwing a log into a waterway while two dog watch



S: Two men playing with a bench in the grass

R: Nine men are playing a game in the park, shirts versus skins



S: Three kids sitting on a road

R: A boy runs in a race while onlookers watch

Table 4.4: Two good examples of output (top), and two examples of poor performance (bottom). Each image has two captions, the system output **S** and a human reference **R**.

Chapter 5

Common Sense

Obtaining common sense knowledge using current information extraction techniques is extremely challenging. In this work, we instead propose to derive simple common sense statements from fully annotated object detection corpora such as the Microsoft Common Objects in Context dataset. We show that many thousands of common sense facts can be extracted from such corpora at high quality. Furthermore, using WordNet and a novel sub-modular k-coverage formulation, we are able to generalize our initial set of common sense assertions to unseen objects and uncover over 400k potentially useful facts.

5.1 Introduction

How can we discover that bowls can hold broccoli, that if a knife touches a cake then a person is probably cutting cake, or that cutlery can be on dining tables? We propose to leverage the effort of computer vision researchers in creating large scale datasets for object detection and use these resources instead to extract symbolic representations of visual common sense. The knowledge we compile is physical, not commonly covered in text and more exhaustive than what people can usually produce.

Our focus is particularly on visual common sense, defined as the information about spatial and functional properties of entities in the world. We propose to extract three types of knowledge from the Microsoft Common Objects in Context dataset [95] (MS-COCO), consisting of 300,000 images, covering 80 objects, with object segments and natural language captions. First, we find spatial relations, e.g. *holds*(bed, dog), from outlines of co-occurring objects. Next, we construct entailment rules like *holds*(bed, dog) $\Rightarrow$ *laying-*

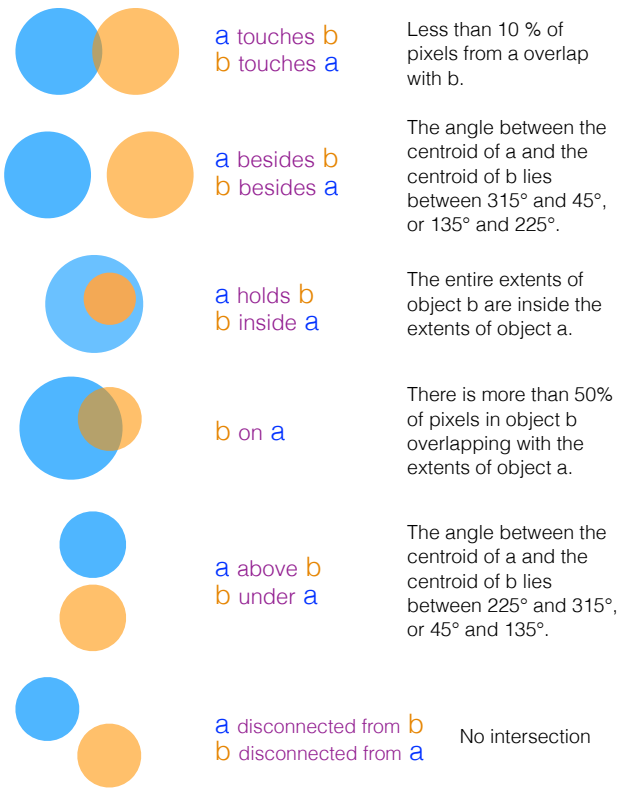


Figure 5.1: We define 6 types of unique relationships: {touches, above, besides, holds, on, disconnected}.

on(dog, bed) by associating spatial relations with text in captions. Finally, we uncover general facts such as *holds*(furniture, domestic animal), applicable to object types not present in MS-COCO by using WordNet [106] and a novel submodular k-coverage formulation.

Evaluations using crowdsourcing show our methods can discover many thousands of high quality explicit statements of visual common sense. While some of this knowledge can be potentially extracted from text [143], we found that from our top 100 extracted spatial relations, e.g. *holds*(bed, dog), only 4 are present in some form in the *AtLocation* relations in the popular ConceptNet [139] knowledge base. This shows that the knowledge we derive provides complimentary information for other more general knowledge bases. Such common sense facts have proved useful for query expansion [78; 9] and could benefit entailment [25], grounded entailment [10], or visual recognition tasks [161].

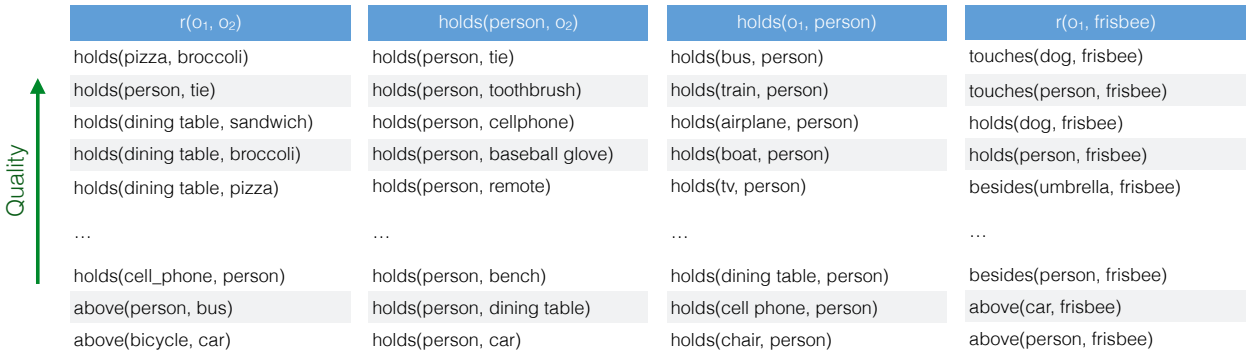


Figure 5.2: Example of our extracted object-object relations. The first column contains the overall 3 best and worst relations ranked by PMI, the following columns show similar results for the queries: what does a person hold? what holds a person?, and what interacts with a frisbee?

5.2 Methods

We assume the availability of an object-level annotated image dataset D containing a set of images with textual descriptions. Each object in an image must be annotated with: (1) a mask or polygon outlining the extents of the object, and (2) the category of the object from a set of categories V and (3) an overall description of the image.

We produce three types of common sense facts, each with an associated scoring function: (1) Object-object relationships implicitly encoded in the relative configurations between objects in the annotated image data, i.e. *on*(bed, dog) (sec 5.2.1) , (2) Entailment relations encoded in the relationships between object-object configurations and textual descriptions i.e. *on*(bed, dog) $\Rightarrow$ *laying-on*(bed, dog) (sec 5.2.2), and (3) Generalized relations induced by using the semantic hierarchy of concepts in WordNet, i.e. *on*(furniture , domestic-animal) (sec 5.2.3).

5.2.1 Mining Object-Object Relations

Our objective in this section is to score and rank a set of relations $S_1 = \{r(o_1, o_2)\}$, where r is a object-object relation and $o_1, o_2 \in V$, using a function $\gamma_1 : S_1 \rightarrow \mathbb{R}$. First, we define a vocabulary R of object-object relations between pairs of annotated objects. Our relations are inspired by Region Connection Calculus [121], and the Visual Dependency Grammar of [38; 36], details in Figure 5.1.

For every image, we record the instances of each of these object-object relations $r(o_1, o_2)$ between all

co-occurring objects in D[1]. We use Point-wise Mutual Information (PMI) to estimate the evidence for each relationship triplet:

$$\gamma_1(r(o_1, o_2)) = log \frac{p[r(o_1, o_2)]}{p[r]p[(o_1, o_2)]}, \tag{5.1}$$

We estimate these probabilities by counting object-object-relation co-ocurrences using existential quantifiers for every image. This means every image can at most contribute one to the count of $r(o_1, o_2)$ so that we do not exacerbate the results by images with many identical object types taken from unusual viewpoints. In Figure 5.2, we provide examples of our extracted object-object relations.

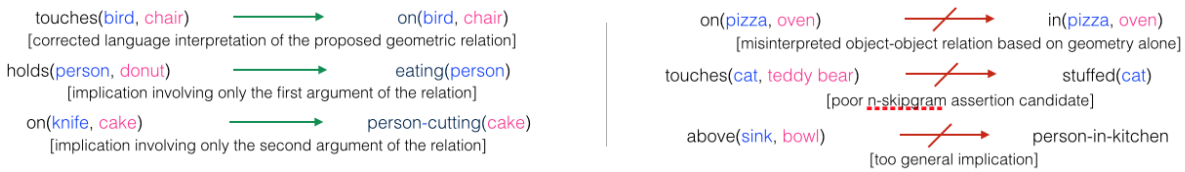


Figure 5.3: Left: correctly identified entailment relations and right: failure cases.

5.2.2 Mining Entailment Relations

In this section we combine our relation-based tuples mined from visual annotations (section **??**) with more than 400k textual descriptions included in MS-COCO. We generate a set of entailments $S_2 = \{r(o_1, o_2) \Rightarrow z\}$, where $r(o_1, o_2)$ is an element from S_1 and z is a consequent obtained from textual descriptions. Similarly as in the previous section, we rank the relations in S_2 using a function $\gamma_2 : S_2 \rightarrow \mathbb{R}$.

We start by generating an exhaustive list of candidate consequents z. We first pre-process the image captions with the part-of-speech tagger and lemmatizer from the Stanford Core NLP toolkit [101], and remove stop words. Then we generate a list of n-length skipgrams in each caption. The set of n-skipgrams are filtered based on predefined lexical patterns[2], and redundancies are removed[3]. Skipgrams, z, are then paired with co-occurring relations, $r(o_1, o_2)$, removing pairs with the disconnected-from spatial relation (see Figure 5.1). Pairs are scored with the conditional probability:

$$\gamma_2(r(o_1, o_2) \Rightarrow z) = \frac{P[z, r(o_1, o_2)]}{P[r(o_1, o_2)]} \tag{5.2}$$

[1] For symmetric relations like *above*(o_1, o_2), and *under*(o_1, o_2) we only record one of the relations.
[2] ⟨noun, verb⟩, ⟨noun,*, verb,*, noun⟩, ⟨noun,*, preposition, *, noun⟩, ⟨noun,*, verb, preposition,*,noun⟩
[3] ⟨noun,*, verb, *, noun⟩ are collapsed to ⟨noun,*, verb, preposition,*,noun⟩.

The consequent z can take the form q, $q(o_1)$, $q(o_2)$, or $q(o_1, o_2)$, by performing a simple alignment with the arguments in the antecedent. We perform this alignment by mapping the object categories in the antecedent $r(o_1, o_2)$ to WordNet synsets, and matching any word in z to any word in the gloss set of the predicate arguments o_1 and o_2. The unmatched words in z form the relation, whereas matched words form arguments. We produce the form q if there are no matches, $q(o_1)$, or $q(o_2)$ when one argument word matches, and $q(o_1, o_2)$ when both match. Examples of discovered entailments are in Figure 5.3.

5.2.3 Generalizing Relations using WordNet

In this section we present an approach to generalize an initial set of relations, S, to objects not found in the original vocabulary V. Using WordNet we construct a superset G containing all possible parent relations for the relations in S by replacing their arguments o_1, o_2 by all their possible hypernyms. Our objective is to select a subset T from G that contains high quality and diverse generalized relations. Note that elements in G can be too general and contradict statements in S while others could be correct but add little new knowledge. To balance these concerns, we formulate the selection as an optimization problem by maximizing a fitness function $\mathcal{L}$:

$$\max_T \mathcal{L}(T), \text{ such that } |T| = k, \text{ and } T \subseteq G, \tag{5.3}$$

$$\mathcal{L}(T) = \lambda \log(1 + \psi(T)) + \sum_{t \in T} \log(1 + \phi(t, S)), \tag{5.4}$$

where ψ is a *coverage* term that computes the total number of facts implied through hyponym relationships by the elements in T. The second term ϕ is a *consistency* term that measures the compatibility of a generalized relation t with the relations in S. We assume that if a relation is missing from S, then it is false (this corresponds to a closed world assumption over the domain of S). Thus, ϕ is the ratio of the scores of relations in S consistent with relation t (i.e. evidence for t based on S), and a value that is proportional to the number of missing relations from S (i.e. the amount of counter-evidence). More concretely:

$$\phi(t, S) = \frac{\sum_{s:t \Rightarrow s \wedge s \in S} \gamma(s)}{\mu \cdot (1 + \sum_{s:t \Rightarrow s \wedge s \notin S} 1) \cdot d(t, S)}, \tag{5.5}$$

where μ is a constant and d is the product of the WordNet distances of the synsets involved in t to their

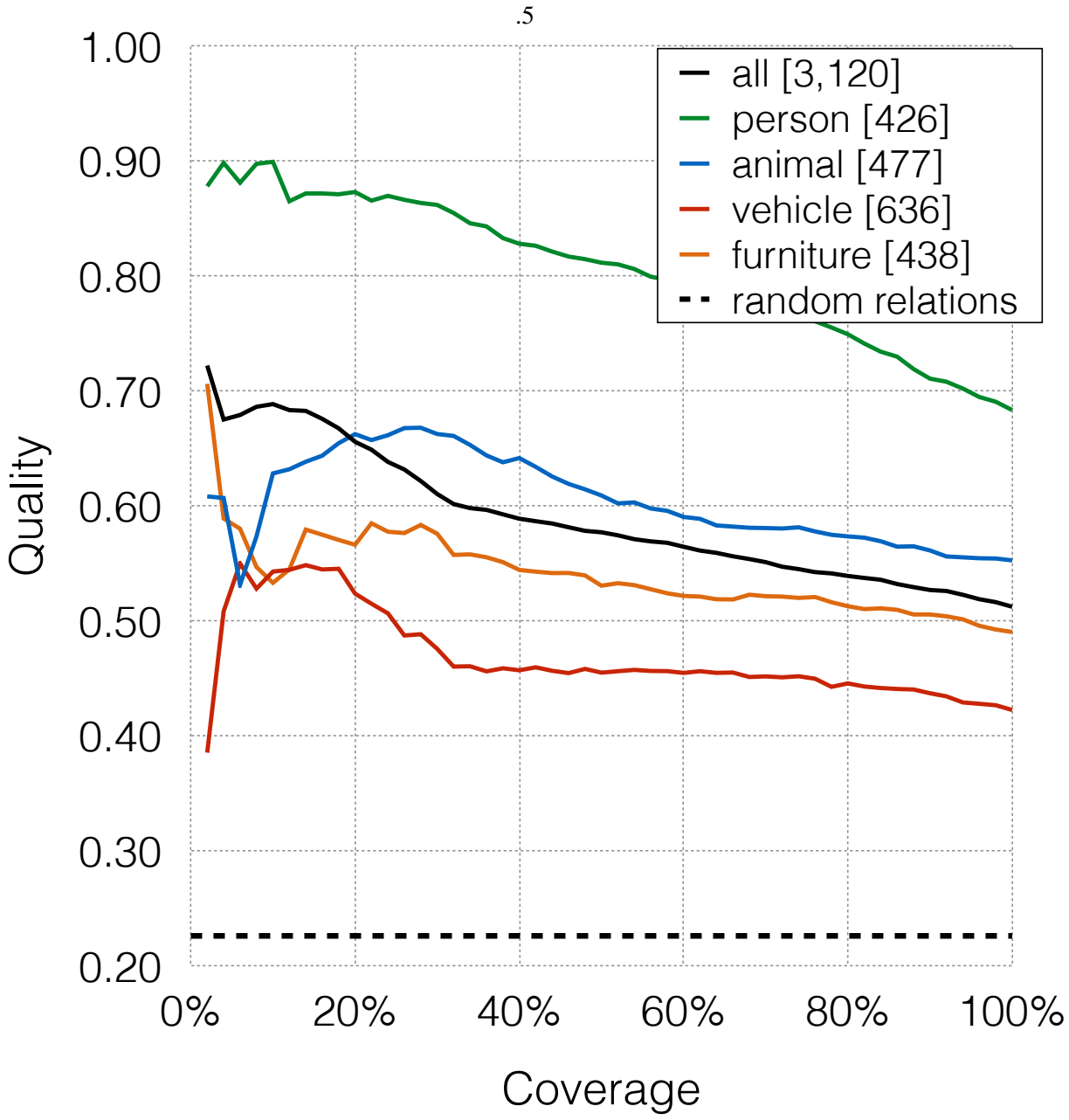


Figure 5.4: Object-object relations; Quality of extracted common sense, as judged by people. Legends show total relations covered at 100% for a few high level types in MS-COCO.

