ISSN:2790-1688 Volume-14-(2025)

Deep Learning for Autonomous Robotics: A Review of Perception and Decision-Making

Zicheng Wang

School of Computer Science, SWU University, Chongqing, China

2205319938@qq.com

Abstract. In recent years, deep learning has revolutionized autonomous robotics by enabling advanced perception and decision-making capabilities. This review paper focuses on the state-of-the-art deep learning methods in autonomous robotics, particularly in the areas of perception and decision-making. We discuss the significant progress made in vision-based perception and reinforcement learning for control. However, challenges such as real-time processing under resource constraints and sample inefficiency in reinforcement learning remain. Additionally, we explore the advancements in human-robot interaction and the ethical concerns associated with autonomous decision-making. We also identify gaps in sim-to-real transfer and multi-agent collaboration. By reviewing the major results and discussions in these areas, this paper aims to provide a comprehensive overview of the current research landscape and highlight important directions for future research. The insights gained from this review could potentially bridge the gap between theoretical research and practical implementation, offering valuable guidance for researchers and practitioners in advancing the field of autonomous robotics. Collectively, these efforts pave the way for more capable, reliable, and human-centric autonomous systems that can serve diverse applications in the real world.

Keywords: Deep learning; Robotics; Robot Interaction.

1. Introduction

In the contemporary technological landscape, autonomous robotics has emerged as a frontier of innovation, with deep learning serving as a powerful driving force. Traditional robotics, limited by reliance on handcrafted features, struggles to adapt to dynamic environments. Deep learning transforms this by enabling end learning from data, allowing robots to develop more adaptive and aware behaviors. This has led to remarkable progress across various logistics, healthcare, and manufacturing sectors.

Based perception, a cornerstone of robotic autonomy, has witnessed breakthroughs thanks to Convolutional Neural Networks (CNNs). These networks achieve remarkable accuracy in object detection and semantic segmentation, tasks critical for robots to understand and interact with their surroundings. Yet, time processing under resource constraints remains a significant hurdle, especially for edge devices with limited computational power.

In the realm of making, RL has gained prominence for enabling robots to make sequential decisions in uncertain environments. Algorithms like Proximal Policy Optimization (PPO) and Soft Actor-Critic (SAC) have shown great potential in robotic manipulation tasks. However, sample inefficiency and safety risks in world applications continue to pose challenges that demand innovative solutions.

Robot interaction further expands the scope of autonomous robotics. Transformer models have enhanced natural language processing capabilities, enabling more intuitive communication between humans and robots. Nevertheless, ethical concerns surrounding autonomous decision-making require careful consideration to ensure that robots' actions align with human values.

This review paper aims to comprehensively examine the latest developments in deep learning for autonomous robotics, with a focus on perception and making. It will highlight key achievements, ongoing challenges, and future research directions, providing valuable insights for researchers and practitioners in the field.

2. Neural Network Foundations for Autonomous Robotics

2.1 Vision-Based Perception

Computer vision is a core component of autonomous robot perception. Over the past decade, CNNs have become the dominant architecture for image classification tasks. In classification tasks, CNNs are trained on large labeled datasets to predict object categories from raw pixel inputs. While CNNs achieve high accuracy, performance can degrade in challenging conditions—such as occlusion, cluttered backgrounds, or class imbalance. Figure 1 presents the output of a CNN-based model's prediction results. The figure displays nine images, each with its predicted label. Correct predictions are shown in green, while incorrect ones are in red. As can be seen, the model performs well on most images, but there are cases where it makes errors. These errors could be due to various factors, such as complex backgrounds, occlusions, or limited training data for certain classes.

Semantic segmentation is another key area of vision-based perception. It aims to classify every pixel in an image, providing a more detailed understanding of the scene. Research has shown that deep learning models can achieve impressive results in this area. By using advanced architectures and training techniques, these models can accurately segment objects in complex environments.

