Jointly Representing Images and Text: Dependency Graphs, Word Senses, and Multimodal Embeddings

Jointly Representing Images and Text: Dependency Graphs, Word Senses, and Multimodal Embeddings

Frank Keller
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
keller@inf.ed.ac.uk

ABSTRACT

The amount of image data available on the web is growing rapidly: on Facebook alone, 350 million new images are uploaded every day. Making sense of this data requires new ways of efficiently indexing, annotating, and querying such enormous collections. Research in computer vision has tackled this problem by developing algorithms for localizing and labeling objects in images. Object classification algorithms have been recently scaled up to work on thousands of object classes [8] based on the ImageNet database [2].

The next frontier in analyzing images is to go beyond classifying objects: to truly understand a visual scene, we need to identify how the objects in that scene relate to each other, which actions and events they are involved in, and ultimately recognize the intentions of the actors depicted in the scene. The key to achieving this goal is to develop methods for parsing images into **structured representations**. A number of approaches have recently been proposed in the literature, including Visual Dependency Representations [4], Scene Graphs [7], and Scene Description Graphs [1]. All of these models represent an image as a structured collection of objects, attributes, and relations between them.

In this presentation, we will focus on Visual Dependency Representations (VDRs), the only approach to image structure that is explicitly **multimodal**. VDRs start from the observation that images typically do not exist in isolation, but co-occur with textual data such as comments, captions, or tags; well-established techniques exist for extracting structure from such textual data. The VDR model exploits this observation by positing an image structure that links objects through geometric relations. Text accompanying the image can be parsed into a syntactic dependency graph [9], and the two representations are aligned, yielding a multimodal graph (see Figure 1). Well-established synchronous parsing techniques from machine translation [11] can be applied to this task, and resulting VDRs are useful for image description and retrieval [5, 3, 10].

Parsing images into multimodal graph structures is an important step towards image understanding. However, for full understanding, representing the **semantics** of the image is also crucial. For example, the images in Figure 2 can all be described using the verb *play* (and presumably are assigned similar VDRs). However, a different meaning (verb sense) of *play* is evoked by each image. This has led to the new task of visual verb sense disambiguation [6]: given an image and a verb, assign the correct sense of the verb, i.e., the one that corresponds to the action depicted in the image. We propose an unsupervised algorithm based on Lesk which performs visual sense disambiguation using textual, visual, and multimodal embeddings. In this presentation, we will discuss how the two tasks of VDR parsing and visual verb disambiguation can be combined to yield more complete syntactico-semantic image representations, which can then underpin applications such as image retrieval, image description, and visual question answering.

Keywords

Language and vision; image description; image parsing; dependency graphs; word-sense disambiguation

Bio

Frank Keller is professor of computational cognitive science in the School of Informatics at the University of Edinburgh. His background includes an undergraduate degree from Stuttgart University, a PhD from Edinburgh, and postdoctoral and visiting positions at Saarland University and MIT. His research focuses on how people solve complex tasks such as understanding language or processing visual information. His work combines experimental techniques with computational modeling to investigate reading, sentence comprehension, translation, and language generation, both in isolation and in the context of visual information such as photographs or diagrams. Prof. Keller serves on the management committee of the European Network on Vision and Language, is a member of governing board of the European Association for Computational Linguistics, and recently completed an ERC starting grant in the area of language and vision.

1. REFERENCES

[1] S. Aditya, Y. Yang, C. Baral, C. Fermuller, and Y. Aloimonos. From images to sentences through scene description graphs using commonsense reasoning and knowledge. arXiv:1511.03292.

iV&L-MM '16, October 16, 2016, Amsterdam, The Netherlands.

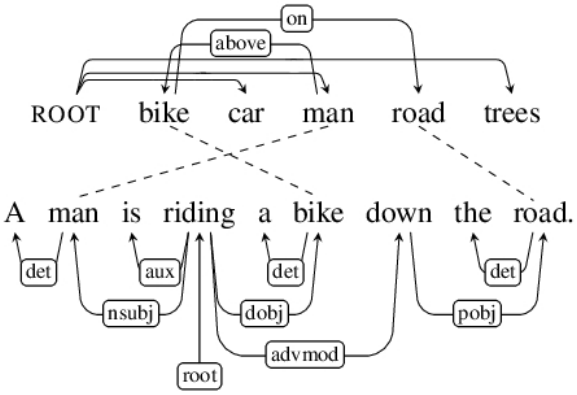
ACM ISBN 978-1-4503-4519-4/16/10..

DOI: http://dx.doi.org/10.1145/2983563.2986050

(a)

A man is riding a bike down the road.
A car and trees are in the background.

(b)



(c)

Figure 1: Example from the Visual and Linguistic Treebank [4]: (a) image annotated with object regions; (b) human-generated image description; (c) visual dependency representation for the image (top) aligned with the linguistic dependency representation for the description (bottom).

(a) (b) (c)

Figure 2: Example from the Verb Sense (VerSe) dataset [6]: three visual senses of the verb *play*: (a) participate in sport, (b) play an instrument, and (c) be engaged in playful activity. The images are taken from MS COCO, which also includes image segmentations and descriptions.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[3] D. Elliott and A. de Vries. Describing images using inferred visual dependency representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 42–52, Beijing, 2015.

[4] D. Elliott and F. Keller. Image description using visual dependency representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Seattle, WA, 2013.

[5] D. Elliott, V. Lavrenko, and F. Keller. Query-by-example image retrieval using visual dependency representations. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 109–120, Dublin, 2014.

[6] S. Gella, M. Lapata, and F. Keller. Unsupervised visual sense disambiguation for verbs using multimodal embedding. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, San Diego, 2016.

[7] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[9] J. Nivre. An efficient algorithm for projective dependency parsing. In *Proceedings of the International Workshop on Parsing Technologies*, pages 149–160, Nancy, 2003.

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

[11] D. A. Smith and J. Eisner. Parser adaptation and projection with quasi-synchronous grammar features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 822–831, Singapore, 2009.