An Embedded Deep Learning Architecture for Real-Time Place Recognition from Omnidirectional Images

Marta Rostkowska

Institute of Robotics and Machine Intelligence,
Poznań University of Technology, ul Piotrowo 3A, 60-965, Poznań, Poland
martarostkowska89@gmail.com

Abstract. This paper presents a real-time solution for place recognition in indoor environments using a convolutional neural network for extracting embeddings from omnidirectional images, which allows the robot to register a description of the entire surroundings. The proposed neural network recognizes places on distorted images from a catadioptric camera, in contrast to the more widely used approach which is based on producing panoramic images from omnidirectional images, which involves many mathematical transformations. The proposed solution achieves robust place recognition results owing to efficient retrieval of embeddings created exploiting transfer learning and fine-tuning on a limited number of actual omnidirectional images. The localization system is implemented on a NVIDIA Jetson TX2 computer with a general purpose graphics processing unit. The proposed neural network architecture makes it possible to process the omnidirectional images in real-time on this embedded hardware, which provides cost and energy efficient means of appearance-based localization for indoor service robots

Keywords: Place recognition · deep learning · omnidirectional vision.

1 Introduction

One of the most important aspects of robot autonomy is the ability to determine agent's location in the environment. Passive cameras are arguably the most popular sensors for robot localization, while particularly interesting are the omnidirectional cameras that enable the whole local scene to be registered in one image. Omnidirectional images are convenient for appearance-based visual localization, called also place recognition. This approach yields information about the similarity of the places observed in the current perception and locations stored in a database [3]. Although appearance-based localization does not provide metric information about the position of the robot in a global reference system, the ability to tell if the robot is close to one of the known locations is often sufficient for indoor navigation.

Therefore, we propose a novel approach that adopts a Convolutional Neural Network (CNN) architecture to process the omnidirectional images for real-time place recognition. The proposed system exploits the concept of global image descriptors, which was already proved to be efficient in place recognition [1]. We employ a CNN to produce the descriptors directly from the omnidirectional

images, thus avoiding the computation overhead required for producing undistorted panoramic images, which are typically used in place recognition systems for catadioptric cameras [6].

2 Localization System Overview

The localization procedure is based on finding the omnidirectional image from a known database that is most similar to the one currently acquired by the robot. As the database images are registered at known locations, finding the one that has the minimal distance (in the sense of appearance similarity) to the current perception makes it possible to roughly localize the robot.

We use a Labbot mobile robot with an integrated omnidirectional vision sensor [4] placed on top (Fig. 1a). The catadioptric sensor consists of a Microsoft Life Cam camera with a hyperbolic mirror which provides a 360° field of view and yields images in 640×480 resolution. The sensor is equipped with a NVIDIA Jetson TX2 computer with an integrated 256-core Pascal architecture General Purpose Graphics Processing Unit (GPGPU). This unit is enough to run our localization system in real-time.

In this research a dataset of 606 images (Fig. 1c and 1d) describing the robot's environment was acquired in one of the Poznan University of Technology buildings (Fig. 1b). In order to remove the areas in the images that do not carry useful information, the raw images are masked, which removes the area outside the hyperbolic mirror, and the area reflecting the camera (Fig. 1e). These images are processed by our CNN to obtain embeddings of the images. Finally, descriptors of 2048×1 size are computed for each image and stored in a database of 2048×n size which is our global map for appearance-based localization over n reference images (n=484 in the experiment). Then the algorithm creates an index from the global map (using Faiss[2] library), which is used for efficient similarity search. All these operations are accomplished off-line.

The main localization task is done on the Jetson platform in real-time. First, the CNN model and the index of images are loaded to the memory, then the candidate images are being found using KNN search in the descriptor space among the descriptors of images from the database. The real-valued descriptors are compared using the L2 distance, which turned out to be more computationally efficient than binarizing the embeddings and using the Hamming distance.