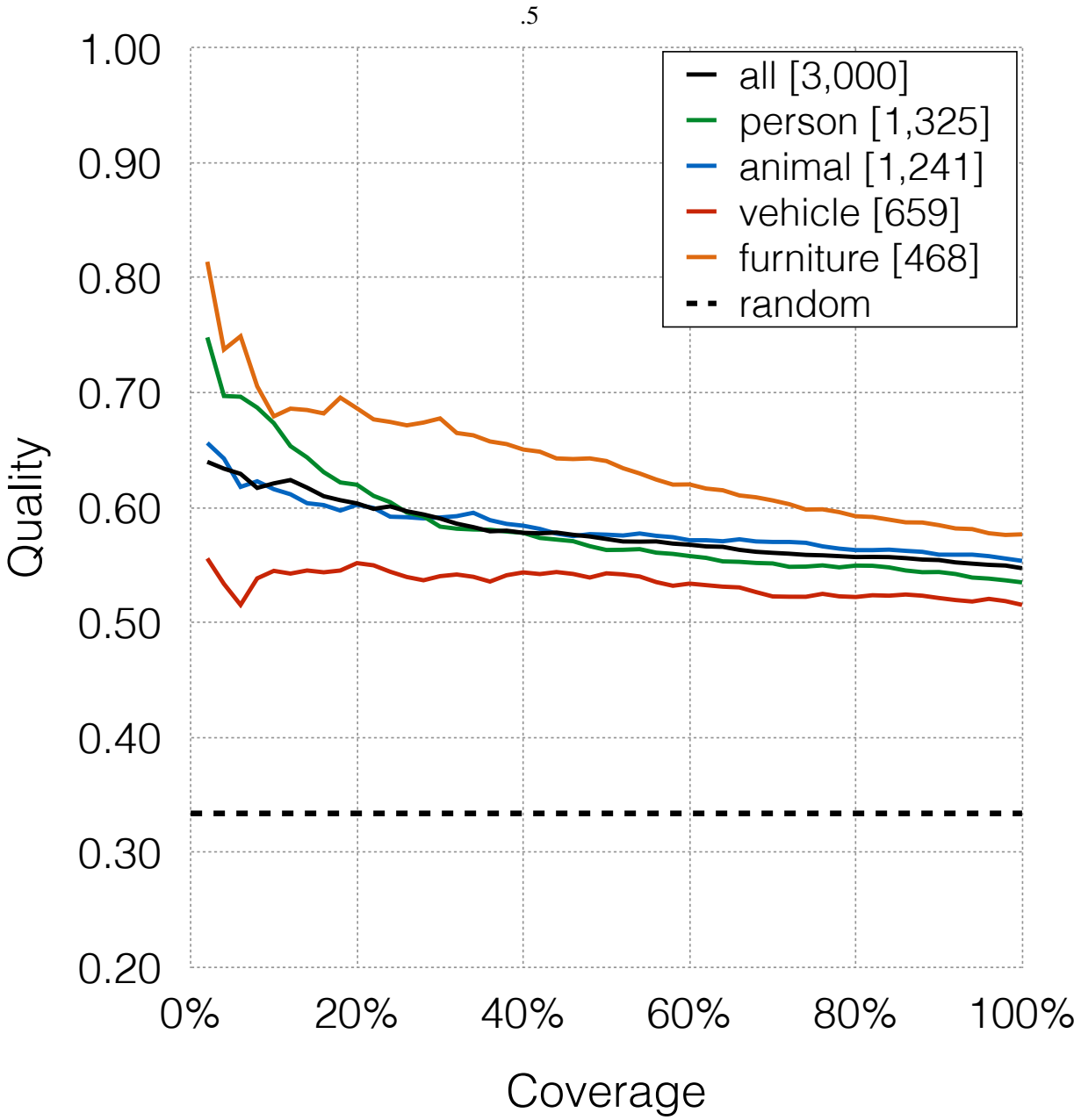


Figure 5.5: Entailment relations; Quality of extracted common sense, as judged by people. Legends show total relations covered at 100% for a few high level types in MS-COCO.

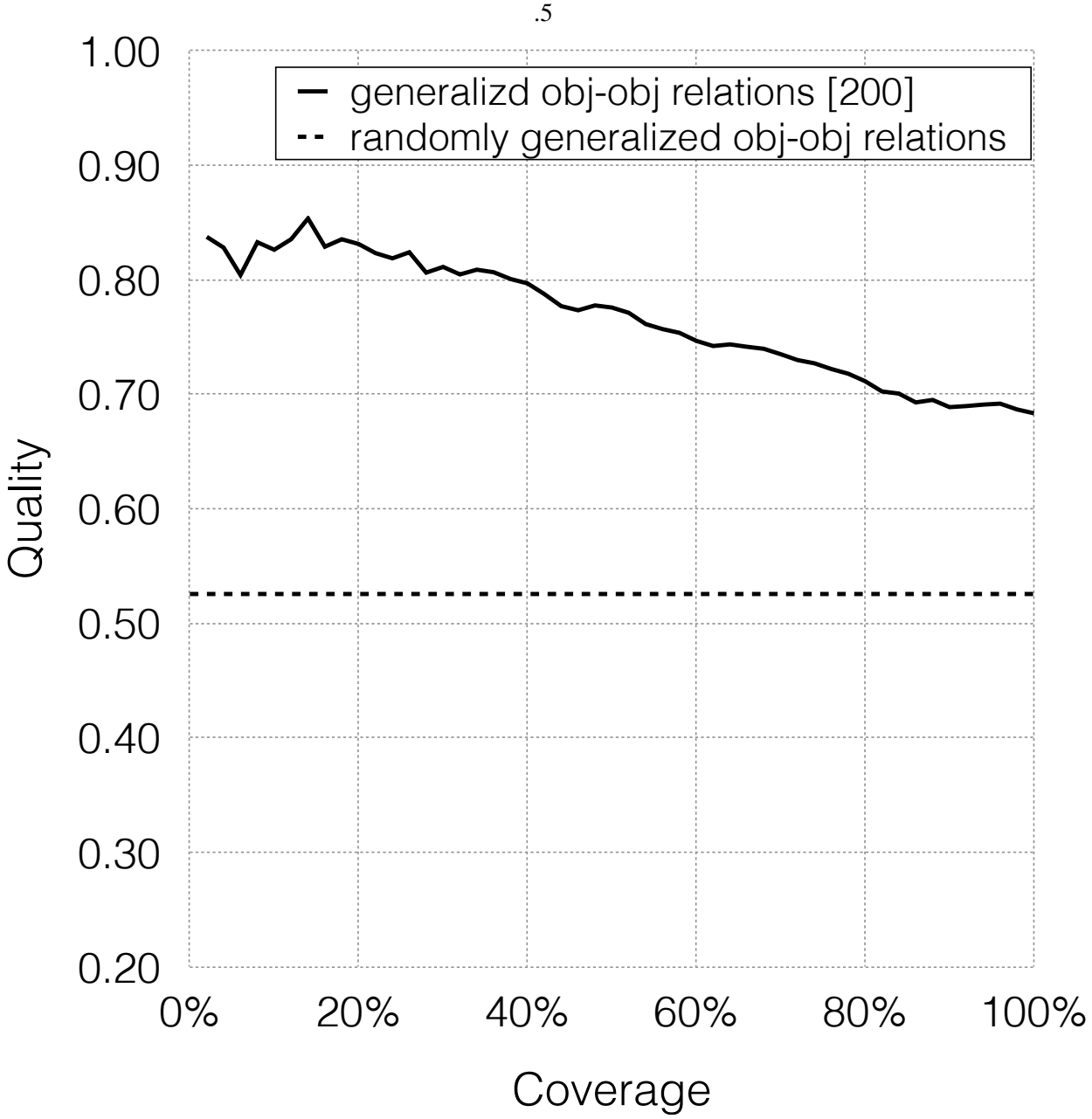


Figure 5.6: Generalizations; Quality of extracted common sense, as judged by people. Legends show total relations covered at 100% for a few high level types in MS-COCO.

nearest synset in S. This penalizes relations that are far away from categories in S. The optimization defined in Equation 5.3 is an instance of the submodular k-coverage problem. We use a greedy algorithm that adds elements to T that maximize $\mathcal{L}$, which due to the submodular nature of the problem approximates the solution up to a constant factor.

5.3 Experimental Setup

Object-Object Relations: We filter out from the initial set of candidate relations the ones that occur less than 20 times. We extract more than 3.1k unique statements (6k including symmetric spatial relations). *Entailment Relations*: We use skipgrams of length 2-6 allowing at most 6 skips, filter candidates such that they occur at least 5 times, and return the top 10 most likely entailments per spatial relation. Overall, 6.3k unique statements are extracted (10k including symmetric relations). *Generalized Relations*: We optimize Equation 5.4 only for object-object relations because the closed world assumption makes counts for implications sparse. The parameter μ is set to the average of the scores, $\lambda = 0.05$ and $k = 200$.

5.4 Evaluation

We evaluated the quality of the common sense we derive on Amazon Mechanical Turk. Annotators are presented with possible facts and asked to grade statements on a five point scale. Each fact was evaluated by 10 workers and we normalize their average responses to a scale from 0 to 1. Figure **??** shows plots of quality vs. coverage, where coverage means the top percent of relations sorted by our predicted quality scores.

Object-Object Relations As a baseline, 1000 randomly sampled relations have a quality of 0.225. Figure 5.4 shows our PMI measure ranks many high quality facts at the top, with the top quintile of the ranking being rated above 0.63 in quality. Facts about persons are higher quality, likely because this category is in over 50% of the images in MS-COCO.

Entailment Relations Turkers were instructed to assign the lowest score when they could not understand the consequent of the entailment relation. As a baseline, 1000 randomly sampled implications that meet our patterns have a quality of 0.33. Figure 5.5 shows that extracting high quality entailment is harder than

object-object relations likely because supposition and consequent need to coordinate. Relations involving furniture are rated higher and manual inspection revealed that many relations about furniture imply stative verbs or spatial terms.

Generalized Relations To evaluate generalizations, Figure 5.6, we also present users with definitions[4]. As a baseline, 200 randomly sampled generalizations from our 3k object-object relations have a quality of 0.53. Generalizations we find are high quality and cover over 400k objects facts not present in MS-COCO. Examples from the 200 we derive include: *holds*(dining-table, cutlery), *holds*(bowl, edible fruit) or *on*(domestic animal, bed).

[4]sometimes rules involve abstract concepts, for example *vessel*, any object that can be used as a container

Chapter 6

Situation Recognition

This paper introduces situation recognition, the problem of producing a concise summary of the situation an image depicts including: (1) the main activity (e.g., clipping), (2) the participating actors, objects, substances, and locations (e.g., man, shears, sheep, wool, and field) and most importantly (3) the roles these participants play in the activity (e.g., the man is clipping, the shears are his tool, the wool is being clipped from the sheep, and the clipping is in a field). We use FrameNet, a verb and role lexicon developed by linguists, to define a large space of possible situations and collect a large-scale dataset containing over 500 activities, 1,700 roles, 11,000 objects, 125,000 images, and 200,000 unique situations. We also introduce structured prediction baselines and show that, in activity-centric images, situation-driven prediction of objects and activities outperforms independent object and activity recognition.

6.1 Introduction

When we look at an image, we instantly and effortlessly recognize not only what is happening (e.g., clipping) but who and what is involved (e.g., a person, shears, a sheep, wool) and how these entities relate to each other, the *roles* that they play (e.g., the person does the clipping, the shears are the clipping tool, and the wool is being clipped from the sheep). In this paper, we argue for explicitly encoding such semantic roles, a key missing ingredient in current paradigms of recognition, in image understanding. We introduce *situation recognition*, a problem that involves predicting activities along with actors, objects, substances, and locations and how these pieces fit together (semantic roles). For example, the leftmost table in Figure 1 shows one

CLIPPING

ROLE	VALUE
AGENT	MAN
SOURCE	SHEEP
TOOL	SHEARS
ITEM	WOOL
PLACE	FIELD

ROLE	VALUE
AGENT	VET
SOURCE	DOG
TOOL	CLIPPER
ITEM	CLAW
PLACE	ROOM

JUMPING

ROLE	VALUE
AGENT	BOY
SOURCE	CLIFF
OBSTACLE	-
DESTINATION	WATER
PLACE	LAKE

ROLE	VALUE
AGENT	BEAR
SOURCE	ICEBERG
OBSTACLE	WATER
DESTINATION	ICEBERG
PLACE	OUTDOOR

SPRAYING

ROLE	VALUE
AGENT	MAN
SOURCE	SPRAY CAN
SUBSTANCE	PAINT
DESTINATION	WALL
PLACE	ALLEYWAY

ROLE	VALUE
AGENT	FIREMAN
SOURCE	HOSE
SUBSTANCE	WATER
DESTINATION	FIRE
PLACE	OUTSIDE

Figure 6.1: Six images that depict situations where actors, objects, substances, and locations play roles in an activity. Below each image is a *realized frame* that summarizes the situation: the left columns (blue) list activity-specific roles (derived from FrameNet, a broad coverage verb lexicon) while the right columns (green) list values (from ImageNet) for each role. Three different activities are shown, highlighting that visual properties can vary widely between role values (e.g., clipping a sheep's wool looks very different from clipping a dog's nails).

such representation: a situation where a man (`agent`) is clipping (`activity`) wool (`item`) from a sheep (`source`) using shears (`tool`) in a field (`place`).

Situation recognition generalizes activity recognition and human-object interaction, using the assignment of roles to define how actors, objects, substances, and locations participate in activities. For example, Figure 6.1 has image pairs that depict the same overall activity but look very different when the participating entities change for the different roles. Previous work has presented models for some aspects of a complete situation, including activity scene models [103] and models of very specific activities paired with a few prototypical objects, such as playing a musical instrument [149]. However, our formulation provides a more complete representation of the different roles that each of the participants can play, and allows us to scale to hundreds of different activities. In essence, we are building representations that support the understanding not just of "What is happening?" but also "Who is doing it?" (the `agent` role), "What are they doing it to?" (`patient`), "What are they doing it with?" (`tool`), "Where did it start?" (`source`), and so on, as appropriate for each activity.

It is difficult to know a priori what roles entities can play in each activity. However, we can draw inspiration from the way verbs are used in the English language by building on FrameNet [53], a linguist-authored verb lexicon. FrameNet pairs every verb with a *frame*, which specifies a set of *semantic roles*. Semantic roles categorize how objects can participate in the activity described by a verb. For example,

the two rightmost images of Figure 6.1 show frames for spraying, which includes semantic roles such as `agent` and `destination`. Such frames have been used to build semantic parsers that match verbs to their arguments in English sentences, for example see [26]. However, here we instead use them to define the space of possible situations, much like how WordNet [106] was used to define ImageNet [126] object classes. For each frame, the verb defines an activity label, and the semantic roles specify how WordNet entities participate in the activity. For example, Figure 6.1 shows situations where the FrameNet verb `spraying` has a semantic role `tool` that is filled with WordNet synsets such as `spray can` or `hose`.

To demonstrate the generality of the situation recognition task, we introduce *imSitu*, a collection of over 125,000 images depicting 200,000 distinct situations. Each situation includes one of 500 possible activities and values for up to 6 activity-specific roles (3.5 on average and 1,700 unique roles in total with 190 types). The images were gathered from Google image search with query expansion techniques and labeled with complete situations on Amazon Mechanical Turk. The annotators specified one of 80,000 possible WordNet sysnets for each role, providing over 11,000 unique values for this image collection. In addition to being large scale, this data is also high quality. For example, even though the space of possible values is very large, 2 out of 3 annotators provided the same synset for over 75% of roles. Sections 6.3 and 6.4 provide the full details of the data collection and statistics.

