Potentials and Limitations of Deep Neural Networks for Cognitive Robots

Doreen Jirak and Stefan Wermter
Knowledge Technology, Department of Informatics, University of Hamburg, Germany

***Abstract*—Although Deep Neural Networks reached remarkable performance on several benchmarks and even gained scientific publicity, they are not able to address the concept of cognition as a whole. In this paper, we argue that those architectures are potentially interesting for cognitive robots regarding their perceptual representation power for audio and vision data. We identify crucial settings for cognitive robotics where deep neural networks have as yet only contributed little compared to the challenges in this area. Finally, we highlight the rather unexplored area of *Reservoir Computing* for sequence learning. This new paradigm of learning recurrent neural networks in a fast and robust way qualifies to be an integral part of cognitive robots and may inspire novel developments.**

I. Introduction

With the advent of both data availability and accelerated GPU computations, deep neural network architectures (DNN) have flourished over the past years and showed tremendous success on several benchmarks for computer vision, and audio and speech. However, DNN models demand high training times for sensitive parameter tuning and network-specific optimization like regularization techniques, which is a rather engineering approach. Although DNNs are motivated by neurobiological principles observed in the visual cortex, the complete picture of the computations boils down to learning filters detecting specific frequency characteristics like edges in images or voice pitch in audio data. As the learning is carried out in a supervised manner, the use of a huge amount of labeled data has a crucial impact on high accuracy and other evaluation measures.

The area of cognitive robotics is highly interdisciplinary and advances in sensor and robot technology allow the integration of human-like capabilities. The aim of cognitive skills in robots is motivated by their increased integration in human's everyday life, whether as a companion in human-robot interaction (HRI) or as an autonomous system in health care and industry.

DNNs can only address part of what actually cognitive systems need and which challenges they face. The focus on supervised benchmarking shifts research on the multiple, interconnected facets of cognitive robots to a rather engineered approach accompanied by constraints on the data and the environment. Critical investigations on DNNs also reveal that due to their high dependence on apriori knowledge, they are easily fooled [1] and fail to cope with, e.g., real-world speech data from children. Recent advances of (deep) unsupervised networks closely related to neural information principles like sparse autoencoders or self organizing maps may stimulate again joining state-of-the-art algorithms successfully applied to robot scenarios with deep architectures.

In the present paper, we will discuss the pros and cons of DNNs in cognitive robotics. We highlight their usefulness but also limitations, discussing also the potential of *Reservoir Computing* in the context of cognitive robots.

II. Lab Conditions vs. The Real World

A challenging aspect of cognitive systems and specifically cognitive robots is the high parallelization of tasks with a vast amount of incoming stimuli which need to be filtered to trigger a suitable (re)action. We are able to focus our *attention* onto a specific object or sound but we can also be easily distracted as is observable in the ventriloquist effect, where humans' sound perception is influenced by a doll being an additional visual stimulus. Researchers thus developed algorithms to capture the most salient cues in the scene but either on datasets [2] or under very constrained lab conditions [3]. Although implementations using deep networks showed promising results in those limited cases, DNN will have problems with changing conditions in real-world environments due to retraining issues inherent in closed set classification system.

A recent robotic experiment on indoor exploration proposed a framework based on DNN to exploit the robot capabilities to learn the unknown terrain in a human-like fashion using the turtlebot platform [4]. For the robot vision, a Kinect device was mounted delivering depth information. Although one may argue that Kinect sensors are error-prone and question their reliability in a robust image capturing, the availability of the authors' code enable other researchers to explore further the advantages and pitfalls of the approach, identifying further problems and possible solutions enriching discussions in the cognitive robotic community for this important task.