3 Deep Learning Architecture

The advantage of CNN in the image description task over traditional descriptors is related to the ability of a CNN to extract rich features. The learned descriptors are more robust to changing lighting or changes in the robot orientation than classic global image descriptors, particularly, if an extensive data augmentation process is applied while learning to disregard these changing factors.

The procedure of extracting image features and storing them in an efficient format is called embedding. It makes possible to access the feature-based description without having to pass the images from a database through the same

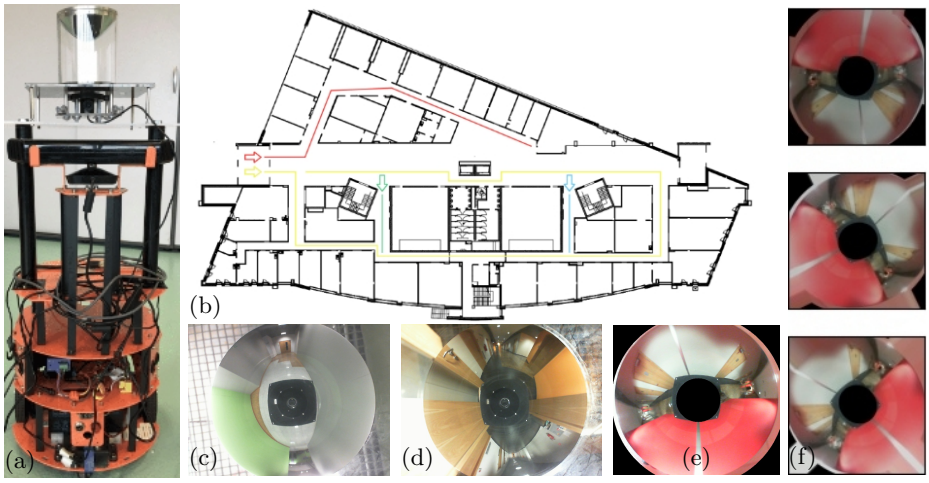

Fig. 1: Labbot robot with the catadioptric vision system (a), robot paths while collecting the images – different colors indicate different paths, then divided into segments (b), omnidirectional images of different locations (c,d), an omnidirectional image after masking (e), and examples of data augmentation (f).

neural model as the query image, as it is done, e.g. by Siamese networks [7] used for image retrieval.

We have tested a number of CNN architectures as the feature extractors in our system, finally choosing the EfficientNet [5] in B5 variant, which has 577 layers, with the input image size defined as (456,456,3). This network has high accuracy with a relatively small number of model parameters, which positively affects the processing speed in our embedded system. Due to the fact that the EfficientNet B5 was pre-trained on images (Imagenet dataset) not related to the target dataset, the network was fine-tuned before use, unfreezing a number of layers and using the categorical crossentropy loss function. This process was implemented using the dataset of around 10000 augmented omnidirectional images, produced from the previously gathered database (Fig. 1f).

A practical problem in the considered scenario was the high self-similarity of the indoor environment. As the images were acquired roughly every 0.5 m along the robot path, the neighboring images in the original database are very similar and often indistinguishable to human being. Therefore, the entire dataset was divided manually into 17 different sections, each section describing a topologically different location. Then, the localization process is executed only with respect to these 17 meaningful locations, while each of them is represented by 30 to 40 acquired images, which are partially overlapping. In the training process, each section was divided into the training (60%), validation (20%) and test (20%) sequences.

The best training results were obtained for unfrozen 50 last layers, learning rate of $1e^{-4}$ and batch size 16, with the resulting training loss: 0.1605, training accuracy: 0.9596, validation loss: 0.1183 and validation accuracy: 0.9796 (Fig. 2).

