Learning Deep Architectures for AI

Yoshua Bengio
Dept. IRO, Université de Montréal
C.P. 6128, Montreal, Qc, H3C 3J7, Canada
Yoshua.Bengio@umontreal.ca
http://www.iro.umontreal.ca/∼bengioy

Technical Report 1312

Abstract

Theoretical results strongly suggest that in order to learn the kind of complicated functions that can represent high-level abstractions (e.g. in vision, language, and other AI-level tasks), one needs *deep architectures*. Deep architectures are composed of multiple levels of non-linear operations, such as in neural nets with many hidden layers or in complicated propositional formulae re-using many sub-formulae. Searching the parameter space of deep architectures is a difficult optimization task, but learning algorithms such as those for Deep Belief Networks have recently been proposed to tackle this problem with notable success, beating the state-of-the-art in certain areas. This paper discusses the motivations and principles regarding learning algorithms for deep architectures, in particular those exploiting as building blocks unsupervised learning of single-layer models such as Restricted Boltzmann Machines, used to construct deeper models such as Deep Belief Networks.

1 Introduction

Allowing computers to model our world well enough to exhibit what we call intelligence has been the focus of more than half a century of research. To achieve this, it is clear that a large quantity of information about our world should somehow be stored, explicitly or implicitly, in the computer. Because it seems daunting to formalize manually all that information in a form that computers can use to answer questions and generalize to new contexts, many researchers have turned to *learning algorithms* to capture a large fraction of that information. Much progress has been made to understand and improve learning algorithms, but the challenge of artificial intelligence (AI) remains. Do we have algorithms that can understand scenes and describe them in natural language? Not really, except in very limited settings. Do we have algorithms that can infer enough semantic concepts to be able to interact with most humans using these concepts? No. If we consider image understanding, one of the best specified of the AI tasks, we realize that we do not yet have learning algorithms that can discover the many visual and semantic concepts that would seem to be necessary to interpret most images. The situation is similar for other AI tasks.

We assume that the computational machinery necessary to express complex behaviors (which one might label "intelligent") requires highly varying mathematical functions, i.e. mathematical functions that are highly non-linear in terms of raw sensory inputs. Consider for example the task of interpreting an input image such as the one in Figure 1. When humans try to solve a particular task in AI (such as machine vision or natural language processing), they often exploit their intuition about how to decompose the problem into sub-problems and multiple levels of representation. A plausible and common way to extract useful information from a natural image involves transforming the raw pixel representation into gradually more abstract representations, e.g., starting from the presence of edges, the detection of more complex but local shapes, up to the identification of abstract categories associated with sub-objects and objects which are parts

of the image, and putting all these together to capture enough understanding of the scene to answer questions about it. We view the raw input to the learning system as a high dimensional entity, made of many observed variables, which are related by unknown intricate statistical relationships. For example, using knowledge of the 3D geometry of solid object and lighting, we can relate small variations in underlying physical and geometric factors (such as position, orientation, lighting of an object) with changes in pixel intensities for all the pixels in an image. In this case, our knowledge of the physical factors involved allows one to get a picture of the mathematical form of these dependencies, and of the shape of the set of images associated with the same 3D object. If a machine captured the factors that explain the statistical variations in the data, and how they interact to generate the kind of data we observe, we would be able to say that the machine *understands* those aspects of the world covered by these factors of variation. Unfortunately, in general and for most factors of variation underlying natural images, we do not have an analytical understanding of these factors of variation. We do not have enough formalized prior knowledge about the world to explain the observed variety of images, even for such an apparently simple abstraction as **MAN**, illustrated in Figure 1. A high-level abstraction such as **MAN** has the property that it corresponds to a very large set of possible images, which might be very different from each other from the point of view of simple Euclidean distance in the space of pixel intensities. The set of images for which that label could be appropriate forms a highly convoluted region in pixel space that is not even necessarily a connected region. The **MAN** category can be seen as a high-level abstraction with respect to the space of images. What we call abstraction here can be a category (such as the **MAN** category) or a **feature**, a function of sensory data, which can be discrete (e.g., the input sentence is at the past tense) or continuous (e.g., the input video shows an object moving at a particular velocity). Many lower level and intermediate level concepts (which we also call abstractions here) would be useful to construct a **MAN**-detector. Lower level abstractions are more directly tied to particular percepts, whereas higher level ones are what we call "more abstract" because their connection to actual percepts is more remote, and through other, intermediate level abstractions.

