

Yue Song · Thomas Anderson Keller ·
Nicu Sebe · Max Welling

Structured Representation Learning

From Homomorphisms and Disentanglement
to Equivariance and Topography

Synthesis Lectures on Computer Vision

Series Editors

Gerard Medioni, University of Southern California, Los Angeles, USA

Sven Dickinson, Department of Computer Science, University of Toronto, Toronto,
Canada

This series publishes on topics pertaining to computer vision and pattern recognition. The scope follows the purview of premier computer science conferences, and includes the science of scene reconstruction, event detection, video tracking, object recognition, 3D pose estimation, learning, indexing, motion estimation, and image restoration. As a scientific discipline, computer vision is concerned with the theory behind artificial systems that extract information from images. The image data can take many forms, such as video sequences, views from multiple cameras, or multi-dimensional data from a medical scanner. As a technological discipline, computer vision seeks to apply its theories and models for the construction of computer vision systems, such as those in self-driving cars/navigation systems, medical image analysis, and industrial robots.

Yue Song · Thomas Anderson Keller ·
Nicu Sebe · Max Welling

Structured Representation Learning

From Homomorphisms and
Disentanglement to Equivariance and
Topography

Yue Song
Department of Computing and Mathematical
Sciences
California Institute of Technology
Milan, Italy

Nicu Sebe
Department of Computer Science
and Information Engineering
University of Trento
Trento, Italy

Thomas Anderson Keller
Kempner Institute
Harvard University
Cambridge, MA, USA

Max Welling
Department of Machine Learning
University of Amsterdam
Amsterdam, The Netherlands

ISSN 2153-1056

ISSN 2153-1064 (electronic)

Synthesis Lectures on Computer Vision

ISBN 978-3-031-88110-7

ISBN 978-3-031-88111-4 (eBook)

<https://doi.org/10.1007/978-3-031-88111-4>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer
Nature Switzerland AG 2026

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface I

The art of machine learning (ML) is to optimally combine inductive bias (a.k.a. priors) with data. With more data, we need to include less prior information and we can let the data speak. This is the modern “scaling” paradigm of LLMs and Foundation models with trillions of parameters, trained on hundreds of thousands of GPUs on the entirety of the internet. For these models, the architecture of choice are usually Transformers, precisely because they scale well.

Are Transformers the final answer? That seems unlikely. In this book, we ask ourselves if there are architectures based on better priors about the world but that also scale to internet-level models. These models will not only be able to learn from fewer data but also exhibit improved scaling laws, ideally with a steeper slope.

What are some interesting priors to build into deep architectures? In this book, we are inspired by both neuroscience and physics. Neuroscience has been ML’s companion right from the start. Early architectures such as Rosenblatt’s Perceptron were already inspired by biological neurons. It’s only more recently that the two fields have gone their own separate ways. But given the huge gap in energy efficiency between artificial and biological neural networks, it may make sense to look at neuroscience again for inspiration.

In this book, we explore the possibility to use oscillators and traveling waves as a new computing paradigm, rather than static representations. Waves have the potential to collect and combine information from long distances, both on space and in time.

Another interesting prior is that in the world around us at the scale that we understand it (objects), things usually don’t change very fast. This is of course different than the nanosecond fluctuation at the level of individual atoms. As deep models build coarse-grained representations in their deeper levels, it seems reasonable to enforce this slowness at the level of abstract (deep) representations.

Since our models often model things in our physical world, we can also contemplate whether the symmetries of that physical world should be represented in our representations. There is now a rich literature on symmetries, such as translational and rotational

symmetries that are exactly enforced through (irreducible or regular) equivariant group representations. However, often we do not know the symmetries present in the data, or some regularities might not be described by groups, or we may not simply know the representations of certain groups. In all these cases we need to generalize the concept of equivariance, on which these hard-coded symmetries are based.

In this book, we consider homomorphisms between latent representations and the world as an appropriate signal to learn such approximate symmetries. That is to say, the dynamics of the latent representations should mirror the corresponding dynamics of the world. What dynamics, or set of transformation of our latent code, do we entertain when we try to learn homomorphisms between the world and our deep representations? Here we are inspired again by neuroscience and physics: we have modeled these representations as collections of interacting oscillators, or in the continuum limit, PDEs, that support wave-like solutions. In some sense, we can think of these representations as a fluid in which waves can develop to perform computations.

