

# Computational Methods for Integrating Vision and Language

# Synthesis Lectures on Computer Vision

## Editor

**Gérard Medioni**, *University of Southern California*

**Sven Dickinson**, *University of Toronto*

Synthesis Lectures on Computer Vision is edited by Gérard Medioni of the University of Southern California and Sven Dickinson of the University of Toronto. The series publishes 50- to 150 page publications on topics pertaining to computer vision and pattern recognition. The scope will largely follow the purview of premier computer science conferences, such as ICCV, CVPR, and ECCV. Potential topics include, but not are limited to:

- Applications and Case Studies for Computer Vision
- Color, Illumination, and Texture
- Computational Photography and Video
- Early and Biologically-inspired Vision
- Face and Gesture Analysis
- Illumination and Reflectance Modeling
- Image-Based Modeling
- Image and Video Retrieval
- Medical Image Analysis
- Motion and Tracking
- Object Detection, Recognition, and Categorization
- Segmentation and Grouping
- Sensors
- Shape-from-X
- Stereo and Structure from Motion
- Shape Representation and Matching

- Statistical Methods and Learning
- Performance Evaluation
- Video Analysis and Event Recognition

### Computational Methods for Integrating Vision and Language

Kobus Barnard

2016

### Background Subtraction: Theory and Practice

Ahmed Elgammal

2014

### Vision-Based Interaction

Matthew Turk and Gang Hua

2013

### Camera Networks: The Acquisition and Analysis of Videos over Wide Areas

Amit K. Roy-Chowdhury and Bi Song

2012

### Deformable Surface 3D Reconstruction from Monocular Images

Mathieu Salzmann and Pascal Fua

2010

### Boosting-Based Face Detection and Adaptation

Cha Zhang and Zhengyou Zhang

2010

### Image-Based Modeling of Plants and Trees

Sing Bing Kang and Long Quan

2009

Copyright © 2016 by Morgan & Claypool

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Computational Methods for Integrating Vision and Language

Kobus Barnard

[www.morganclaypool.com](http://www.morganclaypool.com)

ISBN: 9781608451128      paperback

ISBN: 9781608451135      ebook

DOI 10.2200/S00705ED1V01Y201602COV007

A Publication in the Morgan & Claypool Publishers series

*SYNTHESIS LECTURES ON COMPUTER VISION*

Lecture #7

Series Editors: Gérard Medioni, *University of Southern California*

Sven Dickinson, *University of Toronto*

Series ISSN

Print 2153-1056    Electronic 2153-1064

# Computational Methods for Integrating Vision and Language

Kobus Barnard  
University of Arizona

*SYNTHESIS LECTURES ON COMPUTER VISION #7*



MORGAN & CLAYPOOL PUBLISHERS

## ABSTRACT

Modeling data from visual and linguistic modalities together creates opportunities for better understanding of both, and supports many useful applications. Examples of dual visual-linguistic data includes images with keywords, video with narrative, and figures in documents. We consider two key task-driven themes: translating from one modality to another (e.g., inferring annotations for images) and understanding the data using all modalities, where one modality can help disambiguate information in another. The multiple modalities can either be essentially semantically redundant (e.g., keywords provided by a person looking at the image), or largely complementary (e.g., meta data such as the camera used). Redundancy and complementarity are two endpoints of a scale, and we observe that good performance on translation requires some redundancy, and that joint inference is most useful where some information is complementary.

Computational methods discussed are broadly organized into ones for simple keywords, ones going beyond keywords toward natural language, and ones considering sequential aspects of natural language. Methods for keywords are further organized based on localization of semantics, going from words about the scene taken as whole, to words that apply to specific parts of the scene, to relationships between parts. Methods going beyond keywords are organized by the linguistic roles that are learned, exploited, or generated. These include proper nouns, adjectives, spatial and comparative prepositions, and verbs. More recent developments in dealing with sequential structure include automated captioning of scenes and video, alignment of video and text, and automated answering of questions about scenes depicted in images.

## KEYWORDS

vision, language, loosely labeled data, correspondence ambiguity, auto-annotation, region labeling, multimodal translation, cross-modal disambiguation, image captioning, video captioning, affective visual attributes, aligning visual and linguistic data, auto-illustration, visual question answering

# Contents

	<b>Acknowledgments</b> .....	<b>xiii</b>
	<b>Figure Credits</b> .....	<b>xv</b>
<b>1</b>	<b>Introduction</b> .....	<b>1</b>
1.1	Redundant, Complementary, and Orthogonal Multimodal Data .....	3
1.1.1	Multimodal Mutual Information .....	6
1.1.2	Complementary Multimodal Information .....	7
1.2	Computational Tasks .....	9
1.2.1	Multimodal Translation .....	11
1.2.2	Integrating Complementary Multimodal Data and Cross Modal Disambiguation .....	16
1.2.3	Grounding Language with Sensory Data .....	17
1.3	Multimodal Modeling .....	18
1.3.1	Discriminative Methods .....	20
1.4	Multimodal Inference—Applications to Computational Tasks .....	20
1.4.1	Region Labeling with a Concept Model .....	20
1.4.2	Cross-modal Disambiguation—Region Labeling with Image Keywords .....	21
1.4.3	Cross-modal Disambiguation—Word Sense Disambiguation with Images .....	21
1.5	Learning from Redundant Representations in Loosely Labeled Multimodal Data .....	22
1.5.1	Resolving Region-label Correspondence Ambiguity .....	23
1.5.2	Data Variation and Semantic Grouping .....	24
1.5.3	Simultaneously Learning Models and Reducing Correspondence Ambiguity .....	25
<b>2</b>	<b>The Semantics of Images and Associated Text</b> .....	<b>27</b>
2.1	Lessons from Image Search .....	28
2.1.1	Content-based Image Retrieval (CBIR) .....	28
2.2	Images and Text as Evidence About the World .....	30

2.3	Affective Attributes of Images and Video . . . . .	31
2.3.1	Emotion Induction from Images and Video . . . . .	32
2.3.2	Inferring Emotion in People Depicted in Images and Videos . . . . .	33
<b>3</b>	<b>Sources of Data for Linking Visual and Linguistic Information . . . . .</b>	<b>35</b>
3.1	WordNet for Building Semantic Visual-linguistic Data Sets . . . . .	35
3.2	Visual Data with a Single Objective Label . . . . .	36
3.3	Visual Data with a Single Subjective Label . . . . .	38
3.4	Visual Data with Keywords or Object Labels . . . . .	38
3.4.1	Localized Labels . . . . .	39
3.4.2	Semantic Segmentations with Labels . . . . .	40
3.5	Visual Data with Descriptions . . . . .	41
3.6	Image Data with Questions and Answers . . . . .	43
<b>4</b>	<b>Extracting and Representing Visual Information . . . . .</b>	<b>45</b>
4.1	Low-level Features . . . . .	46
4.1.1	Color . . . . .	46
4.1.2	Edges . . . . .	47
4.1.3	Texture . . . . .	47
4.1.4	Characterizing Neighborhoods Using Histograms of Oriented Gradients . . . . .	48
4.2	Segmentation for Low-level Spatial Grouping . . . . .	50
4.3	Representation of Regions and Patches . . . . .	50
4.3.1	Visual Word Representations . . . . .	51
4.4	Mid-level Representations for Images . . . . .	51
4.4.1	Artificial Neural Network Representations . . . . .	51
4.5	Object Category Recognition and Detection . . . . .	52
<b>5</b>	<b>Text and Speech Processing . . . . .</b>	<b>55</b>
5.1	Text Associated with Audiovisual Data . . . . .	56
5.2	Text Embedded Within Visual Data . . . . .	56
5.3	Basic Natural Language Processing . . . . .	57
5.4	Word Sense Disambiguation . . . . .	57
5.5	Online Lexical Resource for Vision and Language Integration . . . . .	58
5.5.1	WordNet . . . . .	58
5.5.2	Representing Words by Vectors . . . . .	61



<b>6</b>	<b>Modeling Images and Keywords</b>	<b>63</b>
6.1	Scene Semantic–Keywords for Entire Images	63
6.2	Localized Semantics–Keywords for Regions	64
6.3	Generative Models with Independent Multi-modal Concepts	66
6.3.1	Notational Preliminaries	66
6.3.2	Semantic Concepts with Multi-model Evidence	66
6.3.3	Joint Modeling of Images and Keywords (PWRM and IRCM)	68
6.3.4	Inferring Image Keywords and Region Labels	71
6.3.5	Learning Multi-modal Concept Models from Loosely Labeled Data	73
6.3.6	Evaluation of Region Labeling and Image Annotation	77
6.4	Translation Models	78
6.4.1	Notational Preliminaries (continuing §6.3.1)	78
6.4.2	A Simple Region Translation Model (RTM)	79
6.4.3	Visual Translation Models for Broadcast Video	81
6.4.4	A Word Translation Model (WTM)	82
6.4.5	Supervised Multiclass Labeling (SML)	82
6.4.6	Discriminative Models for Translation	86
6.5	Image clustering and Interdependencies Among Concepts	88
6.5.1	Region Concepts with Image Categories (CIRCM)	88
6.5.2	Latent Dirichlet Allocation (LDA)	90
6.5.3	Multiclass Supervised LDA (sLDA) with Annotations	92
6.6	Segmentation, Region Grouping, and Spatial Context	92
6.6.1	Notational Preliminaries (continuing §6.3.1 and §6.4.1)	95
6.6.2	Random Fields for Representing Image Semantics	95
6.6.3	Joint Learning of Translation and Spatial Relationships	97
6.6.4	Multistage Learning and Inference	99
6.6.5	Dense CRFs for General Context	101
6.6.6	Dense CRFs for Multiple Pairwise Relationships	102
6.6.7	Multiscale CRF (mCRF)	103
6.6.8	Relative Location Prior with CRFs	103
6.6.9	Encoding Spatial Patterns into the Unary Potentials with Texture-layout Features	104
6.6.10	Discriminative Region Labeling with Spatial and Scene Information	105
6.6.11	Holistic Integration of Appearance, Object Detection, and Scene Type	106
6.7	Image Annotation Without Localization	106
6.7.1	Nonparametric Generative Models	107

