REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
14-02-2023	Final Report	11-Sep-2015 - 10-Sep-2019

4. TITLE AND SUBTITLE
Final Report: ARO 5.2.3 Actionable Information-Based Inference for Control and Interaction with Dynamic Scenes

5a. CONTRACT NUMBER
W911NF-15-1-0564

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of California - Los Angeles
Office of Contract and Grant Administration
11000 Kinross Avenue, Suite 211
Los Angeles, CA 90095 -1406

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
66731-MI.18

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Stefano Soatto
UU	UU	UU	UU		19b. TELEPHONE NUMBER 310-825-4840

RPPR Final Report
as of 24-Feb-2023

Agency Code: 21XD

Proposal Number: 66731MI **Agreement Number: W911NF-15-1-0564**
INVESTIGATOR(S):

Name: Stefano Soatto Ph.D.
Email: soatto@cs.ucla.edu
Phone Number: 3108254840
Principal: Y

Organization: **University of California - Los Angeles**
Address: Office of Contract and Grant Administration, Los Angeles, CA 900951406
Country: USA

DUNS Number: 092530369 EIN: 956006143
Report Date: 10-Dec-2019 Date Received: 14-Feb-2023
Final Report for Period Beginning 11-Sep-2015 and Ending 10-Sep-2019
Title: ARO 5.2.3 Actionable Information-Based Inference for Control and Interaction with Dynamic Scenes
Begin Performance Period: 11-Sep-2015 **End Performance Period:** 10-Sep-2019
Report Term: 0-Other

Submitted By: Sim-Lin Lau Email: simlin@cs.ucla.edu
Phone: (310) 825-2303

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 5 **STEM Participants:** 5

Major Goals: Objective
Develop information-based methods for the design and learning of representations for visual data for the purpose of interaction with physical scenes. Interaction involves decision and control tasks based on visual data, whereby the underlying scene is complex, uncertain, and dynamic.

Approach
A “representation” is any computable function of the data that is “useful” for a task, or a class of tasks, where “useful” is defined in information-theoretic terms (information measures). An optimal representation is one that maximizes information content (or equivalently entails no information loss, due to the Data Processing Inequality), and at the same time is invariant to all sources of nuisance variability, which can be expressed as a minimality relative to a complexity criterion. Nuisance factors are generative factors in the data that are uninformative of the task. An optimal representation, or a minimal sufficient invariant, would be the ideal memory or “state” of a system for a specific task.
This project has pioneered the formal characterization of optimal representations as Sufficient Invariants (ICLR 2016), and the analysis of the resulting properties, including computability (JMLR 2018). While the project started independently of the wave of Deep Learning, and many of the ideas were rooted in foundational work conducted under the aegis of ARO during the prior decade, Deep Neural Networks proved to be an essential tool for implementing optimal representations at scale. Up to that point, minimality was being sought in the form of minimal dimensionality (ICCV 2009), and the kind of invariances being sought were maximal invariants. For example, the maximal invariants for images under all possible viewpoints and illumination is a zero-measure set (an Attributed Reeb Tree, a topological construction of the image), whereas Deep Learning map images onto activations maps of far higher dimension, where what is minimal is not the dimension, but rather the information in the representation. Dimension is only a rough upper-bound of information.
However, even defining – let alone computing – information in a trained Deep Neural Network (DNN), is non-trivial, for real DNNs used today are deterministic functions, not probabilistic models, so traditional notions of (Shannon) information yield nonsensical results, for instance the fact that the trained weights contain zero information about the dataset (if quantized), or infinite differential mutual information (if considered as continuous). During this project, we have pioneered the notion of Information in the Weights, shown that it is well-defined even for deterministic networks, regardless of whether they are trained with Stochastic Gradient Descent, in the end there is one set of weights and the activation maps are computed deterministically at inference time. We have shown how the Information in the Weights, which can be computed during training, bounds the information in the activations of test data, which of course cannot be computed, but that defines the generalization gap of interest for any inductive inference task.