Another important factor is that learning in humans is not subject to one specific task for a specific point in time and space but is continuous throughout life. Although DNNs produce stable feature representations which may code for more abstract concepts similar to neural codes along the brain hierarchy, yet they are less flexible and do not allow online adaptation nor any memory mechanisms. A new trend, however, to avoid retraining is using transfer learning: weights of a deep network successfully trained on a set of e.g. images are *transfered* to solve a similar task. A fruitful combination for robot applications could be to join those strategies with open set classification algorithms like the growing-when-required networks (GWR), which allow insertion and deletion of labels.

The benefit of a GWR is that it provides a computational method to the question how actually learning is guided in the absence of any labels. The robust representations emerging from a deep neural network can thus complement the flexible classification in environments beyond the restricted lab conditions.

III. Children vs. Adult Learning

Developmental psychology is an important research area to gain an understanding on how we acquire cognitive skills starting from first imitation tasks to controlled motor acts, from first babbling to words and sentences to finally reasoning, consciousness and empathy (theory of mind). In contrast to deep networks being trained on billions of e.g. images for object recognition, infants learn rather slow and are provided only with a very limited excerpt of the world. DNNs tuned for benchmarking as was carried out during the past years is very different from developmental learning and the acquisition of cognitive skills. Although their performance on speech recognition tasks is indubitable, DNN models show low performance when the speech resources are getting noisy (e.g. repetitions) or are from children, since their language differs from adults.

In addition, the way children learn is driven by high intrinsic motivation [5], curiosity [6] and their physical interaction with the world (embodiment). The latter was shown to be essential in exploring affordances for the associative learning between objects and grasp types [7], keeping also into account parental scaffolding [8]. The link between motor activation in the brain and language acquisition speaks in favor of embodied learning which demands to ground experiments in real robots (see e.g. [9]) rather than a set of labeled data.

Until now, these important aspects to acquire cognitive skills are absent in learning systems based on DNNs which may be explained by their difficulty in interpretation. A literature research on this topic revealed that only a few trends emerged towards closing this gap. One approach used a generative deep neural network based on the time course of learning object categories in children [10]. Essentially, the authors demonstrated that despite using linear neurons, their model was able to capture the refinement of semantic labels and with an additional analysis on the temporal dynamics of gradient-based learning, that this differentiation emerges naturally in hierarchical networks.

Another recent paper [11] identified a gap between developmental learning and deep neural networks along the hierarchical processing of object recognition on the perceptual level for the acquisition of action knowledge on the cognitive level where affordances play a crucial role. In the light of embodied cognition, affordances are grounded in sensorimotor experiences. The authors [11] argued in favor of unsupervised learning techniques as labels of sensorimotor experiences do not exist. Though deep networks do not allow for flexible, online adaptation of e.g. grasping when affordances may change (think about the change of perspective of a mug handle), the main argument integrating deep unsupervised models is that they can serve as an enhancement strategy for learned sensorimotor experience as in humans during sleep. In conclusion, although still other developmental learning mechanisms as the mentioned curiosity are missing and can not be captured by DNNs, they might be a useful tool for representational learning, encouraging a novel research area called "deep developmental learning" [11].

Regarding the interpretability issues in DNN, a deep network in a one-shot word learning scenario was employed and the results investigated by means of methods from cognitive psychology, namely the "cognitive bias" which allows children to eliminate word hypotheses from a vast possible word space [12]. The network evaluation, which used training on the popular ImageNet database, underpinned the "shape bias" hypothesis, i.e. that humans tend to label objects similar in shape as one entity. Joining DNNs with cognitive psychology may open the way to further investigate the positive as well as negative aspects of biases in human learning, complementing behavioral studies which found the basis for many cognitive robotic scenarios.

IV. Reservoir Computing for Sequential Learning in Robot Scenarios

Learning sequential tasks is inevitably necessary for the development of cognitive systems, may it be navigation, planning or communication with language and gestures. Deep neural architectures do not possess an explicit time resolution to capture time correlation nor timescales inherent in sentences and actions. A popular technique to overcome the missing temporal link in, e.g., CNNs is to use 3D convolution kernels for image stacks representing videos. However, the benefit of such kernels on the performance compared to the standard 2D kernel is not yet satisfiable answered [13].

