A multiview learning approach for image annotation with visual and textual features

Résumé du projet de recherche (Langue 1)

Most photo sharing sites like Flickr give their users the opportunity to manually assign class label to images. These labels provide descriptive keywords for the image, and play an important role in the organization of the image collection, since they can be used to browse or search in collection of images. Their usage is however limited to the small part of images that are manually labeled. Automating the annotation process is mandatory to extend the categorization to the entire ever-growing image collection. In this project, we consider the problem of automatic image categorization, emphasizing on two specific properties of image collections: (a) for many categories, we may have a very limited number of labeled examples, but a very large number of additional, unlabeled images and (b) images can naturally be represented in several distinct feature spaces. Among these feature spaces, we distinguish visual feature spaces such as bag-of-visual words obtained from SIFT descriptors or color histograms, from a textual feature space a bag-of-words representation extracted from the text surrounding the image posted on a Website. Our approach relies on the various possible representations – also called views – of each image following the multiview learning paradigm. Each visual feature space, as well as the textual representation of the image is used to train a classifier on the labeled training set. After this first step where we obtain as many classifiers as available views, we adopt a consensus-based self-training algorithm to carry out semi-supervised learning: each unlabeled image is classified by the single-view classifiers, and, among these images, those for which the majority of the classifiers predict the same class label are pseudo-labeled (according to the majority vote) and added to the initial training set. The view-specific classifiers are then re-trained, and the procedure is repeated until convergence is achieved. When learning is finished, test images are labeled according to the majority vote of all single-view classifiers. Multiview learning has previously been used for image categorization. We use of textual features and the semi-supervised learning algorithm rather than active learning – the latter requiring a human to label new examples. We consider each type of visual feature as a separate view (in addition to the textual one) to obtain more reliable pseudo-labels during the semi-supervised learning step, but also our use of all views during test using a late fusion scheme (i.e. majority vote). Note that the possibility of using more than two views come from our adoption of consensus-based algorithms instead of the original co-training.