To support future work on the imSitu data, we provide results for a baseline model — a Conditional Random Field (CRF) which includes CNN [136] features (fine tuned by backpropagating the CRF error). This approach significantly outperforms a 5000-way classifier that predicts one of the 10 most frequent situations per verb. The CRF achieves 32.3% top-1 and 58.9% top-5 accuracy for activity prediction and predicts entire situations correctly 14.2% of the time. When compared to independent models trained on the same activity-centric data, the approach improves top-1 accuracy for object recognition by 8.6% and top-1 activity recognition by 1.2%, demonstrating that the model benefits significantly from the context that is provided by jointly predicting the full situation. These results suggest that situation recognition with the *imSitu* dataset has the potential to become a strong benchmark for the study of objects, activities, and their interactions through semantic roles.

6.2 Formal Task Definition

In situation recognition, we assume discrete sets of verbs V, nouns N, and frames F. Each frame $f \in F$ is paired with a discrete set of semantic roles E_f. For example, Figure 6.1 shows six different situations, representing the verbs `clipping`, `jumping`, and `spraying`. While some semantic roles, e.g. `agent`, are shared across all three frames, others (e.g., `tool`) only appear for some. Additionally, each semantic role $e \in E_f$ is paired with a noun value $n_e \in N \cup \{\varnothing\}$, where $\varnothing$ indicates the value is either not known or does not apply. For example, in the first image in Figure 6.1, the semantic role `item` takes the value `wool`. In this paper, the verb set V and frame set F are derived from FrameNet, while the noun set N is drawn from WordNet. We refer to the set of pairs of semantic roles and their values as a realized frame, $R_f = \{(e, n_e) : e \in E_f\}$. In the third image of Figure 6.1, R_f = {(`agent`,`boy`), (`source`,`cliff`), (`obstacle`,$\varnothing$), (`destination`,`water`), (`place`,`lake`) }. Finally, a realized frame is valid if and only if each value $e \in E_f$ is assigned exactly one noun n_e.

Now, given an image, our task is to predict a situation, $S = (v, R_f)$, specified by a verb $v \in V$ and a valid realized frame R_f. For example, in the last image of Figure 6.1, the predicted situations is $S =$ (`spraying`, {(`agent`,`fireman`), (`source`,`hose`), (`substance`,`water`), (`destination`,`fire`), (`place`,`outside`)}).

6.3 Dataset Collection

We introduce imSitu, a dataset of images labeled with situations. Our annotation approach is scalable, the image labeling is done on Mechanical Turk and covers over 500 verbs with 125,000 images, and is relatively affordable, annotation cost approximately $80 per verb.

6.3.1 Filtering and Labeling FrameNet

FrameNet is a rich resource that pairs verbs with frames and semantic roles. It is designed to cover, as much as possible, all English verbs and all roles they can take, not just those that can be visually recognized in an image. For example, it would include verbs such as `attempt` with roles such as `goal` that take other verbs as arguments. To define our recognition task, we manually filtered FrameNet to find verbs and roles that could be reliably recognized in images, and provide English labels for use in the crowdsourcing interface.

This was done by a small set of trusted annotators.

Finding Visual Verbs and Roles We gathered 9683 candidate verbs and asked annotators to determine if they could be reliably recognized in images, and, if so, to provide a support image.[1] Verbs that were not recognizable generally fell into one of a few classes, including: (a) abstract, such as "presuming," (b) representational, such as "thinking," where we could find a supporting image evocative of the verb but did not depict it literally happening (c) technical, including "blanching," where crowd workers were unlikely to know the word's meaning, or otherwise just (d) hard, including "insufflating," where the annotator does not know the word or what it would look like. Annotators were first calibrated to confirm they understood these categories and confusing cases were publicly discussed. In total, 1053 verbs (10.9%) were marked as visually recognizable. To find visual roles, annotators were shown visual verbs and their example images and asked to select the subset of visually recognizable semantic roles, a generally easier task.

Labeling Verbs and Roles To support later crowd sourcing, the annotators also provided simple English descriptions of the visual verbs and roles. They wrote a single sentence that summarizes all of the roles for each verb. For example, for the verb clipping in Figure 6.1, the sentence would be "An AGENT clips an ITEM from a SOURCE using a TOOL in a PLACE." This sentence was shown to crowd workers to define the roles that each verb supports.

Example Creation Finally, to help crowd workers understand how to produce situation annotations, a few example image labels were produced for each verb. Five computer science undergraduates read definitions for all 1053 candidate verbs and retrieved three images that correspond to each verb from Google Image Search. If the annotators were unable to find such images, the verb was removed. Overall, 580 verbs passed this filtering stage.

6.3.2 Image Annotation

The final image annotations were gathered on Amazon Mechanical Turk in a two-stage process, that involved first filtering automatically collected images and then filling in the role values for target frames.

[1] We extended the nearly 5,000 verbs in FrameNet to include additional verbs from PropBank [75], a closely related verb lexicon, that were mapped to FrameNet as part of the SemLink project [117].

Figure 6.2: A word cloud of verbs in imSitu where larger words have a larger rate of unseen value-role combinations. Verbs with a low rate, e.g. "flossing" (0.7%) have specific meaning as compared to verbs such as "putting" (15.5%) or "biting" (7.7%).

Candidate Image Filtering Candidate images were retrieved by searching for phrases related to a target activity in Google Image Search. Phrases were mined from a subset of Google Syntactic N-Grams [63] that focuses on verb-argument structure. The phrases we extracted contain the target verb and include all descendants of the verb in a syntactic parse. We selected 450 such phrases, picking the most frequent 150 that contain "n-subj," ,"d-obj", or "p-obj" dependencies. For example,"cutting" would have the p-obj "scissors." Using dependencies guarantees that the queried words occur in different syntactic positions relative to the target verb. We retrieved 200 full-color medium-sized images that pass safe search and consider all returned images as candidates. Workers were instructed to select images that contain the desired activity and (1) are not modified or computer generated and (2) contain at least some part of the main entity doing the action in the image.

Value Filling Selected images were next presented for value filling. Workers were shown a definition of the target verb, a sentence summarizing the semantic roles associated with verb and example images of realized frames for that verb. They were asked to chose a category from an auto-complete drop-down menu, that also presents synset definitions, to fill slots; to select the most specific WordNet synset, and if more than one could apply, select the most relevant. For groups, they were asked to either find a word that refers to the group (for example, "people," "couple") or simply use the singular ("person"). They were required to annotate at least one value per image and not to fill in values that could not be reasonably inferred from the image.

6.3.3 Diversity and Coverage

The goal of imSitu is to include as many verbs as possible and have samples for all unique combinations of semantic roles and values. This is challenging because situations are structured and there can be a combinatorial number of possible realized frames. We adopted a dynamic strategy to increase diversity while not wasting money on verbs where we have already seen most combinations. First, candidate images from Google Image Search were presented for filtering by uniformly drawing images from query phrases, thus maximizing the diversity of types of images. 200 images were annotated in this way with full structures, providing a lower bound on the number of images per verb in imSitu. Then, we dynamically decided whether to continue to collect more annotations.

The rate at which unseen combinations occur can be approximated by splitting the data into a train and test set and computing how often a value appears in a semantic role in the test set but never appeared in train set. We refer to this as the out of vocabulary (OOV) rate of a verb, and compute it by averaging 1000 random splits of the data. Figure 6.2 visualizes the current OOV rate for a sample of verbs currently in imSitu. If during the collection process the OOV rate of verb was greater than 5%, we continued to collect images, up to a maximum of 400 images. While for some verbs this significantly improved the OOV rate, other verbs will always have a high rate. For example, despite collecting 400 images of the verb "making" and "putting," both have an OOV rate of 15%. This is a fundamental challenge in situation recognition. On the other hand, "baptizing" has an OOV of zero with just 200 image samples. The final global OOV rate in imSitu is 3.5%.

6.3.4 Cost

During the collection process, every verb had a hard constraint of costing no more than $120 and was discontinued when it exceeded this amount. The largest contributor to the cost of collecting imSitu was the true positive rate of candidates retrieved from Google Image Search. Figure 6.3 shows the true positive rates for a sample of verbs currently in imSitu. Over 25% of verbs were cost prohibitive to collect directly from Google Image Search results. In cases when we were able to collect at least 50 images but exceeded a cost threshold before collecting 200 images, we made a second effort. A new set of queries for Google Image Search was constructed by pairing the verb with a noun that occurred in an annotated frame. The returned images were used to reseed the filter phase of our annotation. This second round allowed us to reduce the percentage of failed verbs to 13%. Overall, failed verbs contributed $7 to the cost of annotating each verb.

6.4 Dataset Statistics

Table 6.1 provides summary statistics about imSitu, collected as described in the last section. In this section, we summarize the overall annotator agreement and highlight several interesting aspects of the data.

Agreement Quality control at scale is challenging. We used an automatic algorithm that discards annotations from workers that it estimates to be unreliable. The details are described in the supplementary

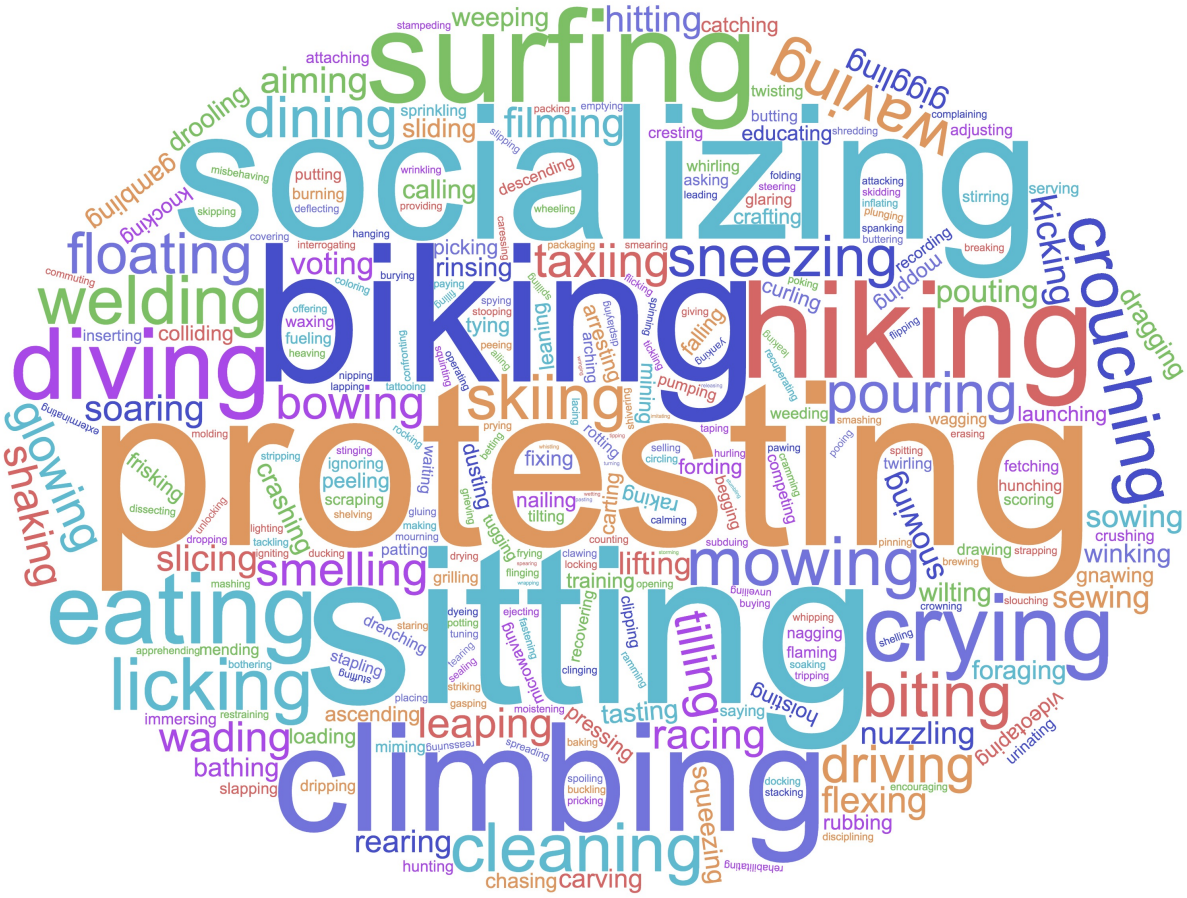


Figure 6.3: A word cloud of verbs in imSitu where larger words have a larger true positive rate for images retrieved from Google Image Search. Verbs with low rates, "fanning" (1%), were cost prohibitive to annotate. For all verbs, the average rate was 6.6%.

verbs	504
images	126,102
realized frames / image	3
total annotations	1,481,851
unique entities (>= 3)	11,538 (6794)
semantic roles / verb (range)	3.55 (1 - 6)
semantic roles (types)	1788 (190)
images / verb (range)	250.2 (200 - 400)
unique realized frames (>= 3)	205,095 (21,505)
out of vocabulary rate (range)	3.5% (0% - 15.8%)
train / dev / test	75,702 / 25,200 / 25,200

Table 6.1: Summary statistics of imSitu.

	Majority	1-link	2-link	3-link
all Roles	76.8	81.5	84.8	86.5
w/o Place	81.5	84.6	88.2	89.9

Table 6.2: Agreement statistics for situation role annotations in imSitu, with and without the Place role. Majority means that at least 2 of 3 Turker annotations agree. N-link means that a majority agree under the relaxed criteria of two synsets matching if they are within N links of each other in the WordNet hierarchy.

material.

All images were annotated by three crowd workers. We measure agreement by comparing the values that workers annotated for semantic roles. We say that two semantic role annotations on a single image agree when they indicate the same WordNet synset (or $\varnothing$). Furthermore, we compute a relaxed version of this criterion, allowing two annotations to match if the synsets are within 1, 2 or 3 links in the WordNet hierarchy. As a point of reference, the following synsets are all 3 links away from each other: "musical instrument" and "trumpet," "child" and "little girl," and "girl" and "person." Table 6.2 summarizes agreement in imSitu.

While the agreement numbers are very high, especially considering Turkers can select one of 80,000 values for each semantic role, there are systematic sources of ambiguity. `Place`, a semantic role present in all frames, is highly ambiguous because it can be identified in three ways: a close interacting object, (e.g., reading at a "desk"), an overall location type (e.g., reading in an "office") or a coarse identifier (e.g., reading "inside"). Table 6.2 demonstrates that `place` is indeed a major contributor to disagreement, accounting for over 25% cases where workers failed to produce a majority. This type of disagreement provides a number of alternative correct answers. Other sources of disagreement are described in the supplementary material.

Entity-Role Relations Figure 6.4 shows a uniform sample of nouns and the number of semantic roles they participate in. As expected there is a large variance; for example,"man" can take up to 798 roles while "basin" only takes 1 role. We also compute the inverse of these statistics: the number of nouns that a role can take, as shown in Figure 6.5.

Entity-Verb Relations Figure 6.6 shows the number of entities a sample of verbs can take. As expected, less structured verbs like "putting" have 653 entities and heavily structured verbs like "flossing" only take 42 nouns.

6.5 Structured Prediction of Frames

Our CRF for predicting a situation, $S = (v, R_f)$, given an image i, decomposes over the verb v and semantic role-value pairs (e, n_e) in the realized frame $R_f = \{(e, n_e) : e \in E_f\}$. The CRF parameters θ can be trained directly from our situation-labeled data. The full distribution, with potentials for verbs ψ_v and semantic roles ψ_e takes the form:

$$p(S|i;\theta) \propto \psi_v(v, i;\theta) \prod_{(e,n_e)\in R_f} \psi_e(v, e, n_e, i;\theta) \tag{6.1}$$

Computing the normalization is efficient: we can enumerate all valid verb-semantic role pairs and then for all pairs sum all possible semantic role values.

