Abstract

The last decade has witnessed an explosion of visual data (images and videos) on the Internet, thanks to the availability of on-line photo-sharing websites and cheap media capturing devices. As a result, it has become necessary to develop new technologies that can help in efficiently archiving, accessing and understanding such large collections of visual data. This dissertation is a small step towards this, where we address the problem of automatically learning semantics of visual data (specifically images) in the form of natural text.

Though the domain of expressing image semantics in the form of text is very wide with several associated sub-problems and applications, we have targeted four specific challenges in this space: (1) To achieve good performance on rare labels that are usually more informative and distinctive than the frequent ones (without compromising on the frequent labels) in the image tagging task. (2) To use data from multiple sources and efficiently integrate them to both generate semantically meaningful descriptions for images, and retrieve images given descriptive textual queries. (3) To propose a structured prediction based technique for cross-modal (image↔text) retrieval task. (4) To design a generic model that can act as a wrapper over existing cross-modal retrieval techniques by making use of additional meta-data, and helps in boosting their performance.

We evaluate the proposed methods on a number of popular and relevant datasets. On the image annotation task, we achieve near state-of-the-art results under multiple evaluation metrics. On the image captioning task, we achieve superior results compared to conventional methods that are mostly based on visual cues and corpus statistics. On the cross-modal retrieval task, both our approaches provide compelling improvements over baseline cross-modal retrieval techniques.