AFRL-AFOSR-VA-TR-2018-0425

Active Sensing Representations for Navigation and Actual Scene Analysis

Stefano Soatto
UNIVERSITY OF CALIFORNIA LOS ANGELES

11/28/2018
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTB1
Arlington, Virginia 22203
Air Force Materiel Command

FINAL REPORT
FA9550-15-1-0229
Active Sensing Representations for Navigation and Visual Scene Analysis

Stefano Soatto
Principal Investigator
University of California, Los Angeles

Abstract:

This project, kicked off in September of 2015, aimed at developing analytical and computational tools to infer optimal representations for decision and control actions based on visual data. Specifically, corresponding to (classes of) tasks, different representations can be designed. For localization tasks, EO imaging and inertial sensors can be used to develop a representation that is minimal-sufficient (an attributed point cloud) and invariant to changes of illumination and partial occlusion. The result is a posterior estimate of the sensor trajectory in SE(3) given all measurements up to the current time, marginalized with respect to all nuisance variability. Semantic understanding of the scene requires more sophisticated representations than point cloud. The project has developed a hierarchy of representations, from low-level (photometric descriptors, point-cloud reconstruction) to mid-level (textures, multi-view descriptor) to high level (optimal representation learned from data directly in an end-to-end fashion). The arc of this project coincided with the ascent of deep learning as a tool to infer representations, and the project has leveraged on empirical progress to develop a coherent theory of deep learning, in synergy with other projects, that connects basic principles of statistical decision theory to the current practice of deep learning, including deriving the first known bound on invariance and minimality of the representation learned by a deep network. Additional breakthroughs, not anticipated in the initial proposal, include analysis of the optimization of deep networks and is described in the body of this report.

Description of results:

A representation is a function of the data that is useful (i.e., informative) for a task. Clearly, that depends on the task. A representation that is suitable for localization of the camera frame in 3D is not suitable for recognizing objects within, and vice-versa. The task informs the classes of nuisance variability affecting the data. While this is exquisitely task-dependent, there are common invariances shared among many tasks. For instance, for most task (but not all, for instance lossless video compression), applying monotonic transformations to the range space of the data should have no effect; for recognition tasks, small diffeomorphic deformations of the domain space of the data also has no effect, but it does for localization tasks, for the sake of example.

In the first aim of this project, we developed low-level geometric and photometric representations. The former is designed to be invariant to photometric variability, the latter to geometric variability. We then have designed more articulate descriptor for mid-level statistics, that are beyond local neighborhood of the image.

However, during the course of this project the power of using convolutional deep neural network architectures as a function class of approximants, with parameters estimated through simple stochastic gradient descent (SGD), has become evident. We have therefore explored whether there are connections between basic principles of statistical decision and information theory guiding the design of

representation, and the result of training deep networks. Surprisingly, there are profound relationships, which we have explored and unraveled, and have led to the Emergence Theory of Deep Learning, which provides the first and only known bound between invariance properties of the learn representation (which is a function of test data, or future data that has not yet seen) and computable functions of past data, or training set. This work has opened the door to a more principled approach to the analysis and designed of representations using deep neural networks, which is the focus of our investigations moving forward beyond the project just completed.

Aim 1: Low-level Photometric Representations

In order to maintain a model of scale and of the relative pose between the sensor platform and objects, a persistent reference frame is needed. To this end, the minimal sufficient invariant representation is an attributed point cloud. Because of occlusions, the number of visible points changes, causing singular perturbations, and because of mismatches, the inference entailed in reconstructing this point cloud is highly non-convex. Nevertheless, in [1], we have provided the first provably convergent algorithms for reconstructing such attributed point cloud, despite adversarial perturbations. This solved a long-standing open problem, albeit for the case when camera orientation is known. In practice, this assumption can be made realistic with the use of inertial and rotational pose estimates, which are well observable and easy to infer along a separate channel (no double-integration, no gravity, no problems with lack of visual parallax).

Designing low-level descriptors is now a mature field, and has been subsumed by representation learning, so we moved on from this line of work.

Aim 2: Mid-Level Photometric Representations: Textures, MV-DPM

Mid-level representations capture non-local statistics such as stationarity. While these should support a wide variety of tasks downstream, the most stringent task is reconstruction of data from the compressed representation. The most challenging case is that of textures, whereby one is interested in capturing ensemble properties of statistically homogeneous regions such as foliage of clouds, but not necessarily the individual pixels, for a variety of tasks. In [2,3] we have developed representations for mid-level descriptions of textures and shown their effectiveness in reconstruction based on perceptual metrics.

Aim 3: Robust Filtering and Fusion

Local and non-local spatial representations must be integrated over time in order to arrive at a persistent model of the scene. The first step is to establish a persistent reference frame, for which we use visual and inertial measurements. In [4,5], we have developed a state-of-the-art visual-inertial fusion system operating in real time and beating commercial applications such as Google Tango on commodity hand-held hardware. Furthermore, we have given a characterization of the observability of pose, an in-depth analytical work that was awarded Best Conference Paper (best overall) at ICRA 2015, the largest robotics conference.

Aim 4: High-level Description: Co-visible surfaces and objects.

