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:
Miguel Vasco. 2020. Multimodal Representation Learning for Robotic Cross-Modality Policy Transfer. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

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].

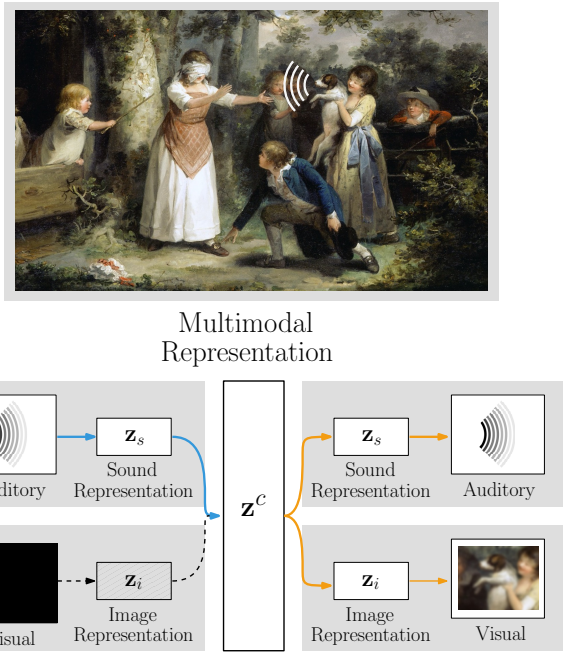

Figure 1: An example of the importance of multimodal representation learning for the execution of modality-independent tasks: in the absence of light, humans can navigate their environment by employing perceptual information from other modalities (such as sound) to generate the absent visual perceptual experience. Image adapted from George Morland's painting "Blind Man's Buff".

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

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

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.

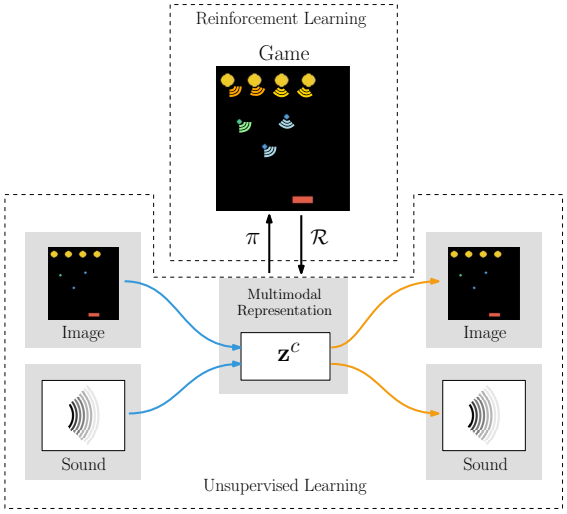

Figure 2: Proposed architecture to address modality transfer in Atari games: using unsupervised learning, we build a multimodal representation $\mathbf{z}^c$ from images and sounds collected from the game. Subsequently, we train a policy π of a reinforcement learning agent using rewards $\mathcal{R}$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

[1] Richard L Canfield and Marshall M Haith. 1991. Young infants' visual expectations for symmetric and asymmetric stimulus sequences. *Developmental Psychology* 27, 2 (1991), 198.

[2] Susan Carey. 2009. *The origin of concepts*. Oxford University Press.

[3] Antonio R Damasio. 1989. Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition* 33, 1-2 (1989), 25–62.

[4] Kenji Doya. 1999. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural networks* 12, 7-8 (1999), 961–974.

[5] Chelsea Finn and Sergey Levine. 2017. Deep visual foresight for planning robot motion. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2786–2793.

[6] Jerry A Fodor. 1981. *Representations: Philosophical essays on the foundations of cognitive science*. MIT Press Cambridge.

[7] David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122* (2018).

[8] Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. 2017. Darla: Improving zero-shot transfer in reinforcement learning. In *International Conference on Machine Learning*. 1480–1490.

[9] Jonathan Kelly and Gaurav S Sukhatme. 2011. Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration. *The International Journal of Robotics Research* 30, 1 (2011), 56–79.

