Multimodal Representation Learning for Robotic Cross-Modality Policy Transfer

Doctoral Consortium

Miguel Vasco
INESC-ID & Instituto Superior Técnico, University of Lisbon
Lisbon, Portugal
miguel.vasco@tecnico.ulisboa.pt

ABSTRACT

In this thesis, we aim at endowing robots with mechanisms to learn multimodal representations from sensory data and to allow them to execute tasks considering different subsets of available perceptions. We address the learning of these representations from supervised, unsupervised and reinforcement learning methodologies in the context of virtual agents and robots. We hope that, by achieving the proposed goals, the contributions of this thesis might prompt future research on applications of multimodal representations in robots and other artificial agents.

KEYWORDS

Deep learning; Deep reinforcement learning; Machine learning for robotics; Knowledge representation and reasoning in robotic systems.

ACM Reference Format:

1 INTRODUCTION

Humans are provided with a remarkable cognitive framework which allows them to create a rich representation of their internal and external reality. These representations may be of a conceptual nature, regarding the categorization and interplay of abstract models of existing (or non-existing) entities, or of a perceptual nature [2, 6]. Perceptual representations are the result of multiple levels of processing of multimodal information provided by the environment, captured by the different sense organs [3, 19].

Perceptual representations play a fundamental role in the planning and execution of tasks [12]. Here we distinguish two different categories of tasks: modality-specific tasks, in which the information of a given modality or subset of modalities is fundamental for its execution (e.g., sorting objects by color), and modality-independent tasks, which can be executed by considering redundant information from different modalities (e.g., navigating within a room). An example of the latter class is shown in Figure 1.

Humans are able to plan and perform modality-independent tasks even if the environment does not provide modality-specific information (e.g., absence of light in a dark room) or if a given sensor is malfunctioning (e.g., blindness), albeit with reduced performance. Indeed, multimodal perceptual representations allow for the inference of the perceptual experiences of missing modalities from available ones [18, 23, 27].

Artificial agents, such as robots, often disregard the relationships between the different modalities that compose their perceptual input. Robots often limit themselves to creating internal representations solely from visual information [5, 21] or from the fusion of different modalities [9, 15]. Such disregard results in the inability of the robot to perform modality-independent tasks when modality-specific information is unavailable, or in the (frequent) case of sensory malfunction. If we aim at having artificial agents, such as service robots or autonomous vehicles, acting reliably in their environments, they must be provided with mechanisms to overcome...
these issues. This thesis aims at endowing robots with mechanisms to learn multimodal representations of their environment and to allow them to execute modality-independent tasks considering different subsets of available perceptions.

2 LEARNING MULTIMODAL REPRESENTATIONS

With that goal in mind, the question of the learning methodology of such representations naturally emerges. Human perceptual representations are continuously learnt and shaped by different learning mechanisms, including supervised and unsupervised and reinforcement learning [4].

In previous work, we addressed the question of creating multimodal representations through supervised learning in the context of human action recognition [25, 26]. Indeed, in human infancy, supervised learning plays a fundamental role in object categorization from few labels provided by a teacher [13]. Our goal was to access if, by considering the multimodal nature of the information provided by a human teacher, the agent could distinguish between different action classes from few training examples. We introduced the notion of motion concept, a representation of the kinematics of the action, along with the contextual background of the action (the location and the objects used during the action). We proposed an online algorithm to learn motion concepts by demonstration and evaluated its performance in both offline [26] and online [25] recognition tasks. The results showed the importance of considering multimodal information in building action representations.

Unsupervised learning also plays a fundamental role in the learning process of human multimodal representations. In particular, infants apply unsupervised learning to leverage statistical regularities in perceptual data to learn the distribution of sounds in their native language [11, 28], to discover simple visual categories [1, 30, 31], and to refine sensory-motor maps [16]. In a work accepted at AAMAS 2020, we address the challenge of building multimodal representations through unsupervised learning in Atari games [22]. In this work, we introduced and formalised the novel problem of modality transfer in deep reinforcement learning. We proposed a three-stage architecture that allows a reinforcement learning agent trained over a given sensory modality to execute its task on a different sensory modality, as presented in Figure 2. In a first stage, we employed multimodal variational auto-encoders [10, 24, 29] to learn a representation of the game scenario in an unsupervised way, considering both the image and sound generated by the game. We evaluated the proposed approach in domains of increasing complexity and showed that the policies learned by our approach were robust to different subsets of available input modalities.

3 FUTURE WORK

So far we addressed how artificial agents can learn multimodal representations through supervised and unsupervised learning. As such, our future work will focus on two different goals:

1. Understand how reinforcement learning can augment multimodal perceptual representation learning of artificial agents in goal-oriented, modality-independent tasks.
2. Extend multimodal perceptual representation learning to robotic agents in a situated environment with humans.

Figure 2: Proposed architecture to address modality transfer in Atari games: using unsupervised learning, we build a multimodal representation $z^c$ from images and sounds collected from the game. Subsequently, we train a policy $\pi$ of a reinforcement learning agent using rewards $R$ obtained considering $z^c$ as states of the world. This allows us apply the same policy $\pi$ for different subsets of available input modalities.

Reinforcement learning plays a significant role in the shaping of human mental representations constructed from perceptual data through unsupervised learning [14]. However, such mechanisms have yet to be translated to a computational setting. Several approaches have been proposed that consider unsupervised algorithms to learn single-modality [7, 8] or multimodal [22] representations of the world and, subsequently, learn a task-policy over that fixed representation. However, none have considered the fundamental importance that reward signals have on the representation of the world itself. We aim at exploring methodologies to enhance unsupervised multimodal representation learning with reward signals obtained from the environment, such as self-attention mechanisms [17, 20].

The second goal of our future work concerns the extension of the multimodal transfer reinforcement learning problem to robotic agents. We are interested in addressing indoor navigation tasks in which a mobile robot is equipped with both camera and laser sensors. Our goal is to access both the limitations of multimodal representation learning from real-life sensory data and the potential for cross-modality policy transfer in robotic agents (e.g., executing a policy trained on visual perceptions when only laser readings are available). We hope that, by achieving the proposed goals, the contributions of this thesis might prompt further research on multimodal representation learning for robots and other artificial agents.

ACKNOWLEDGMENTS

This work was partially supported by national funds through the Portuguese Fundação para a Ciência e a Tecnologia under project UIDB/50021/2020 (INESC-ID multi annual funding). The author acknowledges the PhD grant SFRH/BD/139362/2018.
REFERENCES


