Encoding Dynamic Invariance for Video Understanding

Julienne LaChance

Background:

In recent years, advances in machine learning have revolutionized the field of static image understanding. The use of specialized convolutional architectures allow deep neural networks to learn image features at multiple scales for tasks such as object detection, segmentation and pose estimation. Despite these successes in 2D, extending such capabilities to video understanding has proved a challenging task. In order to best perform human activity recognition, for instance, we must capture dynamic information across multiple frames, thereby motivating questions such as: What is an activity and how should we represent it? Do activities have well-defined spatial and temporal extent? [1], and so on.

Inspired by the use of such tools as Kalman filters in the engineering world, I add a new question to this list: if we have a model for some known dynamics in a video recognition task, is there a way to encourage the network to capture these dynamics in the learning process? Here, this problem is referred to as that of *encoding dynamic invariance* in the network, in correspondence with the goal of fixing the known dynamical model via structures within the network itself.

As a motivating example: consider the task of tracking individuals in video. For instance, a biologist might be interested in monitoring the behavior of individual animals in a flock or herd, or a traffic engineer might be interested in the typical flow of pedestrians at a subway terminal (see Fig. 1). In practice, this task becomes difficult due to challenges such as occlusion, changes in clothing and lighting, severe deformation, rare or novel poses, and in the case of certain species of animal, the lack of strongly distinguishing visual features. Consider now that the researcher has a well-developed model of the gait dynamics of the video subjects, and wishes to encapsulate this model within a deep neural network in a way that allows the network to distinguish between individuals as a function of their gait patterns. Such a scenario is not unreasonable: already, researchers are able to use gait stability analysis to monitor changes in gait with age and injury, and measure differences in male/female gait [2].

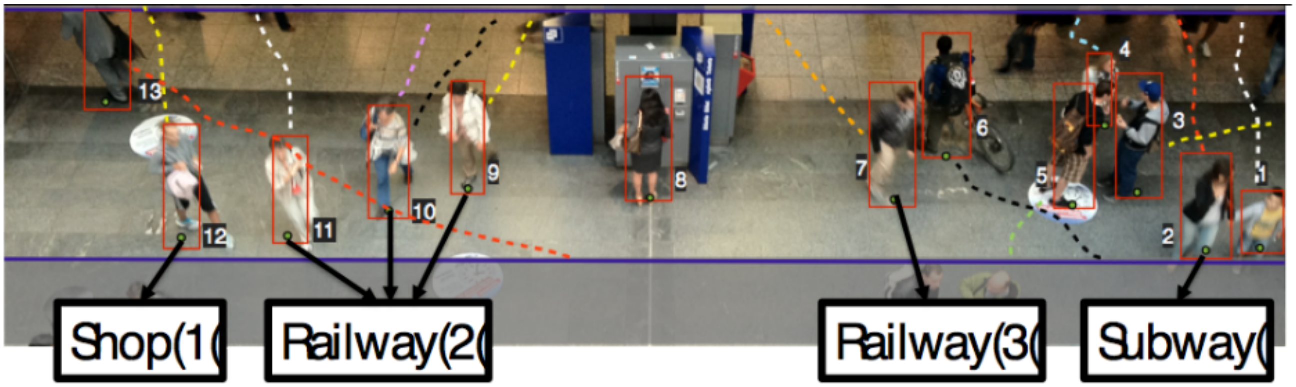


Figure 1. Human activity recognition example: understanding traffic patterns at a subway terminal. Image courtesy of Alahi and Fei-Fei, 2014.

Encoding Dynamic Invariance in Deep Neural Networks:

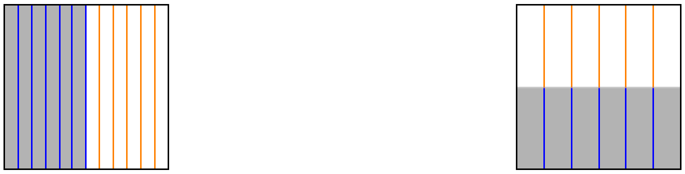
How can dynamical systems researchers exchange tools with the experts in machine learning? The key idea is this: ***any dynamical system may be viewed as a network, and vice versa***. Below, I provide an example which illustrates how the two are related. The consequence of this connection is that we have the potential to use data-driven techniques from dynamical systems- including those which identify invariant sets, coherent sets, and other features directly from observations of the system- in network applications. Some of these connections are intuitive: for example, an invariant set in the dynamical system is equivalent to congestion in the network. Initially, the scope of this work will cover methods of encoding the structure of dynamical systems in deep networks using such network representations of the dynamics. Rather than forcing a fixed model of the dynamics into the network, I hope to capture the topology of the known dynamical systems in phase space (as is done in the Gaio style; see below). In this way, modules with learnable parameters can be constructed which encourage the network to learn the desired system dynamics.

Dynamical system to network representation:

Let us now consider a classic example of a discrete-time dynamical system [3], to demonstrate how we might represent such a system as a network. The system we consider here is the Baker's transformation on the unit square, according to:

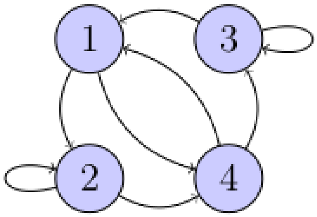
$$T(x_1, x_2) = \begin{cases} (2x_1, \frac{1}{2}x_2), & x_1 \in [0, \frac{1}{2}) \\ (2x_1 - 1, \frac{1}{2}x_2 + \frac{1}{2}), & x_1 \in [\frac{1}{2}, 1), \end{cases}$$

which, as shown here, maps the shaded area at the left into the shaded area at the right, and maps vertical lines to vertical lines.



To represent this as a graph, we may split the unit square into four equal sections, as shown below. Then, an adjacency graph is constructed according to how subsets map to the next respective subset. The graph representation is also shown here, and its corresponding adjacency matrix is binary:

1	3
2	4



For more general dynamical systems, one may refine subsets of the phase space into smaller boxes in regions of interest. Software packages such as Gaio [4]- a toolbox for set-oriented numerics in dynamical systems- use adaptive methods such as this to approximate the Perron-Frobenius operator of a system.

References:

[1] Sigurdsson, Russakovsky and Gupta, "What Actions are Needed for Understanding Human Actions in Videos?", ICCV 2017.

[2] Ihlen et. al., "Phase-dependent changes in local dynamic stability of human gait", Journal of Biomechanics, 2012.

[3] Example and images courtesy of Clarence W. Rowley and Amit Singer, Princeton University.

[4] Dellnitz, M., Froyland, G., Junge, O. "The algorithms behind GAIO - Set oriented numerical methods for dynamical systems", 2001.