Deep Learning Applications in Simultaneous Localization and Mapping

Haoyang Zhang
The School of Electronic Engineering, Xidian University, Xi’an, Shaanxi, 710126
820005318@qq.com

Abstract. Simultaneous Location and Mapping (SLAM) is a research hotspot in the field of intelligent robots in recent years. Its processing object is the visual image. Deep learning has achieved great success in the field of computer vision, which makes the combination of deep learning and slam technology a feasible scheme. This paper summarizes some applications of deep learning in SLAM technology and introduces its latest research results. The advantages and disadvantages of deep-learning-based-SLAM technology are compared with those of traditional SLAM. Finally, the future development direction of SLAM plus deep learning technology is prospected.

1. Introduction
These guidelines, written in the style of a submission to J. Phys.: Conf. Ser., show the best layout for your paper using Microsoft Word. If you don’t wish to use the Word template provided, please use the following page setup measurements[1]. SLAM is a process in which robots are equipped with vision, laser, odometer, and other sensors to realize self-positioning while building maps in unknown environments. It plays a key role in robot autonomous navigation tasks [1]. It can replace humans in some special occasions, such as the military, transportation, service industry, and other fields. For a long time, location is the premise of path planning. In the location problem, the primary task of the machine is to perceive the surrounding environment and then describe it. SLAM can present information more efficiently and intuitively than traditional text, image, and video. In the environment where GPS cannot be used normally, SLAM can also be used as an effective alternative to achieve real-time navigation in unknown environment. SLAM technology plays a more and more important role in many fields, such as service robot, driverless car, augmented reality and so on.

As shown in Figure 1, a complete SLAM framework consists of the following four aspects: front-end tracking, backend optimization, loop detection, and map reconstruction [1]. Visual odometer (VO) is a process of estimating motion information of an agent using input information from only one or more cameras; Back-end optimization is responsible for receiving the posture information measured by VO, and then resolves the optimization of the historical track after the robot detects the closed-loop; Loop detection determines whether a robot has entered the same place in history. It triggers the SLAM back-end global consistency algorithm to optimize the map, eliminating cumulative trajectory errors and map errors; Map reconstruction is responsible for building a map suitable for the task according to the camera pose and images.
Traditional visual SLAM algorithms have not made breakthrough progress since 2017 due to these unavoidable problems below, and there is no perfect solution:

1. The robustness of the traditional algorithm is not very high under adverse conditions such as poor illumination conditions or large illumination changes;
2. If the camera moves too fast, the traditional algorithm is easy to lose the tracking target;
3. Traditional algorithms cannot recognize foreground objects, that is, they can only be regarded as "bad points", and there is no good solution.

Hardware technology develops rapidly according to Moore's law, computing power is greatly improved, and more and more computer vision problems have made great breakthroughs. At present, the performance of deep learning in image classification, recognition, object detection, image segmentation, and other fields is much higher than traditional manually designed algorithms. Visual SLAM also takes images as the processing object. The combination of deep learning and SLAM improves the application limitations caused by manual design features such as visual odometer and scene recognition and potentially improves the learning ability and intelligence level of the robot.

2. Deep learning and visual odometer

To complete autonomous navigation, a mobile robot first needs to determine its own position and attitude, that is, positioning. Visual odometry (VO) estimates the motion of the camera by tracking the feature points between adjacent image frames and reconstructs the environment. VO mostly estimates the pose of the current frame by calculating the motion between frames. The visual odometer based on deep learning does not need complex geometric operations. The end-to-end operation form makes the method based on deep learning more concise.

The network architecture proposed by Daniel and Malisiewicz [2] completes point tracking and obtains the homography between adjacent frames. As shown in Figure 2, the main feature of the model is to use the cooperation of two CNN [3] to generate homography matrix to complete the estimation of camera pose. The first is called magicpoint network, which extracts the salient points of a single image. The second is called magicwarp, which operates on the output of magicpoint.
Although deep learning has become the mainstream method to solve many computer vision problems and achieved good performance, the research on VO is very limited, especially in three-dimensional geometry. According to the theory and practical experience of transfer learning, if the trained neural network model parameters are to be used, the problems must be similar. However, at present, the main solution in the field of computer vision is high-level information such as recognition and classification, which is different from the task of VO. That is also the reason why the traditional VO algorithm relies heavily on geometric features rather than high-level semantic information. At the same time, motion is a continuous change process. The ideal VO algorithm should model the change and connection of a series of images. This requires that we should use time-series data to learn from image sequences.

Compared with MagicPoint and MagicWarp, which only complete homography estimation for the basic geometric features of two frames of images, Wang et al. [4] proposed a DeepVO algorithm of monocular VO based on deep learning for image sequences. For data-driven networks such as DeepVO, it is often sensitive to the rules hidden behind the data. Therefore, DeepVO has good robustness and fitting ability for information interference such as label error, which reflects a significant advantage of the data-driven model.

Although the VO method based on deep learning has obtained some results in camera pose estimation, it cannot replace the method based on geometry. The deep learning method is a feasible supplement, which can combine geometry features with neural network models. Further improving the accuracy and robustness of VO is the foreseeable development direction at present.