Each potential in the CRF is log linear:

$$\psi_v(v, i;\theta) = e^{\phi_v(v,i)\theta} \tag{6.2}$$

$$\psi_e(v, e, n_e, i;\theta) = e^{\phi_e(v,e,n_e,i)\theta} \tag{6.3}$$

where ϕ_e and ϕ_v encode scores from the output of a CNN. To learn this model, we assume that for an image i in dataset D there can, in general, be a set A_i of possible ground truth situations. We optimize the log-likelihood of observing at least one situation $S \in A_i$:

$$\sum_{i\in D} \log\left(1 - \prod_{S\in A_i} (1 - p(S|i;\theta))\right) \tag{6.4}$$

CRF Features In Equation 7.2 and 7.3 we introduce two feature functions that are implemented by adapting a neural network pretrained on the ImageNet Challenge [126]. We use VGG Large Network [136] in Caffe [71] with the final layers reduced to dimensionality 1024. The output of VGG is used as the input to a fully connected layer which predicts potential values in our CRF, similar to neural networks used for semantic role labeling in sentences [55]. At training time, we optimize Equation 7.4 with stochastic gradient ascent using a batch size of 192. We fine tune all layers of VGG for 30 epochs and reduce the initial learning rate of 1e-5 by a factor of ten for every ten epochs.

		top-1 predicted verb				top-5 predicted verbs				ground truth verbs		
		verb	value	value-any	value-full	verb	value	value-all	value-full	value	value-all	value-full
dev	Discrete Classifier	26.4	4.0	0.4	0.2	51.1	7.8	0.6	0.4	14.4	0.9	0.6
	CRF	32.2	24.6	14.3	11.2	58.6	42.7	22.7	17.5	65.9	29.5	22.3
test	Discrete Classifier	26.8	4.1	0.3	0.2	51.2	7.8	0.5	0.4	14.4	0.8	0.6
	CRF	32.3	24.6	14.2	11.2	58.9	42.8	22.5	17.5	65.7	29.0	22.0

Table 6.3: Situation prediction results in imSitu. Structured prediction outperforms classification of ten most common situations per activity.

6.6 Experiments

We present the first results for situation prediction in imSitu and also compare performance to baselines that independently recognize activities and objects.

6.6.1 Situation Recognition

Metrics We measure accuracy for different components of predicted situations. Because the evaluation data has situations provided by multiple annotators, we consider verb predictions (verb) and semantic role-value pair predictions (value) correct if they match any of the annotations. A realized frame is correct if it strictly matches all semantic role-value pairs provided by a single annotation (value-full) or if each pair matches at least one annotation (value-any). We also report accuracy with ground truth verbs.

Systems In addition to the CRF model described in Section 7.3, we also present a simple discrete classification baseline. The classifier selects one of the 10 most frequent realized frames for each verb seen in the training data, producing a 5040-class problem. For training, each realized frame is assigned as a positive example to the classifier output with the fewest number of differences. The classifier uses the same VGG features and fine tuning procedure as the CRF but with an initial learning rate of 1e-3.

Quantitative Results Table 6.3 summarizes our experiments on the imSitu development set. We also ran these experiments once on the imSitu test which confirms our development results. Overall, the CRF outperforms the discrete classifier by large margins. Verb accuracy is 32.5% and rises to 59% in the top-5. We can isolate the performance of assigning values to semantic roles by considering prediction accuracy given ground truth verbs. The discrete classifier is significantly worse in this context at value and full prediction because it cannot assign new combinations of entities to roles at test time.

		activity		object	
		top-1	top-5	top-1	top-5
dev	Activity	30.6	57.4	-	-
	Object	-	-	64.9	94.1
	Situation	32.25	58.6	72.9	95.0
test	Activity	31.1	57.7	-	-
	Object	-	-	64.1	94.2
	Situation	32.3	58.9	72.7	94.8

Table 6.4: Object and activity recognition results in imSitu. Joint prediction of object and activity through situation recognition improves over independently predicting either object or activity.

Qualitative Results Figure 6.7 shows a random selection of predictions from the CRF model on the development images where it predicted the correct verb. Over two thirds of the cases are correct or have only one incorrect role assignment. Furthermore, many of the errors are actually somewhat plausible. For example, the pole vaulter in the bottom right image is going over a horizontal pole. Other cases show similar reasonable errors, including confusing a cow with a horse, in the image second from the top and right.

6.6.2 Activity and Object Recognition

Metrics We evaluate activity and object recognition using top-1 and top-5 accuracy. For activity recognition, we treat the situation activity label as the gold standard. For object recognition, we assume any synset value annotated in a labeled frame is a gold standard object in the image.

Systems For activity recognition, we adapt our situation CRF by maximizing the potential in Equation 7.9 and predicting the corresponding verb. As a baseline, we train a discrete classifier for all verbs in imSitu, using VGG features and an identical fine tuning setup as the CRF but with an initial learning rate of 1e-3. For object recognition, we use our CRF to compute probability of observing any synset in the dataset by marginalizing Equation 7.9 over verbs and predicting the synset with the maximum marginal probability. As a baseline, we train an discrete classifier for all noun synsets in imSitu. We create psuedo-examples for every unique synset associated with an image and train the classifier on this expanded dataset, using identical training setup as the CRF but with an initial learning rate of 1e-3.

Quantitative Results Table 6.4 summarizes our experiments on the imSitu development set. We also ran these experiments once on the imSitu test data, which confirms our development results. Our situation CRF significantly outperforms predicting either activities or objects in isolation, by 1.2% and by 8.6% at top-1,

respectively. Overall, the results are encouraging; the context provided by situations is helping significantly, and improved models that more accurately reason about how objects interact with activities have significant potential to improve all three recognition tasks.

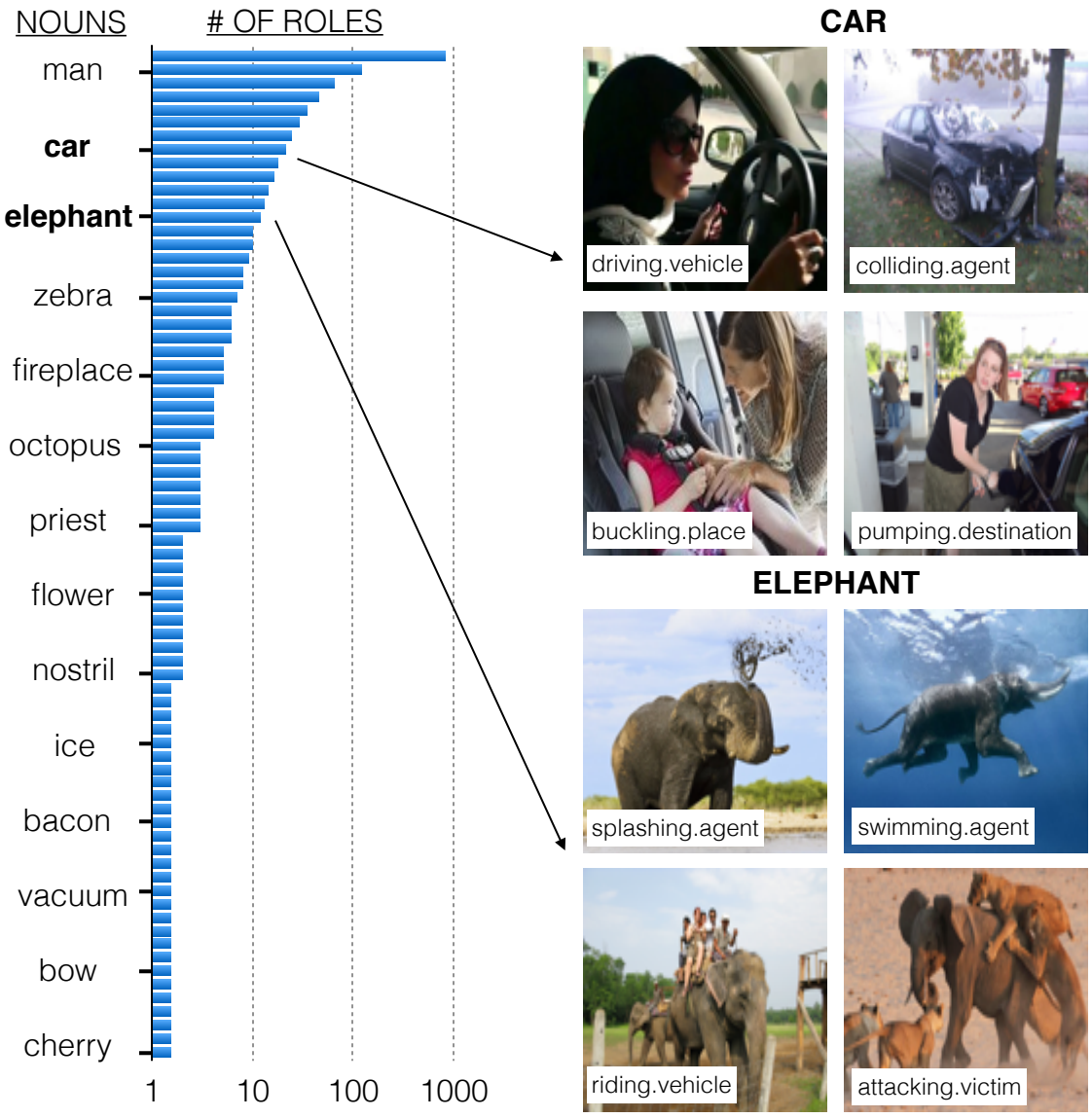


Figure 6.4: The number of semantic roles a noun can participate in, on a log-scale. 62% of nouns in imSitu appear with more than one semantic role. The most frequent noun, "man" appears in 44.6% of the roles. On the right are the different roles the nouns "car" and "elephant" paricipate in. Some roles can define particular viewpoints, such the role "place" being assigned "car" commonly indicates the interior view of the car.

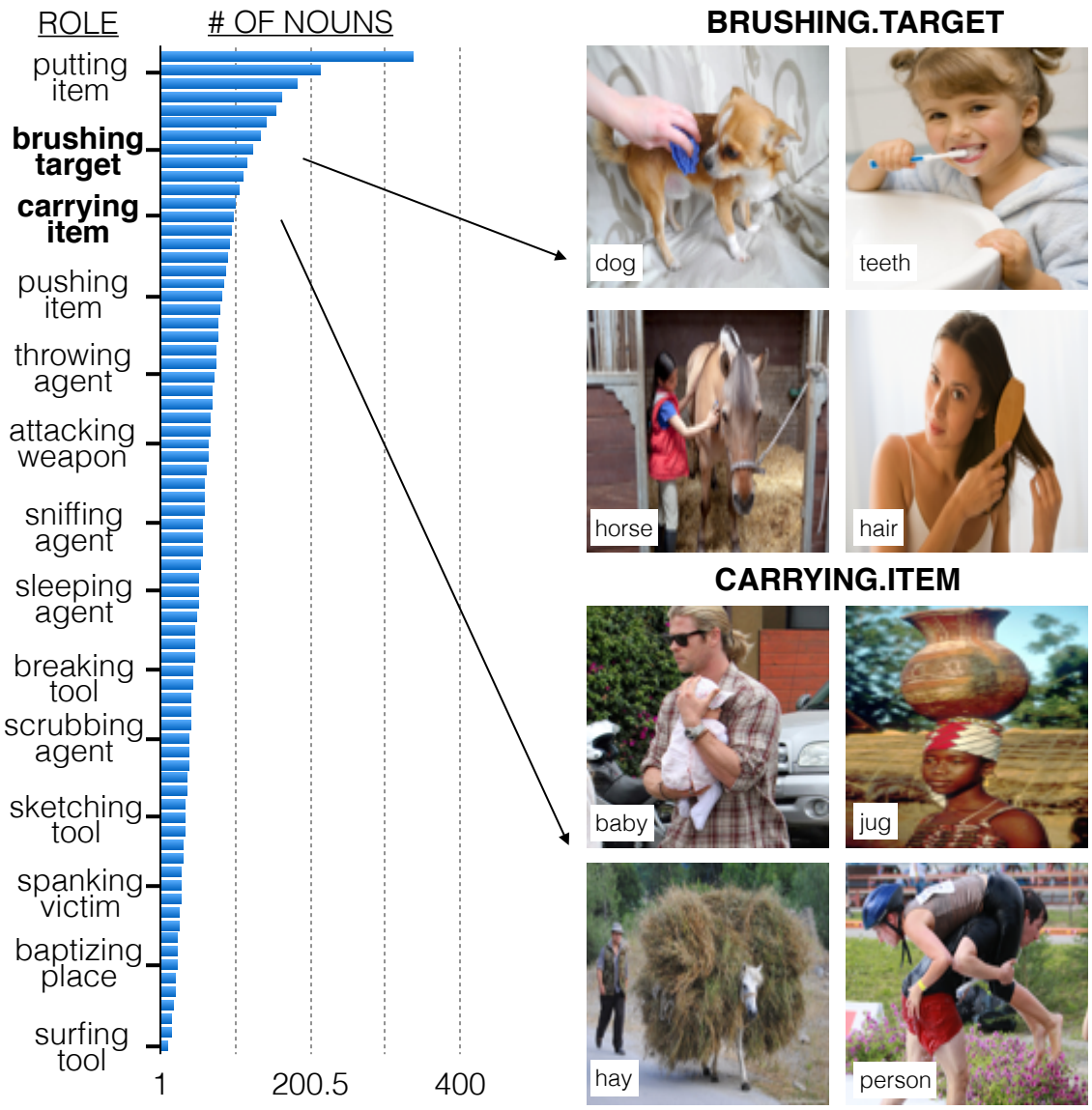


Figure 6.5: On the left, the number of nouns that can participate in a sample of semantic roles (not all labeled). On average 64.7 nouns appear per role. Some roles, such as the "tool" of "surfing" take very few values, indicating the majority of the information about the situation is indicated by the verb. On the right are examples of nouns that fill the "target" of "brushing" (the thing being brushed) and the "item" of carrying (the thing being carried), showing significant visual variation when the values are changed.

