Perspectives on Deep Multimodel Robot Learning

Wolfram Burgard, Abhinav Valada, Noha Radwan, Tayyab Naseer, Jingwei Zhang, Johan Vertens, Oier Mees, Andreas Eitel and Gabriel Oliveira

Abstract In the last decade, deep learning has revolutionized various components of the conventional robot autonomy stack including aspects of perception, navigation and manipulation. There have been numerous advances in perfecting individual tasks such as scene understanding, visual localization, end-to-end navigation and grasping, which has given us a critical understanding on how to create individual architectures for a specific task. This now brings us to the question, as to whether this disjoint learning of models for robotic tasks, effective in the real-world and whether it is scalable? And more generally, is training task specific models on task specific datasets beneficial to architecting robot intelligence as a whole? In this paper, we argue that multimodel learning or joint multi-task learning is an effective strategy for enabling robots to excel across multiple domains. We describe how multimodel learning can facilitate generalization to unseen scenarios by utilizing domain-specific cues from auxiliary tasks and discuss some of the current mechanisms that can be employed to design multimodel frameworks for robot autonomy.

1 Introduction

Robots today have evolved from being able to perform only structured actions to being able to act re-actively based on sensing their environment. Robot learning has played a crucial role in enabling this capability. The classical paradigm involves a pipeline containing modules for perception, world modelling, planning and control, each of which are carefully engineered, incorporating handcrafted features and task-specific structures. A typical modern robot control system is an ensemble of modules, which often contain learning based models, and that are designed to perform dedicated tasks aimed at accomplishing a specific goal. In the last decade, Deep Convolutional Neural Network (DCNN) architectures have achieved remarkable re-

All authors are with the Department of Computer Science, University of Freiburg, Germany · Corresponding author's e-mail: burgard@informatik.uni-freiburg.de

sults across several robotic problems. However, the focus has been on designing individual networks for specific problems including perception, localization, navigation and manipulation. In addition, several disjoint models have been used in conjunction. This limits the overall learning ability of the robot as most models are trained in a supervised fashion and independently, therefore they have no ability to share cross-domain information using training signals from auxiliary tasks. Our vision is a unified multimodel deep learning framework that jointly learns multiple robot tasks across multiple domains including perception, planning and control. We propose a multimodel framework that incorporates soft parameter sharing thereby enabling the network to decide what layers from auxiliary tasks to share and which sub-models can benefit from representations learned by layers in other sub-models. We believe that this will enable robots to learn tasks with limited amount of data by leveraging transfer learning across sub-models and equipping it with the capability to continuously learn from what it experiences and perceives in the real-world.

In the following sections, we first describe the classical pipeline that is commonly employed to enable robots to perform autonomous actions. We then give an overview of deep learning approaches that have demonstrated substantial progress in relevant basic modules for perception, localization and navigation. Finally we discuss our perspectives on how to enable robots to more proficiency learn from the world around them using multimodel frameworks.

2 Classical Paradigms

The classical definition of an agent is anything that can perceive its surroundings and act upon it and in the context of robotics, autonomous agents inhabit a complex dynamic environment. In order to achieve their predefined goals, they first need to sense their surroundings before they can plan actions. Robots are often equipped with multiple sensors that provide complementary information. Extracting information from raw sensor data is in itself a challenging task which requires expert knowledge of both the environment and characteristics of the data produced by the sensor. Over time, several approaches have been developed for feature extraction, some intricately handcrafted and recently even learned from sensor data. The extracted features are then used to infer information about the environment.

Complementary to the perception module, the robot also needs accumulated knowledge about the world in which it is placed. Accordingly, a world module needs to be carefully designed such that static persistent landmarks in the environment, for example walls, poles and trees, are well represented. Furthermore, it defines the possible set of actions for the agent along with the state transition function which defines the state of the world after each action. Instead of providing the world model to the agent, we let the system built its own model using information gathered from the perception module. However, relying solely on the perception module to build the model is a challenging task as the entire environment is not visible in one sensor observation. Hence to build a complete model, the agent needs to explore the en-

vironment over time and correct any inconsistencies occurring during observation. For the agent to achieve its goal, it relies on the output of the perception module along with the world model to formulate a plan. The planning module is responsible to provide a plan that can be executed in the current state space and successfully complete the required task. To this end, the planning module needs to not only formulate the plan but also have the ability to recover from failure and replan in the event of an unexpected situation. The control module is responsible for providing the proper control commands to the actuators of the robot in a way that follows the plan provided to perform a predefined action.

3 Emergence of Deep Models

In the last few years, convolutional neural networks has revolutionized several core components that constitute an autonomous robotic system. They have brought about a significant change in the traditional pipelines employed. We briefly discuss some of these advances in the following sections.

