frontiers | Frontiers in Computer Science

EDITORIAL
published: 01 June 2022
doi: 10.3389/fcomp.2022.930067

Editorial: Language and Vision in Robotics: Emerging Neural and On-Device Approaches

Giovanni Luca Masala[1]*, *Massimo Esposito*[2], *Umberto Maniscalco*[3] and *Andrea Calimera*[4]

[1] School of Computing, University of Kent, Canterbury, United Kingdom, [2] Institute for High Performance Computing and Networking of the National Research Council of Italy, Naples, Italy, [3] Institute for High Performance Computing and Networking of the National Research Council of Italy, Palermo, Italy, [4] Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy

Keywords: brain-inspired architecture, NLP, robotics, computer vision, language development, deep learning

Editorial on the Research Topic

Language and Vision in Robotics: Emerging Neural and On-Device Approaches

INTRODUCTION

Endowing machines with the ability to represent and understand the physical world in which they live is a longstanding challenge in the AI research community. The last years have seen significant advancements in the fields of Natural Language Processing (NLP) and Computer Vision (CV) and the development of robotic hardware and accompanying algorithms.

Deep learning is feeding NLP advances thanks to neural networks models with outstanding achievements in several tasks, such as language modeling (Devlin et al., 2019), sentence classification (Pota et al., 2020), named entity recognition (Catelli et al., 2021), sentiment analysis (Pota et al., 2021), and question answering (Zhang et al., 2019). Deep learning has led to impressive achievements on various Computer Vision tasks and becoming state of the art in computer vision (He et al., 2015; Smith et al., 2021).

Even though these fields are among the most actively developing AI research areas, until recently, they have been treated separately without many ways to benefit from each other. On the contrary, it is fundamental to integrate verbal and no verbal communication to consider the multimodal nature of communication (Maniscalco et al., 2022). With the expansion of deep learning approaches, researchers have started exploring the possibilities of jointly applying both NLP and CV approaches to improve robotic capabilities.

A prevalent artifice to self-organize language-meaning representations in robotic architectures is the use of supervised learning methods amid which some have taken significant steps toward humanlike intelligence demonstrated in learning experiments with real robots (Ogata et al., 2007; Tani, 2016; Yamada et al., 2016).

However, these proposals mainly implicate low-level motor skills and in a way that neglects perspectives from cognitive linguistics and psychology. By contrast, this has inspired the efforts of several authors that model usage-based language acquisition and production, unidirectionally (Golosio et al., 2015; Moulin-Frier et al., 2017; Hinaut and Twiefel, 2019) or bidirectionally (Heinrich et al., 2020), i.e., learn motor meaning from language and emerge language skills from motor exploration. It is important to highlight also the attempt to model multiple language learning (Giorgi et al., 2020).

This Research Topic aims to provide an overview of the research being carried out in both the areas of NLP and CV to allow robots to learn and improve their capabilities for exploring, modeling, and learning about the physical world. As this integration requires an interdisciplinary attitude, the Research Topic aims to gather researchers with broad expertise in various fields—machine learning,

OPEN ACCESS

Edited and reviewed by:
Marcello Pelillo,
Ca' Foscari University of Venice, Italy

***Correspondence:**
Giovanni Luca Masala
g.masala@kent.ac.uk

Specialty section:
This article was submitted to Computer Vision, a section of the journal Frontiers in Computer Science

Received: *27 April 2022*
Accepted: *16 May 2022*
Published: *01 June 2022*

Citation:
Masala GL, Esposito M, Maniscalco U and Calimera A (2022) Editorial: Language and Vision in Robotics: Emerging Neural and On-Device Approaches. Front. Comput. Sci. 4:930067. doi: 10.3389/fcomp.2022.930067

computer vision, natural language, neuroscience, and psychology—to discuss their cutting edge work as well as perspectives on future directions in this exciting space of language, vision and interactions in robots.

PAPERS INCLUDED IN THIS RESEARCH TOPIC

Vargas et al. investigated how salient terms related to verbal communication in robotics have evolved over the years, what are the Research Topics that recur in the related literature, and what are their trends. The study is based on a computational linguistic analysis conducted on a database extracted from the Scopus database using specific keywords. The authors show how relevant terms of verbal communication evolved, which are the main coherent topics and how they have changed over the years. They highlighted positive and negative trends for the most coherent topics and the distribution over the years for the most significant ones. In particular, verbal communication resulted in being highly relevant for social robotics.

Giorgi et al. proposed a novel interactive learning method, using a brain-inspired architecture, to model an appropriate mapping of language with the percept and internal motor representation in humanoid robots. The method grants flexible run-time acquisition of novel linguistic forms and real-world information, without training the cognitive model anew. Furthermore, their approach supports incremental open-ended learning. This research presents the first robotic instantiation of a complex architecture based on Baddeley's Working Memory (WM) model. The authors demonstrated the ability of the robot to understand instructions involving higher-order (abstract) linguistic concepts of developmental complexity, which cannot be directly hooked in the physical world and are not pre-defined in the robot's static self-representation.