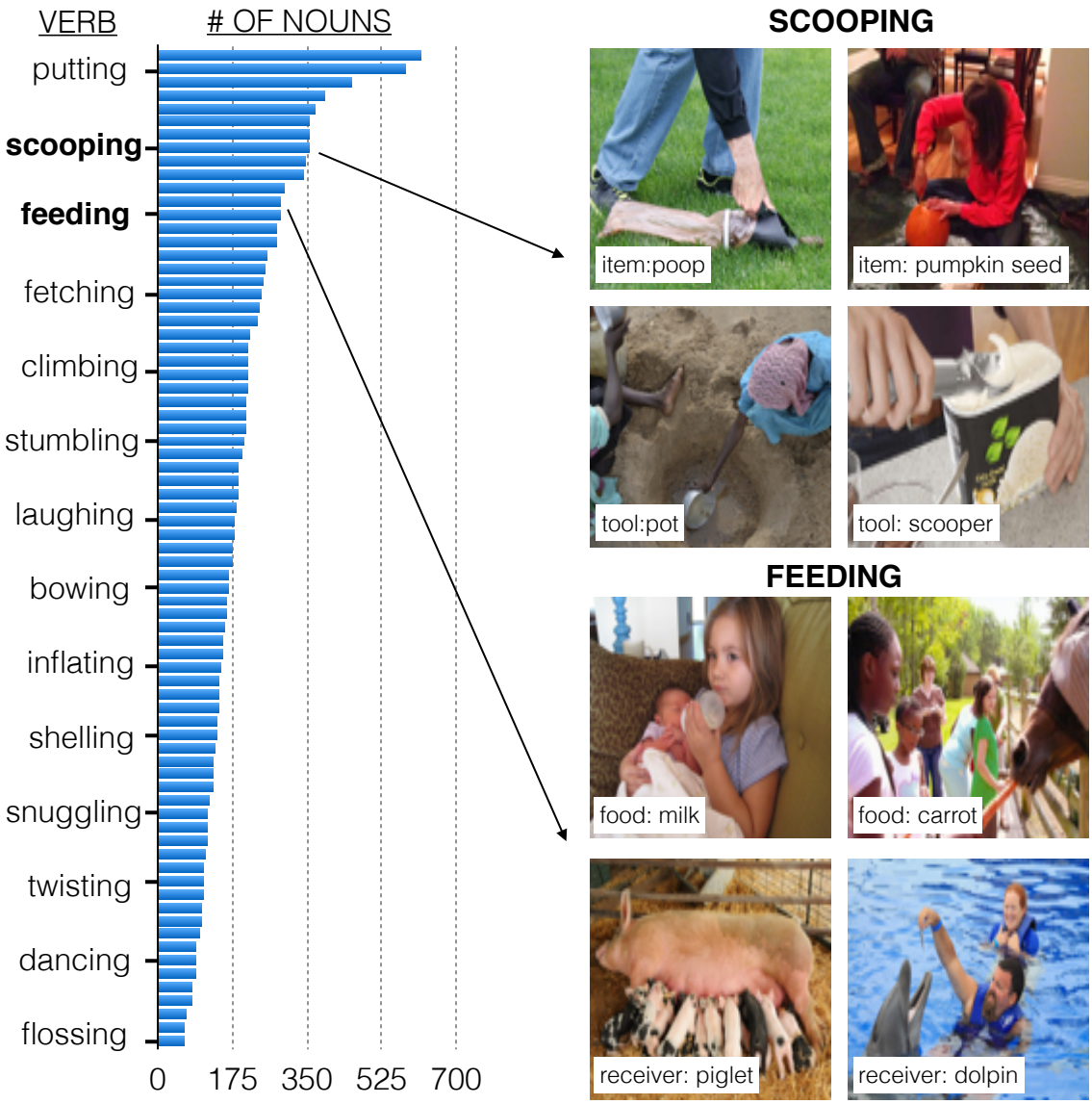


Figure 6.6: On the left, the number of nouns that appear with a sample of verbs (not all labeled). Some verbs (e.g. putting or scooping) require the ability to predict hundreds of noun values, while others (e.g. flossing) can only happen in a few canonical ways. On average, 199 nouns occur with a verb. On the right are example nouns for "scooping" and "feeding" and the roles they play.

Figure 6.7: Example realized situations from imSitu. Below each image is a table where the first row is the activity, the left column is semantic roles, and the right column is values for those roles. On the left outlined in gold are examples of gold standard annotated data. On the right is random output from our CRF model when it correctly predicted the activity. Incorrect semantic role values are highlighted in red, whereas correct ones are green.

Chapter 7

Sparsity in Situation Recognition

Semantic sparsity is a common challenge in structured visual classification problems; when the output space is complex, the vast majority of the possible predictions are rarely, if ever, seen in the training set. This paper studies semantic sparsity in situation recognition, the task of producing structured summaries of what is happening in images, including activities, objects and the roles objects play within the activity. For this problem, we find empirically that most substructures required for prediction are rare, and current state-of-the-art model performance dramatically decreases if even one such rare substructure exists in the target output.We avoid many such errors by (1) introducing a novel tensor composition function that learns to share examples across substructures more effectively and (2) semantically augmenting our training data with automatically gathered examples of rarely observed outputs using web data. When integrated within a complete CRF-based structured prediction model, the tensor-based approach outperforms existing state of the art by a relative improvement of 2.11% and 4.40% on top-5 verb and noun-role accuracy, respectively. Adding 5 million images with our semantic augmentation techniques gives further relative improvements of 6.23% and 9.57% on top-5 verb and noun-role accuracy.

7.1 Introduction

Many visual classification problems, such as image captioning [95], visual question answering [3], referring expressions [74], and situation recognition [152] have structured, semantically interpretable output spaces. In contrast to classification tasks such as ImageNet [126], these problems typically suffer from *semantic*



CARRYING

ROLE	VALUE
AGENT	MAN
ITEM	BABY
AGENTPART	CHEST
PLACE	OUTSIDE

ROLE	VALUE
AGENT	WOMAN
ITEM	BUCKET
AGENTPART	HEAD
PLACE	PATH

ROLE	VALUE
AGENT	MAN
ITEM	TABLE
AGENTPART	BACK
PLACE	STREET

Figure 7.1: Three situations involving `carrying`, with semantic roles `agent`, the carrier, `item`, the carried, `agentpart`, the part of the agent carrying, and `place`, where the situation is happening. For carrying, there are many possible carry-able objects (nouns that can fill the `item` role), which is an example of semantic sparsity. Such rarely occurring substructures are challenging and cause significant errors, affecting not only performance on role-values but also verbs.

sparsity; there is a combinatorial number of possible outputs, no dataset can cover them all, and performance of existing models degrades significantly when evaluated on rare or unseen inputs [6; 160; 32; 152]. In this paper, we consider situation recognition, a prototypical structured classification problem with significant semantic sparsity, and develop new models and semantic data augmentation techniques that significantly improve performance by better modeling the underlying semantic structure of the task.

Situation recognition [152] is the task of producing structured summaries of what is happening in images, including activities, objects and the roles those objects play within the activity. This problem can be challenging because many activities, such as `carrying`, have very open ended semantic roles, such as `item`, the thing being carried (see Figure 7.1); nearly any object can be carried and the training data will never contain all possibilities. This is a prototypical instance of semantic sparsity: rare outputs constitute a large portion of required predictions (35% in the imSitu dataset [152], see Figure 7.2), and current state-of-the-art performance for situation recognition drops significantly when even one participating object has few

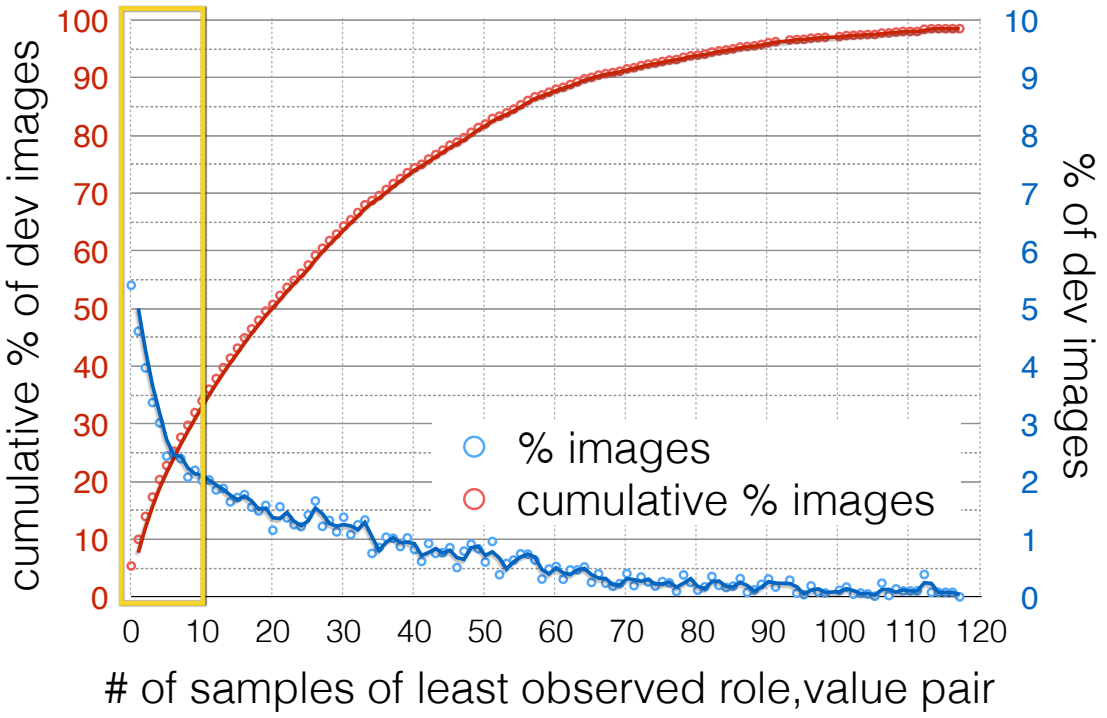


Figure 7.2: The percentage of images in the imSitu development set as a function of the total number of training examples for the least frequent role-noun pair in each situation. Uncommon target outputs, those observed fewer than 10 times in training (yellow box), are common, constituting 35% of all required predictions. Such semantic sparsity is a central challenge for situation recognition.

samples for it's role (see Figure 7.3). We propose to address this challenge in two ways by (1) building models that more effectively share examples of objects between different roles and (2) semantically augmenting our training set to fill in rarely represented noun-role combinations.

We introduce a new compositional Conditional Random Field formulation (CRF) to reduce the effects of semantic sparsity by encouraging sharing between nouns in different roles. Like previous work [152], we use a deep neural network to directly predict factors in the CRF. In such models, required factors for the CRF are predicted using a global image representation through a linear regression unique to each factor. In contrast, we propose a novel tensor composition function that uses low dimensional representations of nouns and roles, and shares weights across all roles and nouns to score combinations. Our model is compositional, independent representations of nouns and roles are combined to predict factors, and allows for a globally shared representation of nouns across the entire CRF.

This model is trained with a new form of semantic data augmentation, to provide extra training samples

for rarely observed noun-role combinations. We show that it is possible to generate short search queries that correspond to partial situations (i.e. "man carrying baby" or "carrying on back" for the situations in Figure 7.1) which can be used for web image retrieval. Such noisy data can then be incorporated in pre-training by optimizing marginal likelihood, effectively performing a soft clustering of values for unlabeled aspects of situations. This data also supports, as we will show, self training where model predictions are used to prune the set of images before training the final predictor.

Experiments on the imSitu dataset [152] demonstrate that our new compositional CRF and semantic augmentation techniques reduce the effects of semantic sparsity, with strong gains for relatively rare configurations. We show that each contribution helps significantly, and that the combined approach improves performance relative to a strong CRF baseline by 6.23% and 9.57% on top-5 verb and noun-role accuracy, respectively. On uncommon predictions, our methods provide a relative improvement of 8.76% on average across all measures. Together, these experiments demonstrate the benefits of effectively targeting semantic sparsity in structured classification tasks.

7.2 Background

Situation Recognition Situation recognition has been recently proposed to model events within images [67; 125; 146; 152], in order to answer questions beyond just "What activity is happening?" such as "Who is doing it?", "What are they doing it to?", "What are they doing it with?". In general, formulations build on semantic role labelling [62], a problem in natural language processing where verbs are automatically paired with their arguments in a sentence (for example, see [26]). Each semantic role corresponds to a question about an event, (for example, in the first image of Figure 7.1, the semantic role `agent` corresponds to "who is doing the carrying?" and `agentpart` corresponds to "how is the item being carried?").

We study situation recognition in imSitu [152], a large-scale dataset of human annotated situations containing over 500 activities, 1,700 roles, 11,000 nouns, 125,000 images. imSitu images are collected to cover a diverse set of situations. For example, as seen in Figure 7.2, 35% of situations annotated in the imSitu development set contain at least one rare role-noun pair. Situation recognition in imSitu is a strong test bed for evaluating methods addressing semantic sparsity: it is large scale, structured, easy to evaluate, and has a clearly measurable range of semantic sparsity across different verbs and roles. Furthermore, as

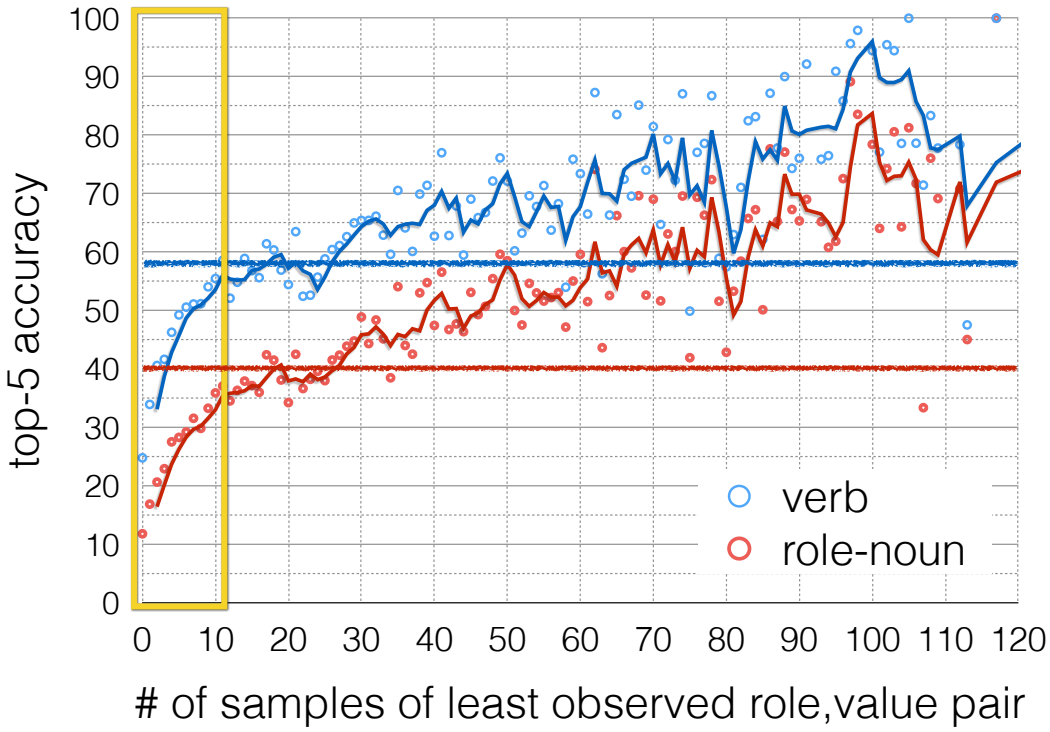


Figure 7.3: Verb and role-noun prediction accuracy of a baseline CRF [152] on the imSitu dev set as a function of the frequency of the least observed role-noun pair in the training set. Solid horizontal lines represent average performance across the whole imSitu dev set, irrespective of frequency. As even one target output becomes uncommon (highlighted in yellow box), accuracy decreases.

seen in Figure 7.3, semantic sparsity is a significant challenge for current situation recognition models.

Formal Definition In situation recognition, we assume a discrete sets of verbs V, nouns N, and frames F. Each frame $f \in F$ is paired with a set of semantic roles E_f. Every element in V is mapped to exactly one f. The verb set V and frame set F are derived from FrameNet [53], a lexicon for semantic role labeling, while the noun set N is drawn from WordNet [106]. Each semantic role $e \in E_f$ is paired with a noun value $n_e \in N \cup \{\varnothing\}$, where $\varnothing$ indicates the value is either not known or does not apply. The set of pairs of semantic roles and their values is called a realized frame, $R_f = \{(e, n_e) : e \in E_f\}$. Realized frames are valid only if each $e \in E_f$ is assigned exactly one noun n_e.

Given an image, the task is to predict a situation, $S = (v, R_f)$, specified by a verb $v \in V$ and a valid realized frame R_f, where f refers to a frame mapped by v . For example, in the first image of Figure 7.1, the predicted situations is $S =$ (`carrying`, {(`agent`,`man`), (`item`,`baby`), (`agentpart`,`chest`), (`place`,`outside`)}).