Scientific barriers that had to be overcome were not just technical, but foundational. As we progressed with the analysis and modeling tasks, the practice of DNNs was maturing, so we were able to quantify information content in trained networks with hundreds of millions of parameters, a scale that looks quant today as large language models are heading to the tens of trillion parameters, but that nonetheless during the progress of this project were out of reach of any technique to quantify and bound information in DNNs.

Significance
This project starts from first principles, defines desirable properties that a representation should have, and then instantiates it for different tasks. The significance of the analysis is that it informs the design and evaluation of a wide variety of methods to detect, localize, classify objects and scene in remote sensing data, which is of obvious significance to applications of importance to the Army: ATR, Persistent ISR as well as mapping, localization, precision-location, formation control.

At the commencement of this project, even defining what an “optimal representation” is, and characterizing its properties, was an open problem. Now a formal characterization is in place. More importantly, since the inception of the project, new computational tools have become available that allow not only to define, but also to compute and, most important, to optimize such information measures.

Accomplishments: Accomplishments
This project pioneered the study of representations from an information perspective. The “Emergence Theory of Deep Learning” defines Information in the Weights, illustrates how it can be computed depending on what prior knowledge was available: In the absence of any information (improper prior), the concept reduced to the classical Fisher Information. In the presence of the most informative prior (the marginal of trained networks over all possible datasets, clearly not a computable one), the Information in the Weights reduces to Shannon’s Mutual Information between the weights and the training data. And in between the user has agency on how to “choose the units” through the choice of prior.
This is an information-theoretic framework that is compatible with PAC-Bayes theory and Kolmogorov complexity, and the first theory to provide bounds on generalization for deep networks that are validated empirically on modern networks. It also explains recently discovered empirical phenomena such as the role of flat minima and overfitting in deep networks.
This development started in 2011, but only with the advent of DNNs the concepts started becoming computable in practice. The starting point was the 2016 paper that formalized the notion of an “optimal representation” based on classical notions of statistical decision and information theory such as sufficiency, minimality, invariance (ICLR 2016). This showed how to define and formalize, but now how to compute and optimize information measures needed to infer optimal representations. This, however, was merely a list of desiderata since, at the time, we did not have means of computing such optimal representations.
The first breakthrough came with Information Dropout, an information-theoretic regularization method that was shown equivalent to injecting multiplicative noise in the activation functions of a deep neural network. The result is surprising: Noisy computation helps (PAMI 2018). This revealed connections between optimal representations, the Information Bottleneck, Variational Autoencoding, and Variational Lower Bound in Bayesian inference. The central tenets of the theory are described in (JMLR 2017), which introduces a novel Information Bottleneck, dual to that described by Tishby and Co-Workers. The relation between these duals provides a bound on generalization based on a key information measure (the information the weights of a deep network contain about the given dataset for the task for which the network is trained).
Moreover, surprising connections between representation (the properties of the function learned, regardless of how it is learned) and optimization (how the loss function is minimized) were discovered: First, it was shown that the stochastic optimization in stochastic gradient descent (SGD) is the solution to a Hamilton-Jacobi partial differential equation (PDE), and that the corresponding distribution evolves according to a linear Fokker-Planck equation, solves an optimal control problem, and can be modified and analyzed using PDE theory (ICML 2017), leading to novel algorithms that beat the state-of-the-art (ICLR 2017), and, finally, shown that, even if explicit regularization is not accounted for, SGD implicitly regularizes the solution in a way that is precisely the same as the information regularization developed in the Emergence Theory (ICLR 2018).
Finally, while an optimal representation is the best one can do for a task with the given data, there is no guarantee that this is any good (it is at most as good as the data!), so the question remains of what can go wrong, based on the data and training we have. In particular, the question of what data can fool a particular representation is known as the problem of “adversarial samples” or “adversarial training”. The first analysis of adversarial perturbations, which are small changes applied to any image in a dataset that will, with high probability, yield the wrong decision, has been developed in this project. They are universal because the perturbation is identical for every image in the

dataset. It is also imperceptible, so to the naked eye, two images that look identical result in completely different decisions (ICLR 2017).