Scene Understanding Scene understanding is an essential component of any robotic system as robots need to first know what and where the elements of the scene are before they can act on them. The advent of DCNNs have brought about several state-of-the-art models for a variety of perception tasks including object recognition [3, 4], detection [4, 10, 14] and semantic segmentation [11, 21, 26]. However, robotic perception models have different requisites than those in computer vision. Robots are often equipped with multiple sensors such as cameras, lidars and radars to perceive their surroundings. Therefore, deep architectures need to efficiently learn a combined representation of the world utilizing these sensors. To this end, multi-stream networks are often used to train each stream on specific modality and fuse them towards the end of the network [3, 24]. Alternatively, architectures have also been designed that fuse feature maps from modality specific streams at intermediate points in the network and converge to a single stream towards the end [8]. As sensor noise is a major hindrance in the real-world, noise augmentation strategies can be employed, either before feeding data to the network [3] or while training [25]. One of the major challenges in real-world robot perception is the ability of models to adapt to changes in appearance due to weather and seasons. In such conditions, incorporating adaptive fusion strategies such as mixture of deep experts has substantially improved the performance and robustness of models for semantic segmentation [26] and pedestrian detection [14]. DCNNs have also been used for specialized classification tasks such of terrains and with unconventional sensors including microphones [25]. Advances such as new pooling strategies that learn statistics of temporal features in the signal enable these approaches to outperform classifiers learned on traditional audio features.

The introduction of fully convolutional neural networks [11] has brought about several state-of-the-art architectures for various robotic applications from segment-

ing roads [19] to human body parts [21]. Incorporating advances such as residual learning and dilated convolutions to learn deep multi-scale features have further pushed the boundaries of achievable performance while maintaining fast inference times [26]. Often in robotics it is also necessary to estimate motion of objects in order to plan future actions. Complementary tasks such as segmentation and motion estimation can be learned using a joint formulation in a unified deep framework [27]. Such networks not only reduce the model complexity but also enable interactive frame rates.

Localization and Odometry Robust place recognition and visual localization of autonomous systems is of paramount importance for relevant robotic applications. Visual localization largely depends on robust and repeatable feature descriptions over large variety of environmental changes. The feature descriptions from deep networks have outperformed the traditional hand-crafted features in this domain due to their ability to learn feature correspondences under different appearances. A model designed for visual localization can also leverage vital information from a model trained on a different task e. g. segmentation [18] or visual similarity. This aspect of joint learning of different tasks enables us to learn a heterogeneous model where subtasks benefit from each other's data. Recently, deep architectures for metric localization have emerged that provide an efficient map representation in addition to demonstrating considerable robustness in challenging perceptual conditions [9, 28]. In contrast to traditional methods, deep models provide a fixed map size and a constant time complexity for camera-based metric localization.

Recently, end-to-end DCNN approaches that estimate visual odometry have also been proposed [29, 15]. Most of these approaches employ a Siamese-type network architecture that take two consecutive images as input and regress the relative transformation between them. In [29], the authors use a AlexNet-based Siamese architecture and an L2-loss layer with equal weights for the translational and rotational components, while in [15] the authors propose a weighting term to balance these components. In order to exploit the advantages of both metric and topological localization, while concurrently reducing the error caused by the accumulation of drift in visual odometry, an optimization technique was proposed that fuses the output of a odometry and topological DCNN [20]. Utilizing the topological information helps in bounding the accumulated drift within consecutive topological nodes, thereby improving the accuracies of such systems by an order of magnitude.

Navigation In the area of navigation, reinforcement learning has been used to investigate the possibility of enabling intelligent agents to learn to navigate through environments without the need for labelled data and without the requirement for explicit localization, mapping or planning procedures as in traditional methods [32]. Deep reinforcement learning methods [17],[7] which originated from solving control problems for playing Atari games are utilized along with deep neural nets as function approximators to represent the Q-value function. In order to ensure that the learned navigation policy can be effectively transferred to new navigation goals and environments, the problem can be framed as a sequence of related reinforcement learning tasks and successor feature based reinforcement learning procedures

are embedded into the network architecture. Unlike the original deep reinforcement learning algorithms that usually result in a black-box function approximator, the successor feature representation of the Q-value function gives us a natural way to transfer learned task solutions to new task instances, while making sure that the solutions to old tasks are preserved after the transfer. Results have demonstrated that the agent is able to learn successful navigation strategies even with sparse supervision from the reward signal it receives and more importantly without any need for human intervention.