Due to the inherent correlations in sequential tasks, Recurrent Neural Networks (RNN) are a widely used tool as they provide local network memory and are thus able to compute different time aspects (e.g. continuous-time RNN). RNNs in their diverse implementations were successfully used in cognitive robotic scenarios (see e.g. [14], [15]). Due to the error computations propagated through the whole network using gradient descent, RNN training suffers from vanishing or exploding gradients, trapping into local minima, and network bifurcations. To overcome these issues, RNNs need a lot of retraining and are thus computationally demanding.

Alternatively, the *Reservoir Computing* (RC) paradigm introduced a simplified training by the conceptual separation of a high-dimensional reservoir of randomly connected neurons providing rich input representations which are read out by simple linear models and thus updating the network weights becomes obsolete. Popular implementations are the Liquid State Machines [16] (LSM) and Echo State Networks [17] (ESN) the latter demonstrating competitive performance in predictions tasks for (chaotic) timeseries [18]. Despite the reduced computational effort, only a few applications for robots were developed until today. Some highlights in the literature using the RC principles include navigation [19], [20]

and language acquisition [21]. Especially the latter research shows that RC is a promising framework combining principles from neuroscience and developmental psychology for the highly complex cognitive task of learning and understanding language. Also, only little is known about the potential of RC algorithms for visual tasks. As images and videos are highly complex in structure and introduce varying lighting conditions and perspectives, a solution combining the robust feature capabilities provided by DNNs with ESNs capturing the inherent temporal structure showed good performance on a gesture recognition task for a set of command gestures [22]. The coupling of DNNs with RC was also highlighted recently regarding future research directions [23]. Methods based on RC can also be exploited for robot control [24] and locomotion [25] but coupling RC with more complex robot scenarios is still in its infancy.

We believe, that the area of cognitive robotics would substantially benefit from further research on the potentials of RC frameworks complementing other computational methods like DNNs for robot perception or even substituting existing RNN methods exploiting the simplified yet robust training. Also, analysis of ESNs regarding their "edge of stability" would unify hypotheses on cognition underlying chaotic computations with performance maximization [26], which can lead to effective network design.

V. Conclusion

The argument in favor of cognitive systems is diametral to the current benchmarking and competition of accuracies using deep neural networks. Although those network architectures are beneficial for perceptual or supervised learning involved in the acquisition of cognitive skills, we highlighted some limitations and suggest alternative computations for future applications. This brief communication focused on recent neural network approaches and *Reservoir Computing* to stimulate discussions on the integration of those methods into cognitive robots. Our intention was not to exclude other important methods like probabilistic models or dynamic neural fields. We rather argue, that advances in technology in humans everyday life demand tighter coupling in the community regarding current progress in the neural network and machine learning area with cognitive models proven to be successful in the development of cognitive robots.

References

[1] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[2] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, pp. 2048–2057.

[3] P. Barros, G. I. Parisi, and S. Wermter, "Emotion-modulated attention improves expression recognition: A deep learning model," *Neurocomputing*, vol. 253, pp. 104–114, Mar 2017.

[4] L. Tai, S. Li, and M. Liu, "Autonomous exploration of mobile robots through deep neural networks," *International Journal of Advanced Robotic Systems*, vol. 14, no. 4, 2017.

[5] P.-Y. Oudeyer and F. Kaplan, "What is intrinsic motivation? a typology of computational approaches," *Frontiers in Neurorobotics*, vol. 1, p. 6, 2009.

[6] J. Gottlieb, P.-Y. Y. Oudeyer, M. Lopes, and A. Baranes, "Information-seeking, curiosity, and attention: computational and neural mechanisms." *Trends in cognitive sciences*, vol. 17, no. 11, pp. 585–593, Nov. 2013.

[7] E. Oztop, N. S. Bradley, and M. A. Arbib, "Infant grasp learning: a computational model," *Experimental Brain Research*, vol. 158, no. 4, pp. 480–503, Oct 2004.