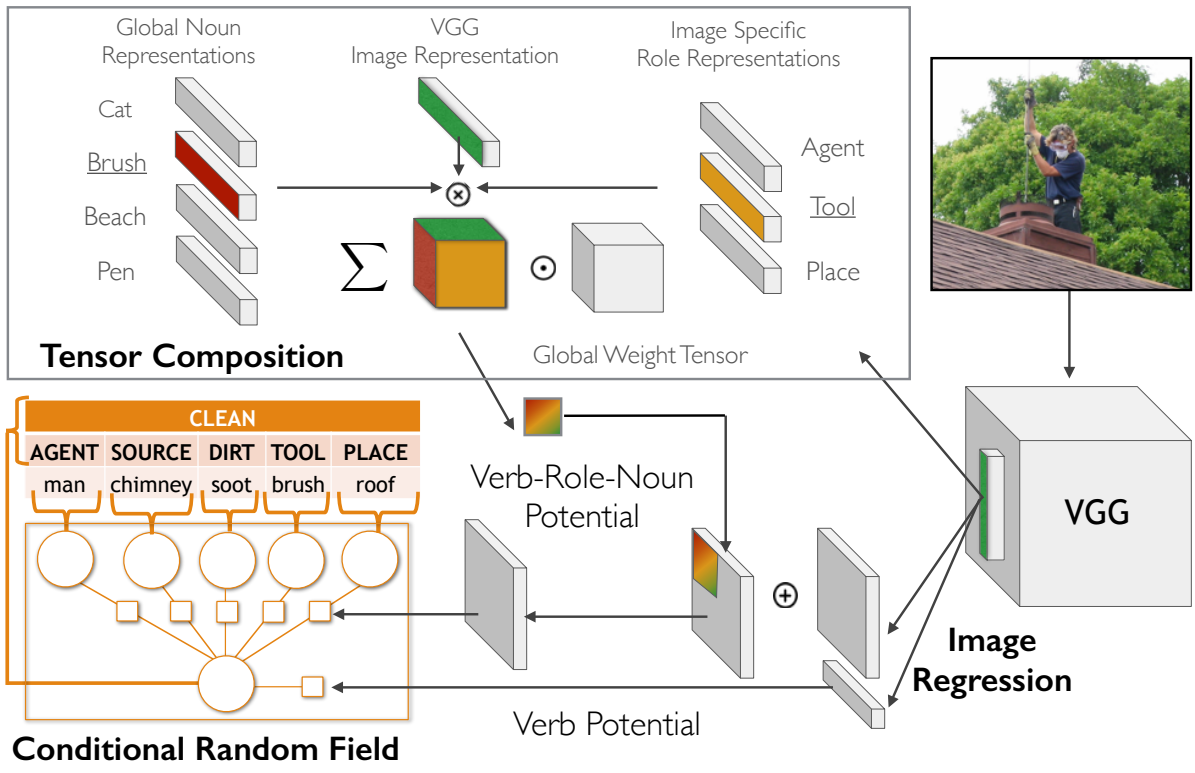


Figure 7.4: An overview of our compositional Conditional Random Field (CRF) for predicting situations. A deep neural network is used to compute potentials in a CRF. The verb-role-noun potential is built from a global bank of noun representations, image specific role representations and a global image representation that are combined with a weighted tensor product. The model allows for sharing among the same nouns in different roles, leading to significant gains, as seen in Section 7.5.

7.3 Methods

This section presents our compositional CRFs and semantic data augmentation techniques.

7.3.1 Compositional Conditional Random Field

Figure 7.4 shows an overview of our compositional conditional random field model, which is described below.

Conditional Random Field Our CRF for predicting a situation, $S = (v, R_f)$, given an image i, decomposes over the verb v and semantic role-value pairs (e, n_e) in the realized frame $R_f = \{(e, n_e) : e \in E_f\}$, similarly to previous work [152]. The full distribution, with potentials for verbs ψ_v and semantic roles ψ_e

takes the form:

$$p(S|i;\theta) \propto \psi_v(v,i;\theta) \prod_{(e,n_e)\in R_f} \psi_e(v,e,n_e,i;\theta) \tag{7.1}$$

The CRF admits efficient inference: we can enumerate all verb-semantic roles that occur and then sum all possible semantic role values that occurred in a dataset.

Each potential in the CRF is log linear:

$$\psi_v(v,i;\theta) = e^{\phi_v(v,i,\theta)} \tag{7.2}$$

$$\psi_e(v,e,n_e,i;\theta) = e^{\phi_e(v,e,n_e,i,\theta)} \tag{7.3}$$

where ϕ_e and ϕ_v encode scores computed by a neural network. To learn this model, we assume that for an image i in dataset Q there can, in general, be a set A_i of possible ground truth situations [1]. We optimize the log-likelihood of observing at least one situation $S \in A_i$:

$$\sum_{i\in Q} \log\Big(1 - \prod_{S\in A_i} (1 - p(S|i;\theta))\Big) \tag{7.4}$$

Compositional Tensor Potential In previous work, the CRF potentials (Equation 7.2 and 7.3) are computed using a global image representation, a p-dimensional image vector $g_i \in \mathcal{R}^p$, derived by the VGG convolutional neural network [136]. Each potential value is computed by a linear regression with parameters, θ, unique for each possible decision of verb and verb-role-noun (we refer to this as image regression in Figure 7.4), for example for the verb-role-noun potential in Equation 7.3:

$$\phi_e(v,e,n_e,i,\theta) = g_i^T \theta_{v,e,n_e} \tag{7.5}$$

Such a model does not directly represent the fact that nouns are reused between different roles, although the underlying neural network could hypothetically learn to encode such reuse during fine tuning. Instead, we introduce compositional potentials that make such reuse explicit.

To formulate our compositional potential, we introduce a set of m-dimensional vectors $D = \{d_n \in \mathcal{R}^m | n \in N\}$, one vector for each noun in N, the set of nouns. We create a set matrices $T = \{H_{(v,e)} \in$

[1] imSitu provides three realized frames per example image.

$\mathcal{R}^{p \times o} | (v, e) \in E_f\}$, one matrix for each verb, semantic role pair occurring in all frames E_f, that map image representations to o-dimensional verb-role representations. Finally, we introduce a tensor of global composition weights, $C \in \mathcal{R}^{m \times o \times p}$. We define a tensor weighting function, T, which takes as input a verb, v, semantic role, e, noun, n, and image representation, g_i as:

$$T(v, e, n, g_i) = C \odot (d_n \otimes g_i^T H_{(v,e)} \otimes g_i) \tag{7.6}$$

The tensor weighting function constructs an image specific verb-role representation by multiplying the global image vector and the verb-role matrix $g_i^T H_{(v,e)}$. Then, it combines a global noun representation, the image specific role representation, and the global image representation with outer products. Finally, it weights each dimension of the outer product with a weight from C. The weights in C indicate which features of the 3-way outer product are important. The final potential is produced by summing up all of the elements of the tensor produced by T:

$$\phi_e(v, e, n_e, i) = \sum_{x=0}^{M} \sum_{y=0}^{O} \sum_{z=0}^{P} T(v, e, n_e, g_i)[x, y, z] \tag{7.7}$$

The tensor produced by T in general will be high dimensional and very expressive. This allows use of small dimensionality representations, making the function more robust to small numbers of samples for each noun.

The potential defined in Equation 7.7 can be equivalently formulated as :

$$\phi_e(v, e, n_e, i) = g_i^T A(d_{n_e} \otimes g_i^T H_{(v,e)}) \tag{7.8}$$

Where A is a matrix with the same parameters as C but flattened to layout the noun and role dimensions together. By aligning terms with Equation 7.5, one can see that tensor potential offers an alternative parametrized to the linear regression that uses many more general purpose parameters, those of C. Furthermore, it eliminates any one parameter from ever being uniquely associated with one regression, instead compositionally using noun and verb-role representations to build up the parameters of the regression.

7.3.2 Semantic Data Augmentation

Situation recognition is strongly connected to language. Each situation can be thought of as simple declarative sentence about an activity happening in an image. For example, the first situation in Figure 7.1 could be expressed as "man carrying baby on chest outside" by knowing the prototypical ordering of semantic roles around verbs and inserting prepositions. This relationship can be used to reduce semantic sparsity by using image search to find images that could contain the elements of a situations.

We convert annotated situations to phrases for semantic augmentation by exhaustively enumerating all possible sub-pieces of realized situations that occur in the imSitu training set (see Section 7.4 for implementation details). For example, in first situation of Figure 7.1, we get the pieces: $(\texttt{carrying}, \{(\texttt{agent}, \texttt{man})\})$, $(\texttt{carrying}, \{(\texttt{agent}, \texttt{man}), (\texttt{item}, \texttt{baby})\})$, ect. Each of these substructures is converted deterministically to a phrase using a template specific for every verb. For example, the template for carrying is "{agent} carrying {item} {with agentpart} {in place}." Partial situations are realized into phrases by taking the first gloss in Wordnet of the synset associated with every noun in the substructure, inserting them into the corresponding slots of the template, and discarding unused slots. For example, the phrases for the sub-pieces above are realized as "man carrying" and "man carrying baby." These phrases are used to retrieve images from Google image search and construct a set, $W = \{(i, v, R_f)\}$, of images annotated with a verb and partially complete realized frames, by assigning retrieved images to the sub-piece that generated the retrieval query.[2]

Pre-training Images retrieved from the web can be incorporated in a pre-training phase. The images retrieved only have partially specified realized situations as labels. To account for this, we instead compute the marginal likelihood, $\hat{p}$, of the partially observed situations in W:

$$\begin{aligned}\hat{p}(S|i;\theta) \propto \psi_v(v, i; \theta) \prod_{(e,n_e)\in R_f} \psi_e(v, e, n_e, i; \theta) \\ \times \prod_{e \notin R_f \wedge e \in E_f} \sum_n \psi_e(v, e, n, i; \theta)\end{aligned} \tag{7.9}$$

[2]While these templates do not generate completely fluent phrases, preliminary experiments found them sufficiently accurate for image search because often no phrase could retrieve correct images. Longer phrases tended to have much lower precision.

During pretraining, we optimize the marginal log-likelihood of W. This objective provides a partial clustering over the unobserved roles left unlabeled during the retrieval process.

Self Training Images retrieved from the web contain significant noise. This is especially true for role-noun combinations that occur infrequently, limiting their utility for pretraining. Therefore, we also consider filtering images in W after a model has already been trained on fully supervised data from imSitu. We rank images in W according to $\hat{p}$ as computed by the trained model and filter all those not in the top-k for every unique R_f in W. We then pretrain on this subset of W, train again on imSitu, and then increase k. We repeat this process until the model no longer improves.

7.4 Experimental Setup

Models All models were implemented in Caffe [71] and use a pretrained VGG network [136] for the base image representation with the final two fully connected layers replaced with two fully connected layers of dimensionality 1024. We finetune all layers of VGG for all models. For our tensor potential we use noun embedding size, $m = 32$, and role embedding size $o = 32$, and the final layer of our VGG network as the global image representation where $p = 1024$. Larger values of m and o did seem to improve results but were too slow to pretrain so we omit them. In experiments where we use the image regression in conjunction with a compositional potential, we remove regression parameters associated with combinations seen fewer than 10 times on the imSitu training set to reduce overfitting.

Baseline We compare our models to two alternative methods for introducing effective sharing between nouns. The first baseline (Noun potential in Table 7.1 and 7.2) adds a potential into the baseline CRF for nouns independent of roles. We modify the probability, from Equation 7.9 of a situation, S, given an image i, to not only decompose by pairs of roles, e and nouns n_e in a realized frame R_f, but also nouns n_e:

$$p(S|i;\theta) \propto \psi_v(v,i;\theta) \prod_{(e,n_e)\in R_f} \psi_e(v,e,n_e,i;\theta)\psi_{n_e}(n_e,i) \tag{7.10}$$

The added potential, ψ_{n_e}, is computed using a regression from a global image representation for each unique n_e.

			top-1 predicted verb			top-5 predicted verbs			ground truth verbs		
			verb	value	value-all	verb	value	value-all	value	value-all	mean
imSitu	1	Baseline: Image Regression [152]	32.25	24.56	14.28	58.64	42.68	22.75	65.90	29.50	36.32
	2	Noun Potential + reg	27.64	21.21	12.21	53.95	39.95	21.45	68.87	32.31	34.70
	3	Inner product composition + reg	32.13	24.77	14.71	58.33	42.93	23.14	66.79	30.2	36.62
	4	Tensor composition	31.73	24.04	13.73	58.06	42.64	22.7	68.73	32.14	36.72
	5	Tensor composition + reg	32.91	25.39	14.87	59.92	44.5	24.04	69.39	33.17	38.02
+ SA	6	Baseline : Image Regression	32.40	24.14	15.17	59.10	44.04	24.40	68.03	31.93	37.53
	7	Tensor composition + reg	34.04	26.47	**15.73**	61.75	46.48	**25.77**	**70.89**	**35.08**	39.53
	8	Tensor composition + reg + self train	**34.20**	**26.56**	15.61	**62.21**	**46.72**	25.66	70.80	34.82	**39.57**

Table 7.1: Situation recognition results on the full imSitu development set. The results are divided by models which were only trained on imSitu data, rows 1-5, and models which use web data through semantic data augmentation, marked as +SA in rows 6-8. Models marked with +reg also include image regression potentials used in the baseline. Our tensor composition model, row 5, significantly outperforms the existing state of the art, row 1, addition of a noun potential, row 2, and a compositional baseline, row 3. The tensor composition model is able to make better use of semantic data augmentation (row 8) than the baseline (row 6).

The second baseline we consider is compositional but does not use a tensor based composition method. The model instead constructs many verb-role representations and combines them with noun representations using inner-products (Inner product composition in Table 7.1 and 7.2). In this model, as in the tensor model in Section 7.3, we use a global image representation $g_i \in \mathcal{R}^p$ and a set noun vectors, $d_n \in \mathcal{R}^m$ for every noun n. We also assume t verb-role matrices $H_{t,v,e} \in \mathcal{R}^{o \times p}$ for every verb-role in E_f. We compute the corresponding potential as in Equation 7.11:

$$\phi_e(v, e, n_e, i) = \sum_k d_{n_e}^T H_{(k,v,e)} q_i \tag{7.11}$$

The model is motivated by compositional models used for semantic role labeling [55] and allows us to trade-off the need to reduce parameters associated with nouns and expressivity. We grid search values of t such that $t \cdot o$ was at most 256, the largest size network we could afford to run and $o = m$, a requirement on the inner product. We found the best setting at $t = 16$, $o = m = 16$.

Decoding We experimented with two decoding methods for finding the best scoring situation under the CRF models. Systems which used the compositional potentials performed better when first predicting a verb v^m using the max-marginal over semantic roles: $v^m = \arg\max_v \sum_{(e,n_e)} p(v, R_f | i)$ and then predict a realized frame, R_f^m, with max score for v^m: $R_f^m = \arg\max_{R_f} p(v^m, R_f | i)$. All other systems performed better maximizing jointly for both verb and realized frame.

			top-1 predicted verb			top-5 predicted verbs			ground truth verbs		
			verb	value	value-all	verb	value	value-all	value	value-all	mean
imSitu	1	Baseline: image regression [152]	19.89	11.68	**2.85**	44.00	24.93	**6.16**	50.80	9.97	19.92
	2	Noun potential + reg	15.88	9.13	1.86	38.22	22.28	5.46	54.65	11.91	19.92
	3	Inner product composition + reg	18.96	10.69	1.89	42.53	23.28	3.69	49.54	6.46	19.63
	4	Tensor composition	19.78	11.28	2.26	42.66	24.42	5.57	54.06	11.47	21.43
	5	Tensor composition + reg	**21.12**	11.89	2.20	45.14	25.51	5.36	53.58	10.62	21.93
+ SA	6	Baseline : image regression	19.95	11.44	2.13	43.08	24.56	4.95	51.55	8.41	20.76
	7	Tensor composition + reg	20.08	11.58	2.22	44.82	26.02	5.55	55.45	11.53	22.16
	8	Tensor composition + reg + self train	20.52	**11.91**	2.34	**45.94**	**26.99**	6.06	**55.90**	**12.04**	**22.71**

Table 7.2: Situation prediction results on the rare portion imSitu development set. The results are divided by models which were only trained on imSitu data, rows 1-5, and models which use web data through semantic data augmentation, marked as +SA in rows 6-8. Models marked with +reg also include image regression potentials used in the baseline. Semantic data augmentation with the baseline hurts for rare cases. Semantic augmentation yields larger relative improvement on rare cases and a composition-based model is required to realize these gains.

Optimization All models were trained with stochastic gradient descent with momentum 0.9 and weight decay 5e-4. Pretraining in semantic augmentation was conducted with initial learning rate of 1e-3, gradient clipping at 100, and batch size 360. When training on imSitu data, we use an initial learning rate of 1e-5. For all models, the learning rate was reduced by a factor of 10 when the model did not improve on the imSitu dev set.