2.2 Reinforcement Learning for Decision-Making

Reinforcement learning (RL) enables robots to autonomously learn control strategies via trial-and-error interactions with their environment. Model-free RL algorithms, such as Proximal Policy Optimization (PPO) and Soft Actor-Critic (SAC), have been widely applied in robotic manipulation tasks. PPO, known for its simplicity and effectiveness, optimizes policies by clipping the probability ratios to ensure stable updates. SAC, on the other hand, focuses on maximizing both the expected reward and the entropy of the policy, leading to more exploratory and robust behaviors.

Despite their strengths, these methods suffer from sample inefficiency, often requiring millions of interactions to converge. They often require a large number of interactions with the environment to learn effective policies, which can be time-consuming and computationally expensive. To address this challenge, researchers have explored various strategies. Off-policy training, for instance, allows the algorithm to learn from previously collected data, improving sample utilization. Experience replay is another technique that stores past experiences in a buffer and samples them randomly during training, breaking the correlation between consecutive samples and stabilizing the learning process.

2.3 Human-Robot Interaction and Ethical Considerations

As robots enter social and domestic environments, natural communication and ethical decision-making become critical. Transformer models have shown great potential in natural language processing tasks, enabling more natural and effective communication between humans and robots. These models can understand and generate human language, allowing robots to follow complex instructions and provide informative feedback.

However, ethical concerns are raised in autonomous decision-making. As robots become more integrated into human society, ensuring that their decisions align with human ethical standards is of paramount importance. Research has suggested incorporating ethical guidelines into the decision-making process of robots, either through explicit rule-based systems or by training models on datasets that reflect human ethical values. This can help robots make decisions that not only achieve their objectives but also adhere to societal norms and values.

Figure that using the model to predict the sort of graphs shows the prediction results of a CNN-based model. This figure shows nine images with their predicted labels. Correct predictions are indicated in green, while incorrect ones are in red. The model demonstrates good performance but also highlights cases where errors occur, such as complex backgrounds or limited training data for certain classes.

