Nicolò Valigi

Software Engineer at Cruise Automation. I write about software and machine learning.

How can Deep Learning help Robotics and SLAM

By now, Deep Learning needs no introduction for most people in the tech community. Powered by massively parallel GPUs and hundreds of research teams around the world, neural networks have taken the machine learning community by storm in the last few years. Convolutional networks have won image classification competitions since 2010, recurrent networks are incredible at modeling and understanding human language, and deep nets can even compose music after listening to Mozart for a while.

Applications of Deep Learning (DL) to autonomous robots are still not as widespread as these more well-known problems listed above. In this article, I'm going to focus on one such area and review the literature on Deep Learning as applied to a crucial problem in perception: Simultaneous Localization and Mapping (SLAM).

State of the (SLAM) art

A majority of SLAM systems share several common components:

- a **feature detector** that finds point of interest within the image (*features*),
- a **feature descriptor** that matches tracks features from one image to the next,
- an **optimization** backend that uses said correspondences to build a geometry of the scene (map) and find the position of the robot,
- a **loop closure detection** algorithm that recognizes previously visited areas and adds constraints to the map.

Due to its similarities to well-studied image classification and retrieval problems, *loop closure* has the most potential to be solved with DL techniques. It's also an important issue, as correct loop closures guarantee the consistency of the SLAM map and improve all-around accuracy. Computational efficiency and robustness to false positives are the most important characteristics of a successful loop closure subsystem.

ORB-SLAM is arguably the most advanced open-source package for visual SLAM. The crucial insight that enables real-time performance is using the same ORB descriptor (Rublee and Bradski (2011)) for both frame-to-frame tracking and loop closure detection. To go from individual features to a global description of the image, ORB uses another common technique, a **bag of words (BoW)** representation.

> *In other words, conventional methods use non-learned feature descriptors and bundle them together using yet another non-learned algorithm (BoW-like).*

Following the same line of reasoning that worked so well for image classification, one could anticipate that *learned* features could be more effective that hand-engineering specific algorithms. That's exactly where the power of deep learning lies: developing better representations for the image data, which is just another form of the heavy-handed compression we need for fast loop closure detection.

The next two sections will be a long-form literature review of some approaches in this field.

A straightforward approach

While the superior accuracy of CNN-based descriptors for **classification** is a well-known fact by now, Fischer et al. (2014) prove that CNNs outperform SIFT even on the matching tasks required for SLAM.

It is thus natural to try to extend this approach from single features to whole-image descriptors as commonly implemented through BoW. In fact, this is exactly the investigation carried out by Hou et al. (2015), who started out with a pre-

trained Caffe model (AlexNet on the Places dataset). The compressed representation of the input image is obtained by pulling data out of the CNN at various layers during the *forward pass*.

This is a very simple operation, almost an "afterthought" in the operation of the network, but still allows quite a bit of flexibility. By comparing representations extracted at different layers, they find that the last pooling layer, just before the final fully connected layers, performs the best.

While the precision/recall tradeoff is comparable to hand-crafted features, CNNs are 1-2 order of magnitude faster to compute.

Chen et al. (2013) have also worked along the same lines, by leveraging the representations embedded in an existing CNN.

End-to-end learning for place recognition

Another interesting work is by Arandjelovic et al. (2015), who went with a fully end-to-end learning approach customized for place recognition.

> *Hou et al. (2015) implement loop closure by piggy-backing on an existing network that was trained using a loss-fuction developed for image classification. The obvious question is whether a network architected and trained for the express purpose of image retrieval can do any better.*

The original contribution in this paper is the *NetVLAD* layer, an aggregation layer trained alongside the main network. The NetVLAD layer simulates the VLAD algorithm for the aggregation of feature descriptors described in Jegou et al. (2010), which is an alternative to BoW-style aggregation. Coupling this old and proven idea with end-to-end learning is one of the main contributions of the paper. To train this network, the authors introduce a *weakly-supervised* loss function based on the (rough) GPS locations available in the Google Street View images used for training.

The authors then show how their network + loss function combination performs better than pre-trained models borrowed from object classification tasks. It seems that the weak supervision is enough to steer the network towards focusing on stable and efficient descriptors.

Lessons learned

1. Training on a scene-heavy dataset, such as Places by MIT CSAIL, or the Google Street View dataset, improves place recognition performance quite a bit compared to object-heavy sets.

2. End-to-end training for place recognition handily beats using pre-trained models focused on classification. It seems that medium-depth representation are quite important here.

3. A feed-forward pass through a CNN is much faster than using hand-engineered features, even on a CPU. Interesting speedups for existing SLAM systems seems within easy reach, with the caveat that they need both feature-level and image-level descriptors at the same time.

Other useful code & references

- an implementation of the VLAD algorithm for hand-engineered features (github)
- the implementation of BoW (github) used in ORB-SLAM
- *scene recognition* is the classification of image *context*. Zhou et al. (2014) introduce a new dataset that improves performance on such task, compared to conventional object-focused image collections. Training loop-closure networks on this dataset yields better performance that

Bibliography

Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. *Arxiv*, 2015. URL: http://arxiv.org/abs/1511.07247, arXiv:1511.07247, doi:10.1109/CVPR.2016.572. ↩

Zetao Chen, Obadiah Lam, Adam Jacobson, and Michael Milford. Convolutional Neural Network-based Place Recognition. *2014 Australasian Conference on Robotics and Automation (ACRA 2014)*, pages 8, 2013. arXiv:1411.1509. ↩

Philipp Fischer, Alexey Dosovitskiy, and Thomas Brox. Descriptor Matching with Convolutional Neural Networks: a Comparison to SIFT. *arXiv*, pages 1–10, 2014. URL: http://arxiv.org/abs/1405.5769, arXiv:1405.5769. ↩

Yi Hou, Hong Zhang, and Shilin Zhou. Convolutional Neural Network-Based Image Representation for Visual Loop Closure Detection. pages 2238–2245, 2015. arXiv:1504.05241. ↩ 1 2

Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3304–3311, 2010. doi:10.1109/CVPR.2010.5540039. ↩

Ethan Rublee and Gary Bradski. ORB - an efficient alternative to SIFT or SURF. 2011. URL: http://www.willowgarage.com/sites/default/files/orb_final.pdf, doi:10.1109/ICCV.2011.6126544. ↩

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning Deep Features for Scene Recognition using Places Database. *Advances in Neural Information Processing Systems 27*, pages 487–495, 2014. URL: http://papers.nips.cc/paper/5349-learning-deep-features-for-scene-recognition-using-places-database.pdf. ↩

Posted on: Thu 15 September 2016

Category: 2016 – Tags: deep learning, robotics