	6.7.2 Label Propagation . . . . .	109
<b>7</b>	<b>Beyond Simple Nouns . . . . .</b>	<b>111</b>
7.1	Reasoning with Proper Nouns . . . . .	112
7.1.1	Names and Faces in the News . . . . .	112
7.1.2	Linking Action Verbs to Pose—Who is Doing What? . . . . .	114
7.1.3	Learning Structured Appearance for Named Objects . . . . .	115
7.2	Learning and Using Adjectives and Attributes . . . . .	116
7.2.1	Learning Visual Attributes for Color Names . . . . .	117
7.2.2	Learning Complex Visual Attributes for Specific Domains . . . . .	118
7.2.3	Inferring Emotional Attributes for Images . . . . .	118
7.2.4	Inferring Emotional Attributes for Video Clips . . . . .	120
7.2.5	Sentiment Analysis in Consumer Photographs and Videos . . . . .	120
7.2.6	Extracting Aesthetic Attributes for Images . . . . .	121
7.2.7	Addressing Subjectivity . . . . .	122
7.3	Noun-Noun Relationships—Spatial Prepositions and Comparative Adjectives . . . . .	122
7.3.1	Learning about Preposition Use in Natural Language . . . . .	123
7.4	Linking Visual Data to Verbs . . . . .	124
7.5	Vision Helping Language Understanding . . . . .	125
7.5.1	Using Vision to Improve Word Sense Disambiguation . . . . .	126
7.5.2	Using Vision to Improve Coreference Resolution . . . . .	126
7.5.3	Discovering Visual-semantic Senses . . . . .	126
7.6	Using Associated Text to Improve Visual Understanding . . . . .	127
7.6.1	Using Captions to Improve Semantic Image Parsing (Cardinality and Prepositions) . . . . .	127
7.7	Using World Knowledge from Text Sources for Visual Understanding . . . . .	127
7.7.1	Seeing What Cannot be Seen? . . . . .	128
7.7.2	World Knowledge for Training Large-scale Fine-grained Visual Models . . . . .	129
<b>8</b>	<b>Sequential Structure . . . . .</b>	<b>131</b>
8.1	Automated Image and Video Captioning . . . . .	131
8.1.1	Captioning by Reusing Existing Sentences and Fragments . . . . .	131
8.1.2	Captioning Using Templates, Schemas, or Simple Grammars . . . . .	132
8.1.3	Captioning Video Using Storyline Models . . . . .	134
8.1.4	Captioning with Learned Sentence Generators . . . . .	134

8.2	Aligning Sentences with Images and Video .....	136
8.3	Automatic Illustration of Text Documents .....	137
8.4	Visual Question and Answering .....	138
<b>A</b>	<b>Additional Definitions and Derivations .....</b>	<b>141</b>
A.1	Basic Definitions from Probability and Information Theory .....	141
A.2	Additional Considerations for Multimodal Evidence for a Concept .....	142
A.3	Loosely Labeled vs. Strongly Labeled Data .....	144
A.4	Pedantic Derivation of Equation (6.13) .....	148
A.5	Derivation of the EM Equations for the Image Region Concept Model (IRCM) .....	150
	<b>Bibliography .....</b>	<b>155</b>
	<b>Author's Biography .....</b>	<b>211</b>



# Acknowledgments

The idea for this book was conceived by Sven Dickinson, who convinced me to undertake the task an embarrassingly long time ago. I am grateful for both Sven's and Gérard Medioni's support and patience. I also appreciate the efforts of two reviewers who provided insightful comments on the manuscript that led to numerous improvements. I am also grateful for Mihai Surdeanu's comments on Chapter 5, and Emily Butler's comments on Sections 2.2 and 7.2. Finally, I thank the editorial staff at Morgan&Claypool, including Diane Cerra, Samantha Draper, C.L. Tondo, Deb Gabriel, and Sara Kreisman and her team, for transforming the manuscript into an actual book.

Kobus Barnard  
February 2016



## Figure Credits

- Figure 1.3c** Sub-image was derived from the Corel™ image data set as permitted within the terms of the user agreement. Copyright © Corel Corporation, all rights reserved.
- Figure 1.4** From the Corel™ image data set as permitted within the user agreement. Copyright © Corel Corporation, all rights reserved.
- Figure 1.6a** From the Corel™ image data set as permitted within the user agreement. Copyright © Corel Corporation, all rights reserved.
- Figure 1.6b** From: K. Barnard, P. Duygulu, N. d. Freitas, D. Forsyth, D. Blei, and M. I. Jordan, “Matching Words and Pictures,” *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003. Used with permission.
- Figure 1.6c** From: G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, “Babytalk: Understanding and generating simple image descriptions,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, pp. 2891–2903, 2013. Copyright © 2013 IEEE. Used with permission.
- Figure 1.7b** From the Corel™ image data set as permitted within the user agreement. Copyright © Corel Corporation, all rights reserved.
- Figure 6.4** From: K. Barnard and D. Forsyth, “Learning the semantics of words and pictures,” *Proc. International Conference on Computer Vision*, pp. II: 408–415, 2001. Copyright © 2001 IEEE. Used with permission.
- Figure 6.6** From: X. He, R. S. Zemel, and M. Carreira-Perpinan, “Multiscale conditional random fields for image labeling,” *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. II: 695–702 Vol. 2, 2004. Copyright © 2004 IEEE. Used with permission.
- Figure 7.1** From: T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth, “Who’s in the Picture,” *Proc. NIPS*, 2004. Used with permission.

- Figure 7.2** From: L. Jie, B. Caputo, and V. Ferrari, “Who’s doing what: Joint modeling of names and verbs for simultaneous face and pose annotation,” *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2009. Used with permission.
- Figure 7.3a** From: K. Yanai and K. Barnard, “Image region entropy: A measure of ‘visualness’ of web images associated with one concept,” *Proc. ACM Multimedia*, Singapore, pp. 419–422, 2005. Copyright © 2005 ACM. Used with permission.
- Figure 7.3b** From: K. Yanai and K. Barnard, “Image region entropy: A measure of ‘visualness’ of web images associated with one concept,” *Proc. ACM Multimedia*, Singapore, pp. 419–422, 2005. Copyright © 2005 ACM. Used with permission.
- Figure 7.3c** From: V. Ferrari and A. Zisserman, “Learning visual attributes,” *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2007. Used with permission.
- Figure 7.4** From: C. R. Dawson, J. Wright, A. Rebguns, M. V. Escarcega, D. Fried, and P. R. Cohen, “A generative probabilistic framework for learning spatial language,” *Proc. IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics*, pp. 1–8, 2013. Copyright © 2013 IEEE. Used with permission.
- Figure 8.1a** From: A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” *Computer Vision and Pattern Recognition*, 2015. Copyright © 2015 IEEE. Used with permission.
- Figure 8.1b** From: S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, “Sequence to sequence - Video to text,” *Proc. ICCV*, 2015. Copyright © 2015 IEEE. Used with permission.



## CHAPTER 1

# Introduction




Knowledge about the world comes to us through multiple modalities, including our primary senses, and also abstractions such as illustrations and language. The different modalities reinforce and complement each other, and provide for more effective understanding of the world around us, provided that we can integrate the information into a common representation or abstract understanding. Similarly, information from multiple modalities can be exploited by intelligent computational systems. This book develops computational approaches for linking and combining the modalities implied by visual and linguistic information. In particular, automatically determining relationships between these modalities can be applied to providing better access to data, training systems to extract semantic content from either visual and linguistic data, and develop machine representations that are indicative of higher level semantics and thus can support intelligent machine behavior.

To make this endeavor concrete, we will develop our discussion in the context of image or video data that has associated text or speech. Figures 1.1 and 1.2 show six examples from this domain. Figure 1.1a shows an image of a mountain goat that has associated keywords provided a human annotator to enable searching for the image by text query. Figure 1.1b shows an image where we imagine a teacher is providing an explicit reference to what is in the image, which is achieved in this case if the learner assumes that the moving object is what is pertinent. Figure 1.1c shows an image with a caption that provides both content-related information that could be inferred from the photograph and meta-data that is meant to complement the visual information. Finally, Figure 1.2 shows three examples where images are chosen to provide key information that is spatial in nature, and thus far more efficient to provide with a figure.

Notice that in Figure 1.1a, the human annotator only had access to the image and their general knowledge about the world. Thus, their annotation is potentially derivable from the image. Of course, the image contains other information that they chose not to write about. The situation in Figure 1.1b is even more restricted where we assume the human teacher is purposefully providing text relevant to both the visual context and the focus of attention of the learner. By contrast, in Figure 1.1c, the annotation specifically provides information that is difficult or impossible to extract from the image.

In these three examples there is overlap in the two sources of information as “mountain goat” is inferable from either the images or the texts, but all the images have information that the associated text does not. Conversely, in (b) the text provides information a learner might not

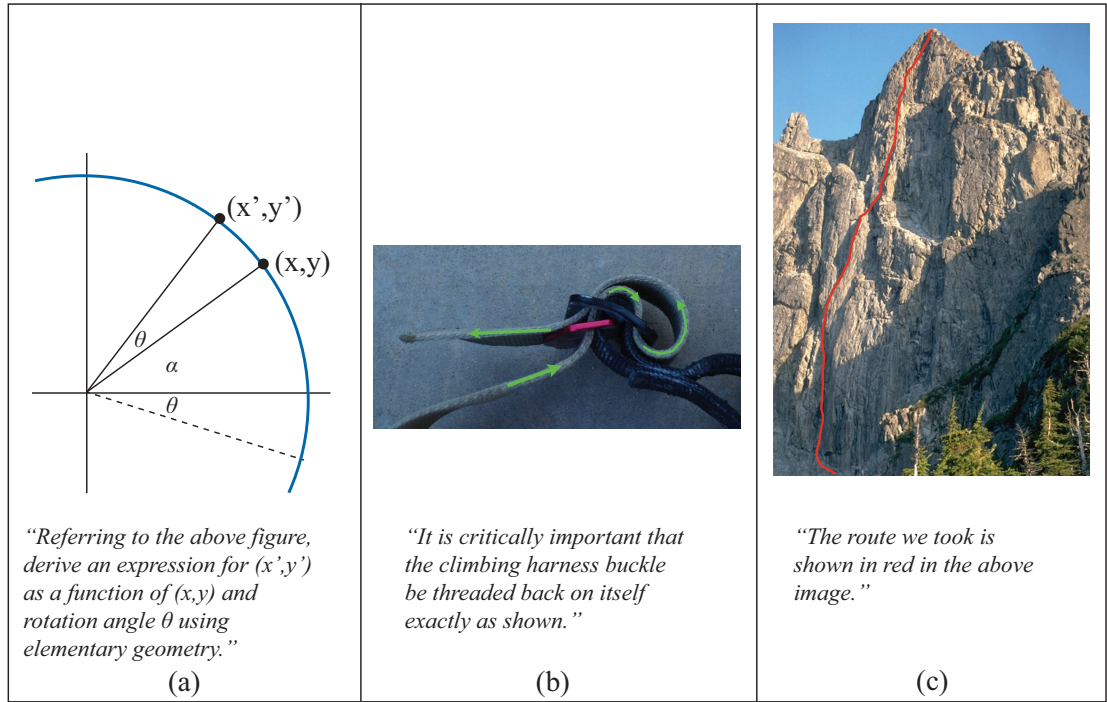
2 1. INTRODUCTION

Visual Data			
Linguistic Data	MOUNTAIN GOAT ROCK FOILAGE SKY	"Look at the mountain goat!"	A mountain goat in the Tantalus Range B.C. showing signs of shedding its winter coat. Photo taken by Kobus Barnard in the summer of 2011.
	(a)	(b)	(c)

**Figure 1.1:** Three examples of aligned visual and linguistic information. (a) Keywords for an image that could be provided by an annotator based on the image content alone. (b) An utterance that refers to the part of the scene that is moving, that could be used to learn “mountain goat” if the learner know the other words in the sentence. Motion enables effective grouping the goat pixels together in space and time (i.e., tracking the movement). This is in contrast to the other examples in this panel, where the parts of the image (if any) that the words refer to are not immediately available. (c) Information provided by the photographer.

know, and in (c) the text has information not available from the image. Further, if the interpreter of the images is a computer program instead of a human, all the texts likely add information.