Further areas for exploration
This project addressed the question of how to define, compute, and analyze optimal representation for a given task, or set of tasks. However, more recently Deep Networks have shown a remarkable emergent ability of transferring knowledge across tasks. This means that a model trained for a task can then easily be repurposed (typically with fine-tuning, but also with a zero-shot modality) for another. This has led some to believe that there is a "representation to rule them all" (a.k.a. a Foundation Model). This is patently simplistic, for in the absence of any knowledge about the task, the best one can do is to just store the data, or any lossless encoding of it (for, trivially, the task may turn out to be to replicate the data). However, the goal of extending representations to span more tasks, characterizing the topology and geometry of the space of learning tasks, and understanding transferability are wide open problems. In a parallel project under separate funding, we have explored how to define a topology in the space of tasks, and defined distances between learning tasks that correlate with ease of transfer learning, but that work is in its infancy, and complementary to work that is just driven by scaling: Train bigger models and see what happens. Both are needed, and will likely be subjects of intense activitiy in years to come.

Training Opportunities: Nothing to Report

Results Dissemination: Results is presented in conferences such as International Conference on Learning Representations (ICLR), International Conference on Machine Learning (ICML), and Journal of Machine Learning Research (JMLR).

Honors and Awards: Stefano Soatto
- Fellow of the ACM.
- Keynote Speaker, ITA Information Theory and Applications, 2018
- Keynote Speaker, LA CTO Forum, 2018
- Keynote Speaker, Tokyo Deep Learning Workshop, 2018
- Keynote Speaker, Workshop on Learning and Adaptation for Sensorimotor Control, Lund, Sweden, 2018
- Distinguished Seminar Speaker, NYU Colloquium on Artificial Intelligence
- Keynote Speaker, CVPR Diff-CVML, 2018
- Keynote Speaker, CVRSUAD, 2017
- Keynote, NIPS workshop on Disentanglement, December 2017.

Protocol Activity Status:

Technology Transfer: No patent, invention and licenses or interaction with DoD laboratories

PARTICIPANTS:

Participant Type: Staff Scientist (doctoral level)
Participant: Sim-Lin Lau
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

ARTICLES:

RPPR Final Report
as of 24-Feb-2023

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: IEEE Transactions on Pattern Analysis and Machine Intelligence
Publication Identifier Type: DOI Publication Identifier: 10.1109/TPAMI.2017.2784440
Volume: 40 Issue: 12 First Page #: 2897
Date Submitted: 2/14/23 12:00AM Date Published:
Publication Location:
Article Title: Information dropout: Learning optimal representations through noisy computatio
Authors: Alessandro Achille, Stefano Soatto
Keywords: deep neural network; inferences
Abstract: The cross-entropy loss commonly used in deep learning is closely related to the defining properties of optimal representations, but does not enforce some of the key properties. We show that this can be solved by adding a regularization term, which is in turn related to injecting multiplicative noise in the activations of a Deep Neural Network, a special case of which is the common practice of dropout. We show that our regularized loss function can be efficiently minimized using Information Dropout, a generalization of dropout rooted in information theoretic principles that automatically adapts to the data and can better exploit architectures of limited capacity. When the task is the reconstruction of the input, we show that our loss function yields a Variational Autoencoder as a special case, thus providing a link between representation learning, information theory and variational inference. Finally, we prove that we can promote the creation of disentangled representations simp
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Journal of Machine Learning Research (JMLR)
Publication Identifier Type: Publication Identifier:
Volume: 19 Issue: 1 First Page #: 1947
Date Submitted: 2/14/23 12:00AM Date Published: 9/13/18 6:41AM
Publication Location:
Article Title: Emergence of invariance and disentangling in deep representations
Authors: Alessadro Achille, Stefano Soatto
Keywords: Representation learning; PAC-Bayes; information bottleneck; flat minima;
Abstract: Using established principles from Statistics and Information Theory, we show that invari- ance to nuisance factors in a deep neural network is equivalent to information minimality of the learned representation, and that stacking layers and injecting noise during training nat- urally bias the network towards learning invariant representations. We then decompose the cross-entropy loss used during training and highlight the presence of an inherent overfitting term. We propose regularizing the loss by bounding such a term in two equivalent ways: One with a Kullbach-Leibler term, which relates to a PAC-Bayes perspective; the other using the information in the weights as a measure of complexity of a learned model, yield- ing a novel Information Bottleneck for the weights. Finally, we show that invariance and independence of the components of the representation learned by the network are bounded above and below by the information in the weights, and therefore are implicitly optimized dur
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 24-Feb-2023