And this brings me to my final point. Due to availability of multi-electrode sensors, waves are now commonly detected in the brain, and neuroscience researchers are starting to ask what its computational benefits might be. Can waves transport and combine information in new ways that we have not yet discovered? This is an intriguing possibility about which I have no doubt we will hear a lot more over the course of the next decade.

I wish you an interesting journey as you travel through the chapters of this book and become inspired to think of new ways to build inductive biases into the next generation of ML models.

Amsterdam, The Netherlands
February 2025

Max Welling

Preface II

The field of machine learning stands at a critical juncture. While recent advances in deep learning have delivered remarkable breakthroughs across domains such as vision, language, and robotics, these successes often come at the cost of massive data requirements, computational inefficiency, a lack of interpretability, and poor generalization to novel scenarios. As machine learning systems are increasingly deployed in real-world applications, these challenges highlight the need for models that go beyond brute force learning and instead can learn more like humans—adaptively, efficiently, and intuitively.

This book, *Structured Representation Learning: From Homomorphisms and Disentanglement to Equivariance and Topography*, offers a timely exploration of how structured approaches can reshape the design and performance of machine learning systems. By embedding principles such as symmetry, topography, and compositionality directly into model architectures, structured representation learning provides a pathway to models that are more robust, efficient, and capable of generalization.

At its core, structured representation learning seeks to address fundamental questions: How can machine learning systems capture the inherent relationships within data, such as symmetries and invariances? How can models decompose complex phenomena into simpler, interpretable components? And how can we align computational representations with the physical and biological principles that govern the real world? This book explores these questions through key concepts such as:

- **Equivariance and Symmetry:** Learning approximately equivariant representations that respect the transformations of the data beyond group theory.
- **Disentanglement:** Designing latent representations that isolate meaningful factors of variation, serving as approximate learned equivariance.
- **Topographic Representations:** Drawing inspiration from biological systems to organize information spatially and temporally in ways that mimic biological neural networks.

- **Physical Priors:** Baking physical principles into machine learning systems to represent the physical relations in the real world.

Structured representation learning represents a paradigm shift. By incorporating the above beneficial inductive biases directly into learning systems, this approach opens the door to machine learning systems that are not only more efficient but also more interpretable and aligned with the complexities of the real world.

This book is written for researchers, practitioners, and students eager to explore the intersection of machine learning, computational neuroscience, and natural sciences. It provides a high-level perspective on the field's foundational ideas while delving into specific techniques and applications that demonstrate the power of structured representation learning. As the demands on machine learning systems continue to grow, structured representations offer a promising direction toward building models that can reason, adapt, and learn with greater data efficiency and generalization abilities. We invite readers to engage with the ideas in this book, explore the rich potential of structured representations, and join in shaping the future of machine intelligence.

Pasadena, CA, USA
Cambridge, MA, USA
November 2024

Yue Song
Thomas Anderson Keller

Contents

Part I Structured Representation Learning: History and State of the Art

1 Introduction	3
1.1 Motivation: The Gap of Sample Efficiency and Generalization	3
1.2 The Promise of Structured Representations	5
1.3 The Way Forward: Naturally Inspired Learned Homomorphisms	7
References	7
2 Background	9
2.1 Structured Representation Learning	9
2.2 Disentangled Representation Learning	10
2.3 Equivariant Neural Networks	11
2.4 Approximately Equivariant and Disentangled Representations	12
2.4.1 Prior Work: Capsule Networks, Homeomorphic VAEs	12
2.4.2 Biological and Physical Inductive Biases for Learning Equivariant Representations	12
References	14

Part II Naturally Inspired Topographically Structured Representation Learning

3 Topographic Variational Autoencoders	21
3.1 Introduction	21
3.2 The Generative Model	22
3.2.1 Topographic Generative Models	23
3.2.2 The Product of Student's-t Model	23
3.2.3 Constructing the Product of Student's-t Distribution	24
3.2.4 Introducing Topography	24
3.2.5 Capsules as Disjoint Topologies	25