In (c) both the visual information and the linguistic information include semantic components that can be dually represented (e.g., the scene contains a mountain goat), and components that are largely restricted to one modality or the other. In the case of the visual information this includes the myriad of visual details such as particular textures and patterns in the rocks, meadow, etc. In the case of linguistic information, while there are can be minor relationships between visual information and time, place, photographer, and cropping, for the most part we do not expect such meta information to be dually represented. Dual representations may have redundant and/or complementary aspects. For example, for someone familiar with mountain goats, “mountain goat” in the caption is redundant. For a computer program that confuses off-white fur coats with cumulus clouds, the two sources of information about the same thing can help each other, reducing the ambiguity about a detected fluffy white region (Figure 1.4 expands on this issue).

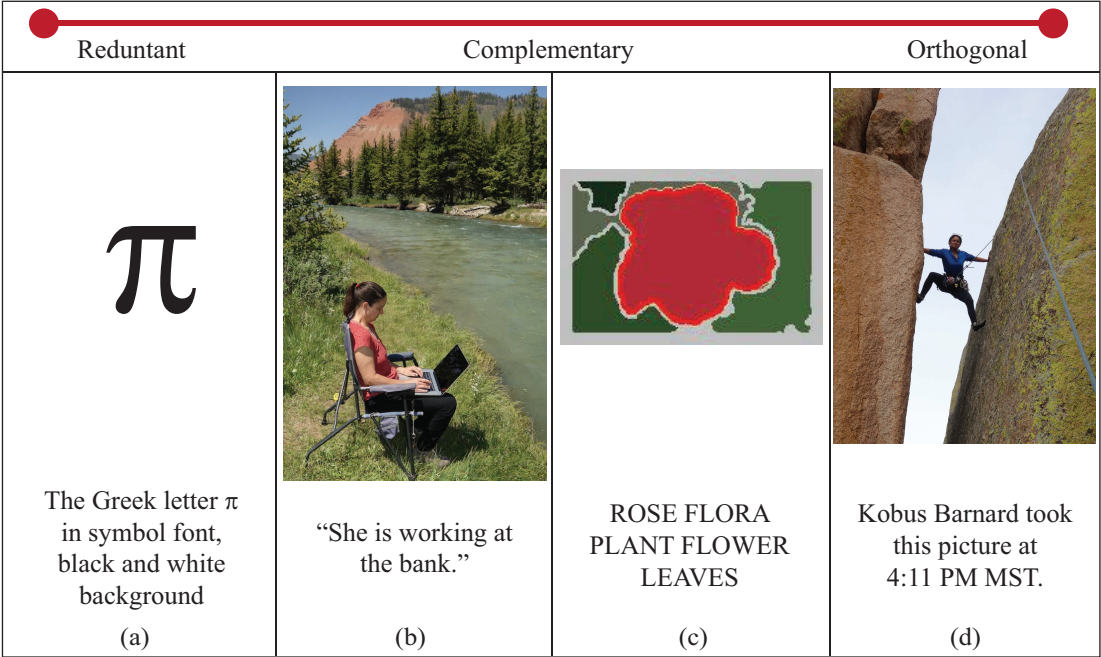


**Figure 1.2:** Examples of text with associated images, where the text drives the need for a figure. This in contrast to Figure 1.1 where text was created in response to what was imaged. In the examples in this figure, spatial information is crucial to what is being communicated and it is more efficient to provide an illustration.

## 1.1 REDUNDANT, COMPLEMENTARY, AND ORTHOGONAL MULTIMODAL DATA

These examples suggest that the relation between visual and linguistic information can be considered with respect to two extremes, ranging from largely intersecting to largely disjoint. Informally we can ask the extent that two sources of information are: (1) informative about each other (*redundant*), through (2) informative about the same thing but not entirely redundant (*complementary*), to (3) completely independent from each other (*orthogonal*). Hence, we consider a continuum from complete redundancy to no interaction among modalities (Figure 1.3).

We will consider the continuum as being *relative to the capabilities of the vision and language processing systems* at our disposal. For example, while a human annotator will have no problem attaching the keyword “rose” to the image in Figure 1.3c, a machine vision system based on the color of blobs might be uncertain whether the red blob should be associated with a red rose or a



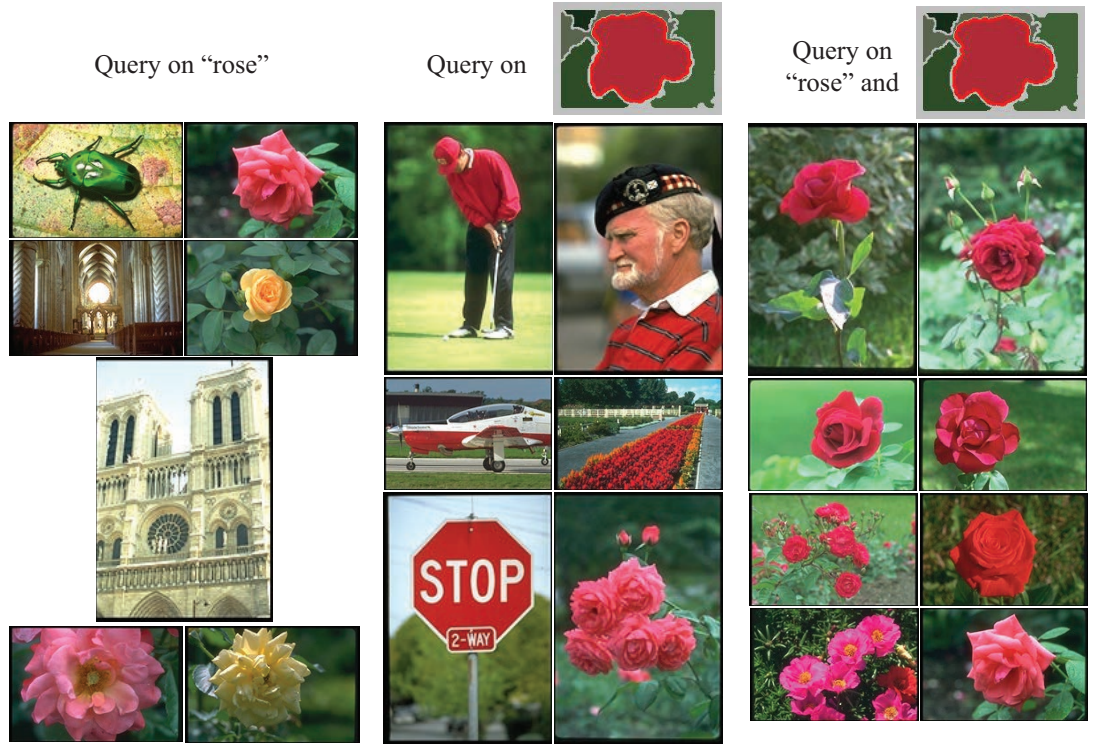
**Figure 1.3:** Given a semantic subspace of interest (e.g., what is in the figure or scene), the degree of overlap between the visual information and the linguistic information lies on a continuum ranging from completely redundant to completely orthogonal. For our purposes we are most interested in information that we can extract with a computer program. Current computer vision and natural language processing are both limited, but often we can alleviate the deficiencies by considering both when the information they provide complement each other. For example, in (b), “bank” is ambiguous as it has multiple senses, but the image removes the ambiguity. In (c), a garden image is represented with colored regions to illustrate the information available to a particular computer program. In general, the red region in the middle has high probability of being a stop sign, but the addition of the keyword can push the probability toward the correct label for that region (rose)—see also Figure 1.4. Sub-image (c) was derived from the Corel<sup>TM</sup> image data set as permitted within the terms of the user agreement.

stop sign. Hence, for this simple machine vision system, the text annotation provides significant additional information (see Figure 1.4). By contrast, from the perspective of the human annotator, the text annotation is derivable from the image, and does not provide more information.

To make these notions more precise, we will take a Bayesian statistical approach where all entities are associated with random variables,<sup>1</sup> and we think in terms of modeling and inference

<sup>1</sup>I will be making use of basic probability and statistics, and elementary concepts from information theory. There are many resources for these topics. Useful refreshers oriented toward machine learning include Bishop [87, Ch. 1] and Koller and





**Figure 1.4:** A simple experiment to demonstrate how keywords and visual features constrain the semantic subspace in complementary ways using the Corel<sup>TM</sup> image database (images are reproduced within the terms of the user agreement). The text query “rose” returns rose bugs, rose windows in cathedral images, as well as some flowers. The query for similar images to one of a red rose, based on color and texture returns a variety of images with prominent red regions. Finally, a query on the conjunction of the two returns images that are both semantically associated with rose, and have regions that are visually similar to the red rose in the query image.

with respect to the joint probability distribution over those variables. In general, we anticipate reasoning about a subset of these variables, and the result will typically be uncertain, as represented by probability distributions. Mathematically, handling the uncertainty is the domain of the probability calculus.

We are particularly interested in probability distributions where the random variables from multiple modalities share information in non-trivial ways. The random variables for the problem domains of interest can include (1) potentially observable data from one or more modalities,

Friedman [331, Ch. 2]. A recommended comprehensive text on information theory, going far beyond what I will use, is Cover and Thomas [147].

## 6 1. INTRODUCTION

and (2) additional, possibly latent, variables representing semantics or other abstract information about the world, which often conceptually span modalities. Such non-observed variables can differ in the degree to which we attach an explicit interpretation to them. For example, in a hand-built system, such variables might have very specific interpretations provided by the builder. On the other hand, in a system based on unsupervised clustering, non-observed variables might represent latent multi-modal clusters. For many tasks, the interpretation of such clusters does not matter. On the other hand, if we believe that abstractions such as semantic concepts are a good explanation of the multimodal structure of the data, then we might be able to learn such concepts. However, in this scenario, evaluation of the learned concepts is then post hoc. Notice that if we were to label the data with concepts, then the concepts would now be considered potentially observable.

Given a random variable representation for a particular domain, we can approach multiple tasks using statistical inference. For a given task we can organize our variables into three roles: (1) evidence—variables that we observe (input); (2) targets of inference—variables that we have chosen to reason about (output); and (3) variables that are neither of these. As a concrete example, we might model joint visual and linguistic data with multimodal clusters. If our task is to infer appropriate words for images, then the observed image data is the evidence, appropriate words (potentially observable) are the targets of inference, and the latent clusters (not observed) support linking images and appropriate words.

### 1.1.1 MULTIMODAL MUTUAL INFORMATION

For what follows, the reader may want to refer to Appendix A.1 for basic definitions.