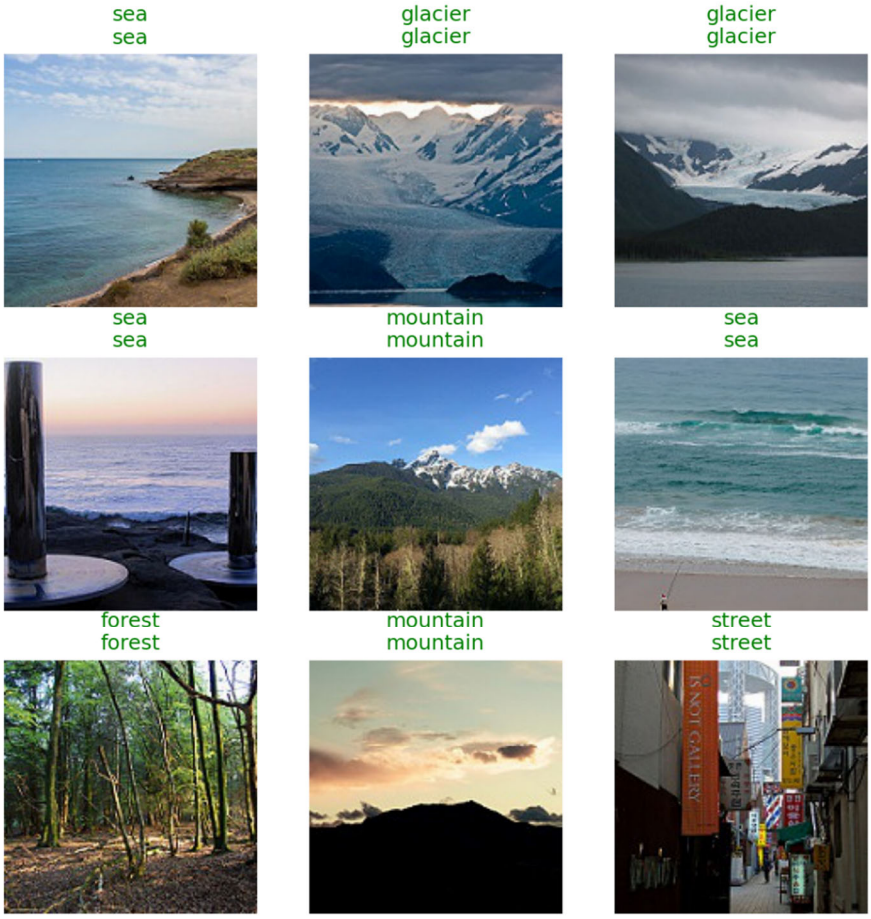

Figure 1. Classification of graphs with CNNs

3. Progress and Challenges in Deep Learning for Autonomous Robotics: A Comparative Review

3.1 Comparative Analysis of Vision-Based Perception Architectures

Deep learning has significantly advanced vision-based perception in autonomous robotics. CNNs have become the dominant architecture for image classification and object detection tasks. They automatically learn hierarchical features from raw pixel data, eliminating the need for manual feature engineering. For example, the AlexNet model achieved breakthrough performance in the ImageNet competition, ushering in a new era of CNN-based vision systems. Subsequent models like VGGNet, ResNet, and, more recently, YOLOv7 have further improved accuracy and efficiency. When comparing these architectures, VGGNet offers a simple and deep structure with uniform filter sizes, which is beneficial for feature extraction but may require more computational resources. ResNet introduces residual connections to ease the training of deep networks, reducing the vanishing gradient problem and improving performance on complex tasks. YOLOv7, on the other hand, excels in real-time object detection scenarios, offering a balance between speed and accuracy that makes it suitable for applications requiring rapid processing, such as real-time surveillance or autonomous vehicle navigation.

However, high computational complexity limits deployment in edge computing contexts, such as mobile robotics and UAVs. To address this, researchers employ model compression techniques (e.g., pruning, quantization) and neural architecture search to reduce inference costs while maintaining

accuracy. These approaches have enabled the deployment of vision-based systems in resource-constrained environments, such as mobile robots and drones.

Moreover, vision-based perception extends beyond object detection to include semantic segmentation and depth estimation. Semantic segmentation models like U-Net and Mask R-CNN can accurately classify every pixel in an image, providing robots with a detailed understanding of their surroundings. U-Net, with its symmetric encoder-decoder architecture, is particularly effective for biomedical image segmentation tasks where precise localization is crucial. Mask R-CNN builds upon Faster R-CNN by adding a branch for predicting segmentation masks, making it suitable for applications requiring both object detection and instance segmentation, such as robotic manipulation in cluttered environments. Depth estimation models enable robots to perceive the 3D structure of the environment, which is crucial for tasks like obstacle avoidance and navigation.

3.2 Sample Efficiency and Safety in Reinforcement Learning Algorithms

Reinforcement learning provides autonomous agents with the ability to learn optimal behaviors through interaction. Model-free RL algorithms like PPO and SAC have demonstrated impressive performance in various robotic manipulation tasks. PPO, with its clipped surrogate objective function, ensures stable policy updates and has been successfully applied in training robots for tasks such as grasping and locomotion. SAC, by incorporating entropy regularization into the reward function, encourages exploration and leads to more robust policies.

When comparing the sample efficiency of these algorithms, model-free methods generally require a large number of interactions with the environment, which can be time-consuming and computationally expensive. To address this issue, researchers have explored off-policy training and experience replay. Off-policy algorithms like Deep Deterministic Policy Gradient (DDPG) and Twin Delayed DDPG (TD3) can learn from previously collected data, improving sample utilization. Experience replay further enhances this by storing and randomly sampling past experiences during training, breaking the correlation between consecutive samples and stabilizing the learning process.

Model-based RL algorithms have also gained attention. These algorithms learn a model of the environment dynamics and use it to generate synthetic experiences for training. While model-based methods can be more sample-efficient, they often face the challenge of model bias. To overcome this, hybrid approaches that combine model-based and model-free techniques have been proposed. For example, the Dyna-Q algorithm uses a learned environment model to generate imagined experiences, which are then used by a model-free RL algorithm to update the policy.