3.2.6	Temporal Coherence and Learned Equivariance	25
3.3	Topographic VAE	27
3.4	Experiments	27
3.4.1	Evaluation Methods	28
3.4.2	Topographic VAE Without Temporal Coherence	29
3.4.3	Learning Equivariant Capsules	29
3.5	Future Work and Limitations	32
3.6	Conclusion	32
3.7	Experiment Details	33
3.7.1	Optimizer Parameters	33
3.7.2	Initialization	33
3.7.3	Model Architectures	33
3.7.4	Choices of \mathbf{W} , \mathbf{W}_δ , and L	34
3.7.5	Hyperparameter Selection	35
3.7.6	MNIST Transformations	35
3.7.7	dSprites Transformations	35
3.7.8	Capsule Correlation Metric (CapCorr)	36
3.7.9	Definition of Roll for Capsules	37
3.8	Extended Results	37
3.8.1	Extended Tables 3.1 and 3.2	38
3.8.2	Impact of \mathbf{W}_δ	39
3.8.3	Generalization to Combined Transformations at Test Time	39
3.9	Proposed Model Extensions	41
3.9.1	Extensions to Roll and CapCorr	41
3.9.2	Non-cyclic Capsules	43
3.9.3	Multi-dimensional Temporally Coherent Capsules	44
3.9.4	Causal Temporal Coherence	44
3.10	Capsule Traversals	45
	References	47
4	Neural Wave Machines	49
4.1	Introduction	49
4.1.1	Traveling Waves in Neuroscience	51
4.1.2	Computational Models of Traveling Waves	52
4.2	Neural Wave Machines	52
4.2.1	Coupled Oscillatory Recurrent Neural Networks	52
4.2.2	Local Connectivity	53
4.3	Experiments	54
4.3.1	Methods	54
4.3.2	Datasets	55
4.3.3	Measuring Spatiotemporal Structure	56
4.3.4	Topographic Orientation Selectivity	56

4.3.5	General Topographic Organization	57
4.3.6	Instantaneous Phase and Velocity	57
4.3.7	Controlled Generation with Induced Traveling Waves	58
4.3.8	Computational Implications of Structure	60
4.3.9	An Inductive Bias for Simple Physical Dynamics	60
4.3.10	Efficiency	60
4.4	Discussion	62
4.4.1	Related Work	62
4.4.2	Limitations	62
4.4.3	Conclusion	63
4.5	Experiment Details	63
4.5.1	Sequence Classification	64
4.5.2	Rotating MNIST and Sine Waves	64
4.5.3	Hamiltonian Dynamics Suite	66
4.5.4	Hardware Details	67
4.6	Analytical Treatment of Neural Wave Machines	67
4.6.1	Bounds on Hidden State Energy	68
4.6.2	Sensitivity to Inputs	68
4.6.3	Bounds on Hidden State Gradient	68
4.6.4	Assumptions	69
4.7	Extended Results	69
4.7.1	Impact of Δt Parameter	69
4.7.2	Additional Efficient Sequence Modeling Results	70
4.7.3	Additional Hamiltonian Dynamics Results	70
4.7.4	On Modeling Chaotic Dynamics	71
4.7.5	On the Formation of Orientation Maps	72
4.7.6	Full Rotating MNIST Topographic Organization	73
4.7.7	Visualizing Traveling Waves on MNIST	74
	References	77

Part III Learned Homomorphisms and Disentangled Representations

5	Latent Traversal as Potential Flows	83
5.1	Motivations	83
5.1.1	Traveling Waves in Neuroscience	83
5.1.2	Fluid Mechanics as Optimal Transport	83
5.2	Learned Potential Flows for Traversal	85
5.2.1	Learning the Potential PDEs	85
5.2.2	Integration with Generative Models	86
5.3	Experiments	88
5.3.1	Evaluation Methods	88
5.3.2	Results with Pre-trained Networks	89

