Int J Comput Vis (2017) 123:1–3
DOI 10.1007/s11263-017-0993-y

Guest Editorial: Image and Language Understanding

Margaret Mitchell[1] · John C. Platt[1] · Kate Saenko[2]

Published online: 31 March 2017

We are pleased to present this special issue of IJCV on combined image and language understanding. It contains some of the latest work in a long line of research into problems at the intersection of computer vision and natural language processing.

Research on language and vision has been stimulated by recent advances in object recognition. While multi-layer (or "deep") models have been applied for more than twenty years (Lawrence et al. 1997; LeCun et al. 1989; Nowlan and Platt 1995), recently they have been shown to be extremely effective at large-vocabulary object recognition (Krizhevsky et al. 2012) and at text generation (Mikolov et al. 2010). The next logical step was to combine these two tasks to enable image captioning: generating a short language description based on an image (Kulkarni et al. 2013; Mitchell et al. 2012). In 2015, deep models produced state-of-the-art results in image captioning (Donahue et al. 2015; Fang et al. 2015; Karpathy and Fei-Fei 2015; Vinyals et al. 2015). These results were facilitated by the MSCOCO data set, which provided multiple crowd-sourced labels for thousands of images (Lin et al. 2014).

The success of deep image captioning initially seemed promising. Had we finally solved combined image and language understanding? A closer inspection, however, revealed that such understanding was far from solved: Approaches that learned to repeat common patterns and relationships in the dataset were capable of doing well (Devlin et al. 2015; Zitnick et al. 2016), while not necessarily generalizing to new images. The success of parroting captions posed a challenge to researchers, who became inspired to create new tasks that push the boundaries of what computers can see and say.

This special issue of the International Journal of Computer Vision captures some of this new research. We received a number of submissions covering a wide variety of new problems in vision and language, a testament to the expanding possibilities in this newly burgeoning field. We selected an assortment of papers that capture some of this diverse research, but it is far from exhaustive. The papers in this issue introduce new problems, solutions, and datasets in vision and language, advancing vision and language research towards deeper image understanding and the ability to communicate about that understanding. The topics in this issue include:

- *Visual Question Answering* (VQA)—instead of generating free-form text, one can build a system that answers intelligent questions about an image. The task of VQA for testing image understanding was first proposed in 2014 (Malinowski and Fritz 2014), which stimulated a number of subsequent papers (Antol et al. 2015; Jabri et al. 2016; Ren et al. 2015). In this issue, the paper titled "VQA: Visual Question Answering" follows up on (Antol et al. 2015) and presents improved models and the results of a VQA challenge held in 2016.
- *Relationships between Objects*—many image understanding tasks require modeling not only the objects in the image, but also the relationships between them (Malisiewicz and Efros 2009; Parikh et al. 2008). However, existing image datasets lack appropriate annotations. In the paper "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations,"

✉ Kate Saenko
saenko@bu.edu

Margaret Mitchell
mmitchellai@google.com

John C. Platt
platt@google.com

[1] Google, 601 N 34th St, Seattle, WA 98103, USA

[2] Computer Science Department, Boston University,
111 Cummington Mall, Boston, MA 02215, USA

the authors introduce a large-scale dataset to advance research in this area. The Visual Genome dataset represents the largest densely-annotated collection of objects, relationships and attributes to date. The paper also introduces a formalized representation of entities and their relationships in the form of a scene graph, and connects the entities to the widely used WordNet semantic database.

- *Region-to-Phrase Correspondence*—going beyond producing a single caption for the entire image, this task establishes correspondences between descriptive phrases and image sub-regions (Sadeghi and Farhadi 2011). In the paper "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models," the authors extend the popular Flickr30k dataset to link mentions of objects in its captions to appropriate regions. They also add co-reference chains linking different mentions of the same region across multiple captions. This work enables a new benchmark for localizing text in images and presents a strong baseline for this problem.
- *Movie Captioning* Similar to how captions may be generated from a single frame, they may also be generated from many frames—such as the many frames per second in a video. Such work adds information about the context in which the image was taken, leveraging the temporal stream of events that give rise to a description (Guadarrama et al. 2013). Movies represent a potential source of data, as they often contain an Audio Description channel that allows visually impaired audiences to follow along by describing what is happening on screen. The paper titled "Movie Description" leverages this data source and contributes a large scale benchmark for movie captioning. The dataset contains clips from 200 movies as well as transcriptions of the audio captions aligned to the clips. The paper also describes the results of two challenges held in 2015 and 2016.

We hope that readers enjoy this deep dive into modern vision and language research.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2425–2433).

Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., & Mitchell, M. (2015). Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing* (Vol. 2: short papers, pp. 100–105). Association for Computational Linguistics, Beijing, China. http://www.aclweb.org/anthology/P15-2017.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625–2634).

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., & Platt, J. C., et al. (2015). From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1473–1482).

Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., & Saenko, K. (2013). Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *2013 IEEE international conference on computer vision (ICCV)* (pp. 2712–2719). IEEE.

Jabri, A., Joulin, A., & van der Maaten, L. (2016). Revisiting visual question answering baselines. In *European conference on computer vision* (pp. 727–739). Berlin: Springer.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3128–3137).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., et al. (2013). Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2891–2903.

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, *8*(1), 98–113.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European conference on computer vision* (pp. 740–755). Berlin: Springer.

Malinowski, M., & Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems* (pp. 1682–1690).

Malisiewicz, T., & Efros, A. (2009). Beyond categories: The visual memex model for reasoning about object relationships. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22, pp. 1222–1230).

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Interspeech* (Vol. 2, pp. 1045–1048). Makuhari, Chiba: ISCA.

Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., & Daumé III, H. (2012). Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th conference of the European chapter of the association for computational linguistics* (pp. 747–756). Association for Computational Linguistics.

Nowlan, S. J., & Platt, J. C. (1995). A convolutional neural network hand tracker. In *Advances in neural information processing systems* (pp. 901–908).

Parikh, D., Zitnick, C. L., & Chen, T. (2008). From appearance to context-based recognition: Dense labeling in small images. In *IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008* (pp. 1–8). IEEE.

Ren, M., Kiros, R., & Zemel, R. (2015). Exploring models and data for image question answering. In *Advances in neural information processing systems* (pp. 2953–2961).

Roberts, L. G. (1963). Machine perception of three-dimensional solids. Ph.D. thesis, MIT.

Sadeghi, M. A., & Farhadi, A. (2011). Recognition using visual phrases. In *2011 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1745–1752). IEEE.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156–3164).

Zitnick, C. L., Agrawal, A., Antol, S., Mitchell, M., Batra, D., & Parikh, D. (2016). Measuring machine intelligence through visual question answering. *AI Magazine*, *37*(1).