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Annual Reviews of Robotics and Autonomous Systems
Publication Identifier Type: Publication Identifier:
Volume: 1 Issue: First Page #: 287
Date Submitted: 2/14/23 12:00AM Date Published: 5/29/18 6:43AM
Publication Location:
Article Title: A separation principle for control in the age of deep learning
Authors: Alessandro Achille, Stefano Soatto
Keywords: deep learning, autonomy, neural network, prediction, representation
Abstract: We review the problem of defining and inferring a state for a control sys- tem based on complex, high-dimensional, highly uncertain measurement streams, such as videos. Such a state, or representation, should contain all and only the information needed for control and discount nuisance vari- ability in the data. It should also have finite complexity, ideally modulated depending on available resources. This representation is what we want to store in memory in lieu of the data, as it separates the control task from the measurement process. For the trivial case with no dynamics, a representation can be inferred by minimizing the information bottleneck Lagrangian in a function class realized by deep neural networks. The resulting representa- tion has much higher dimension than the data (already in the millions) but is smaller in the sense of information content, retaining only what is needed for the task. This process also yields representations that are invariant to nuisance factors and h
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Research in the Mathematical Sciences
Publication Identifier Type: DOI Publication Identifier: 10.1007/s40687-018-0148-y
Volume: 5 Issue: First Page #: 1
Date Submitted: 2/14/23 12:00AM Date Published: 6/29/18 2:44AM
Publication Location:
Article Title: Deep relaxation: partial differential equations for optimizing deep neural networks
Authors: Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto, Guillaume Carlier
Keywords: deep learning, partial differential equations, stochastic gradient descent, neural networks
Abstract: Entropy-SGD is a first-order optimization method which has been used successfully to train deep neural networks. This algorithm, which was motivated by statistical physics, is now interpreted as gradient descent on a modified loss function. The modified, or relaxed, loss function is the solution of a viscous Hamilton–Jacobi partial differential equation (PDE). Experimental results on modern, high-dimensional neural networks demonstrate that the algorithm converges faster than the benchmark stochastic gradient descent (SGD). Well-established PDE regularity results allow us to analyze the geometry of the relaxed energy landscape, confirming empirical evidence. Stochastic homogenization theory allows us to better understand the convergence of the algorithm. A stochastic control interpretation is used to prove that a modified algorithm converges faster than SGD in expectation.
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 24-Feb-2023