3. Deep learning and loop detection

Loop detection is to judge that the robot has returned to its original position and reasonably allocate the cumulative error to the loop trajectory. The description and matching between images are the key technology of loop detection. In traditional methods, researchers usually use hand-crafted features to describe an image. Artificial features are divided into local features and global features. Local features include ORB, SIFT, SURF, etc. Global features include GIST, fisher vector and others describe the features of the whole image in different calculation methods. The development of deep learning technology provides a new way to solve the loop detection problem in SLAM. In the past two years, many researchers have extracted the features of 3D point cloud by learning, which makes it possible for loop detection based on 3D point cloud.

Deep VSLAM adopts the way of learning, and its future development will be more similar to human perception and thinking. The input of the model constructed in reference is an image sequence. This method performs well in the simulation environment, which proves the role of soft attention [5] model in closed-loop detection. According to the experimental performance of the model, it does not achieve the expected effect in the actual environment. But the end-to-end learning mode and the data processing process of the whole model are in line with the human perception process, and there is a lot of room for development in the future.

4. Comparison between the deep learning and the traditional methods

Traditional SLAM design based on image information feature representation first needs to solve the problem of salient feature selection. Robust features should be invariant in different viewing angles, changing illumination intensity, and changing background. Feature extraction can be divided into two parts: feature detection and feature description. Whether the extracted feature descriptor has good invariance directly affects the inter frame estimation accuracy of SLAM system, the efficiency and accuracy of loop closure detection, and the functionality of semantic knowledge base.

Local features based on the directional histogram, such as HOG (histogram of oriented gradient), SIFT, and SURF have occupied the traditional SLAM algorithm for a long time. Such features need to be carefully designed by experts with domain expertise, which is also called feature engineering. At the same time, the performance of manually designed features decreases in the face of changes in light intensity and object motion, especially in the field of object recognition, which has become the
bottleneck of improving the performance of SLAM system. On the other hand, the feature extraction of traditional SLAM algorithm is separated from classifier design. As a result, the matching accuracy of SIFT and other features is not high enough. When building semantic map, traditional SLAM not only needs to build object feature description database but also needs to train decision forest and other classifiers for object classification.

End-to-end SLAM using deep learning can bypass many of the most difficult parts of traditional SLAM systems, such as external parameter calibration and multi-sensor frequency matching. It avoids some of the challenges in front-end and back-end algorithms. Combining deep learning as a new SLAM implementation method has strong theoretical significance, but the drawbacks of end-to-end SLAM are obvious.

SLAM uses a data-driven approach to learning rules, which is groundless in principle and has no reason to obtain high-precision solutions. Another problem is that the generalization of the model cannot be guaranteed. The traditional SLAM system is usually a very complex structure. Every step of operation from the front-end to the back-end has a clear purpose. The traditional method has detailed mathematical theory as its support and has a strong explanatory ability. However, the parameters need to be carefully selected. While approximating SLAM systems with highly data-dependent deep learning may produce good results for some datasets, changing scenarios may be less sensitive, but if the dataset is large enough, the neural network can still show strong data adaptability, so the volume of the dataset is an important factor affecting the accuracy of the neural network.

5. Conclusion
In terms of human perception, when facing the objects in the scene, we cannot only obtain the location information (3D) but also determine the color data (3D). Human beings even can also obtain semantic information, such as surface hardness, instance contour, whether it can be touched, and so on. However, it is far not enough to build a 3D point cloud-only by deep SLAM. Therefore, it is necessary to build a high-level map with richer content in a higher dimension to meet various needs. Positioning and perception are not the ultimate goals of SLAM. VSLAM completes complex tasks on the premise of accurate positioning and perception. This puts forward higher requirements for the combination of deep learning and SLAM. When training deep SLAM, we take the completion of tasks as the standard. Deep learning simulates the human brain structure, constructs a complex neural network model, and uses a large amount of data for training to simulate the human learning process. It has achieved great success in the fields of semantic segmentation, object recognition, and action recognition. The combination of deep learning and SLAM improves the application limitations caused by manual design features such as visual odometer and scene recognition. Thus, the learning ability and intelligence level of the robot can be potentially improved.

However, the information features learned through deep learning technology still lack intuitive significance and clear theoretical guidance. Therefore, on the one hand, in the future, the performance of SLAM system will also benefit from the development of deep learning theory for a long time; On the other hand, deep learning is mostly applied to local sub-modules of SLAM, such as positioning module or loop closure detection module. How to use deep learning technology to realize the whole slam process is still a great challenge.

References
[1] Fuentes-Pacheco J, Ruiz-Ascencio J, Rendón-Mancha J M. (2015) Visual simultaneous localization and mapping: a survey. Artificial intelligence review, 43(1): 55-81.
[2] DeTone D, Malisiewicz T, Rabinovich A. (2017) Toward geometric deep slam. arXiv preprint arXiv:1707.07410.
[3] Sharif Razavian A, Azizpour H, Sullivan J, et al. (2014) CNN features off-the-shelf: an astounding baseline for recognition//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 806-813.
[4] Wang S, Clark R, Wen H, et al. (2017) Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks//2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2043-2050.

[5] Southall C, Stables R, Hockman J. (2017) Automatic drum transcription for polyphonic recordings using soft attention mechanisms and convolutional neural networks.

[6] Handa A, Bloesch M, Pătrăucean V, et al. (2016) gvnn: Neural network library for geometric computer vision//European Conference on Computer Vision. Springer, Cham, 67-82.