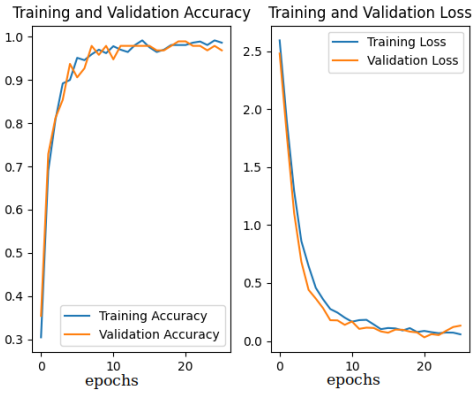

Fig. 2: Model training results

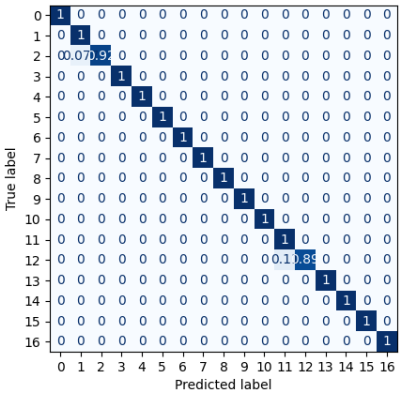

Fig. 3: Confusion matrix for 17 sections

4 Results

On the test dataset containing 122 pictures, the average accuracy of place recognition was 98% (Fig. 3), while the average processing time of a single picture was 480ms, with standard deviation of 83ms and max time of 1313ms, which allows localization at frame rate of the robot's camera. An example of correct place recognition is given in Fig. 4. The most often sections mismatching is related to a situation where the same place is the beginning of a new section and the end of the previous one. Errors are also caused by blurred images and light spots.

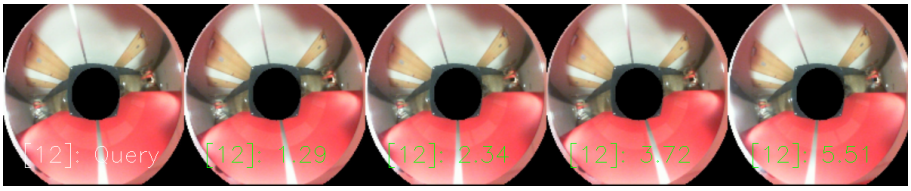

Fig. 4: Results of sample section predictions. The first image is a query, the others are the four closest neighbors. In square brackets there is the section number, and next to it is the L2 distances between the query and presented image.

5 Conclusion

This short paper demonstrated that a CNN can be trained efficiently, using transfer learning and fine-tuning approach, to produce embeddings that describe distorted omnidirectional images in an appearance-based localization system. The proposed architecture makes it feasible to run the entire process in real-time on-board of an integrated sensor with an embedded Jetson TX2 computer. Further research concerns applying spherical representations to the omnidirectional images to avoid the inactive areas, and employing a more advanced learning technique, such as triplet loss.

References

1. Arroyo, R., Alcantarilla, P. F., Bergasa, L. M., Romera, E.: Towards life-long visual localization using an efficient matching of binary sequences from images. IEEE Int. Conf. on Robotics and Automation, (2015), 6328–6335.
2. Facebook AI Research: Faiss, https://github.com/facebookresearch/faiss. Last accessed 28 March 2022.
3. Lowry, S., Sunderhauf, N., Newman, P., Leonard, J., Cox, D., Corke, P., Milford, M.: Visual place recognition: A survey. IEEE Transactions on Robotics, 32, (2016), 1–19.
4. Rostkowska, M., Skrzypczyński, P.: Hybrid field of view vision: From biological inspirations to integrated sensor design. IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI), Baden-Baden, (2016), 653–658.
5. Tan, M., Le, Q, V.: EfficientNet: Rethinking model scaling for convolutional neural networks. Proc. 36th Int. Conf. on Machine Learning, PMLR 97:6105-6114 (2019).
6. Wang, T., Huang, H., Lin, J., Hu, C., Zeng, K., Sun, M.: Omnidirectional CNN for visual place recognition and navigation. IEEE Int. Conf. on Robotics and Automation (ICRA), Brisbane, (2018), 2341—2348.
7. Zemel, R., Koch, G., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. ICML Deep Learning Workshop, Lille, France, (2015).