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Journal of Statistical Mechanics: Theory and Experiment
Publication Identifier Type: DOI Publication Identifier: 10.1088/1742-5468/ab39d9
Volume: 2019 Issue: 12 First Page #: 124018
Date Submitted: 2/14/23 12:00AM Date Published: 12/1/19 8:00AM
Publication Location:
Article Title: Entropy-SGD: biasing gradient descent into wide valleys
Authors: Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Je
Keywords: machine learning
Abstract: This paper proposes a new optimization algorithm called Entropy- SGD for training deep neural networks that is motivated by the local geometry of the energy landscape. Local extrema with low generalization error have a large proportion of almost-zero eigenvalues in the Hessian with very few positive or negative eigenvalues. We leverage upon this observation to construct a local- entropy-based objective function that favors well-generalizable solutions lying in large flat regions of the energy landscape, while avoiding poorly-generalizable solutions located in the sharp valleys. Conceptually, our algorithm resembles two nested loops of SGD where we use Langevin dynamics in the inner loop to compute the gradient of the local entropy before each update of the weights. We show that the new objective has a smoother energy landscape and show improved generalization over SGD using uniform stability, under certain assumptions. Our experiments on convolutional and recurrent networks demonstrate
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 3-Accepted
Conference Name: Conference on Computer Vision and Pattern Recognition (CVPR) 2016
Date Received: 20-Jun-2016 Conference Date: 26-Jun-2016 Date Published: 26-Jun-1962
Conference Location: Las Vegas, NV
Paper Title: An Empirical Evaluation of Current Convolutional Architecture's Ability to Mange Nuisance Location and Scale Variability
Authors: Karianakis, N, Dong, J, Soatto, S
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 3-Accepted
Conference Name: International Conference on Learning Representations 2016 (ICLR)
Date Received: 20-Jun-2016 Conference Date: 03-May-2016 Date Published: 05-May-2016
Conference Location: San Juan, Puerto Rico
Paper Title: Visual Representations: Defining properties and deep approximation
Authors: Soatto, S., Chiuso, A
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 5-Submitted
Conference Name: 2016 IEEE International Conference on Robotics and Automation (ICRA)
Date Received: 20-Jun-2016 Conference Date: 16-May-2016 Date Published:
Conference Location: Stockholm, Sweden
Paper Title: Intent-aware long-term prediction of pedestrian motion
Authors: Karasev, V, Ayvaci, A, Heisele, B, Soatto, S
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 24-Feb-2023

Publication Type: Conference Paper or Presentation **Publication Status:** 3-Accepted
Conference Name: 2015 IEEE International Conference on Computer Vision (ICCV)
Date Received: 20-Jun-2016 Conference Date: 07-Dec-2015 Date Published:
Conference Location: Santiago, Chile
Paper Title: Self-Occlusions and Disocclusions in Causal Video Object Segmentation
Authors: Yang, Y, Sundaramoorthi, G, Soatto, S
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 0-Other
Conference Name: nternational Conference on Machine Learning, Workshop on Principal Approaches to Deep Learning (ICML PADL 2017)
Date Received: 02-Oct-2017 Conference Date: 10-Aug-2017 Date Published:
Conference Location: Sydney, Australia
Paper Title: On the emergence of invariance and disentangling in deep representations
Authors: Alessandro, Achille and Stefano, Soatto
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 2-Awaiting Publicat
Conference Name: Computer Vision and Pattern Recognition (CVPR) conference
Date Received: 02-Oct-2017 Conference Date: 21-Jul-2017 Date Published:
Conference Location: Honolulu, Hawaii
Paper Title: Visual-Inertial-Semantic Scene Representation for 3D Object Detection
Authors: Jingming, Dong, Xiaohan, Fei, Stefano, soatto
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 2-Awaiting Publicat
Conference Name: Computer Vision and Pattern Recognition (CVPR) conference
Date Received: 02-Oct-2017 Conference Date: 21-Jul-2017 Date Published:
Conference Location: Honolulu, Hawaii
Paper Title: Zero Shot Learning via Multi-Scale Manifold Regularization
Authors: Shay Deutsch, Soheil Kolouri, Kyungnam Kim, Yuri Owechko, Stefano Soatto,
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: European Conference on Computer Vision (ECCV)
Date Received: 14-Feb-2023 Conference Date: 08-Sep-2018 Date Published:
Conference Location: Munich, Germany
Paper Title: Conditional Prior Networks for Optical Flow
Authors: Yanchao Yang, Stefano Soatto
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 0-Other
Conference Name: Conference on Computer Vision and Pattern Recognition (CVPR) and
Date Received: 14-Feb-2023 Conference Date: 21-Jul-2017 Date Published:
Conference Location: Honolulu, Hawaii
Paper Title: Classification regions of deep neural networks
Authors: Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, Stefano Soatto
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 24-Feb-2023

