A Sentence is Worth a Thousand Pixels

Sanja Fidler
TTI Chicago
fidler@ttic.edu

Abhishek Sharma
University of Maryland
bhokaal@cs.umd.edu

Raquel Urtasun
TTI Chicago
rurtasun@ttic.edu

Abstract

We are interested in holistic scene understanding where images are accompanied with text in the form of sentential descriptions. We propose a conditional random field model for semantic parsing which reasons jointly about which objects are present in the scene, their spatial extent as well as semantic segmentation, and employs text as well as image information as input. We automatically parse the sentences and extract objects and their relationships, and incorporate them into the model. We demonstrate the effectiveness of our approach on the UIUC dataset and show segmentation improvements of 12.5% *over the visual only model.*

1. Introduction

Images rarely appear in isolation. Photo albums are usually equipped with brief textual descriptions, while images on the web are usually surrounded by related text. In robotics, language is the most convenient way to teach an autonomous agent novel concepts or to communicate the mistakes it is making.

In the past decade, we have witnessed an increasing interest in leveraging text and image information in order to improve image retrieval [12] or generate brief description of images [2, 6, 8]. However, very few approaches [7, 11] try to use text to improve semantic understanding of images beyond simple image classification, or tag generation [1]. This is perhaps surprising, as image descriptions can resolve a lot of ambiguities inherent to visual recognition tasks. If we were able to retrieve the objects and stuff present in the scene, their relations and the actions they perform from textual descriptions, we should be able to do a much better job at automatically parsing those images.

Here we are interested in exploiting textual information for semantic scene understanding. In particular, our goal is to reason jointly about the scene type, objects, their location and spatial extent in an image, while exploiting textual information in the form of complex sentential image descriptions generated by humans.

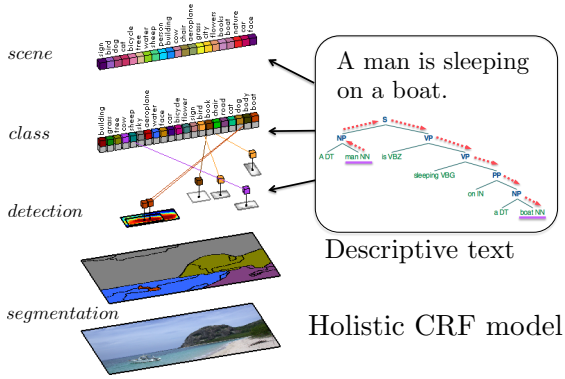

Figure 1. Our holistic model which employs visual information as well as text in the form of complex sentences.

2. Holistic Scene Understanding

We briefly describe our approach to holistic scene understanding. Our setup is the following: we have a set of images we want to parse, each of which is accompanied by a few descriptive sentences.

We formulate the problem as the one of inference in a CRF. The random field contains variables representing the class labels of image segments at two levels in a segmentation hierarchy (smaller and larger segments) as well as binary variables indicating the correctness of candidate object detections. In addition, binary variables encode the presence/absence of a class in the scene. Fig. 1 gives an overview of our model. We employ potentials which utilize both image information as well as text. We automatically parse the sentences and extract objects and their relationships, and incorporate those into the model, both via potentials as well as by re-ranking the candidate bounding boxes.

Parsing Text We extract part of speech tags (POS) of all sentences using the Stanford POS Tagger for English language [13]. We syntactically parse the sentences using the Stanford Parser with factored model [5]. Given the POS, parse trees and type dependencies, we extract information about whether an object class was mentioned as well as its cardinality (number of instances). We also extract the relationships between the objects, e.g., object A is near or on top of object B, by extracting the prepositions from text.