We now consider the continuum more formally, focusing on the shared information between modalities. For simplicity, I will assume that all variables of a modality are considered together.<sup>2</sup> Also, to begin, I will restrict attention to observable data (extended shortly, §1.1.2). With these simplifications, the defining property of the continuum amounts to the degree of independence between the two modalities.

On the left extreme of Figure 1.3, the information provided by at least one of the modalities could be predicted from the other. As a very simple example, suppose that we have two images, one clearly of a dog and one clearly of a cat. Also suppose we have a vocabulary consisting of “dog” and “cat,” and that we can classify the images accurately. Then, choosing the dog image at random, we necessarily get the “dog” label, and conversely, choosing the “dog” label at random automatically leads to the dog image. Of course, the cat image and label behave similarly. We get

---

<sup>2</sup>One can easily consider data components within a modality with the same analysis. For example, we can consider color and texture information as two different modalities that share some degree of information with each other and/or other modalities, despite both being visual.

(for dogs):

$$\begin{aligned} p(\text{dog}, \text{"dog"}) &= p(\text{dog}) = p(\text{"dog"}) = \frac{1}{2}, \\ \text{and thus } \frac{p(\text{dog}) p(\text{"dog"})}{p(\text{dog}, \text{"dog"})} &= p(\text{dog}) = \frac{1}{2}. \end{aligned} \quad (1.1)$$

The mutual information is then given by:

$$\begin{aligned} I[\text{label}, \text{image}] &= - \sum_{x \in (\text{"dog"}, \text{"cat"})} \sum_{y \in (\text{dog}, \text{cat})} p(x, y) \log_2 \left( \frac{p(x) p(y)}{p(x, y)} \right) \\ &= - \left( \frac{1}{2} \bullet \log_2 \left( \frac{1}{2} \right) + \frac{1}{2} \bullet \log_2 \left( \frac{1}{2} \right) \right), \\ &= 1 \end{aligned} \quad (1.2)$$

which is distinctly greater than zero, reflecting the fact that the two data sources share information and that they are not independent.

On the other hand, for the right extreme of Figure 1.3, the information from each of the two modalities is independent. Observing one of them does not tell us anything about the other. For example, data such as the time of day, camera used, or photographer name, typically provides very little information about image content, and knowing the values of these variables only slightly changes our estimate for the distribution of the object in the center of the image. Notationally, this means that

$$p(\text{object} | \text{time}, \text{camera}, \text{photographer}) \cong p(\text{object}), \quad (1.3)$$

which is the definition that *object* is independent of the three other variables. In terms of information theory, the mutual information between the two variable sets is close to zero. Specifically, if we denote  $\text{meta} = (\text{time}, \text{camera}, \text{photographer})$  then  $p(\text{object}, \text{meta}) = p(\text{object}) p(\text{meta})$  and

$$\begin{aligned} I[\text{object}, \text{meta}] &= \sum p(\text{object}, \text{meta}) \log_2 \left( \frac{p(\text{object}, \text{meta})}{p(\text{object}) p(\text{meta})} \right) \\ &\cong \sum p(\text{object}, \text{meta}) \log_2(1) \\ &= 0. \end{aligned} \quad (1.4)$$

### 1.1.2 COMPLEMENTARY MULTIMODAL INFORMATION

The continuum would be of limited interest if we only used it to represent the degree of independence between two observable variable sets. We now consider adding random variables that represent abstractions such as semantics, which are informed by multiple modalities. It is an interesting empirical fact that different modalities can often work together in a complementary fashion, which is tied to the fact that multimodal data is common in documents. We can view

## 8 1. INTRODUCTION

abstract semantics as explaining the partial dependence (non negligible mutual information) between observables. However, we can also choose to be agnostic as to any such explanation, as long as the performance on a specific task is satisfactory.

To make the discussion more concrete, consider a random variable over semantic concepts,  $c$ , within scenes. For example, a concept could be that of a red rose. We will reason about which concept is most applicable for a particular image region,<sup>3</sup> and we will represent our uncertain understanding of which concepts are promising with a probably distribution over the finite set of semantic concepts. Note that while we often use text to indicate semantics, semantic entities are different from the observed text, which only gives us a partial indication of meaning. This will manifest in a number of ways in this book, including the two following. First, words in observed text are often ambiguous because words often have multiple meanings referred to as *senses*. For example, “bank” has a number of meanings including a financial institution, or a slope, or an edge. To distinguish senses of words, it is common to append a sense number to the word as in “bank\_1” or “bank\_2.”<sup>4</sup> Second, words in text may be incorrect or misleading if they are naively expected to link to visual data, especially if they are extracted from captions using imperfect language processing. For example, if we simply extract nouns from “marmots are often seen in mountain goat terrain,” then we might expect that the image contains depictions for both a marmot and a mountain goat, which is possible, but not likely the intent of the caption.

To simplify reasoning about semantic concepts given multiple data sources, let us suppose that the observables ( $A$ ,  $B$ ,  $C$ , ...) are conditionally independent, given the concept,  $c$ . We are then hopeful for two properties. First, we would like each data source to be informative about  $c$ , and thus have the potential to reduce our uncertainty about it. Second, we would like different data sources to reduce the uncertainty differently so that they all can improve our estimates of what is in the scene. In the general case, this means that any pair of data sources,  $A$  and  $B$ , tends to be in the complementary (middle) part of the scale in Figure 1.3. In particular, on the left extreme of the continuum, one of the variables (say  $A$ ) implies the other ( $B$ ). Hence, knowing  $A$  means that we know  $B$ , and hence subsequently learning  $B$  does not provide any more information. Thus, if all variables are helpful when considered together, we are not operating in the left extreme.

Analyzing the right extreme is more challenging. One might expect that if  $A$  and  $B$  are both informative about  $c$ , then they must be informative about each other (and thus cannot be independent). However, this is not always the case as they may be informative about different aspects of  $c$ . Hence, in this case, complementary data can also be independent. Briefly digressing, the contrapositive of this (informal) proposition claims that if  $A$  and  $B$  are independent, and we know one of them (say  $A$ ) then the other ( $B$ ) would not tell us anything more about  $c$ . Formally, this says that  $c$  and  $B$  can be conditionally independent given  $A$ . Then we would have

<sup>3</sup>We could also consider the probability that  $K$  labels are the  $K$ -best for the image considered as a whole (more complicated), or the probability that each label is independently appropriate for the image (simpler, but less illustrative, as it amounts to a collection of separate simple cases).

<sup>4</sup>This is the convention followed by WordNet [30, 210, 425, 426], which is the source of all sense-disambiguated words in this book. See also §3.1.



$p(c | A, B) = p(c | A)$  and then  $B$  would not be helpful once we have  $A$ . However, while we expect this to be often true, it is possible that  $p(c | A, B) \neq p(c | A)$ —see Appendix A.2. In other words, it is possible that independent data sources are not necessarily quite on the right extreme of the continuum, as they could work together to estimate  $c$ . Hence, the term “orthogonal” is better than “independent” for the right side of the continuum.

Regardless, in the ideal case, each data modality has some shared information with our inference target, and they are also diverse so that there is an advantage to including all of them. Consider getting the values of these variables one at a time. Before we know any of the observations, the entropy of the inference target ( $c$ , in our example) is simply that of its prior distribution. Then, if we are told the value of  $A$ , which is informative about  $c$ , the decrease in entropy is given by the mutual information  $H[c, A] = I[c] - I[c | A]$ . This corresponds (informally) to the degree that the posterior distribution  $p(c | A)$  is more peaked than the prior distribution  $p(c)$ . Now, if we are told the value of  $B$ , we can hope for further reduction in entropy, and thus even more certainty in the posterior,  $p(c | A, B)$ . Since we assume that the observables are conditionally independent given concepts, we have

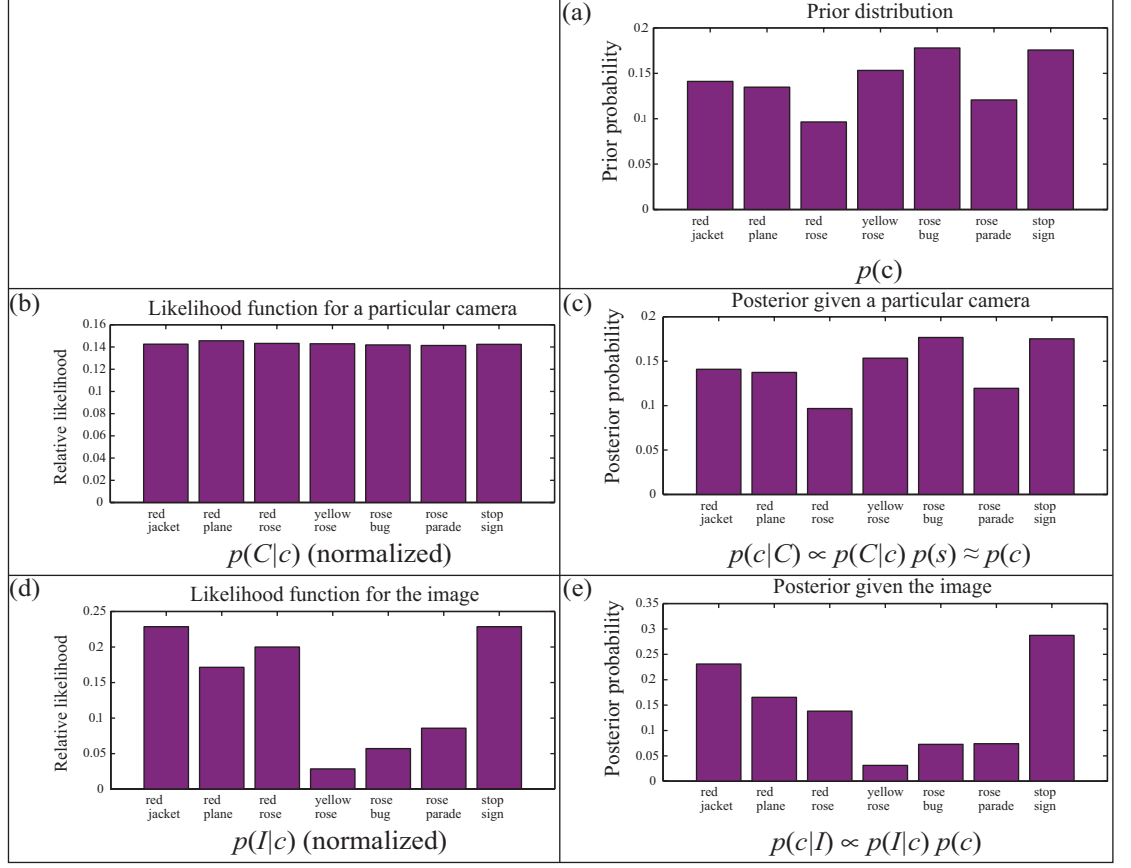
$$p(c | A, B) \propto p(A | c) p(B | c) p(c) \propto p(c | A) p(B | c), \quad (1.5)$$

applying Bayes rule twice. In other words, ignoring constants, the posterior  $p(c | A)$  is revised by the additional information in the likelihood,  $p(B | c)$ , which, informally, is peaked (lower entropy) in accordance with how informative it is.