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Conference on Computer Vision and Pattern Recognition (CVPR)
Date Received: 14-Feb-2023 Conference Date: 22-Jul-2017 Date Published:
Conference Location: Honolulu, Hawaii
Paper Title: Analysis of universal adversarial perturbations
Authors: Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, Stefano Soatto
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 3-Accepted
Conference Name: Conference on Computer Vision and Pattern Recognition (CVPR)
Date Received: 14-Feb-2023 Conference Date: 29-Jun-2018 Date Published:
Conference Location: Salt Lake City, Utah
Paper Title: OATM: Occlusion Aware Template Matching by Consensus Set Maximization
Authors: Simon Korman, Mark Milam, Stefano Soatto
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: 2018 Information Theory and Applications Workshop (ITA)
Date Received: 14-Feb-2023 Conference Date: 11-Feb-2018 Date Published:
Conference Location: San Diego, CA
Paper Title: Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks
Authors: Pratik Chaudhari, Stefano Soatto
Acknowledged Federal Support: **Y**

Partners

,

I certify that the information in the report is complete and accurate:
Signature: Sim-Lin Lau
Signature Date: 2/14/23 3:00AM

Final Report
Award # W911NF-15-1-0564
Actionable Information-Based Inference for Control and Interaction with Dynamic Scenes

Stefano Soatto
University of California, Los Angeles
Email: soatto@ucla.edu Tel. (310)825-4840

Objective

Develop information-based methods for the design and learning of representations for visual data for the purpose of interaction with physical scenes. Interaction involves decision and control tasks based on visual data, whereby the underlying scene is complex, uncertain, and dynamic.

Approach: A "representation" is any computable function of the data that is "useful" for a task, or a class of tasks, where "useful" is defined in information-theoretic terms (information measures). An *optimal* representation is one that maximizes information content (or equivalently entails no information loss, due to the Data Processing Inequality), and at the same time is invariant to all sources of nuisance variability, which can be expressed as a *minimality* relative to a complexity criterion. Nuisance factors are generative factors in the data that are uninformative of the task. An optimal representation, or a *minimal sufficient invariant*, would be the ideal memory or "state" of a system for a specific task.

This project has pioneered the formal characterization of optimal representations as *Sufficient Invariants* (Soatto and Chiuso, 2016), and the analysis of the resulting properties, including *computability* (Achille and Soatto, 2018). While the project started independently of the wave of Deep Learning, and many of the ideas were rooted in foundational work conducted under the aegis of ARO during the prior decade, Deep Neural Networks proved to be an essential tool for implementing optimal representations at scale. Up to that point, *minimality* was being sought in the form of minimal *dimensionality* (ICCV 2009), and the kind of invariances being sought were *maximal invariants*. For example, the maximal invariants for images under all possible viewpoints and illumination is a zero-measure set (an Attributed Reeb Tree, a topological construction of the image), whereas Deep Learning map images onto activations maps of far higher dimension, where *what is minimal is not the dimension, but rather the information in the representation.* Dimension is only a rough upper-bound of information.

However, even defining – let alone computing – information in a trained Deep Neural Network (DNN), is non-trivial, for real DNNs used today are deterministic functions, not probabilistic models, so traditional notions of (Shannon) information yield nonsensical results, for instance the fact that the trained weights contain zero information about the dataset (if quantized), or infinite differential mutual information (if considered as continuous). During this project, we have pioneered the notion of *Information in the Weights*, shown that it is well-defined even for deterministic networks, regardless of whether they are trained with Stochastic Gradient Descent, in the end there is one set of weights and the activation maps are computed deterministically at

inference time. We have shown how the Information in the Weights, which can be computed during training, bounds the information in the activations of test data, which of course cannot be computed, but that defines the generalization gap of interest for any inductive inference task.