In addition to improving sample efficiency, ensuring safety in RL-driven decision-making is of paramount importance. Constrained policy optimization, safe exploration, and shielded RL introduce safety constraints into the learning process. Techniques like constrained policy optimization and shielded RL aim to prevent the robot from taking unsafe actions during training and deployment. These methods provide theoretical guarantees on safety while still allowing the robot to learn effective policies.

3.3 Ethical Considerations and Current Limitations in Human-Robot Interaction

As autonomous robots become prevalent in human environments, ethical considerations and natural communication become increasingly important. Transformer models have revolutionized natural language processing and have been successfully applied in enabling robots to understand and generate human language. Models like BERT and GPT have demonstrated remarkable performance in language understanding and generation tasks, enabling more natural and effective communication between humans and robots.

Despite this progress, autonomous decision-making in social contexts raises ethical challenges. Robots may face dilemmas involving conflicting human values, such as prioritizing treatment efficacy over patient comfort in healthcare. Encoding ethical principles into robot behavior remains nontrivial, particularly given the abstract and contextual nature of ethics. To address these concerns, researchers have proposed incorporating ethical guidelines into the decision-making process of robots.

One approach is to develop explicit rule-based systems that encode ethical principles as constraints or objectives. Another approach is to train models on datasets that reflect human ethical values, allowing the robot to learn appropriate behavior through imitation or reinforcement learning. Additionally, explainable AI techniques can be employed to make the robot's decision-making process transparent to humans, enabling better understanding and trust.

While significant progress has been made in vision-based perception, reinforcement learning, and human-robot interaction, there are still numerous challenges and open questions that need to be addressed. Bridging the sim-to-real gap, improving sample efficiency and safety in RL, and ensuring ethical decision-making are some of the key areas where further research is needed. By continuing to advance these fields, we can pave the way for more capable, reliable, and trustworthy autonomous robotic systems.

4. Future Directions

4.1 Bridging the Sim-to-Real Gap

A persistent challenge in autonomous robotics is the discrepancy between simulated environments and real-world deployment, commonly referred to as the sim-to-real gap. Future research could focus on developing more advanced domain adaptation techniques. For example, researchers could explore the use of advanced generative models to create more realistic synthetic data or develop better sim-to-real transfer learning algorithms that can effectively generalize from simulations to real-world scenarios. Recent work, such as the "real-is-sim" framework, proposes maintaining a dynamically corrected physics simulator that continuously aligns with real-world observations. This approach treats the simulator as the execution environment throughout data collection, training, and deployment, thereby eliminating the traditional divide between simulation and reality and shifting the challenge of domain alignment to the simulator itself.

4.2 Improving Sample Efficiency in Reinforcement Learning

Reinforcement learning algorithms typically require a large amount of interaction data with the environment. Future work could focus on improving sample efficiency by developing better exploration strategies. For instance, curiosity-driven exploration methods could be refined to encourage robots to explore their environment more effectively. Additionally, leveraging prior knowledge or transferring knowledge from related tasks could help reduce the amount of data needed for training. Research suggests that combining imitation learning (IL) and reinforcement learning (RL) can enhance sample efficiency. IL facilitates rapid skill acquisition through expert demonstrations, while RL allows for self-discovery of optimal strategies. Integrating these methods could yield more data-efficient and generalizable robotic policies, especially in real-world scenarios where data collection is costly or risky.

4.3 Ensuring Safe Autonomous Decision-Making

As autonomous robots are increasingly deployed in safety-critical applications, ensuring their decisions are safe and reliable is of paramount importance. Future research could integrate formal verification methods into learning algorithms and develop reward functions that explicitly encode safety constraints. Techniques such as incorporating formal verification methods into the learning process or designing better reward functions that explicitly account for safety could be explored. Furthermore, creating better safe mechanisms that allow robots to recognize and handle situations beyond their training could enhance safety. For example, in autonomous delivery robots operating in crowded urban areas, integrating safety constraints and verification mechanisms into DRL can help robots learn to avoid risky behaviors while maintaining reliable paths to their destinations.

