Cross-Modal Learning for Scene Understanding

Karteek Alahari: karteek.alahari@inria.fr
Cordelia Schmid: cordelia.schmid@inria.fr
http://thoth.inriaples.fr

Location of the internship: The internship will be in the Thoth team at Inria Grenoble, and will be co-supervised by Karteek Alahari (Inria researcher) and Cordelia Schmid (Inria Research Director). The team is specialized in computer vision, in particular visual recognition.

Topic: The main goal of this work is the development of a model learned from different sources, including image, video and text data. This goal will be achieved through addressing the problem of cross-model learning in the context of CNNs. The internship will focus on advancing the state of the art in learning representations for scene understanding from multiple data modalities [6, 8]. In particular, the objective is to use text data as a form of weak supervision and also to exploit its well-grounded formalism through linguistic models [4]. The formulation will be based on graph convolution networks [2,5] and graph matching [7,10]. The nodes of the graph represent individual data items, and each edge in the graph denotes the relationship between the two data items of the nodes it connects. For instance, in the case of two image regions, it will represent their similarity or relationships such as *is on*, *part of*, *occurs in* (e.g., a cup is on the table) [9]. The challenge now is to not only learn these relationships, but also the representation of each graph node and the structure of the graph, formulated as a joint learning problem. Our initial work in learning graph structure for matching [1] and learning semantic correspondence [3] will form the basis for this.

Skills and profile: The student must have solid programming skills as well as solid mathematics knowledge (especially linear algebra and statistics). Knowledge of deep learning tools is a strong plus.

References

[1] M. Cho, K. Alahari, and J. Ponce. Learning graphs to match. In *ICCV*, 2013.

[2] M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS*, 2016.

[3] K. Han, R. Rezende, B. Ham, K.-Y. Wong, M. Cho, C. Schmid, and J. Ponce. SCNet: Learning Semantic Correspondence. In *ICCV*, 2017.

[4] J. Hockenmaier. *Data and Models for Statistical Parsing with Combinatory Categorial Grammar*. PhD thesis, University of Edinburgh, 2003.

[5] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[6] H. Kuehne, A. Richard, and J. Gall. Weakly supervised learning of actions from transcripts. *CVIU*, 2017.

[7] L. Lovász and M. D. Plummer. *Matching theory*. North-Holland, Amsterdam, New York, 1986.

[8] A. Miech, J.-B. Alayrac, P. Bojanowski, I. Laptev, and J. Sivic. Learning from video and text via large-scale discriminative clustering. In *ICCV*, 2017.

[9] J. Peyre, I. Laptev, C. Schmid, and J. Sivic. Weakly-supervised learning of visual relations. In *ICCV*, 2017.

[10] J. Yan, M. Cho, H. Zha, X. Yang, and S. M. Chu. Multi-graph matching via affinity optimization with graduated consistency regularization. *PAMI*, 2016.