Scientific barriers that had to be overcome were not just technical, but foundational. As we progressed with the analysis and modeling tasks, the practice of DNNs was maturing, so we were able to quantify information content in trained networks with hundreds of millions of parameters, a scale that looks quant today as large language models are heading to the tens of trillion parameters, but that nonetheless during the progress of this project were out of reach of any technique to quantify and bound information in DNNs.

Significance: This project starts from first principles, defines desirable properties that a representation should have, and then instantiates it for different tasks. The significance of the analysis is that it informs the design and evaluation of a wide variety of methods to detect, localize, classify objects and scene in remote sensing data, which is of obvious significance to applications of importance to the Army: ATR, Persistent ISR as well as mapping, localization, precision-location, formation control.

At the commencement of this project, even defining what an "optimal representation" is, and characterizing its properties, was an open problem. Now a formal characterization is in place. More importantly, since the inception of the project, new computational tools have become available that allow not only to define, but also to compute and, most important, to optimize such information measures.

Accomplishments:

This project pioneered the study of representations from an information perspective. The "Emergence Theory of Deep Learning" defines Information in the Weights, illustrates how it can be computed depending on what prior knowledge was available: In the absence of any information (improper prior), the concept reduced to the classical Fisher Information. In the presence of the most informative prior (the marginal of trained networks over all possible datasets, clearly not a computable one), the Information in the Weights reduces to Shannon's Mutual Information between the weights and the training data. And in between the user has agency on how to "choose the units" through the choice of prior.

This is an information-theoretic framework that is compatible with PAC-Bayes theory and Kolmogorov complexity, and the first theory to provide bounds on generalization for deep networks that are validated empirically on modern networks. It also explains recently discovered empirical phenomena such as the role of flat minima and overfitting in deep networks.

This development started in 2011, but only with the advent of DNNs the concepts started becoming computable in practice. The starting point was the 2016 paper that formalized the notion of an "optimal representation" based on classical notions of statistical decision and information theory such as sufficiency, minimality, invariance (Soatto and Chiuso, 2016). This showed how to define and formalize, but now how to compute and optimize information measures needed to infer optimal representations. This, however, was merely a list of desiderata since, at the time, we did not have means of computing such optimal representations.

The first breakthrough came with Information Dropout, an information-theoretic regularization method that was shown equivalent to injecting multiplicative noise in the activation functions of a deep neural network. The result is surprising: Noisy computation helps (Achille and Soatto, 2018a). This revealed connections between optimal representations, the Information Bottleneck, Variational Autoencoding, and Variational Lower Bound in Bayesian inference. The central tenets of the theory are described in (Achille and Soatto, 2018b), which introduces a novel Information Bottleneck, dual to that described by Tishby and Co-Workers. The relation between these duals provides a bound on generalization based on a key information measure (the information the weights of a deep network contain about the given dataset for the task for which the network is trained).

Moreover, surprising connections between representation (the properties of the function learned, regardless of how it is learned) and optimization (how the loss function is minimized) were discovered: First, it was shown that the stochastic optimization in stochastic gradient descent (SGD) is the solution to a Hamilton-Jacobi partial differential equation (PDE), and that the corresponding distribution evolves according to a linear Fokker-Planck equation, solves an optimal control problem, and can be modified and analyzed using PDE theory (Chaudhari et al, 2018), leading to novel algorithms that beat the state-of-the-art (Chaudhari et al, 2019), and, finally, shown that, even if explicit regularization is not accounted for, SGD implicitly regularizes the solution in a way that is precisely the same as the information regularization developed in the Emergence Theory (Chaudhari and Soatto, 2018).