Moving from a spatially local sparse representation to a global persistent one also requires spatial integration, and in particular going from points to surfaces. This is an inverse problem the most difficult part of which is dealing with changes of topology due to occlusions and reconstruction errors. In [6,7],

we have developed solid modeling tools for topology estimation and surface fitting in sparse point cloud, leveraging occlusions.
In [8], we have presented the first 3D object detector to operate in real time with knowledge of scale and occlusion. While there is a lot of talk about ``image recognition'', most of the (thousands) of papers published every years refer to recognition of *images* of objects. Since there are no objects in images, just pixels, none of these methods are cognizant of scale and occlusion, so they cannot distinguish, for instance, a real car from a toy car, and as soon as an object disappears from views, it ceases to exist (there is no memory nor temporal coherence).

We have shown empirically that, contrary to popular belief, these deep networks are not particularly effective at marginalizing scale and visibility. These should, therefore, be represented and modeled explicitly. The system we have developed, analyzed and implemented, exploits both visual AND inertial sensors, so it can discriminate between objects of different size (e.g. a real car from a toy car). It also has memory, so when an object becomes occluded, it remains in the memory of the system, which can predict when it will return into view, and perform long-term data association. This work re-interprets modern deep convolutional networks as likelihood functions, an interpretation put forth in an ICLR paper in 2016 [9], that functions as an implicit measurement equation in a nonlinear Bayesian filter, implemented with particles.

Unanticipated Breakthroughs

During the course of this project we have been able to make significant advances in a theoretical framework for deep learning. In [10], we have shown that one can define optimal representations starting from first principles (minimality, sufficiency, invariance, independence), arriving at a variational optimization problem that is, at face value, intractable. However, in [11], we have shown that one can, rather than compute and optimize the regularization functionals of this variational functional, directly control it, by injecting noise during the learning process. This is highly unintuitive but can be proven, and empirically verified, to yield representations that are invariant and with maximally independent components.

The complete theory relates these first principles to the practice of Deep Learning. In [10], we show that the information in the weights, which is the regularizer just describes, bounds from above the minimality of the representation, and therefore its invariance, as proven in the paper. This program has been extended from decision to control tasks in [12].

In order for this program to be enacted, first-order optimization needs to be developed to converge to so-called "flat minima", or minima with low information content. In [13] chaudhariAl17 we have shown that a simple modification of SGD, which has deeply-rooted connections to statistical physics, converges to such flat minima with high probability and represents now the state-of-the-art.

This has shown to significantly speed up convergence compared to current methods. The paper has also inspired a number of spin-offs and copycats that are currently being explored by practitioners. In addition, connections of this method to PDEs has triggered much interest. Although the paper has just been submitted, its technical report version has gathered attention and we expect to continue developing it in the year to come. We have also summarized these contributions in the context of the current literature in [14].

References

[1] Thomas Goldstein, Paul Hand, Choongbum Lee, Vladislav Voroninski, Stefano Soatto (2016). "ShapeFit and ShapeKick for Robust, Scalable Structure from Motion," Proceeding of European Conference on Computer Vision (ECCV), pp. 289-304.

[2] Georgios Georgiadis, Stefano Soatto (2016). "A Mid-level Representation of Visual Structures for Video Compression," Workshop on Applications of Computer Vision (WACV), pp. 1-8.

[3] Georgios Georgiadis, Alessandro Chiuso, Stefano Soatto (2015). "Texture Representations for Image and Video Texture Synthesis," Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2058-2066.

[4] Konstantine Tsotsos, Alessandro Chiuso, Stefano Soatto (2015). "Robust Filtering for Visual Inertial Sensor Fusion," Proceeding of 2015 IEEE Conference on Robotics and Automation (ICRA), pp. 5203-5210.

[5] Joshua Hernandez, Konstantine Tsotsos, Stefano Soatto (2015). "Observability, Identifiability and Sensitivity of Vision-assisted Inertial Navigation," Proceeding of 2015 IEEE Conference on Robotics and Automation (ICRA), pp. 2319-2325.

[6] Virginia Estellers, Stefano Soatto (2016). "Detecting Occlusions as an Inverse Problem," Journal of Mathematical Imaging and Vision, Vol. 54, No. 2, pp. 181-198.

[7] Virginia Estellers, Stefano Soatto, Xavier Bresson (2015). "Adaptive Regularization with the Structure Tensor," IEEE Transactions on Image Processing, Vol. 24, No. 6. Pp. 1777-1790.

[8] Jingming Dong, Xiaohan Fei, Stefano Soatto (2017). "Visual Inertial Semantic Scene Representation for 3D Object Detection," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-27, 2017, Honolulu, HI.

[9] Nikolaos Karianakis, Jingming Dong, Stefano Soatto (2016). "An Empirical Evaluation of Current Convolutional Architecture's Ability to Manage Nuisance Location and Scale Variability," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 26- July 1, 2016, Las Vegas, NV.

[10] Alessandro Achille, Stefano Soatto (2017). "Emergence of Invariance and Disentangling in Deep Representations," Journal of Machine Learning Research (JMLR), in press.

[11] Alessandro Achille, Stefano Soatto (2018). "Information Dropout: Learning Optimal Representations Through Noisy Computation," IEEE Transactions on Pattern Analysis and Machine Intelligence," DOI: 10.1109/TPAMI.2017.2784440

[12] Alessandro Achille, Stefano Soatto (2018). "A Separation Principle for Control in the Age of Deep Learning," Annual Review of Control, Robotics, and Autonomous Systems, Vol. 1, pp. 287-307.

[13] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, Riccardo Zecchina (2016). "Entropy-SGD: Biasing Gradient Descent into Wide Valleys," Proceeding of the International Conference on Learning Representations (ICLR), 2016.

[14] Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto, Guillaume Carlier (2017). "Deep Relaxation: partial differential equations for optimizing deep neural networks," ArXiv 1704.04932.