Semantic Augmentation In experiments with semantic augmentation, images were retrieved using Google image search. We retrieved 200 medium sized, full-color, safe search filtered images per query phrase. We produced over 1.5 million possible query phrases from the imSitu training set, the majority extremely rare. We limited the phrases to any that occur between 10 and 100 times in imSitu and for phrases that occur between 3 and 10 times we accepted only those containing at most one noun. Roughly 40k phrases were used to retrieve 5 million images from the web. All duplicate images occurring in imSitu were removed. For pretraining, we ran all experiments up to 50k updates (roughly 4 epochs). For self training, we only self train on rare realized frames (those 10 or fewer times in imSitu train set). Self training yielded diminishing gains after two iterations and we ran the first iteration at k=10 and the second at k=20.

Evaluation We use the standard data split for imSitu[152] with 75k train, 25k development, and 25k test images. We follow the evaluation setup defined for imSitu, evaluating verb predictions (verb) and semantic role-value pair predictions (value) and full structure correctness (value-all). We report accuracy at top-1, top-5 and given the ground truth verb and the average across all measures (mean). We also report performance for examples requiring rare (10 or fewer examples in the imSitu training set) predictions.

		top-1 predicted verb			top-5 predicted verbs			ground truth verbs		
		verb	value	value-all	verb	value	value-all	value	value-all	mean
imSitu	Baseline: Image Regression [152]	32.34	24.64	14.19	58.88	42.76	22.55	65.66	28.96	36.25
	Tensor composition + reg	32.96	25.32	14.57	60.12	44.64	24.00	69.2	32.97	37.97
+ SA	Baseline : Image Regression	32.3	24.95	14.77	59.52	44.08	23.99	67.82	31.46	37.36
	Tensor composition + reg + self train	**34.12**	**26.45**	**15.51**	**62.59**	**46.88**	**25.46**	**70.44**	**34.38**	**39.48**

Table 7.3: Situation prediction results on the full imSitu test set. Models were run exactly once on the test set. General trends are identical to experiments run on development set.

		top-1 predicted verb			top-5 predicted verbs			ground truth verbs		
		verb	value	value-all	verb	value	value-all	value	value-all	mean
imSitu	Baseline: Image Regression [152]	**20.61**	11.79	**3.07**	44.75	24.85	5.98	50.37	9.31	21.34
	Tensor composition + reg	19.96	11.57	2.30	44.89	25.26	4.87	53.39	10.15	21.55
+ SA	Baseline : Image Regression	19.46	11.15	2.13	43.52	24.14	4.65	51.21	8.26	20.57
	Tensor composition + reg + self train	20.32	**11.87**	2.52	**47.07**	**27.50**	**6.35**	**55.72**	**12.28**	**22.95**

Table 7.4: Situation prediction results on the rare portion of imSitu test set. Models were run exactly once on the test set. General trends established on the development set are supported.

7.5 Results

Compositional Tensor Potential Our results on the full imSitu dev set are presented in Table 7.1 in rows 1-5. Overall results demonstrate that adding a noun potential (row 2) and our baseline composition model (row 3) are ineffective and perform worse than the baseline CRF (row 1). We hypothesize that systematic variation in object appearance between roles is challenging for these models. Our tensor composition model (row 4) is able to better capture such variation and effectively share information among nouns, reflected by improvements in value and value-all accuracy given ground truth verbs while maintaining high top-1 and top-5 verb accuracy. However, as expected, many situations cannot be predicted only compositionally based on nouns (consider that a horse sleeping looks very different than a horse swimming and nothing like a person sleeping). Combination of the image regression potential and our tensor composition potential (row 5) yields the best performance, indicating they are modeling complementary aspects of the problem. Our final model (row 5) only trained on imSitu data outperforms the baseline on every measure, improving over 1.70 points overall.

Results on the rare portion of the imSitu dataset are presented in Table 7.2 in rows 1-5. Our final model (row 5) provides the best overall performance (mean column) on rare cases among models trained only on imSitu data, improving by 0.64 points on average. All models struggle to get correctly entire structures (value-all columns), indicating rare predictions are extremely hard to get completely correct while the baseline model which only uses image regression potentials performs the best. We hypothesize that image

regression potentials may allow the model to more easily coordinate predictions across roles simultaneously because role-noun combinations that always co-occur will always have the same set of regression weights.

Semantic Data Augmentation Our results on the full imSitu development set are presented in Table 7.1 in rows 6-8. Overall results indicate that semantic data augmentation helps all models, while our tensor model (row 7) benefits more than the baseline (row 6). Self training improves the tensor model slightly (row 8), making it perform better on top-1 and top-5 predictions but hurting performance given gold verbs. On average, our final model outperforms the baseline CRF trained on identical data by 2.04 points.

Results on the rare portion of the imSitu dataset are presented in Table 7.2 in rows 6-8. Surprisingly, on rare cases semantic augmentation hurts the baseline CRF (line 6). Rare instance image search results are extremely noisy. On close inspection, many of the returned results do not contain the target activity at all but instead contain target nouns. We hypothesize that without an effective global noun representation, the baseline CRF cannot extract meaningful information from such extra data. On the other hand, our tensor model (line 7) improves on these rare cases overall and with self training improves further (line 8).

Overall Results Experiments show that (a) our tensor model is able perform better in comparable data settings, (b) our semantic augmentation techniques largely benefit all models, and (c) our tensor model benefits more from semantic augmentation. We also present our full performance on top-5 verb across all numbers of samples in Figure 7.5. While our compositional CRF with semantic augmentation outperforms the baseline CRF, both models continue to struggle on uncommon cases. Our techniques seem to give most benefit for examples requiring predictions of structures seen between 5 and 35 times, while providing somewhat less benefit to even rarer ones. It is challenging future work to make further improvements for extremely rare outputs.

We also evaluated our models on the imSitu test set exactly once. The results are summarized in Table 7.3 for the full imSitu test set and in Table 7.4 for the rare portion. General trends established on the imSitu dev set are supported. We provide examples in Figure 7.6 of predictions our final system made on rare examples from the development set.

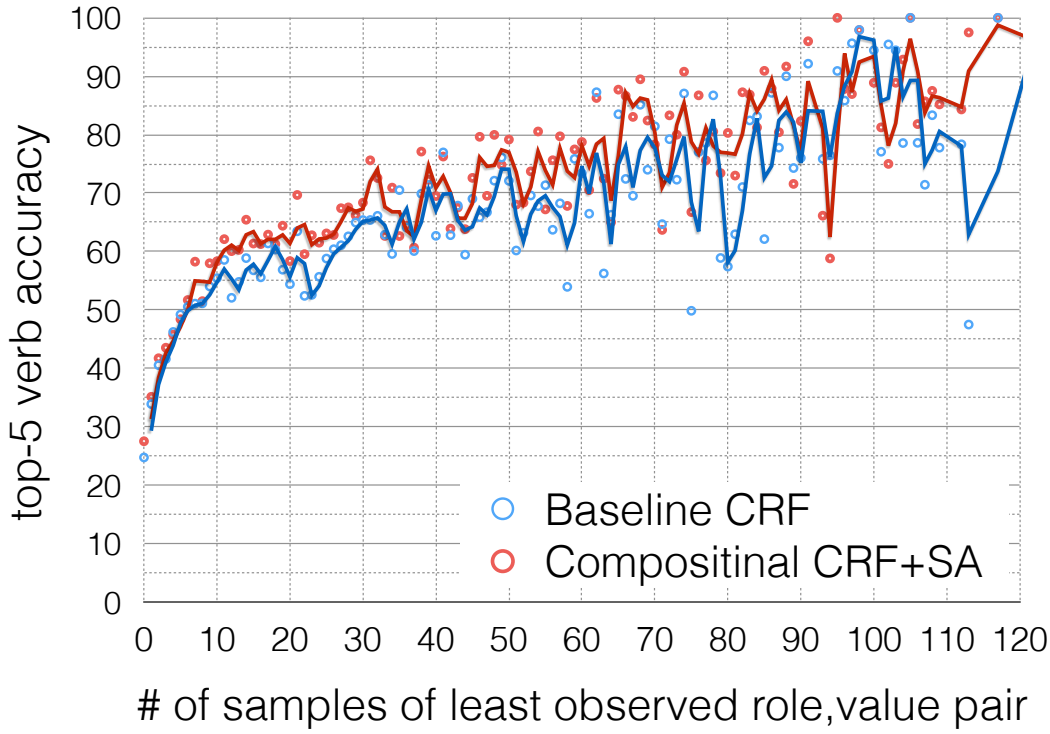


Figure 7.5: Top-5 verb accuracy on the imSitu development set. Our final compositional CRF with semantic data augmentation outperforms the baseline CRF on rare cases (fewer than 10 training examples), but both models continue to struggle with semantic sparsity. For our final model, the largest improvement relative to the baseline are for cases with 5-35 examples on the training set.

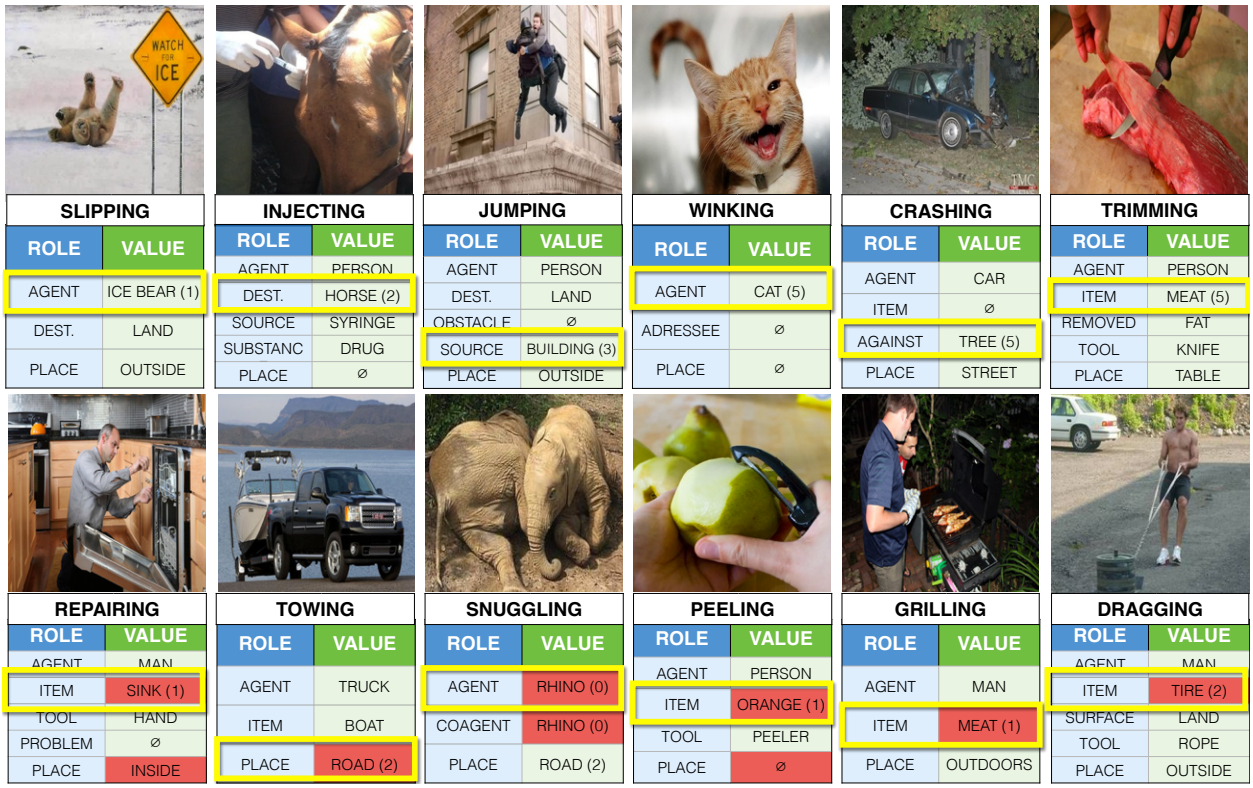


Figure 7.6: Output from our final model on development examples containing rare role-noun pairs. The first row contains examples where the model correctly predicts the entire structures in the top-5 (top-5, value-all). We highlight the particular role-noun pairs that make the examples rare with a yellow box and put the number occurrences of it in the imSitu training set. The second row contains examples where the verb was correctly predicted in the top-5 but not all the values were predicted correctly. We highlight incorrect predictions in red. Many such predictions occur zero times in the training set (ex. the third image on the second row). All systems struggle with such cases.

Chapter 8

Conclusion

In this thesis we explored whether natural language can be used to define what information is extracted from an image. We introduced several proposals that use unconstrained language to define recognition targets and showed even highly ambiguous sentimental language can be grounded to produce well defined predictions. We showed how to use natural language ontologies to (a) to expand knowledge that can be extracted from labeled images and (b) to define novel visual recognition problems.

In Chapter 3, we explored how visual language, both literal and sentimental, maps to the overall physical appearance and style of virtual characters. While the section focused on avatar design, our approach has implications for a broad class of natural language-driven dialog scenarios. In many situations, a user may be perfectly able to formulate a high-level description of their intent ("Make my resume look cleaner" "Buy me clothes for a summer wedding," or "Play something more danceable") while having little or no understanding of the complex parameter space that the underlying software must manipulate in order to achieve this result. We demonstrated that these high-level sentimental specifications can have a strong relationship to literal aspects of a problem space and showed that sentimental language is a concise, yet noisy, way of specifying high level characteristics. Sentimental language is an unexplored avenue for improving natural language systems that operate in situated settings. It has the potential to bridge the gap between lay and expert understandings of a problem domain.

In Chapter 4 we used dense annotations of images to study description generation. The annotations allowed us to not only develop new models, better capable of generating human-like sentences, but also

to explore what visual information is crucial for description generation. Experiments showed that activity and bounding-box information is important and demonstrated areas of future work. In images that are more complex, for example multiple sentient objects, object grouping and reference will be important to generating good descriptions.

In Chapter 5 we use an existing object detection dataset to extract 16k common sense statements about annotated categories. We also show how to generalize using WordNet and induced hundreds of thousands of facts about *unseen* objects. The information we extracted is visual, large scale and good quality. It has the potential to be useful for both visual recognition and entailment applications.

In Chapter 6 we introduced the problem of situation recognition and described the construction of imSitu, a large new situation recognition data set. Key to the formulation was the use of semantic roles to represent how objects, actors, and other entities participate in different activities. The situation recognition task is challenging but provides strong context for recognizing activities and objects. Situation recognition significantly improves over existing formalisms for activities in images, expanding on both the coverage and richness of the representation.

In Chapter 7 we studied semantic sparsity in situation recognition. Despite the fact that the vast majority of the possible output configurations are rarely observed in the training data, we showed it was possible in introduce new compositional models that effectively share examples among required outputs and semantic data augmentation techniques that significantly improved performance.

8.1 Future Work

Future work in language based representations of image is broad, including high level questions about further expanding visual representations and low level technical questions such as generalization to unobserved configurations in structured output.

Scene Graphs The work presented in this thesis explored on a small scale what labeling everything in an image might mean and at a large scale defined a framework for relations in images, situation recognition. One potential avenue to explore is combining the ontology of relations defined in imSitu with other core computer vision problems to form a more complete representation for an entire image. This would include

labeling regions for arguments of relations and labeling multiple relationships within an image. While there have been attempts at such proposals, such as the Visual Genome[81] project, these efforts do not benefit from a tight coupling with language based resources and as such have inconsistent coverage within image and have poor coverage of semantic units across image.

Structured Zero-shot Learning The work in this thesis revealed that semantic sparsity in structured predictions are the central challenge for situation recognition. An extreme version of this issue is to observe no examples of a required sub-part of a prediction. This case is effectively captured in Chapter 7, but has many special features unique to it. For example, it is currently not possible to train a model that can score all possible unobserved, in training data, subparts of situations, and thus a model must explicitly make choices about what zero-shot configurations should be possible and not. This zero-shot scenario will be increasingly important as annotation efforts on structured output, whether situations or graph structured representation, can only cover a small subset of the combinatorial options.