Finally, while an optimal representation is the best one can do for a task with the given data, there is no guarantee that this is any good (it is at most as good as the data!), so the question remains of what can go wrong, based on the data and training we have. In particular, the question of what data can fool a particular representation is known as the problem of "adversarial samples" or "adversarial training". The first analysis of adversarial perturbations, which are small changes applied to *any image* in a dataset that will, with high probability, yield the wrong decision, has been developed in this project. They are universal because the perturbation is identical for every image in the dataset. It is also imperceptible, so to the naked eye, two images that look identical result in completely different decisions (Chaudhari et al, 2019).

Further areas for exploration: This project addressed the question of how to define, compute, and analyze *optimal representation for a given task, or set of tasks*. However, more recently Deep Networks have shown a remarkable *emergent ability of transferring knowledge across tasks*. This means that a model trained for a task can then easily be repurposed (typically with fine-tuning, but also with a zero-shot modality) for another. This has led some to believe that there is a "representation to rule them all" (a.k.a. a Foundation Model). This is patently simplistic, for in the absence of any knowledge about the task, the best one can do is to just store the data, or any lossless encoding of it (for, trivially, the task may turn out to be to replicate the data). However, the goal of extending representations to span more tasks, characterizing the topology and geometry of the space of learning tasks, and understanding transferability are wide open problems. In a parallel project under separate funding, we have explored how to define a topology in the space of tasks, and defined distances between learning tasks that correlate with ease of transfer learning, but that work is in its infancy, and complementary to work that is just

driven by scaling: Train bigger models and see what happens. Both are needed, and will likely be subjects of intense activity in years to come.

Awards, honors (Stefano Soatto)

- Fellow of the ACM.
- Keynote Speaker, ITA Information Theory and Applications, 2018
- Keynote Speaker, LA CTO Forum, 2018
- Keynote Speaker, Tokyo Deep Learning Workshop, 2018
- Keynote Speaker, Workshop on Learning and Adaptation for Sensorimotor Control, Lund, Sweden, 2018
- Distinguished Seminar Speaker, NYU Colloquium on Artificial Intelligence
- Keynote Speaker, CVPR Diff-CVML, 2018
- Keynote Speaker, CVRSUAD, 2017
- Keynote, NIPS workshop on Disentanglement, December 2017.

Publications

S. Soatto and A. Chiuso (2016). "Modeling visual representations: Defining properties and deep approximations", *International Conference on Learning Representations (LCLR),* San Juan, Puerto Rico, poster.

A. Achille and S. Soatto (2018a). "Information dropout: Learning optimal representations through noisy computation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.14, No. 12, pp. 2897-2905. DOI: 10.119/TPAMI.2017.2784440.

A. Achille and S. Soatto (2018b). "Emergence of invariance and disentanglement in deep representations", *The Journal of Machine Learning Research (JMLR)*, Vol. 19, No. 1, pp. 1947-1980.

P. Chaudhari, A. Oberman, S. Osher, S. Soatto, G. Carlier (2018). "Deep relaxation: Partial differential equations for optimizing deep neural networks", *Research in the Mathematical Sciences*, Vol. 5, pp. 1-30. DOI: 10.1007/s40687-018-0148-y

P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun and R. Zecchina (2019). "Entropy-SGD: Biasing gradient descent into wide valleys", *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2019, No. 12, p. 124018. DOI: 10.1088/1742-5468/ab39d9.

P. Chaudhari and S. Soatto (2018). "Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks", *2018 Information Theory and Applications Workshop (ITA)*, ICLR, San Diego, CA. DOI: 10.1109/ITA.2018.8503224

S-M Moosavi-Dezfooli, A. Fawzi, P. Frossard and S. Soatto (2017). “Analysis of universal adversarial perturbations”, *International Conference on Learning Representations (LCLR).* arXiv:1705.09554v2