What are you talking about? Text-to-Image Coreference

Chen Kong[1] Dahua Lin[3] Mohit Bansal[3] Raquel Urtasun[2,3] Sanja Fidler[2,3]
[1]Tsinghua University, [2]University of Toronto, [3]TTI Chicago
kc10@mails.tsinghua.edu.cn, {dhlin,mbansal}@ttic.edu,{fidler,urtasun}@cs.toronto.edu

Abstract

In this paper we exploit natural sentential descriptions of RGB-D scenes in order to improve 3D semantic parsing. Importantly, in doing so, we reason about which particular object each noun/pronoun is referring to in the image. This allows us to utilize visual information in order to disambiguate the so-called coreference resolution problem that arises in text. Towards this goal, we propose a structure prediction model that exploits potentials computed from text and RGB-D imagery to reason about the class of the 3D objects, the scene type, as well as to align the nouns/pronouns with the referred visual objects. We demonstrate the effectiveness of our approach on the challenging NYU-RGBD v2 dataset, which we enrich with natural lingual descriptions. We show that our approach significantly improves 3D detection and scene classification accuracy, and is able to reliably estimate the text-to-image alignment. Furthermore, by using textual and visual information, we are also able to successfully deal with coreference in text, improving upon the state-of-the-art Stanford coreference system [7].

1. Introduction

Imagine a scenario where you wake up late on a Saturday morning and all you want is for your personal robot to bring you a shot of bloody mary. You could say *"It is in the upper cabinet in the kitchen just above the stove. I think it is hidden behind the box of cookies, which, please, bring to me as well."* For a human, finding the mentioned items based on this information should be an easy task. For autonomous systems, sentential descriptions can serve as a rich source of information. Text can help us parse the visual scene in a more informed way, and can facilitate for example new ways of active labeling and learning.

Understanding descriptions and linking them to visual content is fundamental to enable applications such as semantic visual search and human-robot interaction. To date, attempts to utilize more complex natural descriptions are rare. Most approaches that employ text and images focus on generation tasks, where given an image one is interested in generating a lingual description of the scene [3, 6, 10, 1], or given a sentence, retrieving related images/videos [12, 8]. An exception is [4], which employed nouns and prepositions extracted from short sentences to boost the performance of object detection and semantic segmentation.

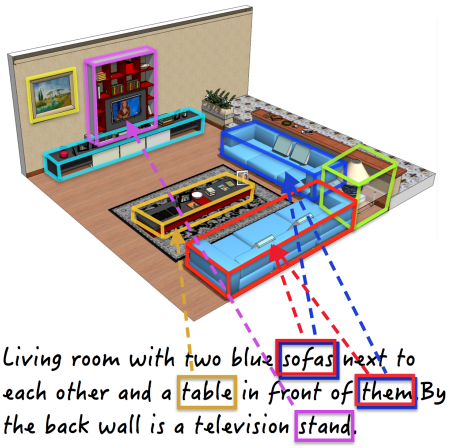

Figure 1. Our model uses lingual descriptions (a string of dependent sentences) to improve visual scene parsing as well as to determine which visual objects the text is referring to. We also deal with coreference within text (e.g., pronouns like "it" or "them").

In this paper we are interested in exploiting lingual descriptions of RGB-D scenes in order to improve 3D object detection as well as to determine which particular object each (pro)noun is referring to in the image. In order to do so, we need to solve the text to image alignment problem (Fig. 1). We propose a holistic model that reasons jointly about the visual scene as well as accompanying text. We demonstrate the effectiveness of our approach in the challenging NYUv2 dataset [11] which we enrich with natural lingual descriptions. A longer version of this paper is in [5].

2. Text to Image Alignement Model

Our input is an RGB-D image of an indoor scene as well as its multi-sentence description. Out goal is to jointly parse the scene and text, and to match text to the visual concepts, performing text to image alignment. We frame the problem as inference in a Markov Random Field (MRF) which reasons about scene type, objects as well as for each (pro)noun which visual concept it describes. To cope with exponentially many object candidates we use bottom-up grouping to generate a smaller set of "objectness" cuboid hypothesis, and restrict the MRF to reason about those.