[8] E. Ugur, Y. Nagai, H. Celikkanat, and E. Oztop, "Parental scaffolding as a bootstrapping mechanism for learning grasp affordances and imitation skills," *Robotica*, vol. 33, no. 5, p. 11631180, 2015.

[9] D. Marocco, A. Cangelosi, K. Fischer, and T. Belpaeme, "Grounding action words in the sensorimotor interaction with the world: experiments with a simulated icub humanoid robot," *Frontiers in Neurorobotics*, vol. 4, 2010.

[10] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Learning hierarchical category structure in deep neural networks," in *Proceedings of the 35th annual meeting of the Cognitive Science Society*, 2013, pp. 1271–1276.

[11] O. Sigaud and A. Droniou, "Towards deep developmental learning," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 2, pp. 99–114, June 2016.

[12] S. Ritter, D. G. T. Barrett, A. Santoro, and M. M. Botvinick, "Cognitive psychology for deep neural networks: A shape bias case study," 2017.

[13] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1725–1732.

[14] M. Oubbati, B. Kord, P. Koprinkova-Hristova, and G. Palm, "Learning of embodied interaction dynamics with recurrent neural networks: some exploratory experiments," *Journal of Neural Engineering*, vol. 11, no. 2, 2014.

[15] S. Murata, Y. Yamashita, H. Arie, T. Ogata, S. Sugano, and J. Tani, "Learning to perceive the world as probabilistic or deterministic via interaction with others: A neuro-robotics experiment," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 4, pp. 830–848, April 2017.

[16] W. Maass, T. Natschläger, and H. Markram, "Real-time computing without stable states: A new framework for neural computation based on perturbations," *Neural Computation*, vol. 14, no. 11, pp. 2531–2560, 11 2002.

[17] H. Jaeger, "The echo state approach to analysing and training recurrent neural networks - with an erratum note," German National Research Center for Information Technology, Tech. Rep., 2001.

[18] M. Lukoševičius, "A practical guide to applying echo state networks," in *Neural Networks: Tricks of the Trade: Second Edition*, G. Montavon, G. B. Orr, and K.-R. Müller, Eds. Springer Berlin Heidelberg, 2012, pp. 659–686.

[19] S. Dasgupta, F. Wörgötter, and P. Manoonpong, "Information dynamics based self-adaptive reservoir for delay temporal memory tasks," *Evolving Systems*, vol. 4, no. 4, pp. 235–249, 2013.

[20] E. A. Antonelo and B. Schrauwen, "On learning navigation behaviors for small mobile robots with reservoir computing architectures," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 4, pp. 763–780, April 2015.

[21] X. Hinaut, M. Petit, G. Pointeau, and P. F. Dominey, "Exploring the acquisition and production of grammatical constructions through human-robot interaction with echo state networks," *Frontiers in Neurorobotics*, vol. 8, p. 16, 2014.

[22] D. Jirak, P. Barros, and S. Wermter, "Dynamic gesture recognition using echo state networks," in *Proceedings of the European Symposium of Artificial Neural Networks and Machine Learning*, 2015, pp. 475–480.

[23] A. Goudarzi and C. Teuscher, "Reservoir computing: Quo vadis?" in *Proceedings of the 3rd ACM International Conference on Nanoscale Computing and Communication*, 2016, pp. 13:1–13:6.

[24] A. Polydoros, L. Nalpantidis, and V. Krger, "Advantages and limitations of reservoir computing on model learning for robot control," 2015.

[25] F. Wyffels and B. Schrauwen, "Design of a central pattern generator using reservoir computing for learning human motion," in *2009 Advanced Technologies for Enhanced Quality of Life*, July 2009, pp. 118–122.

[26] F. M. Bianchi, L. Livi, and C. Alippi, "Investigating echo-state networks dynamics by means of recurrence analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–13, 2016.