Bibliography

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016.

[2] G. Angeli, P. Liang, and D. Klein. A simple domain-independent probabilistic approach to generation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2010.

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision*, 2015.

[4] S. Antol et al. Vqa: Visual question answering. *arXiv preprint arXiv:1505.00468*, 2015.

[5] Y. Artzi and L. Zettlemoyer. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1(1):49–62, 2013.

[6] Y. Atzmon, J. Berant, V. Kezami, A. Globerson, and G. Chechik. Learning to generalize to new compositions in image understanding. *arXiv preprint arXiv:1608.07639*, 2016.

[7] M. Baroni and A. Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

[8] J. Berant and P. Liang. Semantic parsing via paraphrasing. In *Association for Computational Linguistics (ACL)*, 2014.

[9] A. Bouchoucha, J. He, and J.-Y. Nie. Diversified query expansion using conceptnet. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, CIKM '13, pages 1861–1864, New York, NY, USA, 2013. ACM.

[10] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[11] S. Branavan, H. Chen, L. Zettlemoyer, and R. Barzilay. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 82–90, 2009.

[12] S. Branavan, D. Silver, and R. Barzilay. Learning to win by reading manuals in a monte-carlo framework. In

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 268–277, 2011.

[13] C. Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *EMNLP*, pages 286–295, August 2009.

[14] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell. Toward an architecture for never-ending language learning. In *AAAI Conference on Artificial Intelligence*, volume 5, page 3, 2010.

[15] D. Chen and R. Mooney. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI-2011)*, pages 859–865, 2011.

[16] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 190–200, 2011.

[17] D. L. Chen, J. Kim, and R. J. Mooney. Training a multilingual sportscaster: Using perceptual context to learn language. *JAIR*, 37:397–435, 2010.

[18] D. L. Chen and R. J. Mooney. Learning to sportscast: a test of grounded language acquisition. In *ICML*, 2008.

[19] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1431–1439, 2015.

[20] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *International Conference on Computer Vision (ICCV)*, pages 1409–1416. IEEE, 2013.

[21] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1409–1416, 2013.

[22] X. Chen et al. Learning a recurrent visual representation for image caption generation. *arXiv:1411.5654*, 2014.

[23] M. Collins. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 1–8, 2002.

[24] B. Coyne and R. Sproat. Wordseye: an automatic text-to-scene conversion system. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 487–496, 2001.

[25] I. Dagan, B. Dolan, B. Magnini, and D. Roth. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(01):105–105, 2010.

[26] D. Das. *Semi-Supervised and Latent-Variable Models of Natural Language Semantics*. PhD thesis, CMU, 2012.

[27] M.-C. de Marneffe, B. MacCartney, C. D. Manning, et al. Generating typed dependency parses from phrase structure parses. In *LREC*, volume 6, pages 449–454, 2006.

[28] V. Delaitre et al. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010.

[29] J. Deng, J. Krause, A. C. Berg, and L. Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in

large scale visual recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 3450–3457. IEEE, 2012.

[30] J. Deng et al. Construction and analysis of a large scale image ontology. Vision Sciences Society, 2009.

[31] J. Deng et al. Large-scale object classification using label relation graphs. In *ECCV*. 2014.

[32] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*, 2015.

[33] S. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3277, 2014.

[34] S. Divvala et al. An empirical study of context in object detection. In *CVPR*, 2009.

[35] J. Eisenstein, J. Clarke, D. Goldwasser, and D. Roth. Reading to learn: Constructing features from semantic abstracts. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 958–967, 2009.

[36] D. Elliott and A. P. de Vries. Describing images using inferred visual dependency representations. *Association for Computational Linguistics (ACL)*, 2015.

[37] D. Elliott and F. Keller. Image description generation from structured image representations. In *EMNLP*, 2013.

[38] D. Elliott, V. Lavrenko, and F. Keller. Query-by-example image retrieval using visual dependency representations. In *International Conference on Computational Linguistics (COLING)*, pages 109–120, August 2014.

[39] D. Elliott et al. Comparing automatic evaluation measures for image description. In *ACL*, 2014.

[40] V. R. et al. Linking people with "their" names using coreference resolution. In *ECCV*, 2014.

[41] Z. et al. Building a large-scale multimodal knowledge base for visual question answering. *arXiv preprint arXiv:1507.05670*, 2015.

[42] M. Everingham et al. The pascal visual object classes challenge 2009. In *2th PASCAL Challenge Workshop*, 2009.

[43] H. Fang et al. From captions to visual concepts and back. *arXiv:1411.4952*, 2014.

[44] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[45] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009.

[46] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*. Springer, 2010.

[47] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European conference on Computer*

Vision, ECCV'10, pages 15–29, 2010.

[48] A. Farhadi et al. Describing objects by their attributes. In *CVPR*, 2009.

[49] A. Farhadi et al. Every picture tells a story: Generating sentences from images. In *ECCV 2010*, pages 15–29. 2010.

[50] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[51] Y. Feng and M. Lapata. How many words is a picture worth? Automatic caption generation for news images. In *ACL*, pages 1239–1249, 2010.

[52] Y. Feng and M. Lapata. Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 831–839, 2010.

[53] C. J. Fillmore et al. Background to framenet. *International Journal of lexicography*, 2003.

[54] N. FitzGerald, Y. Artzi, and L. Zettlemoyer. Learning distributions over logical forms for referring expression generation. In *EMNLP*, 2013.

[55] N. FitzGerald et al. Semantic role labelling with neural network factors. In *EMNLP*, 2015.

[56] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*. 2013.

[57] A. Frome et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[58] S. R. Fussell and M. M. Moss. Figurative language in emotional communication. *Social and cognitive approaches to interpersonal communication*, page 113, 1998.

[59] C. Galleguillos et. al. Context based object categorization: A critical survey. *CVIU*, 2010.

[60] H. e. a. Gao. Are you talking to a machine? dataset and methods for multilingual image question answering. *arXiv preprint arXiv:1505.05612*, 2015.

[61] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. *arXiv preprint arXiv:1511.06062*, 2015.

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

[63] Y. Goldberg et al. A dataset of syntactic-ngrams over time from a very large corpus of english books. In **SEM*, 2013.

[64] S. Guadarrama et al. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013.

[65] G. Guo et al. A survey on still image based human action recognition. *Pattern Recognition*, 2014.

[66] A. Gupta and P. Mannem. From image annotation to image description. In *NIPS*, volume 7667, pages 196–204, 2012.

[67] S. Gupta and J. Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[68] A. Gupta et al. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*. 2008.

[69] S. Gupta et al. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[70] M. Hodosh et al. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 2013.

[71] Y. Jia et al. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

[72] C. JÃűrgensen. Attributes of images in describing tasks. *Information Processing & Management*, 34(2âĂŞ3):161 – 174, 1998.

[73] A. Karpathy et al. Deep visual-semantic alignments for generating image descriptions. *arXiv:1412.2306*, 2014.

[74] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. 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, 2014.

[75] P. Kingsbury and M. Palmer. From treebank to propbank. In *LREC*. Citeseer, 2002.

[76] C. Kong et al. What are you talking about? text-to-image coreference. In *CVPR*, 2014.

[77] I. Konstas and M. Lapata. Concept-to-text generation via discriminative reranking. In *ACL*, pages 369–378, 2012.

[78] A. Kotov and C. Zhai. Tapping into knowledge base for concept feedback: Leveraging conceptnet to improve search results for difficult queries. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 403–412, New York, NY, USA, 2012. ACM.

[79] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2973–2980, 2012.

[80] E. Krahmer and K. Van Deemter. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218, 2012.

[81] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016.

[82] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg. Baby talk: Understanding and generating simple image descriptions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1601–1608, 2011.

[83] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In *ACL*, pages 359–368, 2012.

[84] T. Kwiatkowski, E. Choi, Y. Artzi, and L. Zettlemoyer. Scaling semantic parsers with on-the-fly ontology matching. 2013.

[85] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[86] C. H. Lampert et al. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[87] A. Lazaridou et al. Is this a wampimuk? In *ACL*, 2014.

[88] D.-T. Le et al. Tuhoi: Trento universal human object interaction dataset. *V&L Net 2014*, 2014.

[89] T. Lei, Y. Zhang, R. Barzilay, and T. Jaakkola. Low-rank tensors for scoring dependency structures. Association for Computational Linguistics, 2014.

[90] J. Lei Ba, K. Swersky, S. Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4247–4255, 2015.

[91] D. B. Lenat, R. V. Guha, K. Pittman, D. Pratt, and M. Shepherd. Cyc: toward programs with common sense. *Communications of the ACM*, 33(8):30–49, 1990.

[92] L.-J. Li and L. Fei-Fei. What, where and who? Classifying events by scene and object recognition. In *ICCV*, pages 1–8. IEEE, 2007.

[93] L.-J. Li et al. What, where and who? classifying events by scene and object recognition. In *CVPR*, 2007.

[94] D. Lin and P. Pantel. Dirt@ sbt@ discovery of inference rules from text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–328. ACM, 2001.

[95] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. 2014.

[96] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015.

[97] T.-Y. Lin et al. Microsoft coco: Common objects in context. In *ECCV*. 2014.

[98] H. Liu and P. Singh. Conceptnet: A practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.

[99] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016.

[100] S. Maji et al. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011.

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

[102] J. Mao et al. Explain images with multimodal recurrent neural networks. *arXiv:1410.1090*, 2014.

[103] M. Marszalek et al. Actions in context. In *CVPR*, 2009.

[104] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proc. of the 2012 International Conference on Machine Learning*, 2012.

[105] K. McRae, G. S. Cree, M. S. Seidenberg, and C. Mcnorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559, 2005.

[106] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[107] G. A. Miller. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[108] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé, III. Midge: Generating image descriptions from computer vision detections. In *EACL*, pages 747–756, 2012.

[109] M. Mitchell, K. van Deemter, and E. Reiter. Natural reference to objects in a visual domain. In *Proceedings of the 6th International Natural Language Generation Conference*, INLG '10, pages 95–104, 2010.

[110] M. Mitchell, K. van Deemter, and E. Reiter. Generating expressions that refer to visible objects. In *Proceedings of NAACL-HLT*, pages 1174–1184, 2013.

[111] F. Och, C. Tillmann, and H. Ney. Improved alignment models for statistical machine translation. In *Proc. of the Joint Conf. of Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, 1999.

[112] F. J. Och. Minimum error rate training in statistical machine translation. In *ACL*, pages 160–167, 2003.

[113] V. Ordonez, J. Deng, Y. Choi, A. C. Berg, and T. Berg. From large scale image categorization to entry-level categories. In *International Conference on Computer Vision (ICCV)*, pages 2768–2775. IEEE, 2013.

[114] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2Text: Describing images using 1 million captioned photographs. In *NIPS*, pages 1143–1151, 2011.

[115] V. Ordonez, W. Liu, J. Deng, Y. Choi, A. C. Berg, and T. L. Berg. Predicting entry-level categories. *International Journal of Computer Vision*, pages 1–15, 2015.

[116] V. Ordonez et al. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

[117] M. Palmer. Semlink: Linking propbank, verbnet and framenet. In *GLC*, pages 9–15, 2009.

[118] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, 2002.

[119] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*, 2001.

[120] A. Rabinovich et al. Objects in context. In *ICCV*, 2007.

[121] D. A. Randell, Z. Cui, and A. G. Cohn. A spatial logic based on regions and connection. In *International Conference on Knowledge Representation and Reasoning*, 1992.

[122] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, 2010.

[123] M. Ren et al. Image question answering: A visual semantic embedding model and a new dataset. *arXiv preprint arXiv:1505.02074*, 2015.

[124] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[125] M. Ronchi and P. Perona. Describing common human visual actions in images. In *British Machine Vision Conference (BMVC)*, 2015.

[126] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2014.

[127] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 1–42, 2014.

[128] F. Sadeghi, S. K. Divvala, and A. Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Conference on Computer Vision and Pattern Recognition*, pages 1456–1464, 2015.

[129] S. Schoenmackers, O. Etzioni, D. S. Weld, and J. Davis. Learning first-order horn clauses from web text. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1088–1098, 2010.

[130] G. Sharma, F. Jurie, C. Schmid, et al. Expanded parts model for human attribute and action recognition in still images. In *CVPR*, 2013.

[131] E. Shutova. Automatic metaphor interpretation as a paraphrasing task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 1029–1037, 2010.

[132] E. Shutova. Models of metaphor in nlp. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 688–697, 2010.

[133] C. Silberer, V. Ferrari, and M. Lapata. Models of semantic representation with visual attributes. In *ACL*, pages 572–582, 2013.

[134] C. Silberer and M. Lapata. Grounded models of semantic representation. In *EMNLP*, July 2012.

[135] C. Silberer et al. Grounded models of semantic representation. In *EMNLP*, 2012.

[136] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*,

abs/1409.1556, 2014.

[137] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[138] K. Soomro et al. Ucf101: A dataset of 101 human actions classes from videos in the wild. 2012.

[139] R. Speer and C. Havasi. Conceptnet 5: A large semantic network for relational knowledge. In *The PeopleâĂŹs Web Meets NLP*, pages 161–176. Springer, 2013.

[140] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy. Approaching the symbol grounding problem with probabilistic graphical models. *AI magazine*, 32(4):64–76, 2011.

[141] A. Torralba, B. C. Russell, and J. Yuen. LabelMe: Online image annotation and applications. *Proceedings of the IEEE*, 98(8):1467–1484, 2010.

[142] A.-M. Tousch, S. Herbin, and J.-Y. Audibert. Semantic hierarchies for image annotation: A survey. *Pattern Recognition*, 45(1):333 – 345, 2012.

[143] L. Vanderwende. Volunteers created the web. In *Proceedings of the 2005 AAAI Spring Symposium, Knowledge Collection from Volunteer Contributors*. American Association for Artificial Intelligence, March 2005.

[144] R. Vedantam et al. Cider: Consensus-based image description evaluation. *arXiv:1411.5726*, 2014.

[145] O. Vinyals et al. Show and tell: A neural image caption generator. *arXiv:1411.4555*, 2014.

[146] S. Yang, Q. Gao, C. Liu, C. Xiong, S.-C. Zhu, and Y. J. Chai. Grounded semantic role labeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 149–159. Association for Computational Linguistics, 2016.

[147] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *Empirical Methods in Natural Language Processing*, 2011.

[148] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and L. Fei-Fei. Action recognition by learning bases of action attributes and parts. In *ICCV*, Barcelona, Spain, November 2011.

[149] B. Yao et al. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010.

[150] B. Yao et al. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.

[151] B. Yao et al. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011.

[152] M. Yatskar, L. Zettlemoyer, and A. Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Conference on Computer Vision and Pattern Recognition*, 2016.

[153] M. Yatskar et al. See no evil, say no evil: Description generation from densely labeled images. **SEM*, 2014.

[154] L. e. a. Yu. Visual madlibs: Fill in the blank image generation and question answering. *arXiv preprint*

arXiv:1506.00278, 2015.

[155] A. Zaenen, J. Carletta, G. Garretson, J. Bresnan, A. Koontz-Garboden, T. Nikitina, M. C. O'Connor, and T. Wasow. Animacy encoding in English: why and how. In *ACL Workshop on Discourse Annotation*, pages 118–125, 2004.

[156] O. F. Zaidan. Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91:79–88, 2009.

[157] J. M. Zelle and R. J. Mooney. Learning to parse database queries using inductive logic programming. In *AAAI/IAAI, Vol. 2*, 1996.

[158] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[159] L. S. Zettlemoyer and M. Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *UAI*, 2005.

[160] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015.

[161] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about Object Affordances in a Knowledge Base Representation. In *European Conference on Computer Vision*, 2014.

[162] Y. Zhu et al. Reasoning about object affordances in a knowledge base representation. In *ECCV*. 2014.

[163] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *CVPR*, 2013.