4 Towards Predictive Multimodel Learning

While we have seen tremendous amount of progress in robot learning these recent years, robots are still far away from being able to self sufficiently learn and execute tasks as efficiently as humans. This is perhaps because research thus far has been focused on learning models for small subtasks without considering that each of these subtasks might have complex interactions that our conjoint supervised system fails to capture. For example, consider the task of semantic segmentation and visual localization, the model trained for segmentation has strong priors about objects and structures in the scene, which can provide an inductive bias to the model being trained for visual localization. Thereby enabling the localization model to generalize better due to the inductive transfer. If these models are trained in a disjoint fashion, it not only affects the scalability but also restricts the transfer of cues and parameter sharing that could potentially occur. Each of these subtasks have been studied for several decades and numerous deep learning architectures have emerged after months of crafting and tuning. Often this effort is reiterated for different subtasks, limiting the overall learning capability of a robotic system as a whole. While this modular paradigm may be effective in accomplishing a task, it will often break down in unforeseen scenarios that occur in this complex real-world. Moreover, in terms of feasibility, individual models require a large amount of specialized labelled training data, deploying a robot with multiple models demands a substantial amount of GPU hardware and the inability of these models to interact and update their weights online based on current observations, make this impractical to use over longer periods. In contrast, a complete end-to-end approach to a multitask problem such as robot autonomy, forces the model to squeeze enormous amount of information about disjoint tasks into the same parameter space, which is infeasible.

In order for robots to be able to effectively learn, they should be able to perceive the states of the world, plan and perform actions based on these observed states, remember outcomes and be able to make predictions based on these for future actions. At present, robots have some of these components in them but they are disjoint and are not learned in a coherent framework. We think that in a multitask learning scenario, models can not only benefit from the transfer of inductive bias from models in multiple domains but in addition, models trained for tasks with a large amount of examples can self supervise training of models with a small number of train-

ing examples. Multitask learning can be defined as a transfer learning mechanism that improves generalization by using domain specific information contained in the training signals of related tasks [1]. Specifically in convolutional neural networks, multitask learning is generally achieved using either hard or soft parameter sharing. In hard parameter sharing, the hidden layers are shared between all the subtasks, while having task-specific output layers. Hard parameter sharing has recently been used for several tasks including facial landmark detection [23, 34], grasping [22] and face recognition [31]. In the aforementioned works, a core DCNN architecture is employed, followed by task-specific inner-product layers. The advantage of incorporating hard parameter sharing is the reduced risk of overfitting, while the disadvantage being the potential corruption of low-level features in the core architecture due to noise from a related subtask. Soft parameter sharing on the other hand, overcomes this drawback by having task-specific sub-networks with separate hidden layers, while using a sharing mechanism such as regularizing the distance between the parameters of the sub-networks using the l_2 norm [2], trace norm [30] or tensor normal priors [12]. However, the main challenge in soft parameter sharing is developing an appropriate sharing mechanism for the tasks at hand. Recently, Misra *et. al.* proposed cross-stitch units [16] for multitask networks that are in soft parameter sharing configuration. These units learn a combination of shared and task-specific representations from multiple sub-networks and has demonstrated improved performance for tasks with limited training data. However, the placement of these cross-stitch units still remains on open research problem. Some of the mechanisms that enable the aforementioned multitask networks to generalize better include regularization, representation bias, eavesdropping, attribute selection and data augmentation. The effect of these mechanisms in multitask networks are discussed in detail in the work of Caruna *et. al.* [1].

Multimodality is another important characteristic that can enable models to learn the most comprehensive information about the scene or situation which can in turn help them reason more effectively. Multimodal learning can be defined as learning from multiple sensory modes such as cameras, lasers, sound and etc. Learning from multimodal sensory data can help robots enhance their perception of the environment and reduce perceptual ambiguity in challenging conditions. Consider a robot that has to grasp an object; if equipped with only a monocular camera, the robot will have to perform millions of grasps to identify the properties of the object and the right strategy, but if the robot is also equipped with a tactile sensor then robot can more efficiently learn the properties and thus correlate this tactile sense to visual features and use it for future inference, even in a different domain. Moreover, by utilizing multimodal data in a multimodel framework, we also enable better representational learning, as task specific sub-models learn a particular noise pattern with respect to a modality and by inductive transfer, all the sub-models implicitly learn a combined representation of several noise patterns.

What we envision is a unified neural network architecture that is able to perform perception, localization, planning and control, not in a completely end-to-end fashion going from visual input to action, but having individual learnable sub-models in a soft parameter sharing configuration for each of these tasks that enable continuous

update of their weights from each others experiences. This requires each of the sub-models to have a network memory, for example, neural turing machines [5] or differential neural computers [6] so that they can quickly store information and reason from it when required. By joint learning of tasks across these multiple domains, we can not only improve the performance of models by more coherent understanding but also in domains with limited amount of data. In our work of Neural SLAM [33], we give intelligent agents long-term memory capabilities, through the integration of an external memory architercure with a deep reinforcement learning framework. Identifying that cognitive mapping is essential for agents to make comprehensive navigation and exploration decisions, we embed the procedures mimicing that of traditional SLAM algorithms, into a completely differentiable deep neural network. The proposed agent is able to learn to map into its external memory and perform effective exploration behaviors. Finally, a critical trait that is important to our proposed multimodel framework is for models to be able to predict future states. Recent work with adversarial training enables models to predict intermediate actions or future frames of a video sequence using unsupervised learning [13]. While this is only the initial stages of being able to create models that can predict what is unknown, instilling this into the multimodel framework will enable us to create robots that can self-sufficiently learn across domains with limited amount of labelled data.

