Résumé du projet de recherche (Langue 1)

Imagine you are asked to answer the following questions. How can we design a computer-aided diagnosis tool for neurological disorders from multiple brain images acquired with different medical imaging devices? How can a computer identify the emotion felt by a person from his/her face and voice? How can such tools remain efficient when the data quality is poor or with missing data? This thesis aims at addressing these questions from the representation learning point of view.

Informations complémentaires (Langue 1)
Many application fields naturally deal with multiview data. For example, in various multimedia learning tasks [2], text, audio and images (different framings from videos) are starting points for these views. Each starting view can then produce other different views through specific descriptor computations (e.g., a digital image can be represented by Gabor filters, SURF, color histograms, etc.). Besides, multimedia learning tasks led to a large amount of fusion-based ad-hoc approaches and experimental results. The success of multiview supervised learning in the multimedia community seems to rely on the ability of the systems to deal with the complementarity of the information carried by each modality. Comparable studies are of importance in many other domains, such as bioinformatics [19], speech recognition [1,12], signal-base multimodal integration [23], gesture recognition [22], etc.

However, despite the numerous published successes of multiview learning regardless of the application domain, no common or formal evidences of multiview learning (dis-)advantages have been proved so far, neither about inter-view specificities nor on performance gains.

Moreover, multiview learning has been theoretically studied mainly under the semi-supervised setting, but only with two facing views [5, 20, 21, 11]. In parallel, ensemble-based learning approaches have been theoretically studied in the supervised setting yielding many results also interesting from a multiview learning perspective since ensembles can be built upon multiple views [17, 24].

Inspired by the emerging field of Representation Learning, we propose to develop these techniques for the specific case of multiview prediction models. The thesis will be decomposed in three related tasks:

• A first one where the goal is to associate to each input x a representation z in a vectorial latent space \( \mathbb{R}^n \). It is denominated as explicit representations in the following since the vector z has to be explicitly computed in order to make predictions.

• A second one where the goal is to learn a complex metric in the input space. It is denominated as implicit representations since the learned metric can be seen as a "classical" distance computed in an unknown latent space. The representations do not need to be computed, as predictions can be made based only on the discovered metric.

• Lastly, the third task has for objective to study the relationships and differences between the methods developed in the two previous tasks. We also plan to work on approaches able to combine them in a joint formulation.

The two different types of approaches are interesting since each of them is adapted to particular learning configurations. Explicit representations are particularly well suited for missing views problem while metric learning naturally fit with the relational views setup. Combining both approaches is a more exploratory but promising direction to develop more general and accurate representation models.