Towards structured representations in deep models and policies: symmetries, compactness, and vector homomophisms

Authors' Names Omitted Intentionally[1]

[1]Microsoft Research, New York City, [2]Microsoft Research, Cambridge, UK, [3]Microsoft Research AI4Science, Amsterdam

Deep learning algorithms are now universally accepted as the state-of-the-art in large visual recognition, sequence modeling, and transduction learning tasks. As data scales in size, they have proven useful for learning causal mid-level *latent* representations from large observational spaces (such as RGB or depth images, and point clouds). However, there are timely research questions that have emerged pertaining to fundamental computational and representation issues; these include computational complexity, data permutation, representation invariance, and a causal realization that is minimal in some sense. Equivariant deep learning is an emerging line of research inquiry that studies means of alleviating these issues from a theory of finite simple groups perspective. In this workshop, we aim to shine light on this emerging topic — addressing recent developments in the research community, and pointing out potential opportunities that can engender non-cumbersome representations that are amenable to embedded computing applications, as well as efficient imitation, privileged and reinforcement learning. Emphasizing open challenges in this direction, we shall foster participation from the Microsoft R&I community on actions necessary to quell current drawbacks in deep representation learning broadly.

1. Significance

Learning compact representations from high-dimensional temporally-varying signal and observation spaces is a key challenge in computer vision [1], signal processing [2], and control-from-pixels [3, 1, 4]. Learned representations from neural architectures such as convolution [5], recurrent [6], and multilayer perceptron networks, provide *ad-hoc* prediction efficacy in classification and statistical learning tasks. However, they tend to be overparameterized for reliable control in large observation spaces. As a result, this slakes their performance in long-range prediction and control tasks. Further drawbacks exist: *convolution networks only process local neighborhoods*, either in space or time [7]; and *recurrent networks suffer from exploding and diminishing gradients* over long-separated interactions. As such, their representations do not provide the *irreducible* features useful for *computationally efficient* downstream tasks in the original space from which the features are learned. Lastly, with their characteristic non-parametric outputs, they are not interpretable; that is, in the sense of e.g. an affine-in-control, $u(t)$, nonlinear system, $\dot{x}(t)$, with a readily discernible state space representation parameterization $\dot{x}(t) = f(x(t), u(t)) + g(x(t))u(t)$ where $f(\cdot)$ and $g(\cdot)$ are $\mathcal{C}^n$ nonlinear functions. In this example, the parameters of interest could evolve on a smooth manifold such as a differentiable Lie group, whereupon the functions $f(\cdot)$ and $g(\cdot)$ are respectively deemed Lie algebra representations $\mathfrak{f}$ and $\mathfrak{g}$ of a manifold's Lie group equation.

It is therefore paramount to devise representations from large datasets in pixel-space that are compact, greedy, and preserve group properties whilst providing the stability for prediction without the need for

⋆ Corresponding author. Email: lekanmolu@microsoft.com

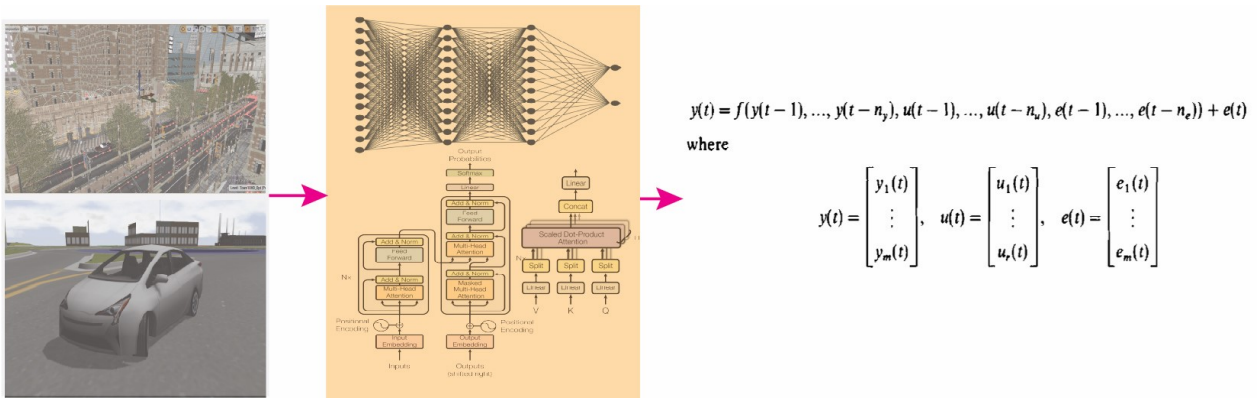

Figure 1: The *minimal* input-state-output realization learning framework: A sequence of observations is passed on to a group-structured neural network that learns a smooth representation of the observation. This is then followed by a nonlinear regression layer that generalizes the learned invariant manifold to a set of nonlinear difference equations.

much post-processing.