Some of the possibilities are illustrated in Figure 1.5, with modalities  $C$  (camera model meta data),  $A$  (image features), and  $B$  (associated text). In this example the camera model is assumed to be at best only slightly informative about what is in the scene. By contrast, extracted image features and associative text are both substantively informative in the example. In addition, image features and associated text are complementary. The likelihood due to learning about the camera is relatively flat, whereas the likelihoods for image features and image text are more informative, and noticeably, have different shapes from each other. Even if the shapes were similar (but not uniform) across semantics, they could still mutually reinforce the same conclusion to reduce the overall uncertainty.

## 1.2 COMPUTATIONAL TASKS

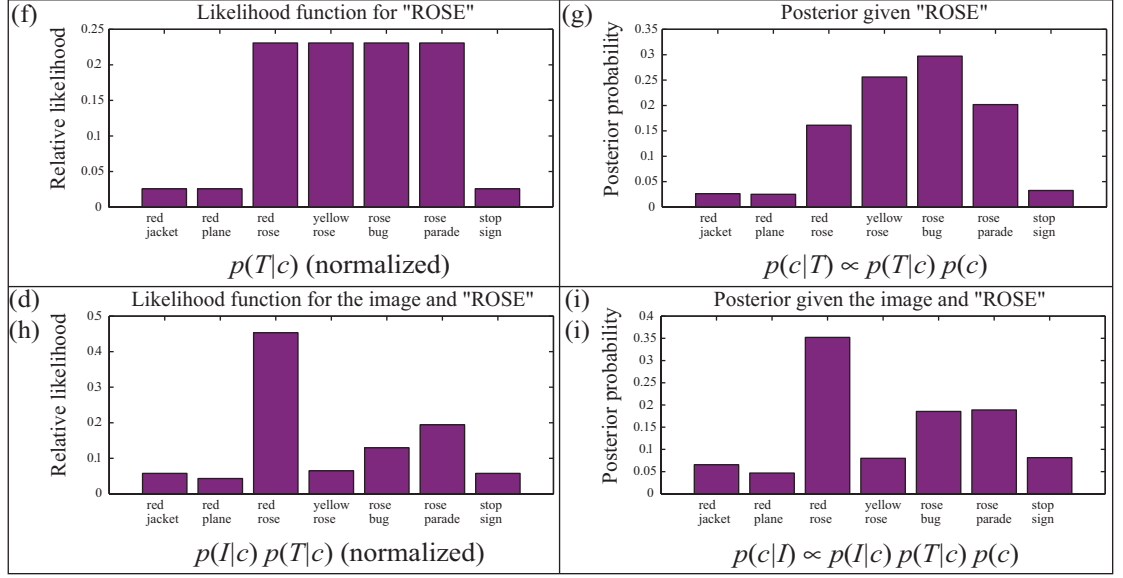
Both joint representation and complementary representation enable interesting and important computational goals. Pragmatically, which of these aspects we focus on depends on the particular task and available data. In this section I introduce three kinds of task, with the first (translation, §1.2.1) relying on sufficient redundancy between data types, the second (cross-modal disambiguation, §1.2.2) relying on overlapping information, as does the third (grounding language, §1.2.3).



**Figure 1.5:** Plots illustrating overlapping information of multimodal data, inspired by the retrieval example in Figure 1.4. All curves are normalized so that quantities sum to one, as the key issue is the shape of the curves. Different evidence modalities reduce the uncertainty among semantic concepts differently. Without any evidence, the distribution over semantic concepts,  $c$ , is the prior distribution illustrated in (a). The left-hand column shows likelihood factors for concepts due to observing (b) the camera make and model,  $C$ , (d) the query image  $I$  (Figure 1.4), (f) associated text,  $T$  (“ROSE”), and (d) both  $I$  and  $T$ , assuming that these modalities are conditionally independent given  $c$ . (*Continues.*)

Note that much of what we consider in this book applies to multimodal data in general<sup>5</sup> although specific examples will be about visual and linguistic data.

<sup>5</sup>Examples of multimodal analysis along the lines of this book, but outside of the scope of vision and language include sounds and text [529, 544] and speech recognition and video [226–228].



**Figure 1.5:** (*Continued.*) Here the choice of camera is minimally affected by the semantics reflected by the relatively flat curve (b). On the other hand, image content and associated words both significantly restrict the semantics, and do so differently, so that together they restrict the semantics more than doing so individually. The right-hand column shows the posterior distribution without any evidence (the prior, (a)), and updated beliefs after observing  $C$ ,  $I$ ,  $T$ , and both  $I$ , and  $T$ .

### 1.2.1 MULTIMODAL TRANSLATION

Figure 1.6 illustrates a significant theme for this book, namely *multimodal translation*, which concerns computationally producing linguistic representations of images, and vice versa. We refer to this task as translation in analogy with the task of automatically translating from one language (e.g., French) to another (e.g., English)<sup>6</sup>—see Figure 1.7. In its simplest form this amounts to having a cross-lingual dictionary that maps words from one language to words in another. The learning task then becomes automatically learning the dictionary. By analogy, each region (French word) is mapped to a label (English word), and the collection of such words becomes a set of automatically generated keywords for the image. This then supports image search by keywords, even when human provided keywords are not available. This is helpful because humans typically search for images based on semantics (§2.1), which are most efficiently provided by language. Image-to-text translation thus links the scientific goal of being able to understand images, to the practical application of image retrieval.

<sup>6</sup>This analogy was introduced by Duygulu et al. [182].


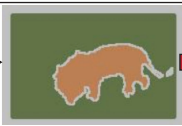


(a) Image auto annotation			tiger cat grass people water bengal buildings ocean forest reef	GRASS TIGER CAT FOREST
			water hippos rhino river grass reflection one-horned head plain sand	HIPPO BULL mouth walk



Image auto annotation results from Barnard and Forsyth [64, Figure 6] showing the regions used in preprocessing. Region features include position, size, color, and texture. The predicted keywords are in rank order of probability. For comparison, the human annotation for this image from the Corel™ database (reproduced under the terms of the user agreement) is shown to the right, with lowercase words being excluded from the vocabulary.

**Figure 1.6:** Variations on the translation task. These examples are meant to be indicative of the goal, i.e., what is to be achieved, rather than how it is done in the particular cases. Multiple methods exist for each task, some of which are discussed in this book. Notice that despite being selected among better results, there are errors in every example. All figures reprinted with permission. (*Continues.*)

We will consider two simple forms of translating image data into a keyword representation. First, we can provide keywords for the image without specifying what parts of the image the keywords refer to, which is known as “auto-annotation” (Figure 1.6a). Second, we can provide words for image regions, which I will refer to as “region-labeling” (Figure 1.6b). A human annotator had no difficulty “translating” the image in Figure 1.6a into keywords, but building a computational system to do the same thing is very challenging. Chapter 6 covers machine learning approaches that use data like that illustrated in Figure 1.8 to build models or classifiers that embody relationships between image features and words. Such a system can then be applied to images that are not in the training data to generate appropriate text for these images.

Going beyond simple keywords, a full translation system would be able to encode much more about the semantics of the scene including relative sizes and locations of the objects as well as the appearance characteristics of the objects and backgrounds. Here, some labels would refer to image regions (“human face”), groups of them (“young boy”), the entire image (e.g., “birthday party”), as well as concepts that speak to object relationships, appearance, actions, and activities. Doing so would demonstrate capability on the key task in machine vision, which is to extract semantic representations from image data, as linguistic representations are close to what we commonly consider semantic information.

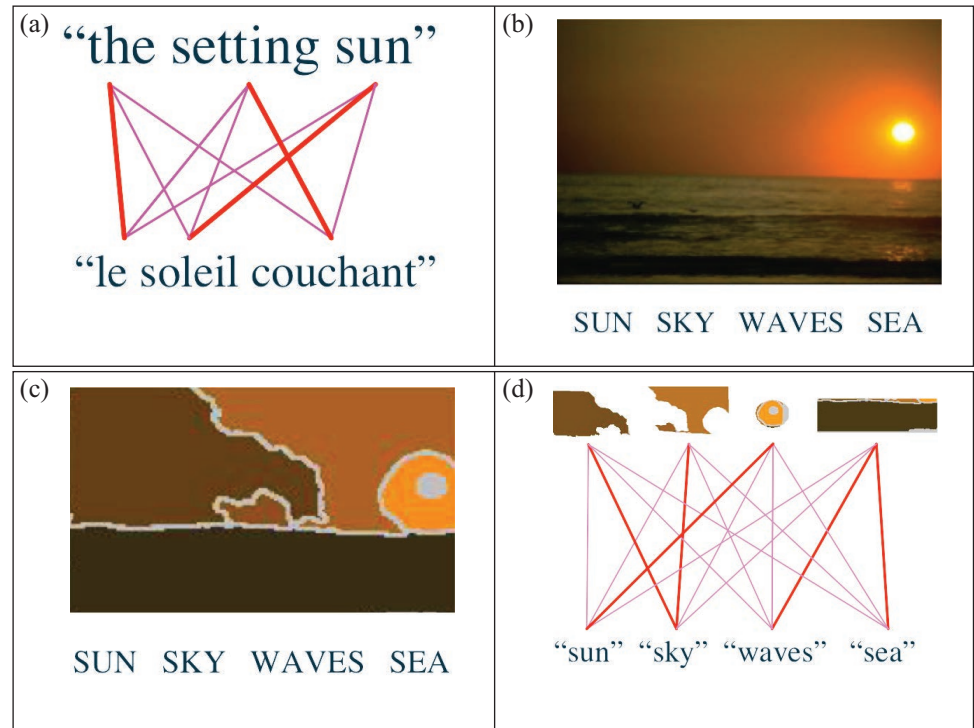
Chapter 7 covers methods using language models beyond keyword. One example is shown in Figure 1.6c. In organizing this quickly growing body of work, it is helpful to distinguish between the tactic of using parts of speech other than nouns to help identify and/or disambiguate objects (e.g., red car) vs. the goal of providing richer translations. The former is often most useful in

(b)	
	Region labeling results from Barnard et al. [60, Figure 6].
(c)	<div><div><p>This is a photograph of one sky, one road and one bus. The blue sky is above the gray road. The gray road is near the shiny bus. The shiny bus is near the blue sky.</p><p>Automatically generated descriptions for images from Kulkarni et al. [338]</p></div><div><p>There are two aeroplanes. The first shiny aeroplane is near the second shiny aeroplane.</p></div></div>
Generating captions	

**Figure 1.6:** (Continued.) Variations on the translation task. These examples are meant to be indicative of the goal, i.e., what is to be achieved, rather than how it is done in the particular cases. Multiple methods exist for each task, some of which are discussed in this volume. Notice that despite being selected among better results, there are errors in every example. All figures reprinted with permission.

