Proposition de thèse
LABEX SMART

Titre : Deep learning for multi-media recognition / Apprentissage profond pour la reconnaissance multi-modale texte/image

Direction de thèse : Nicolas Thome et Matthieu Cord
Laboratoire d'accueil : **L**aboratoire d'**I**nformatique de **P**aris **6** UMR 7606
École doctorale : EDITE

Labex SMART : Ce sujet se positionne dans l'axe "Le développement des services numériques pour l'accès à la connaissance et à l'information, le traitement des données numériques". En effet, c'est un sujet de recherche orienté sur la modélisation et l'interprétation de données visuelles et le développement de nouvelles techniques d'apprentissage pour le traitement de données massives multimédia (image et texte) à large échelle. Notons également que les aspects big data sont présents dans les axes de l'IUIS.

Context

After the huge success of large Convolutional Neural Networks (CNNs) at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 [KSH12], deep learning nowadays appears as the dominant technique for many visual data understanding tasks. Beyond the outstanding performances for large-scale recognition, representations extracted from CNNs trained on ImageNet ("deep features") prove to be very effective for transfer learning, so that state-of-the-art results for any visual recognition tasks are now obtained with deep features as input, *e.g.* object detection [GDDM14], image segmentation [PCMY15], pose estimation [SRY+15], *etc.* Learning deep representations from various modalities, *e.g.* images and text, also recently emerged into the computer vision and machine learning communities.

The goal of this Ph.D. proposal is to further study such deep architectures for unified image and textual representations. In this context, we plan to investigate several applications related to Multimedia processing for human related purposes (see example in Fig 1) :

— Image-to-image search for large-scale visual content retrieval.
— Tag-to-image and image-to-tag search for large-scale Internet multimedia database.
— Image-to-caption search on large-scale multimodal collection. Indeed, the possibility of automatically describing the content of an image using text caption (few sentences) is emerging.

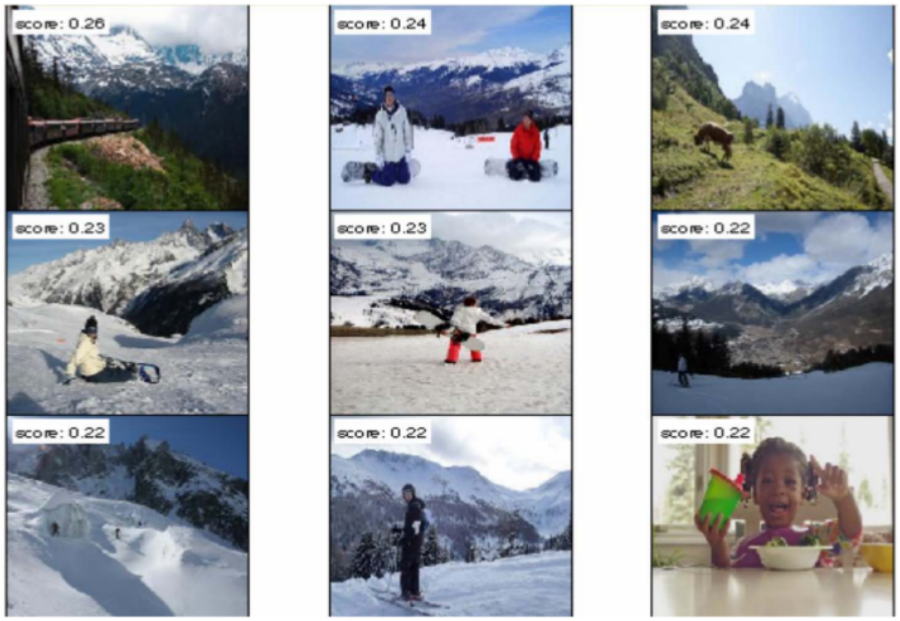

FIGURE 1 – Result on COCO dataset for query : girl playing in snow near mountain.

Objectives

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, *e.g.* Canonical Correlation Analysis (CCA) [And84] and its recent deep extension [AABL13]. From this perspective, several applications given in Fig 1 will be addressed, especially Visual Query answering (VQA).

A second aspect of this thesis is 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 [OBLS15, 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.

Références

[AABL13] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 1247–1255, May 2013.

[And84] T. W. Anderson, editor. *An Introduction to Multivariate Statistical Analysis*. Wiley, 1984.

[DTC15] Thibaut Durand, Nicolas Thome, and Matthieu Cord. MANTRA : Minimum Maximum Latent Structural SVM for Image Classification and Ranking. In *International Conference on Computer Vision (ICCV)*, 2015.

[DTC16] Thibaut Durand, Nicolas Thome, and Matthieu Cord. WELDON : Weakly Supervised Learning of Deep Convolutional Neural Networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[GDDM14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[OBLS15] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free ? – weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015.

[PCMY15] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *IEEE International Conference on Computer Vision, ICCV 2015*, pages 1742–1750, 2015.

[RVH$^+$15] Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder network. *CoRR*, abs/1507.02672, 2015.

[SRY$^+$15] James Steven Supancic, Grégory Rogez, Yi Yang, Jamie Shotton, and Deva Ramanan. Depth-based hand pose estimation : data, methods, and challenges. In *ICCV*, Santiago, Chile, December 2015.