5.3.3	Results with Pre-trained VAEs	90
5.3.4	Results with Networks Trained from Scratch	91
5.4	Discussions	93
5.4.1	Flow Path Properties	93
5.4.2	Limitations and Future Extensions	95
	References	96
6	Flow Factorized Representation Learning	99
6.1	Factorized Representation Learning	99
6.1.1	Learned Equivariance	99
6.1.2	Disentanglement	99
6.2	The Generative Model	100
6.2.1	Flow Factorized Sequence Distributions	100
6.2.2	Prior Time Evolution	101
6.3	Flow Factorized Variational Autoencoders	102
6.3.1	Inference with Observed Transformation Categories	102
6.3.2	Inference with Unknown Transformation Categories	103
6.3.3	Posterior Time Evolution	104
6.3.4	Optimal Transport for Posterior Flow	104
6.4	Experiments	106
6.4.1	Evaluation Methods	106
6.4.2	Learning Equivariant Latent Flows	107
6.4.3	Results on Complex Real-World Datasets	110
6.5	Discussions	110
6.5.1	Extrapolation to Switching/Superposing Transformation	110
6.5.2	Equivariance Generalization to New Data	113
	References	114
7	Unsupervised Factorized Representation Learning Based on Sparse Transformation Analysis	117
7.1	Motivations	117
7.1.1	Sparsity in Natural Videos	117
7.1.2	Helmholtz Decomposition	118
7.2	The Generative Model	118
7.2.1	Factorized Sequence Distributions	118
7.2.2	Spike and Slab Priors	119
7.2.3	Prior Time Evolution	121
7.3	Helmholtz Flow Variational Autoencoders	121
7.3.1	Helmholtz Decomposed Latent Flows	121
7.3.2	Evidence Lower Bound and Inference	122
7.3.3	Posterior Time Evolution	123
7.4	Experiments	124

7.4.1	Evaluation Methods	124
7.4.2	Learning Composable Equivariant Latent Flows	125
7.4.3	Analysis of Real-World Videos	129
7.5	Discussions	134
7.5.1	Switchability and Composability	134
7.5.2	Handling Periodic Transformations	134
7.5.3	Learning Separate Controls	135
	References	136
8	Conclusion	139
8.1	Naturally Inspired Structured Representations	139
8.2	Learned Homomorphisms and Disentangled Representations	140
	Reference	140

Acronyms

AI	Artificial Intelligence
AR	AutoRegressive Model
ELBO	Evidence Lower BOund
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
HJE	Hamilton-Jacobi Equation
ICA	Independent Component Analysis
LLM	Large Language Models
LN	Layer Normalization
LSTM	Long Short-Term Memory
MLP	Multi-layer Perceptron
ODE	Ordinary Differential Equations
OoD	Out-of-Distribution
OT	Optimal Transport
PDE	Partial Differential Equation
PINN	Physics-Informed Neural Network
RNN	Recurrent Neural Network
SE(N)	Special Euclidean Group
SFA	Slow Feature Analysis
SO(N)	Special Orthogonal Group
SOM	Self-Organizing Map
VAE	Variational Auto-Encoder

List of Figures

Fig. 1.1	Samples from text-to-image generation program DALLE-2. The prompts are: (top) “a teddy bear on the moon”, (middle) “blue cube on top of a red cube”, & (bottom) “a banana holding a monkey”. We see that while the model can generate incredibly realistic images of relatively novel objects, it often fails to understand basic relations between objects	4
Fig. 3.1	Overview of the Topographic VAE with shifting temporal coherence. The combined color/rotation transformation in input space τ_g becomes encoded as a Roll within the capsule dimension. The model is thus able decode unseen sequence elements by encoding a partial sequence and Rolling activations within the capsules. We see this resembles a commutative diagram	22
Fig. 3.2	An example of a neighborhood structure which induces disjoint topologies (A.K.A. capsules). Lines between variables T_i indicate that sharing of U_i , and thus correlation	26
Fig. 3.3	Maximum activating images for a topographic VAE trained with a 2D torus topography on MNIST	30
Fig. 3.4	Capsule Traversals for TVAE models on dSprites and MNIST. The top rows show the encoded sequences (with greyed-out images held-out), and the bottom rows show the images generated by decoding sequentially Rolled copies of the initial activation \mathbf{t}_0 (indicated by a grey border)	30