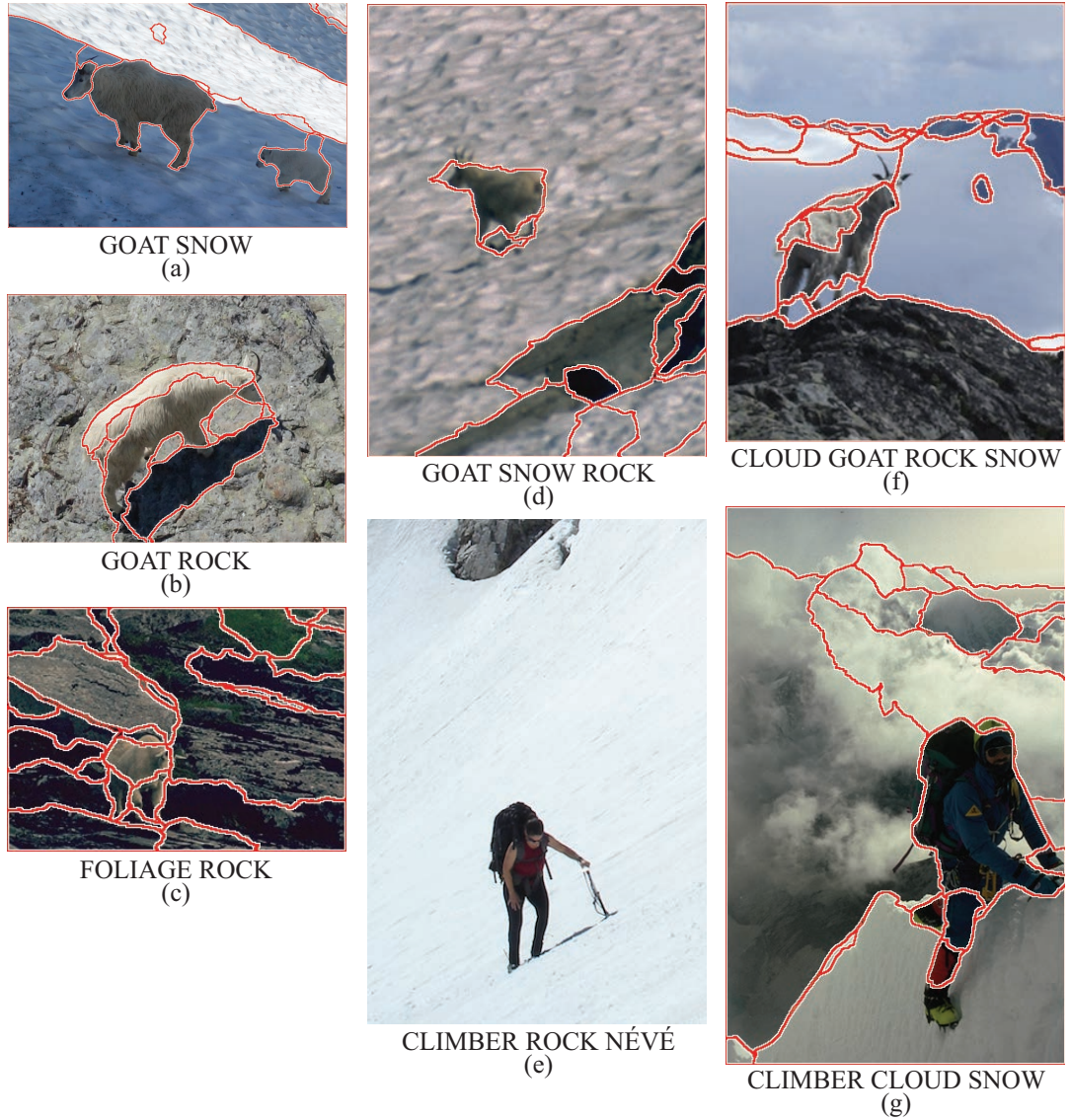
model learning, where the additional information provided by adjectives and spatial prepositions help reduce correspondence ambiguity between visual and linguistic elements during training.<sup>7</sup> The visual meanings of adjectives and spatial prepositions are either supplied by system developers

<sup>7</sup>For example, Gupta and Davis [259] proposed an effective system for learning the translations dictionary which augmented nouns annotations with propositions over noun pairs for comparative prepositions (e.g., above) and adjectives (e.g., taller). They learned nouns visual characteristics and relationship jointly, as well as relationship priors that helped annotate new data. This is discussed further in §7.3.



**Figure 1.7:** The analogy between machine translation in natural language processing, and learning to label regions in loosely labeled data: (a) translating a simple phrase from English to French. Because of differences in word ordering (as in this example), or multiple words mapping to a single word, or some words not mapping to any words in the other language, there is little information in the local ordering and the correspondence between elements is ambiguous. Each English word can conceivably map to any of the French words, i.e., any of the lines may be correct. However, with additional examples (e.g., seeing “sun” in different contexts), our belief increases that certain matches are likely to be correct (red lines), thereby automatically creating a probabilistic dictionary. (b) A Corel<sup>TM</sup> image (reproduced under the terms of the user agreement) with associated keywords. (c) Image regions produced by a simple segmentation program. Notice that the mapping between regions and keywords is not, a priori, known, and thus the situation is similar to that in (a) as emphasized in (d). A system for image-to-text translation will typically use an extensive set of aligned data (e.g., (c), Figure 1.8) to learn a dictionary. The dictionary can then be used to attach keyword probabilities to regions in images without keywords, which then can be distilled into a list of annotation words for the image.





**Figure 1.8:** Example images that are annotated with appropriate keywords. The red lines outline machine segmentations achieved automatically by a version of normalize cuts [518] with the cost matrix taking into account color, texture, and edges [395]. Such data is referred to as loosely labeled because, while the words are usually relevant, we do not know which part of the image should be associated with each word.

or learned together with noun meanings. Regardless, having visual meanings for non-noun words can help provide richer translations and finer grained image search.

### 1.2.2 INTEGRATING COMPLEMENTARY MULTIMODAL DATA AND CROSS MODAL DISAMBIGUATION

The opposite of translating from one modality to another is using both modalities together to understand their combined story. In such tasks, a computer program plays a similar role to a person who uses both the text and the figures of a document to understand it. For example, we might consider building an application that analyzes archives of news photos with captions to mine historic data for interesting trends.

Recall that a particular data component might provide completely orthogonal information to the other parts of the data (extreme right of the scale in Figure 1.3), or it might be semantically linked to other data components, but each one provides additional, non-overlapping, information (middle of the scale in Figure 1.3), or it could be highly redundant with other data. Unlike the translation task where only one modality is present, with both modalities the redundant information is superfluous. At the other extreme (orthogonal information), then there is little to do from an inference perspective, other than record the independent piece of data. For example, if the name of the photographer cannot help interpret the image, then we are limited to making the photographer name available when needed, as we would do to support image search by photographer.

Thus, from a computational perspective, we focus on the case when the two modalities provide partially overlapping information. Specifically, at least one of the modalities provides information about the other that is not otherwise available. For example, consider the image in Figure 1.3c. The region shown is likely to be one of a limited set of possibilities including a rose or stop sign. Adding the caption “rose” helps distinguish among them. Note that from the perspective of image retrieval, the keyword “rose” does not invariably lead to flowers (see Figure 1.4), and further, does not alone specify a red one. Hence, the caption and the image data complement each other. Using the complementary data to distinguish among possibilities illustrates a second important kind of task, namely *cross-modal disambiguation*.

The rose/stop-sign example illustrates that images and text can be ambiguous when considered separately, but are not necessarily ambiguous when they are considered together. This reflects how humans often arrange jointly occurring images and text—captions tend to omit what is visually obvious, and images are chosen to provide information that is awkward to provide using language. Thus, we can loosely interpret the task of building computational systems to understand such data as extracting a non-ambiguous multimodal representation of what the document creator intended, thereby reversing the creation of those artifacts. To solve this inverse problem we observe that each modality constrains the underlying semantics of the scene. When we put these complementary constraints together, they can constrain the semantics sufficiently to be useful to an information system user. In particular, ambiguities at the category level (e.g., “rose” vs. “stop



sign”) can be removed, thereby leading to an interpretation of the data that will be judged more semantically correct.

Figure 1.3b shows an example of a different problem, namely that words have multiple meanings, again illustrated using “bank” as an example. Human readers resolve such ambiguities using a variety of contextual information ranging from nearby words to the entire document and cultural background. The example shows how image data can also play this role.<sup>8</sup> In particular, the image depicting an outdoor scene can tell us that the second sense of bank is likely meant.

### 1.2.3 GROUNDING LANGUAGE WITH SENSORY DATA

Language is extremely valuable for agents collaborating to survive in the world, but the abstractions in a symbolic system are not inherently connected to the physical environment. Defining all symbols in terms of other symbols is circular, and results in a closed system that is not informative about anything physical. Endowing cognitive systems with language that can refer to the physical world can be achieved by associating words with sensory data (e.g., images), thereby *grounding* their meaning. For example, while a word like “snow” can be defined abstractly using other words, its physical meaning requires shared experience of snow, or the words used to define it. Without meanings eventually linking to the world through sensor input, there is no way to use language to communicate about interacting with the world.

The need for computational approaches for grounding language arises naturally in human-robot communication. For example, Roy and colleagues have developed a robot, Ripley, that learns from language/visual association, and can respond to verbal requests grounded in the world [409, 410, 491]. Consider a simple directive to a robot like Ripley: “put the red cone on top of the blue block.” For the robot to accomplish this task, it needs to connect words in the directive to the visual data from the world in front of it.

These kinds of human-robot interactions can have elements of both exploiting redundant information and integrating complementary information. The existence of a “red cone” and “blue block” in the scene is implicit in the text, and is redundant information that the robot can use to learn the visual representation of the words. On the other hand, both vision and language are needed to establish the physical coordinates of the object to be moved, and where it should be moved. In other words, executing the directive requires both processing the sentence and tying the appropriated parts to the physical world through the sensory system.

The robot interaction domain shares with image understanding applications the notion of referential meaning (words referring to the world), but also relies on function meaning, which is described by Roy as “agents using language to achieve goals” [490]. In the example of moving the red cone, the intent is clearly specified as a directive. In the general case, inferring intent is very difficult. An interesting example from Roy [490] is translating the statement “this coffee is cold” into “bring me hot coffee” based on the appropriate context. In making use of multimodal data to understand such scenarios, Roy categorizes “signs,” which we can define as evidence available in

<sup>8</sup>This was first studied computationally by Barnard et al. [66] (see also [67]).

data, into three groups: *natural*, *indexical*, and *intentional*. Natural signs are evidence of what is in the world. Indexical signs are also about the world, but are specific to spatiotemporal coordinates of those entities, such as object position and pose. Finally, intentional signs are evidence about the intent of agents.