We do not know exactly how to build robust **MAN** detectors or even intermediate abstractions that would be appropriate. Furthermore, the number of visual and semantic categories (such as **MAN**) that we would like an "intelligent" machine to capture is large. The focus of deep architecture learning is to automatically discover such abstractions, from the lowest level features to the highest level concepts. Ideally, we would like learning algorithms that enable this discovery with as little human effort as possible, i.e., without having to manually define all necessary abstractions or having to provide a huge set of relevant hand-labeled examples. If these algorithms could tap into the huge resource of text and images on the web, it would certainly help to transfer much of human knowledge into machine-interpretable form.

One of the important points we argue in the first part of this paper is that the functions learned should have a structure composed of multiple levels, analogous to the multiple levels of abstraction that humans naturally envision when they describe an aspect of their world. The arguments rest both on intuition and on theoretical results about the representational limitations of functions defined with an insufficient number of levels. Since most current work in machine learning is based on shallow architectures, these results suggest investigating learning algorithms for deep architectures, which is the subject of the second part of this paper.

In much of machine vision systems, learning algorithms have been limited to specific parts of such a processing chain. The rest of of design remains labor-intensive, which might limit the scale of such systems. On the other hand, a hallmark of what we would consider intelligent includes a large enough vocabulary of concepts. Recognizing **MAN** is not enough. We need algorithms that can tackle a very large set of such tasks and concepts. It seems daunting to manually define that many tasks, and learning becomes essential in this context. It would seem foolish not to exploit the underlying commonalities between these these tasks and between the concepts they require. This has been the focus of research on *multi-task learning* (Caruana, 1993; Baxter, 1995; Intrator & Edelman, 1996; Baxter, 1997). Architectures with multiple levels naturally provide such sharing and re-use of components: the low-level visual features (like edge detectors) and intermediate-level visual features (like object parts) that are useful to detect **MAN** are also useful for a large group of other visual tasks. In addition, learning about a large set of interrelated concepts might provide a key to the kind of broad generalizations that humans appear able to do, which we would not expect from

separately trained object detectors, with one detector per visual category. If each high-level category is itself represented through a particular configuration of abstract features, generalization to unseen categories could follow naturally from new configurations of these features. Even though only some configurations of these features would be present in the training examples, if they represent different aspects of the data, new examples could meaningfully be represented by new configurations of these features. This idea underlies the concept of *distributed representation* that is at the core of many of the learning algorithms described in this paper, and discussed in Section 4.

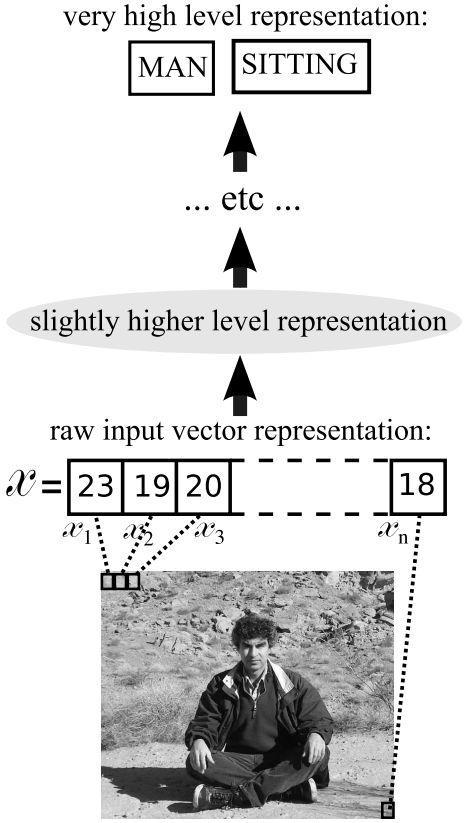

Figure 1: We would like the raw input image to be transformed into gradually higher levels of representation, representing more and more abstract functions of the raw input, e.g., edges, local shapes, object parts, etc. In practice, we do not know in advance what the "right" representation should be for all these levels of abstractions, although linguistic concepts might help us imagine what the higher levels might implicitly represent.

This paper has two main parts which can be read almost independently. In the first part, Sections 2, 3 and 4 use mathematical arguments to motivate deep architectures, in which each level is associated with a distributed representation of the input. The second part (in the remaining sections) covers current learning algorithms for deep architectures, with a focus on Deep Belief Networks, and their component layer, the Restricted Boltzmann Machine.

The next two sections of this paper review mathematical results that suggest limitations of many existing learning algorithms. Two aspects of these limitations are considered: insufficient *depth of architectures*, and *locality of estimators*. To understand the notion of **depth of architecture**, one must introduce the notion of a **set of computational elements**. An example of such a set is the set of computations performed by an