Learning models for visual 3D localization with implicit mapping

Björn Bebensee (2019–21343)
Topics in Artificial Intelligence

October 17, 2019

A key limitation in 3D localization is the need for hand-engineered representations and it may benefit from more abstract representations. The optimal representation of maps is not always clear yet with traditional methods it needs to be defined in advance in advance which may limit the algorithm's performance. Rosenbaum et al. [1] propose a new approach for visual localization which instead of defining these maps explicitly learns an implicit, abstract representation of the environment. To achieve this they propose using Generative Query Networks (GQNs) which have shown promise in learning representations of 3D scenes.

In this context they consider the re-localization task, where, given a collection of context images (with known camera poses), the goal is to find the camera pose of a new image. The dataset used is a collection of random walks in Minecraft. The localization problem can be described as an inference task: given observed images X in an environment E with camera pose P (see figure 2a) predict $Pr(P|X, E)$. The GQN model they use for this consists of a representation network which processes each context image along with its camera pose to obtain a single scene representation vector r and a generation network which, given this scene representation and the target camera pose, computes a probability distribution of the target image. As it is unlikely that the global scene representation vector r can capture all the information from a complex scene, the authors introduce an attention mechanism to their model which replaces the representation network in the GQN model and computes a patch dictionary containing all the image patches. The generation network then performs soft attention using this patch dictionary (in place of the global representation vector r).

Rosenbaum et al. compare this generative approach with a discriminate approach that, instead of learning in the $P \to X$ direction, learns in the $X \to P$ direction and thus learns to predict a distribution over the camera pose conditioned on the target image. They find that the models with the attention mechanism perform better (as expected). Furthermore the generative model captures the uncertainty better than the discriminative variant which relies more on the prior of poses it has learned rather than the provided context images. However, the discriminative model is much more efficient as it only requires a single forward pass during inference while the generative model needs to solve an optimization problem.

References

[1] Rosenbaum, Dan, et al. "Learning models for visual 3D localization with implicit mapping." *arXiv preprint arXiv:1807.03149* (2018).