1.1. Motivation

Modern machine learning algorithms are informed by neural architectures that accept inputs characterized by vectors, matrices, or tensors. These have had great success in modeling sequential data, signal processing [2], and image-processing problems. Owing to the natural design of these networks' structural layout, however, data augmentation is a prerequisite for efficiently capturing world invariance during training. This data augmentation is a laborious process that increases the size of training data; and this data size invariably necessitates deep and large networks in order to learn effective representations. Even so, learned representations in these situations are characterized by high sample complexity, are non-modular and non-portable for use in many downstream tasks such as spatial AI for embodied devices, robotics, or real-time inference on mobile devices. Recent efforts in the physics-informed machine learning research community is increasingly showing the effectiveness of non-vector inputs at capturing efficient representations despite less training data size, and zero/minimal augmentation. These emerging network architectures exploit sets as inputs by re-expressing network layers as differentiable modules that capture desirable group symmetry properties and representations.

In the last few years, the research on modular neural architectures that can process sets with permutation invariance [8], and equivariance [9], whilst capturing nice structural properties inherent in the data has taken higher priority in effective representation learning across multiple lines of research inquiry [10, 11]. In this workshop, we will explore recent advancements in the applications of finite simple groups, Lie theory, and and sets in general, for dealing with complexity that can exploit **structure** and compositionality with minimal computations. This is toward improving training time, sampling complexity, and reducing the necessity for voluminous data before researchers and practitioners can learn compact representations with deep learning neural architectures.

2. Goals of the Workshop

To overcome the associated downsides that are complementary to learning representations with vector-based inputs, we propose a radical departure from modern representation learning settings that involves projections from pixel space to an arbitrarily-sized latent space. **The goals of this workshop are** to bring

researchers together across geography, and missions towards exploring next-generation in-situ learning of object classes, image texture characteristics, orientation, mechanism configuration, and relationships among different components in an image by bringing to bear the classical concepts of differential geometry, modern MDP homomorphic representation in policy gradients, and Lie group theory to structured deep learning.

This is a natural evolution of the research in latent states discovery from pixels [12], [13] that has been championed in our New York City lab whereby vector-input and encoder projection-based representations suffered from three characteristic drawbacks namely

1. lack of an appropriate dimensioning of the state space;
2. a lack of a natural injective property from the learned representation to the representation's original domain; and
3. the computational complexity that arises from such learned representations given the overparameterization of the minimal realization of the state information.

Under the hypothesis that by using structured filters that capture fundamental natural properties such as symmetry in (special) Euclidean-3 spaces, permutation invariance, periodicity, time evolutions of objects (from the viewpoint of a static observer), **the goals of this workshop will be an exposition for participants on how one can obtain**

- representations that are intuitively meaningful; that is, with an injective homomorphic property to the natural environment in which they are learned;
- "state"-based vector spaces from representation manifolds and accurate dimensions of finite state spaces learned from abstract manifold representations of pixel-based inputs; and
- that these *somewhat parsimonious representations* can be further decomposed into minimal realizations that aid stable predictions, control applications, and planning capabilities in downstream tasks.

This hypothesis and goal statement is depicted in Figure 1.

3. Intended Outcomes

This section of the proposal discusses the content of the proposed workshop and highlights tentative and confirmed speakers for the keynote talks and spotlight presentations. **We conclude the section by highlighting the expectated outcomes** from the organization of this workshop.

3.1. Workshop Format

We propose an excitingly engaging workshop format where audience participation will be engaged via panel debates amongst key stakeholders, hybrid mode meeting option (in-person and online) with contemporary interactive software tools (including Teams and Gather.Town). We are being intentional about the diversity of organizers and speakers list, including seeking representation of women in the external partnerships of Microsoft and underrepresented members of the Microsoft research community.

A technical portion of the program shall be discharged by the organizers highlighting the current state-of-the-art with respect to numerical algorithms, benchmarks, spotlight talks on previous impactful journals. To avoid talk fatigue and facilitate the rich exchange of information and ideas at the workshop, we propose ten-minute tea breaks every fifteen minutes before the top of the hour per session.

For the hybrid mode operation, we shall employ Teams and GatherTown for online registered participants; and we shall tightly integrate the online efforts with those of the in-person attendees. We will coordinate