References

1. Caruana, R.: Multitask learning. Machine Learning **28**(1), 41–75 (1997)
2. Duong, L., Cohn, T., Bird, S., Cook, P.: Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (2015)
3. Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M., Burgard, W.: Multimodal deep learning for robust rgb-d object recognition. In: International Confonference on Intelligent Robots and Systems (2015)
4. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)
5. Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. arXiv preprint arXiv:1410.5401 (2014)
6. Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S.G., Grefenstette, E., Ramalho, T., Agapiou, J., et al.: Hybrid computing using a neural network with dynamic external memory. Nature **538**(7626), 471–476 (2016)
7. van Hasselt, H., Guez, A., Silver, D.: Deep reinforcement learning with double q-learning. In: Proc. of the Thirtieth AAAI Conference on Artificial Intelligence (2016)
8. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: Asian Conference on Computer Vision (2016)
9. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: International Conference on Computer Vision (2015)
10. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single shot multibox detector. In: European Conference on Computer Vision (2016)

11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
12. Long, M., Wang, J.: Learning multiple tasks with deep relationship networks. arXiv preprint arXiv:1506.02117 (2015)
13. Lotter, W., Kreiman, G., Cox, D.D.: Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint arXiv:1605.08104 (2016)
14. Mees, O., Eitel, A., Burgard, W.: Choosing smartly: Adaptive multimodal fusion for object detection in changing environments. In: International Confonference on Intelligent Robots and Systems (2016)
15. Melekhov, I., Kannala, J., Rahtu, E.: Relative camera pose estimation using convolutional neural networks. arXiv preprint arXiv: 1702.01381 (2017)
16. Misra, I., Shrivastava, A., Gupta, A., Hebert, M.: Cross-stitch Networks for Multi-task Learning. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
17. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature **518** (2015)
18. Naseer, T., Oliveira, G., Brox, T., Burgard, W.: Semantics-aware visual localization under challenging perceptual conditions. In: International Confonference on Robotics and Automation (2017)
19. Oliveira, G., Burgard, W., Brox, T.: Efficient deep models for monocular road segmentation. In: International Confonference on Intelligent Robots and Systems (2016)
20. Oliveira, G., Radwan, N., Burgard, W., Brox, T.: Topometric localization with deep learning. In: arXiv preprint arXiv:1706.08775 (2017)
21. Oliveira, G., Valada, A., Bollen, C., Burgard, W., Brox, T.: Deep learning for human part discovery in images. In: International Confonference on Robotics and Automation (2016)
22. Pinto, L., Gupta, A.: Learning to push by grasping: Using multiple tasks for effective learning. In: International Confonference on Robotics and Automation (2017)
23. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. arXiv preprint arXiv:1603.01249 (2016)
24. Valada, A., Oliveira, G., Brox, T., Burgard, W.: Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In: International Symposium on Experimental Robotics (2016)
25. Valada, A., Spinello, L., Burgard, W.: Deep feature learning for acoustic-based terrain classification. In: International Symposium on Robotics Research (2015)
26. Valada, A., Vertens, J., Dhall, A., Burgard, W.: Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In: International Confonference on Robotics and Automation (2017)
27. Vertens, J., Valada, A., Burgard, W.: Smsnet: Semantic motion segmentation using deep convolutional neural networks. In: International Confonference on Intelligent Robots and Systems (2017)
28. Walch, F., Hazirbas, C., Leal-Taix, L., Sattler, T., Hilsenbeck, S., Cremers, D.: Image-based localization using lstms for structured feature correlation. In: International Conference on Computer Vision (2017)
29. Wang, S., Clark, R., Wen, H., Trigoni, N.: Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In: International Confonference on Robotics and Automation (2017)
30. Yang, Y., Hospedales, T.M.: Trace norm regularised deep multi-task learning. arXiv preprint arXiv:1606.04038 (2016)
31. Yin, X., Liu, X.: Multi-task convolutional neural network for face recognition. arXiv preprint arXiv:1702.04710 (2017)
32. Zhang, J., Springenberg, J.T., Boedecker, J., Burgard, W.: Deep reinforcement learning with successor features for navigation across similar environments. In: International Confonference on Intelligent Robots and Systems (2017)
33. Zhang, J., Tai, L., Boedecker, J., Burgard, W., Liu, M.: Neural slam. arXiv preprint arXiv:1706.09520 (2017)
34. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision (2014)