Titre du sujet de thèse :
Deep Learning for unified visual and textual representation: from images to captions / Apprentissage profond de représentations conjointes visuelle et textuelle : de l'image à la légende

Direction de thèse : M. Cord et P. Gallinari

Equipe / Laboratoire de l'UPMC : équipe MLIA , LIP6

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 a large Convolutional Neural Networks (CNN) architecture [1] at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012, deep learning appears nowadays as the dominant technique for many visual data understanding tasks. These deep approaches involve learning every step in the process, from the pixels to the final label assignment passing through a set of features, from very local to high level ones. One of the keys of this success is the learning of a powerful embedding of the visual data through different layers of the network before getting more specific to the final task in the last layers.

This problem is related to the generic representation learning problem in Machine Learning [2]. The basic idea is to learn an embedding of raw data into a semantical high dimensional continuous space. Many different tasks have emerged very recently based on this idea of having a common latent space where heterogeneous data may be embedded and from where any king of reconstruction may be generated [3].

Objectives

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 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.

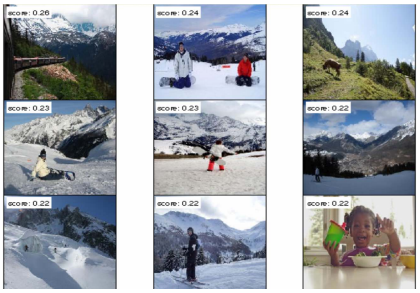

Figure 1: Result on COCO dataset for query: girl playing in snow near mountain

To achieve our objectives, we propose to structure the PhD program with two main parts:

- building useful visual representation. We will consider new components or improved existing components of deep visual architectures, and optimizing the hyper-parameter learning procedures in specific supervised models, focusing specially on the compression of information in such networks. Despite achieving state-of-the-art results for many problems, CNNs suffer

from intrinsic limitations related to geometric transformations in the input. The only invariance truly treated by such models is the invariance to translation (strongly localized), which is coded by pooling (aggregation) operations in deep architectures. We intend to explore clues in the architectures that would allow us to directly encode invariance to scale, as in [4], and detection problems related to small objects in large images (with strategies as R-CNN [5]). We also plan to investigate learning problems in the network, as the reduction of the number of parameters following the recent proposition of mimic deep nets with shallow ones [6].

- learning unified latent space for visual and textual information. Starting with the deep visual space previously learnt and a deep textual data representation (as word2vect), a new joint embedding may be learned. We will consider joint optimization to align the outputs, as CCA and deep CCA [7]. If relevant, others techniques for learning representations will be considered. Ultimately, this final representation allows to easily associate images and words. Extensions to sentence generation may be studied from this unified representation (following the recent Google strategy [3]). We will develop several applications as mentioned about visual retrieval using different types of queries.

Experimental validation will be performed on large-scale datasets, ImageNet and multimedia COCO dataset. Several libraries offer support to convolutional networks (Caffe [8], MatConvNet [9], FAIR for Torch [1]). We will consider the latter one because of its flexibility. This research will benefit from all of the computational power available at LIP6 (many GPU nodes), and UPMC.

References

[1] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *NIPS*, pp. 1–9, 2012.

[2] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1798–1828, 2013.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *arXiv preprint arXiv:1411.4555*, 2014.

[4] A. Kanazawa, A. Sharma, and D. Jacobs, "Locally scale-invariant convolutional neural network," *Deep Learning and Representation Learning Workshop: NIPS*, 2014.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*. IEEE, 2014, pp. 580–587.

[6] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Advances in Neural Information Processing Systems*, 2014, pp. 2654–2662.

[7] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML*, 2013, pp. 1247–1255.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[9] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *British Machine Vision Conference*, 2014.

[1] https://research.facebook.com/blog/879898285375829/fair-open-sources-deep-learning-modules-for-torch/