Procedia Computer Science

Volume 88, 2016, Pages 211–216

7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2016

The Enacted KOAN – An Agent's Knowledge of Agency

Justin Brody

Goucher College

Abstract

We present Knowledge Of Action Networks, which provide an enactive machine learning model for *knowledge of agency* in artificial intelligence. These networks, which are expected to be part of embodied intelligences existing in dynamic environments, learn to represent their environment while simultaneously learning to represent their own actions and bodies within that environment. Thus self and world are intricately coupled in their basic representations. We will also explore some of the (many) expected contributions of such networks for implementing *minimal self-models*, which are basic models of self-aware agents.

Keywords: Enaction, Sense of agency, Deep learning, Embodied cognition, Self

1 Introduction

In this paper we will present a machine learning model that allows an artificial agent to develop a *knowledge of agency* based on its actions and sensory feedback from its environment. Knowledge of agency refers to an agent's knowledge that it is undergoing a particular action. This is generally called *sense* of agency [3] in the literature, but we prefer to use the term knowledge to avoid any implication that our implementation will have any qualitative feeling associated with it. Crucially, the knowledge that we will describe will be **immediate** in the sense of being unmediated (or at least minimally mediated), **grounded** in that it interacts dynamically with the "real" world (rather than being an abstract symbol), **situated** in that it functions as a representation of agency at a particular time and in a particular environment, and **open** in that it can dynamically alter in new situations (while representing the same thing.) These properties will each be discussed more fully in a later section. I see the current work as grounding many (but not all) uses of the indexical "I"; future work will integrate this into a broader model of a *minimal self*.

A minimal self is perhaps most simply understood as a minimal model of of an agent's subjectivity; see [8] and [12] for two different but overlapping approaches to this idea. In [2] it was argued that an account of subjectivity is a fundamental notion in need of a good computational model and that such a model will enable progress on a number of difficult questions in cognitive science. While there is no consensual set of necessary and sufficient conditions for a minimal self-model, the broader program is predicated on endowing a model

Selection and peer-review under responsibility of the Scientific Programme Committee of BICA 2016

doi:10.1016/j.procs.2016.07.427

with a *knowledge of agency, knowledge of ownership, situatedness, temporal awareness* and *self-reference*. I will argue that all but the last two are achieved (at least partially) by the current work; the final properties will be explored in future work.

The network presented here, termed a KOAN (Knowledge Of Agency Network) develops a representation of its agency by learning to predict the sensory consequences of its actions. In particular, the network will represent its environment in a hierarchical way and while it's learning that representation will try to predict the output of one high-level feature based on the current sensory information and the action it is about to undertake. Since the prediction and representation are learned simultaneously, the result of this will be that the high level feature learned will have to represent a kind of *sensory locus of action*. For example, if an agent is capable of moving its body in several discrete directions, then this learning paradigm will force the representation to be a representation of the body in a way that covaries with the appropriate actions.

The model is enactive (see, e.g., [6]) in the sense that the prediction forms a simultaneous representation of self and action. Indeed, while the self is modeled in the sensory data it is modeled inherently as an *actor* rather than a thing. Although the action is modeled with an explicit representation, it will ultimately be represented as a dynamic interaction between self and environment.

In a review article [5], Limanowski and Blakenburg argued for using predictive coding as a way of obtaining an enactive minimal self-model. While the language we use here is different, the current model should translate well into the vocabulary of predictive coding; thus the machine learning model described here can be thought of as providing an implementation of the kind of enactive self-model they describe.

I will present the machine-learning model in the next section and follow this by a discussion of its properties and some of the potential technologies I see arising from it (many of these ideas are present in [5], but with a specific model at hand we can give somewhat more implementational detail).

The work in this paper is a piece of a larger program of implementing a minimal self-model in collaboration with Donald Perlis and his research group at the University of Maryland. In particular, the importance of the initial concepts discussed in Section 3 was pointed out to me by Perlis and many of the other ideas in this paper either directly arose from or else were inspired by conversations with him.

2 The Model

In this section I will define a KOAN as a general scheme that relies on an environmental encoding module and a self-action prediction module. Given an agent situated in a changing environment, we will denote the totality of its sensory input at time t by S_t (conceived as a vector in some high-dimensional Euclidean space). If an agent has a fixed (and finite) set of possible actions it can take, say $\alpha_1, \ldots, \alpha_n$, then we will denote the action that it does take at time t by a_t. Our basic approach is to learn a high level representation of the environment and force that representation to pick out the effects of the agent's actions.

To that end we will define a KOAN as consisting of two modules: a hierarchical representational network $\mathcal{H}$ (for example a deep convolutional network) and a predictive network $\mathcal{P}$ (e.g. a multi-layer perceptron). The network $\mathcal{H}$ functions to represent the environment at a given time while $\mathcal{P}$ functions to make predictions about a part that representation based on the current environment and the action taken. In order for $\mathcal{P}$ to successfully make such a prediction, the part of the representation in question must covary with the action, forcing the network to