[10] Timo Korthals, Daniel Rudolph, Jürgen Leitner, Marc Hesse, and Ulrich Rückert. 2019. Multi-Modal Generative Models for Learning Epistemic Active Sensing. In *IEEE International Conference on Robotics and Automation*.

[11] Patricia K Kuhl, Karen A Williams, Francisco Lacerda, Kenneth N Stevens, and Björn Lindblom. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255, 5044 (1992), 606–608.

[12] William Land, Dima Volchenkov, Bettina E Bläsing, and Thomas Schack. 2013. From action representation to action execution: exploring the links between cognitive and biomechanical levels of motor control. *Frontiers in Computational Neuroscience* 7 (2013), 127.

[13] Alexander LaTourrette and Sandra R Waxman. 2019. A little labeling goes a long way: Semi-supervised learning in infancy. *Developmental science* 22, 1 (2019), e12736.

[14] Chi-Tat Law and Joshua I Gold. 2009. Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nature neuroscience* 12, 5 (2009), 655.

[15] Ren C Luo and Chun Chi Lai. 2013. Multisensor fusion-based concurrent environment mapping and moving object detection for intelligent service robotics. *IEEE transactions on industrial electronics* 61, 8 (2013), 4043–4051.

[16] Hiroshi Makino, Eun Jung Hwang, Nathan G Hedrick, and Takaki Komiyama. 2016. Circuit mechanisms of sensorimotor learning. *Neuron* 92, 4 (2016), 705–721.

[17] Anthony Manchin, Ehsan Abbasnejad, and Anton van den Hengel. 2019. Reinforcement learning with attention that works: A self-supervised approach. In *International Conference on Neural Information Processing*. Springer, 223–230.

[18] Daphne Maurer, Thanujeni Pathman, and Catherine J Mondloch. 2006. The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental science* 9, 3 (2006), 316–322.

[19] Kaspar Meyer and Antonio Damasio. 2009. Convergence and divergence in a neural architecture for recognition and memory. *Trends in neurosciences* 32, 7 (2009), 376–382.

[20] Alexander Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, and Danilo Jimenez Rezende. 2019. Towards interpretable reinforcement learning using attention augmented agents. In *Advances in Neural Information Processing Systems*. 12329–12338.

[21] Lerrel Pinto, Dhiraj Gandhi, Yuanfeng Han, Yong-Lae Park, and Abhinav Gupta. 2016. The curious robot: Learning visual representations via physical interactions. In *European Conference on Computer Vision*. Springer, 3–18.

[22] Rui Silva, Miguel Vasco, Francisco S. Melo, Ana Paiva, and Manuela Veloso. 2019. Playing Games in the Dark: An approach for cross-modality transfer in reinforcement learning. *arXiv preprint arXiv:1911.12851* (2019).

[23] Charles Spence. 2011. Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics* 73, 4 (2011), 971–995.

[24] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. 2016. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891* (2016).

[25] Miguel Vasco, Francisco Melo, David Matos, Ana Paiva, and Tetsunari Inamura. 2019. Online Motion Concept Learning: A Novel Algorithm for Sample-Efficient Learning and Recognition of Human Actions. In *International Conference on Autonomous Agents and MultiAgent Systems*. Montreal, Canada, 2244–2446.

[26] Miguel Vasco, Francisco S. Melo, David Martins de Matos, Ana Paiva, and Tetsunari Inamura. 2019. Learning multimodal representations for sample-efficient recognition of human actions. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[27] Peter Walker, J Gavin Bremner, Uschi Mason, Jo Spring, Karen Mattock, Alan Slater, and Scott P Johnson. 2010. Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science* 21, 1 (2010), 21–25.

[28] Janet F Werker and Richard C Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development* 7, 1 (1984), 49–63.

[29] Hang Yin, Francisco S Melo, Aude Billard, and Ana Paiva. 2017. Associate latent encodings in learning from demonstrations. In *AAAI Conference on Artificial Intelligence*.

[30] Barbara A Younger. 1985. The segregation of items into categories by ten-month-old infants. *Child Development* (1985), 1574–1583.

[31] Barbara A Younger and Leslie B Cohen. 1986. Developmental change in infants' perception of correlations among attributes. *Child development* (1986), 803–815.