Intentional signs can come from both language and the physical world. In particular, what is possible in the physical world can reduce ambiguity about what directives means [246]. For example, going back to putting cones on blocks, one can imagine a reasoning system that understands that putting cone shapes on block shapes works better than the reverse, which can further reduce ambiguity between blocks and cones, both for executing the task and learning what they are more generally.

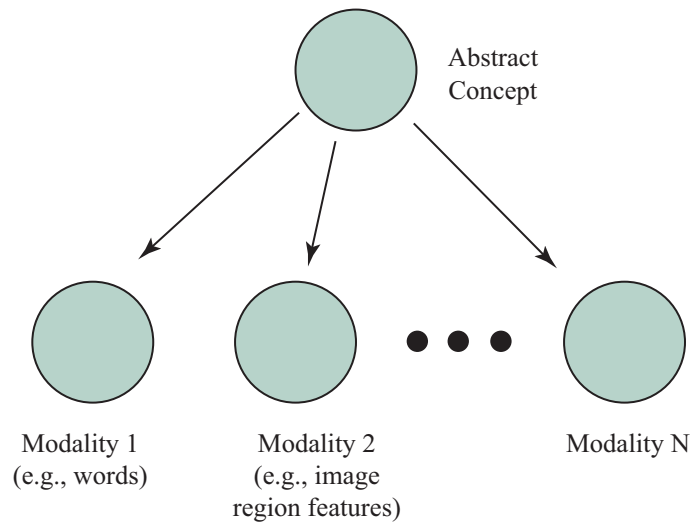
In summary, language enables agents to cooperate by being able to represent goals compactly. However, actions needed to fulfill the goal typically are relative to the environment, which must be perceived and analyzed with respect to the goal. Hence, grounding language is an important computational task within our domain that brings in different aspects than the previous two tasks.

### 1.3 MULTIMODAL MODELING

Many of the models considered in this book build upon the basic model suggested by Figure 1.9. Here, concepts are treated as underlying *latent* variables that generate data for multiple modalities (e.g., words and image region features). For example, the multi-modal concept of a tiger is both associated with the label “tiger” (a human categorization) and what it looks like. Clearly, this can be extended to other text information, and other modalities. Regardless, following the graphical modeling paradigm,<sup>9</sup> the distribution of the data for each modality is independent, conditioned on the concept. Thus, the linkage between modalities is only through the concept. Variations within this approach include different particulars of how concepts generate data for each modality, and how the concepts themselves arise. For example, outdoor scenes have a different mix of concepts than indoor scenes, which can be modeled by adding a higher-level clustering for scene types which provides for different distributions of concepts (see §6.5).

Binding data from different modalities through underlying concepts nicely handles arbitrary sets of modalities, and does not give any of them special status. For example, visual appearance is handled symmetrically with text labels, and similarly with other modalities that one might add, such as how an object moves or how it feels to the touch. Additional modalities provide additional opportunities to learn from data. For example, consider a learned concept linked both to “giraffe” and a particular spotted pattern. Text-based collections of semantic information such as WordNet (§3.1) could then connect the underlying concept to that of a quadruped that has mobility and whose environment is often “savanna” which may be a word not yet connected to visual

<sup>9</sup>This book relies on modest familiarity with graphical models as well as basic probability and statistics. For graphical models I suggest Ch. 8 of Bishop [87], which is available on-line, or the very comprehensive book by Koller and Friedman [331]. Chapter 2 of Koller and Friedman also provides a concise review of probability for applications such as the ones in this book.



**Figure 1.9:** A simple model for multi-modal data. Here the multimodal data connects to a common abstract (latent) concept, rather than being directly connected. For example, images with captions could be modeled as a collection of multiple concepts, each of which gives rise to words and image region characteristics. Further, image region characteristics could be broken into sub-modalities such as color, and texture. Readers familiar with graphical models will recognize concepts as latent cluster variables and the observations of different modalities being conditionally independent given the cluster. More complex structures such as hierarchical concept hierarchies and context dependent concepts and/or data models can be constructed within this general paradigm. An alternative approach, suitable for some applications, is to build discriminative models for predicting one modality directly from the other, e.g., predicting words from image data.

information. The giraffe’s spotted pattern could also be linked to video data, where segmenting out the giraffe based on its movement is easier, and where its movement reveals its articulated structure.<sup>10</sup> Similarly, linking the spotted pattern to video or image data could provide further understanding of the appearance of giraffes, as well as a visual understanding of its typical environment and thus the newly learned word “savanna.” While a system to integrate all these sources of data in a general-purpose system has yet to be built, examples of many of the components have been explored in limited domains.

To summarize, when visual and linguistic information provide alternative representations of the same underlying concepts, we can have access to rich joint representations of those concepts. These representations can provide computational systems with some of the attributes that we associate with deeper understanding. For example, consider the concept suggested by “car.”

<sup>10</sup>This strategy was first explored by Ramanan et al. [477].

Certainly we can describe cars with words, but we can also visualize them and connect them with movement and functional attributes. Specifically, the abstract semantic concept is *grounded* by sensory data, which can be strictly observational, or can include interacting with the concept (e.g., driving the car) as discussed in the human-robot interaction example.

### 1.3.1 DISCRIMINATIVE METHODS

The generative approach just described is intuitively appealing, as it is explicit about how our modeling constructs connect to the distribution of the observed data. The focus on the joint probability makes the model suitable for multiple tasks, and the accompanying distributional answers provide uncertainty estimates and can be useful for subsequent processes. However, providing general capabilities can come at the expense of reduced performance on specific tasks. Intuitively, if performance on a specific task such as auto-annotation is paramount, then anything in the approach that is not optimized for that task is potentially a liability. This has led researchers to consider discriminative methods for tasks that can be expressed as a classification problem. Generally, the focus of such methods is to establish what differentiates the classes in the space of features (e.g., color and texture descriptors), with less interest on why they are differentiable, or what the confidence might be. Because auto-annotation is important for practical reasons that are agnostic to how it is achieved, I include some discriminative approaches in this book (§6.4.6, §6.6.10, §6.7.2).

## 1.4 MUTIMODAL INFERENCE–APPLICATIONS TO COMPUTATIONAL TASKS

I now provide a taste of how we can approach some of the computational tasks introduced in §1.2. For each of the three examples that follow, we assume that we already have the model. Learning model parameters from data is covered in §1.5.

### 1.4.1 REGION LABELING WITH A CONCEPT MODEL

Region labeling, covered in depth in Chapter 6, maps vectors of image region features,  $\mathbf{r}$ , to words,  $w$ . We can imagine generating regions with labels by first sampling a concept,  $c$ , from a distribution  $p(c)$ , and then generating a word,  $w$ , and a region feature vector,  $\mathbf{r}$ , based on the concept. Formally,  $\mathbf{r}$  and  $w$  conditionally independent given the latent concept,  $c$ . Our generative model for a region-word pair is thus

$$p(w, \mathbf{r}, c) = p(c) p(w | c) p(\mathbf{r} | c). \quad (1.6)$$

To do inference on the observed variables  $w$  and  $\mathbf{r}$  marginalize out the unobserved concept:

$$p(w, \mathbf{r}) = \sum_c p(c) p(w | c) p(\mathbf{r} | c) = \sum_c p(w | c) p(\mathbf{r}, c). \quad (1.7)$$

Thus,

$$p(w|\mathbf{r}) = \sum_c p(c|\mathbf{r}) p(w|c), \quad (1.8)$$

where

$$p(c|\mathbf{r}) = \frac{p(c) p(\mathbf{r}|c)}{\sum_{c'} p(c') p(\mathbf{r}|c')}. \quad (1.9)$$

This provides a distribution over words for each region. To label a region we can choose the word with maximal probability, or one might take advantage of the full distribution for further processing.

The result (1.8) is very intuitive. It says that for each semantic concept,  $c$ , consider how likely it is responsible for the region under consideration,  $\mathbf{r}$ . To the extent that it is as measured by  $p(c|\mathbf{r})$ , we accordingly weight the word distribution,  $p(w|c)$ , associated with that concept.

#### 1.4.2 CROSS-MODAL DISAMBIGUATION—REGION LABELING WITH IMAGE KEYWORDS

Now consider region labeling when we have the luxury of image keywords,  $\mathbf{k}$ . Assuming that image keywords are informative about region labels, we can improve the labeling by using this additional source of information. We consider that region labels come from  $\mathbf{k}$  with high probability, but there is ambiguity in which word links to which region. Alternatively, the region labeler provides word probabilities for each region, but restricting them to  $\mathbf{k}$  further disambiguates the meaning. In short, this task is an example of cross-modal disambiguation.

Notice that  $\mathbf{k}$  and  $w$  are **not** conditionally independent given the abstract concept,  $c$ , and thus Figure 1.9 does not apply.<sup>11</sup> But we can improve the estimate of  $p(w|\mathbf{r})$  using the conditional independence of  $\mathbf{k}$  and  $\mathbf{r}$  given  $w$  as follows:

$$\begin{aligned} p(w|\mathbf{k}, \mathbf{r}) &\propto p(\mathbf{k}, \mathbf{r}|w) p(w) \\ &\propto p(\mathbf{k}|w) p(\mathbf{r}|w) p(w) \\ &\propto p(\mathbf{k}|w) p(w|\mathbf{r}). \end{aligned} \quad (1.10)$$

A simple choice for  $p(\mathbf{k}|w)$  is to set it to one if  $w \in \mathbf{k}$ , and zero otherwise. We can also consider learning  $p(\mathbf{k}|w)$  from data to better reflect how annotators choose keywords.

#### 1.4.3 CROSS-MODAL DISAMBIGUATION—WORD SENSE DISAMBIGUATION WITH IMAGES

Finally, consider the task of word sense disambiguation with images [66, 67].<sup>12</sup> Here, we assume that we have distributions over senses for each word,  $w$ , based on the collection of words,  $W$ ,

<sup>11</sup>To see that conditional independence does not hold, consider a tiger concept that is associated with the words “cat” and “tiger.” Either could be used as the region label, but having decided on a choice, the image keyword choice is fixed.

<sup>12</sup>The formulation in Barnard and Johnson [66] implicitly assumed that the prior over senses  $p(s)$  is uniform, which is sub-optimal. Algebraically this assumption leads to dropping the final division by  $p(s)$  in (1.12).

## 22 1. INTRODUCTION

from natural language processing. Denote this distribution by  $p(s|w, W)$ . We wish to improve this using an illustrative image,  $I$ , by computing  $p(s|w, W, I)$ . To take a similar approach to the previous two examples, we assume our learned concept model has distributions over words senses,  $s$ , instead of words themselves. Then we can compute distributions over senses, given a region,  $\mathbf{r}$ , by (1.8), with  $w$  replaced by  $s$ . Further, we consider that images taken as a whole—not regions—influence caption senses. Thus, we need to estimate  $p(s|I)$ , where the image,  $I$ , is represented by the set of regions,  $\{\mathbf{r}_i\}$ . To estimate  $p(s|I)$  we assume that words senses for  $I$  are independently drawn by first choosing a region at random, and then choosing a sense based on the region. This leads to the image word sense distributions being the average of those for the regions. Formally,

$$p(s|I) = \frac{1}{N} \sum_{i=1}^N p(s|\mathbf{r}_i). \quad (1.11)$$

Then,

$$\begin{aligned} p(s|w, W, I) &\propto p(s, w, W, I) \\ &\propto p(I|s, w, W) p(s|w, W) \\ &\propto p(I|s) p(s|w, W) \\ &\propto p(s|I) p(s|w, W) / p(s), \end{aligned} \quad (1.12)$$

ignoring probabilities over subsets of  $(w, W, T)$  as these variables are given, and using  $w, W \perp I | s \Rightarrow p(I|w, W, s) = p(I|s)$ .