Nichols et al. presented a collaborative story generation system that works with a human storyteller to create a story by generating new utterances based on the story so far. Authors introduced the task of *collaborative story generation*, where an artificial intelligence agent or a robot, and a person collaborate to create a unique story by taking turns adding to it. They evaluated the system by having human participants play the collaborative story generation game and comparing the stories they create with our system to a naive baseline. Their evaluation shows that participants have a positive view of collaborative story generation with a social robot and consider rich, emotive capabilities to be key to an enjoyable experience.

To reliably interact with humans in the physical world, robots must learn to execute commands that are extended in time while being responsive to changes in their environments. This requires the robot to jointly represent the symbolic knowledge in language and the perceptual information from the environment as well as generalize to different commands and maps. A popular representation to encode complex commands is linear temporal logic (LTL). In this context, Kuo et al. demonstrated how a reinforcement learning (RL) agent can use compositional recurrent neural networks to learn to carry out commands specified in LTL. The compositional structures presented by the authors are not specific to LTL, thus opening the path to RL agents that perform zero-shot generalization in other compositional domains. Furthermore, the authors developed a novel form of multi-task learning for RL agents that allows them to learn from a diverse set of tasks and generalize to a new set of diverse tasks without any additional training.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

Catelli, R., Casola, V., De Pietro, G., Fujita, H., and Esposito, M. (2021). Combining contextualized word representation and sub-document level analysis through Bi-LSTM+ CRF architecture for clinical de-identification. *Knowl. Bas. Syst.* 213:106649. doi: 10.1016/j.knosys.2020.106649

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. *Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguist.* 1, 4171–4186. doi: 10.48550/arXiv.1810.0480

Giorgi, I., Golosio, B., Esposito, M., Cangelosi, A., and Masala, G. L. (2020). "Modelling multiple language learning in a developmental cognitive architecture," in *IEEE Transactions On Cognitive And Developmental Systems*. Piscataway, NJ: IEEE. doi: 10.1109/TCDS.2020.3033963

Golosio, B., Cangelosi, A., Gamotina, O., and Masala, G. L. (2015). A cognitive neural architecture able to learn and communicate through natural language. *PLoS ONE* 10:e0140866. doi: 10.1371/journal.pone.0140866

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*. Piscataway, NJ: IEEE, 1026–1034. doi: 10.1109/ICCV.2015.123

Heinrich, S., Yao, Y., Hinz, T., Liu, Z., Hummel, T., Kerzel, M., et al. (2020). Crossmodal language grounding in an embodied neurocognitive model. *Front. Neurorobot.* 14:52. doi: 10.3389/fnbot.2020.00052

Hinaut, X., and Twiefel, J. (2019). Teach Your Robot Your Language! trainable neural parser for modeling human sentence processing: examples for 15 languages. *IEEE Trans. Cogn. Dev. Syst.* 12, 179–188. doi: 10.1109/TCDS.2019.2957006

Maniscalco, U., Storniolo, P., and Messina, A. (2022). Bidirectional multi-modal signs of checking human-robot engagement and interaction. *Int. J. Soc. Robot.* 21:855. doi: 10.1007/s12369-021-00855-w

Moulin-Frier, C., Fischer, T., Petit, M., Pointeau, G., Puigbo, J. Y., Pattacini, U., et al. (2017). DAC-h3: a proactive robot cognitive architecture to acquire and express knowledge about the world and the self. *IEEE Trans. Cogn. Dev. Syst.* 10, 1005–1022. doi: 10.1109/TCDS.2017.2754143

Ogata, T., Murase, M., Tani, J., Komatani, K., and Okuno, H. G. (2007). "Two-way translation of compound sentences and arm motions by recurrent neural networks," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Piscataway, NJ: IEEE), 1858–1863. doi: 10.1109/IROS.2007.4399265

Pota, M., Esposito, M., De Pietro, G., and Fujita, H. (2020). Best practices of convolutional neural networks for question classification. *Appl. Sci.* 10:4710. doi: 10.3390/app10144710

Pota, M., Ventura, M., Fujita, H., and Esposito, M. (2021). Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets. *Expert Syst. Appl.* 181:115119. doi: 10.1016/j.eswa.2021.115119

Smith, M. L., Smith, L. N., and Hansen, M. F. (2021). The quiet revolution in machine vision - a state-of-the-art survey paper,

including historical review, perspectives, and future directions. *Comput. Indus.* 130:103472. doi: 10.1016/j.compind.2021.103472

Tani, J. (2016). *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780190281069.001.0001

Yamada, T., Murata, S., Arie, H., and Ogata, T. (2016). Dynamical integration of language and behaviour in a recurrent neural network for human–robot interaction. *Front. Neurorobot.* 10:5. doi: 10.3389/fnbot.2016.00005

Zhang, Z., Wu, Y., Zhou, J., Duan, S., and Zhao, H. (2019). "SG-Net: syntax-guided machine reading comprehension," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI, 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020* (New York, NY), 9636–9643. doi: 10.1609/aaai.v34i05.6511

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.