Multi-modal Learning

Danijel Skočaj¹, Matej Kristan¹, Alen Vrečko¹, Aleš Leonardis¹, Mario Fritz², Michael Stark², Bernt Schiele², Somboon Hongeng³, and Jeremy L. Wyatt³

¹ VICOS Lab, University of Ljubljana, Ljubljana, Slovenia
first.last@fri.uni-lj.si

² Technische Universität Darmstadt, Darmstadt, Germany
lastname@informatik.tu-darmstadt.de

³ Intelligent Robotics Laboratory, School of Computer Science, University of Birmingham, Birmingham, UK
{jlw,sxh}@cs.bham.ac.uk

7.1 Introduction

The main topic of this chapter is learning, more specifically, multimodal learning.

In biological systems, learning occurs in various forms and at various developmental stages facilitating adaptation to the ever changing environment. Learning is also one of the most fundamental capabilities of an artificial cognitive system, thus significant efforts have been dedicated in CoSy to researching a variety of issues related to it.

Learning involves numerous deep problems that stretch far beyond straightforward applications of current statistical methods. In particular, learning should take place in a rich interaction of an artificial cognitive system with the environment or with a human (tutor) exploiting multiple modalities to arrive at new concepts and/or extend the current ontologies or merely to adapt to different circumstances.

One of the main topics which was addressed within the CoSy project was the selection of appropriate representations that allow for their efficient creation and modification as the new data is acquired, and the levels and types of supervision that guide the learning processes. In this context, the exploitation of multiple modalities is crucial as they provide means for robust and efficient learning that is not possible within a single modality. For example, the data of one modality may act as a means of supervision for learning within another modality, or, through a simple dialogue inherent ambiguities present in a visual signal can be resolved. Similarly, an artificial cognitive system can learn important insights into how the environment is structured, including causal relations, by observing activities or exercising certain actions.
Therefore, the learning we are addressing in this chapter involves information from multiple modalities. One modality helps another modality by providing additional information, i.e., one modality supervises another modality. In Sections 7.2 and 7.3 we address the multimodal learning in the context of interaction between the visual and communication modalities. The PlayMate system we have developed has the capability of relating the visual information and the information obtained through the communication subsystem. In this way, the learning of visual concepts can be supervised by language. In Sections 7.4 and 7.5 the pure visual information is interacting with the information produced by performing certain actions (pushing, grasping). Broadly speaking we can view visual information and the information obtained from actions as two separate sub-modalities. In this view, sections 7.4 and 7.5 address interaction between the visual sub-modality and the action sub-modality. Again, visual cues are learnt that could not have been learnt without having access to information from different origins (sub-modalities). Note that in the literature, the multimodal learning, as defined here, is also frequently referred to as cross-modal learning. Both terms emphasize the interaction between different modalities during the learning process and are often used interchangeably; so they are in this book.

In terms of the levels of supervision, a variety of different modes have been applied, ranging from fully supervised, weakly supervised, to unsupervised. As for the types of supervision they were achieved through dialogues, perception of affordance cues and exploratory strategies. The approaches presented in Sections 7.2 and 7.3 focus on analysing different levels of supervision, which is achieved through a dialogue with a human teacher. The integrated system we have developed facilitates such kind of research and provides means for applying different learning strategies in an interactive learning setting.

The remainder of the chapter contains four sections which address the issues mentioned above in different contexts.

Section 7.2 presents an interactive framework for continuous learning of visual concepts. The main goal that was set is to learn associations between automatically extracted visual features and words describing the scene (visual attributes and spatial relations) in a user friendly, natural, open-ended, and continuous manner. The system facilitates interactive learning of basic visual concepts in a dialogue with a human tutor. The learning can be performed with different levels of supervision. The developed framework also supports unlearning, which enables recovery from any errors accumulated in the learned models. The entire learning process is fully integrated with other parts of the system that provide information needed by the learner and use the information produced by the learner. The entire system thus provides visual input and enables verbal and non-verbal communication with a tutor facilitating continuous and interactive cross-modal learning.

In Section 7.3 crossmodal learning of visual categories is performed in a way that combines supervised and unsupervised training methods. While supervised methods tend to produce more accurate results, unsupervised methods