### 1.5 LEARNING FROM REDUNDANT REPRESENTATIONS IN LOOSELY LABELED MULTIMODAL DATA

To learn parameters for the concept model from §1.4.1, we nominally need examples of the concepts together with their visual features and associated words. One way to do this would be manually segment images into regions corresponding to concepts, and provide words for each region, which implicitly provide concept examples. Notice that this does not establish what the concepts actually are, as we only have a collection of examples of different ones. To proceed, we could cluster the concepts based on both their word attributes and their visual features, thereby getting concepts defined by groups of examples that are similar both in the distribution of words used to describe them and in visual appearance.

While this approach is sound, the need for a large amount of manual segmentation and annotation is problematic. As research into various aspects of image understanding has progressed, various stores of such data have accrued through significant human effort. In addition, methods for semi-supervised semantic labeling have been developed, and crowdsourcing through Amazon Mechanical Turk [3] has become increasingly popular. However, despite the hope that eventually there will be enough high quality training data, most research into linking vision and language

has focused on using *loosely labeled* data (see Figure 1.8), as such data is much more plentiful and easier to create or collect from the web. In such data, images are annotated by keywords or captions, but we do not know which aspects of the image are relevant to which words. Each word is somewhat informative, but it is also ambiguous. This is an example of *correspondence ambiguity*.

Correspondence ambiguity is similar to the *data association* problem, which we would have if we take a clustering approach. Here, we do not know *a priori* which concept a word or a region comes from, and the number of possible assignments is exponential in the size of the data set, making it potentially a far worse computational issue. With loosely labeled data, the assumption that the words and regions come from the set of clusters constrains the possibilities. For example, if we assign  $R$  regions to  $R$  concepts out of  $C$  possible concepts, then the words are now constrained to come from one of  $R$  concepts instead of one of  $C$ , which is typically a much larger number. In practice, we often focus on the data-to-concept correspondence, with the data-data correspondence being implicit—if a word and a region in the same document come from the same concept, then they are linked.

Notice that these correspondence problems arise only during inference and, generally, as in this example, when we want to learn models from data. One of the strengths of the probabilistic graphical modeling paradigm is that it helps us keep modeling and inference (which includes learning) separate. Generative models for the data (e.g., (1.6)) do not need to represent the association of the data and the model because the model considers each document as an independent sample. Correspondence problems arise when we see the data, but we do not know which part of the model (e.g., which concept) is responsible for it.

### 1.5.1 RESOLVING REGION-LABEL CORRESPONDENCE AMBIGUITY

To develop intuition about why the region-label correspondence ambiguity can be overcome, consider Figure 1.8a. Based on this single image we are not able to judge whether the word “snow” should be interpreted as a label for one or more goat regions, one or more snow regions, or noise (not visually relevant). Without additional sources of information, this correspondence ambiguity cannot be resolved. Interestingly, multiple examples can help reduce the ambiguity. If we were able to link the goat regions in Figure 1.8a with those in Figure 1.8b based on appearance, then this would suggest that they are linked to the same concept. Since we assume that the data from multiple modalities in a single image share concepts, this would lead to the conjecture that those regions also likely link to the only common word keyword (“goat”). Further, we would also have some handle on the assignment of the snow regions in Figure 1.8a to “snow” because we expect that it is quite likely “snow” refers to some part of the image, and further that it is less likely to also refer to the goat. We refer to this kind of argument as *exclusion reasoning*.<sup>13</sup> This notion is important because concepts that are relatively common and are easy to link to visual features (e.g.,

<sup>13</sup>This was the term used by Barnard and Fan [62] who considered the constraint computationally. The concept was referred to as the exclusion principle by Siskind [528] who discusses it in the context of language learning in humans.



water, sky, snow) can reduce the labeling ambiguity for rare or otherwise difficult concepts (e.g., climber or mountain goat).

We see that the key leverage that we have on correspondence ambiguity is the fortuitous occurrences of subsets of concepts in various data examples. While correspondence ambiguity is generally an impediment, it is interesting to compare having  $N$  labeled regions compared with  $N$  loosely labeled images with  $R$  regions and  $R$  words. Notice that these are the same if  $R = 1$ . In other words, we can compare having more information that is ambiguous (e.g., if  $R = 3$ , we have three words that pertain to one of three regions) vs. less information without ambiguity (e.g., if  $R = 1$ , the data only speaks to one region, but it does so unambiguously). Whether or not we would prefer more regions at the expense of more ambiguity is difficult to analyze in general, but it is certainly a function of the statistics of the pairings. On the one extreme, if, for example, “skiers” and “snow” always occur as a pair, then we cannot distinguish them and we would prefer some single region examples of them. On the other extreme, if we assume that regions can be tokenized precisely (e.g., a snow region always matches snow and never anything else), and pairings occur with uniform probability, then it is possible that multiple regions are more informative (see Appendix A.3; also of relevance is Cour et al. [146]). Of course, many concepts commonly co-occur, and so pairing with uniform probability is not a realistic assumption. Pragmatically, it is a happy accident that loosely labeled data is easy to come by in quantity, as it improves our chances of having the diverse sets of concepts in images needed to deal with the correspondence ambiguity.

In addition to linking visually similar entities across loosely labeled instances, there are several other ways to reduce correspondence ambiguity. First, rich text descriptions such as “a red car on gray pavement” (see §7.1, §7.2, §7.3) can help localize labels. Second, detectors for specific entities (e.g., “faces”) can localize labels that can be associated with them (e.g., “person”) [166]. Third, keyword lists can be processed using tools such as WordNet (§3.1) to merge redundant words. For example, keyword lists often include a specific terms (e.g., “f-17”) and more general terms (e.g., “jet”) which typically map to the same image features. Finally, words can be processed to ignore ones unlikely to have identifiable visual features because they are abstract (e.g., “religious”) (see §7.2).

### 1.5.2 DATA VARIATION AND SEMANTIC GROUPING

Learning from data sets that have many examples of each concept is also warranted because both the linguistic and the visual evidence for concepts vary substantively. This variation, especially in the case of visual features of objects, is one reason why recognition of object categories is very difficult. With large data sets we can learn the statistics of features for concepts that are valid over a broad domain, and thus are potentially indicative of the world at large (i.e., ecologically valid). However, we also must recognize the limitations of our data domains, which are invari-



ably biased.<sup>14</sup> For example, we often learn from stock photo collections or web images, which are typically chosen as canonical or aesthetic visual representations of something in the world. Images from such datasets have *photographer bias*, and what is learned from these domains may be less useful in other domains. For example, images taken by competent photographers are quite different from the images taken by a robot navigating a building.

Given data variation, to establish the word and region models associated with concepts (e.g.,  $p(w|c)$  and  $(p(r|c))$  in (1.6)), a learning process often (at least implicitly) needs to group textual and visual features in the training data.

**Grouping text elements.** If one assumes that keywords are synonymous with concepts, then grouping text elements is not needed (see §6.4.4 and §6.4.5 for approaches based on this assumption). However, words are not the same as concepts for several reasons. First, as I have already noted, many words have multiple senses. Second, a concept might be reasonably connected to multiple words, either because they are near synonyms (e.g., “snow” vs. “névé”) or the concept needs multiple words to describe it. Hence, words in the training corpus need to be grouped, and the same word might be in different groups depending on the context. Notice that determining that two words are close in meaning might be possible using tools such as WordNet (§3.1), or can potentially be learned based on co-occurring visual similarity.

**Grouping regions.** Similarly, if we are using regions for visual representation, then they (at least implicitly) need to be grouped to form a representation of the visual appearance of the concept. Here features are typically represented using continuous values, and so there is at least a notion of closeness assuming that the relative weights of the features are known or learned. Since the underlying assumption is that concepts are somewhat visually homogenous, large variation might be captured using multiple related concepts (e.g., red roses and white roses). Again, the learning algorithms will at least implicitly group features in the training data into concepts, and again, similar features might be grouped differently depending on context. For example, a white region in a skiing image (with the word “snow”) might be grouped with a snow concept during learning, but a second similar region in a cloud image (without the word “snow”) might be grouped with a cloud concept.

### 1.5.3 SIMULTANEOUSLY LEARNING MODELS AND REDUCING CORRESPONDENCE AMBIGUITY

The previous discussion suggests two key processes for learning from generic multimodal data: (1) learning models for the underlying structure; and (2) learning the correspondence between data and model elements. While some methods do not break the problem down in this way, this is often the case for generative models. A key observation is that solving one of these problems

<sup>14</sup>Data set bias in object recognition is considered by Torralba and Efros [557]. See also [317]. Also relevant is mapping what is learned using from one dataset so that it is more applicable for use on another data set, which is often referred to domain adaptation or transfer learning (e.g., [245, 337, 499]).

helps solve the other. For the first, knowing which data elements correspond to a given concept enables the building the textual and visual model for the concept. For example, if we know that the snow regions in the images in Figure 1.8 are connected to the snow concept, we can aggregate all these observations into a visual model for snow. Similarly, if we know the word instances that correspond to the concept (e.g., “snow,” “névé”), then we can learn a distribution for the use of the words given the concept. For example, since “névé” is more specific than “snow,” we expect that it might be used less on average.

For the second, if we know the model, we can compute the data-model correspondences. For example, given a good visual representation of the snow concept, we can link regions in Figure 1.8a to snow, and, given a language representation of snow, we can then attach the word “snow” to the concept as well. Exclusion reasoning could then further reduce the correspondence ambiguity between “snow” and “goat.”

Since neither the correspondences nor the models are typically known, but either one can be used to estimate the other, we have a classic “chicken-and-egg” situation that is common in computer vision. This suggests the algorithmic approach of learning them iteratively. For example, we could compute initial noisy models ignoring correspondence ambiguity, and then use those models to reduce the ambiguity. On subsequent iterations, improved models can be computed from the data with reduced ambiguity, and those models can further reduce the ambiguity. This simple iterative approach can be formally defined as an instance of the Expectation-Maximization (EM) algorithm,<sup>15</sup> which is a common method for learning models for images and text (first discussed in detail in §6.3.5).

<sup>15</sup>Expectation Maximization (e.g., [111, 113, 122, 169, 412]) is commonly used for missing value problems. In computer vision, the missing values are often the correspondence between data elements and model components during learning.