4.4 Enhancing Multi-Robot Collaboration

In many real-world scenarios, robots need to work together as a team to accomplish complex tasks. Future research could focus on improving robot collaboration. Federated learning offers a promising approach for enabling robots to learn collaboratively without sharing raw data. Researchers could further develop federated learning algorithms tailored for robotic applications. Additionally, exploring better communication protocols and coordination strategies would be crucial for enhancing the efficiency of robot systems. Recent studies have shown the potential of multi-agent reinforcement learning in promoting collaboration among robots. For instance, in search-and-rescue missions, teams of drones and ground robots can work together to navigate and locate survivors in disaster-stricken areas, improving the efficiency of rescue operations.

4.5 Addressing Ethical and Legal Challenges

The deployment of autonomous robots raises significant ethical and legal questions. Future research could focus on developing ethical frameworks for robot decision-making. This could involve creating guidelines and regulations that ensure robots' actions align with human values and societal norms. Additionally, addressing legal liability issues surrounding autonomous systems will be crucial for their broader adoption. As robots become more integrated into human society, ensuring their decisions adhere to ethical principles is vital. Legal research should also address questions of accountability, liability, and transparency, especially in sectors like healthcare, where decisions may have high-stakes outcomes. Embedding ethical reasoning into decision architectures will be essential for ensuring societal alignment.

4.6 Advancing Human-Robot Interaction

Human-robot interaction will play an increasingly important role as robots become more integrated into human society. Future research could focus on making robot behavior more interpretable and transparent to humans. For example, developing better explainable AI techniques that allow robots to explain their decisions and actions understandably could improve trust and acceptance. Additionally, designing more natural and intuitive interaction interfaces would be beneficial for enhancing human-robot collaboration. Recent research has demonstrated the potential of transformer models in enabling more natural and effective communication between humans and robots. These models can understand and generate human language, allowing robots to follow complex instructions and provide informative feedback, thereby enhancing human-robot interaction.

4.7 Exploring New Architectures and Algorithms

The field of deep learning is constantly evolving, with new architectures and algorithms emerging regularly. Future research could explore the application of these new developments to autonomous robotics. For example, transformer-based models have shown great potential in various domains and could be further adapted for robotic vision and decision-making tasks. Similarly, neuromorphic computing architectures inspired by the human brain could offer new possibilities for efficient and low-power robotic systems. Recent advancements in 3D physically grounded world models combine particle-based simulation with visual rendering via Gaussian Splatting, enabling real-time tracking and simulated interaction of physical scenes without violating physical and structural constraints. These models can serve as natural simulation environments for robotics, providing a promising direction for future research.

5. Conclusion

In this review paper, we have comprehensively examined the art of deep learning methods for autonomous robotics with a focus on perception and making. We highlighted the transformative impact of CNNs on perception, enabling robots to perform complex tasks like object detection and semantic segmentation. Despite their success, challenges in time processing and the real gap remain.

In the realm of making, we explored the advancements in reinforcement learning and the efforts to improve sample efficiency and safety. The potential of federated learning for robot collaboration was also discussed. For robot interaction, we emphasized the role of transformer models and the critical need to address ethical concerns in autonomous systems.

The review concluded by outlining key future research directions, including bridging the real gap, enhancing the safety and sample efficiency of RL algorithms, and promoting ethical and transparent robot interaction. These directions point the way forward for the continued advancement of autonomous robotics, aiming to create systems that are not only technically capable but also reliable and aligned with human values.

References

[1] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[2] Silver, D., et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489.

[3] Mnih, V., et al. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.

[4] Vinyals, O., et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 575(7782), 370-374.

[5] Gu, S., et al. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. IEEE International Conference on Robotics and Automation (ICRA), 3386-3393.

[6] Silver, D., Schrittwieser, J., Simonyan, K., et al. (2017). Mastering the game of Go without human knowledge. Nature, 550(7676), 354-359.

[7] Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction. The MIT Press.

[8] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[9] Vaswani, A., et al. (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 6000-6010).