Language Guided Visual Perception

Mohamed H. Elhoseiny
Dept. of Computer Science

9/9/2016 at 12:00 pm
CoRE A (301)

Abstract

People typically learn through exposure to visual stimuli associated with linguistic descriptions. For instance, teaching visual concepts to children is often accompanied by descriptions in text or speech. This motivates the question of how this learning process could be computationally modeled. In this dissertation we explored three settings, where we showed that combining language and vision is useful for machine perception using images and videos.

First, we addressed the question of how to utilize purely textual description of visual classes with no training images, to learn explicit visual classifiers for them. We propose and investigate two baseline formulations, based on regression and domain transfer that predict a classifier. Then, we propose a new constrained optimization formulation that combines a regression function and a knowledge transfer function with additional constraints to predict the classifier parameters for new classes. We also proposed kernelized models, which allow utilizing any kernel functions in the visual space and text space. We applied the models to predict visual classifiers for two fine-grained categorization datasets, and the results indicate successful predictions against several baselines.

Second, we modeled searching for events in videos as a language and vision problem, where we proposed a zero-shot event detection method using multi-modal distributional semantic embedding of videos. Our zero-shot event detection model is built on top of distributional semantics and extends it in the following directions: (a) semantic embedding of multimodal information in videos (with focus on the visual modalities), (b) automatically determining relevance of concepts/attributes to a free text query, which could be useful for other applications, and (c) retrieving videos by free text event query based on their content. We validated our method on the TRECVID MED (Multimedia Event Detection) challenge. Using only the event title as a query, our method outperformed the state-of-the-art that uses big descriptions.

Third, and motivated by the aforementioned results, we proposed a uniform and scalable setting to learn unbounded number of visual facts. We proposed models that can learn, not only objects, but also their actions, attributes and interactions with other objects, in one unified learning framework and in a never-ending way. The training data comes as structured facts in images, including (1) objects (e.g., <boy>), (2) attributes (e.g.,<boy, tall>), (3) actions (e.g., <boy, playing>, and (4) interactions (e.g., <boy, riding, a horse >). We have worked on the scale of 814,000 images and 202,000 unique visual facts. Our experiments show the advantage of relating facts by the structure in the proposed models compared to four designed baselines on bidirectional fact retrieval.

Defense Committee: Prof. Ahmed Elgammal (Chair), Prof. Casimir Kulikowski, Prof. Abdeslam Boularias,
Prof. Abhinav Gupta (Carnegie Mellon University)