Parsing Textual Descriptions We extract part of speech tags (POS) of all sentences in a description using [13]. Type dependencies were obtained with [2]. We extract nouns and their attributes from POS. Since the sentences in our de-

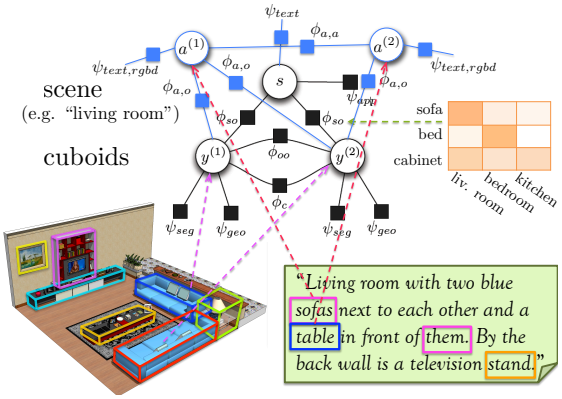

Figure 2. Our model. Black nodes and arrows represent visual information [9], blue are text and alignment related variables.

scriptions are not independent but typically refer to the same entity multiple times, we are faced with the so-called *coreference resolution* problem. For example, in "A table is in the room. Next to it is a chair.", both *table* and *it* refer to the same object and thus form a coreference. To address this, we use the Stanford coreference system [7] to predict clusters of coreferrent mentions.

Visual Parsing Our approach works with a set of object candidates represented as 3D cuboids. We follow [9] to get cuboid candidates by generating ranked 3D "objectness" regions that respect intensity and occlusion boundaries in 3D.

Our Joint Visual and Textual Model We define a Markov Random Field (MRF) reasoning about scene type, 3D objects and which object each (pro)noun refers to. Let s be a random variable for scene type, and y_i be a r.v. associated with a candidate cuboid, encoding its class, where $y_i = 0$ denotes a false positive. For each (pro)noun for a class of interest we generate an indexing r.v. $a_j \in \{0, \cdots, K\}$ (where K is the number of cuboids) that *selects* the cuboid that the noun refers to. The role of $\{a_j\}$ is thus to align text with visual objects. Here $a_j = 0$ means that there is no cuboid corresponding to the noun. For plural forms we generate as many a variables as the cardinality of the (pro)noun. Our MRF energy sums energy terms exploiting image and textual information. Graphical model is in Fig. 2.

We use several potentials: a unary for scene appearance that uses RGB-D and text information. For cuboid unary and pairwise potentials we follow [9]. A sentence describing an object can carry rich information about its properties, as well as 3D relations within the scene. For example, a sentence "There is a wide wooden table by the left wall." provides size and color information about the table as well as roughly where it can be found in the room. To encode this in the model, we use a unary for the alignment variable a trained with RGB-D and text features. We further form a pairwise compatibility term between a and each y_i ensuring that the noun for a agrees with the class of y_i.

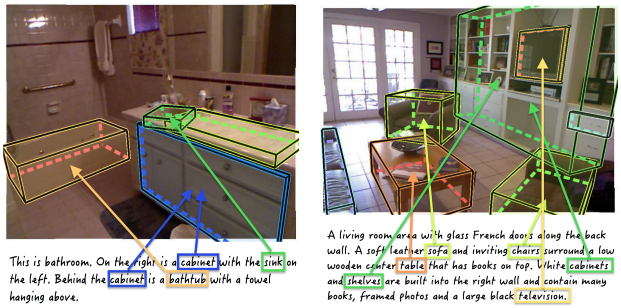

Figure 3. Text to visual object alignment using GT cuboids.

3. Experimental Evaluation

We test our model on NYUv2 which we augment with descriptions. For 3D object detection we use the class set of 21 objects as in [9] and 13 scene classes. We evaluate 3D detection performance, scene classification accuracy, alignment of text to image, and accuracy of the coreference resolution. When using GT cuboids we show a 6.4% improvement over the visual-only model [9] for 3D detection and 14.4% for scene. We improve 2.3% improvement for text-to-image alignment when using our holistic CRF over a unary-only approach. For real cuboids we improve 6.8% for objects and 7.2% for scene. We improve 4.7% over a unary-only approach for text-to-image alignment. A few qualitative examples for GT cuboids are in Fig. 3.

References

[1] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang. Video-in-sentences out. In *UAI*, 2012. 1

[2] M. de Marneffe, B. MacCartney, and C. Manning. Generating typed dependency parses from phrase structure parses. In *LREC*, 2006. 1

[3] 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

[4] S. Fidler, A. Sharma, and R. Urtasun. A sentence is worth a thousand pixels. In *CVPR*, 2013. 1

[5] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *CVPR*, 2014. 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] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916, 2013. 1, 2

[8] D. Lin, S. Fidler, C. Kong, and R. Urtasun. Visual semantic search: Retrieving videos via complex textual queries. In *CVPR*, 2014. 1

[9] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV*, 2013. 2

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

[11] N. Silberman, P. Kohli, D. Hoiem, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 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