Résumé du projet de recherche (Langue 1)

After the huge success of large Convolutional Neural Networks (CNNs) at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012, deep learning nowadays appears as the dominant technique for many visual data understanding tasks. Applications which were previously out of reach, such as generating sentences from images, or visual query answering, recently emerged into the computer vision and machine learning communities.

Résumé du projet de recherche (Langue 2)

The goal of this Ph.D. proposal is to further study such deep architectures for unified image and textual representations. Firstly, we aim at exploring multi-modal embeddings, with the goal to learn a joint representation from heterogeneous modalities, e.g. image and text. A particular interest will be given to training schemes based on aligning representations in the joint space. From this perspective, several applications will be addressed: Visual Query answering -- Tag-to-image and image-to-tag search -- Image-to-caption search on large-scale multimodal collection. A second aspect of this thesis is to go deeper toward alignment between image and text modalities, especially by incorporating spatial information. Basically, we aim at matching image and text regions based on their semantics. In this context, having precise annotations for large scale datasets in not a viable solution, due to the expensiveness of the labeling. To overcome this issue, weakly supervised learning strategies dedicated to automatically selecting relevant visual and textual locations from coarse annotations will be studied [DTC15,DTC16]. Finally, to push forward the relaxation of annotations, unsupervised learning methods will be explored. In particular, we want to extend recent work on ladder network [RVH+15], where the reconstruction scheme is questioned by not asking to the internal representation to do the job alone, but adding skip connections coming from the input. One interesting option would be to explicitly model specific representations for each example, which are irrelevant for a given supervised task. Basically, the idea of the training scheme is to separate the extraction of invariant representations, useful for the supervised task, and variant features (i.e. specific to each example), needed to reconstruct each training sample. The underlying assumption is that the explicit decomposition of variant and invariant features drives the learning towards more effective (robust) representations. [DTC15] Thibaut Durand, Nicolas Thome, Matthieu Cord. MANTRA: Minimum Maximum LSSVM for Image Classification and Ranking, ICCV 2015. [DTC16] Thibaut Durand, Nicolas Thome, Matthieu Cord. WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks, CVPR 2016. [RVH+15] Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, Tapani Raiko. Semi-Supervised Learning with Ladder Networks, NIPS 2015.