Data-efficient representation learning for visual place recognition

Leyva Vallina, María

DOI:
10.33612/diss.736449452

Chapter 6

Summary and Outlook

6.1 Summary

In the first half of this thesis, we introduced a novel visual place recognition dataset, namely TB-Places. This dataset provides a challenging and realistic benchmark for evaluating and comparing different place recognition algorithms. It includes a variety of visual challenges, such as changes in lighting conditions, seasonal variations and drastic changes in viewpoint. Moreover, the dataset includes over 90K images taken in two garden environments over three years, which allows for a comprehensive evaluation of different algorithms. Each image has an associated 6DOF camera pose and a vector with a similarity ground truth for each possible match image. Moreover, we evaluated several methods on this dataset and concluded that existing methods do not generalize well for garden environments.

We devoted the second half of this dissertation to explore the possibility of learning an explicitly continuous measure of similarity for visual place recognition, not only for garden environments but also for urban and indoor environments. For this, we automatically re-annotated the TB-Places, MSLS, and 7Scenes dataset with a graded ground truth that we derived from the already provided geometrical information. We used this new ground truth to optimize our novel Generalized Contrastive Loss, a reformulation of the contrastive loss that is defined for a ground truth with values $\in [0, 1]$, thus enabling the trained model to explicitly learn measures of partial similarity. This reformulation also dispensed with the need to perform descriptor-based pair mining to ensure convergence, making our approach much less resource hungry than previous binary methods. Our models trained with a GCL function outperformed the state-of-the-art methods while simplifying the training pipeline and reducing the training time.

This contrastive approach led to excellent results, but the loss to optimize consists of two terms: the first one pushes representations of similar images closer in the latent space, while the second one pushes the representations of dissimilar images to be farther apart. This division assumes a certain binarization of the problem, which is not coherent with the fact that image similarity is a continuous at-

tribute. From this insight we overcame the contrastive optimization approach and re-defined visual place recognition as a regression problem, introducing an inductive bias to map the graded ground truth to the Euclidean distance between the representations. This approach takes the same time to train as the GCL one, but it reaches a performance competitive with the state-of-the-art in fewer iterations, making it very data-efficient.

In conclusion, in this thesis we presented a shift of paradigm for visual place recognition, redefining it as a regression problem that predicts a continuous measure of image similarity, in contrast to previous approaches that treated the problem as a binary one. This redefinition allows us to train larger models in a faster way, both in terms of time and iterations, while outperforming more complex approaches that are trained for longer and with more complex pipelines.

6.2 Outlook

This dissertation opens an interesting research path which can take several directions. Although we have introduced TB-Places, a dataset for visual place recognition in gardens and obtained good results with several methods, there is still a lot of room for improvement. One path could be exploring algorithms for visual localization and pose refinement in repetitive structures, following a path similar to Torii et al. (2015a), which is tested only for indoor environments.

From a more general approach, although we have demonstrated that visual place recognition can be tackled as a regression problem, and we have introduced data-efficient methods to learn good representations, all those methods are supervised, so they still rely upon a large amount of labelled data. One could explore automatic labelling strategies, for instance, based on unsupervised local features, such as HOG, or keypoints, and proceed with a supervised learning pipeline, but with a ground truth generated without manual intervention or available pose metadata. Another alternative would be exploring self-supervised approaches such as Chen et al. (2020); Grill et al. (2020); Chen and He (2021), although those have the limitation of requiring a multi-GPU environment (the standard configuration is 8 GPUs), something that is not achievable by many research teams.