	back.	aerop.	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dtable	dog	horse	mbike	person	pplant	sheep	sofa	train	monitor	**averg.**
Textonboost (unary) [10]	77.8	14.1	3.4	0.7	11.3	3.3	25.5	30.9	10.3	0.7	13.2	10.8	5.2	15.1	31.8	41.0	0.0	3.7	2.4	17.1	33.7	16.8
Holistic Scene Understanding [14]	77.3	25.6	12.9	14.2	19.2	31.0	34.6	38.6	16.1	7.4	11.9	9.0	13.9	25.4	31.7	38.1	11.2	18.8	6.2	23.6	34.4	23.9
ours	76.9	31.3	29.7	37.3	27.7	29.5	52.1	40.0	38.0	6.6	55.9	25.2	33.2	38.2	44.3	42.5	15.2	32.0	20.2	40.7	48.4	36.4

Table 1. Segmentation results on UIUC sentence dataset. By leveraging text information our approach improves 12.5% AP.

Segmentation Potentials We use [10] to compute unary segmentation potentials for each segment.

Class Presence from Text: We use two types of unary potentials, depending on whether a class was mentioned or not in the text. When a *class is mentioned*, we use the average cardinality (across all sentences) for each class. When a *class is not mentioned* we simply use a bias. We also use a pairwise potential between class presence and segment variable that ensures that the classes that are inferred to be present in the scene are compatible with the classes that are chosen at the segment level.

Detection Potentials We use [3] to generate object hypotheses. For each image, we use boxes that exceed DPM thresholds, unless the object class is specifically mentioned in text. In this case, we add as many boxes as dictated by the extracted object cardinality. We utilize both text and images to compute the score for each detection. We also have a pairwise potential between the box and the class presence variable to ensure compatibility at the scene level.

Cardinality potential: We use a high-order potential to exploit the cardinality estimated from text. Our potential penalizes all box configurations that have cardinality smaller than the estimated cardinality from text.

Using prepositions: People tend to describe the objects in relation to each other, e.g., "the cat is on the sofa". This additional information should help boost certain box configurations that are spatially consistent with the relation. In order to exploit this fact, we extract prepositions from text and use them to score pairs of boxes.

Text Scene Potential: We train a classifier based on bag-of-words from text, and use the output as a unary for the scene variable in the model. Following [14], we also use scene-class co-occurrence as a pairwise potential between the scene and class-presence variable.

Learning and Inference We employ the distributed convex belief propagation algorithm of [9] for inference. For learning, we employ the primal-dual algorithm of [4].

3. Experimental Evaluation

For evaluation, we use the UIUC dataset [2], which contains 1000 images taken from PASCAL VOC 2008. As evaluation measure, we employ the standard IOU meassure. Our baselines consists of [10] as well as the holistic model of [14], which only employ visual information. As shown in Table 1, the unary alone performs poorly (17%). The holistic model of [14] achieves 23.9%. In contrast, by leveraging text, our approach performs very well, achieving 36.4%. Fig. 3 shows some examples of our inference.

sent 1: "Passengers at a station waiting to board a train pulled by a green locomotive engine." **sent 2:** "Passengers loading onto a train with a green and black steam engine." **sent 3:** "Several people waiting to board the train."

sent 1: "Man using computer on a table." **sent 2:** "The man sitting at a messy table and using a laptop." **sent 3:** "Young man sitting at a table staring at laptop."

Figure 2. Results as a function of the # of sentences employed.

References

[1] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002. 1

[2] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences for images. In *ECCV*, 2010. 1, 2

[3] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9), 2010. 2

[4] T. Hazan and R. Urtasun. A primal-dual message-passing algorithm for approximated large scale structured prediction. In *NIPS*, 2010. 2

[5] D. Klein and C. Manning. Fast exact inference with a factored model for natural language parsing. In *NIPS'03*. 1

[6] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, 2011. 1

[7] L. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding:classification, annotation and segmentation in an automatic framework. In *CVPR*, 2009. 1

[8] V. Ordonez, G. Kulkarni, and T. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 1

[9] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Distributed message passing for large scale graphical models. In *CVPR*, 2011. 2

[10] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 2

[11] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*, 2010. 1

[12] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, 2012. 1

[13] K. Toutanova, D. Klein, and C. Manning. Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*, 2003. 1

[14] Y. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 2