Compression-Driven Progress in Science

Leo Pape
Utrecht University
Utrecht, The Netherlands
l.pape@geo.uu.nl

Abstract

The construction of an *artificial scientist*, a machine that discovers and describes the general rules governing a variety of complex environments, can be considered an important challenge for artificial general intelligence. Recently, a computational framework for scientific investigation has been postulated in the theory of compression-driven progress. Here, I propose an implementation of an artificial scientist based on the compression principle, and explore the possibilities and challenges for its application in scientific research.

Introduction

Human beings reckon scientific understanding of the world among the most powerful of their abilities (e.g. Rorty, 1991), and it is therefore not surprising that researchers try to simulate this capability with computer programs and robots (e.g. King et al., 2009; Schmidt and Lipson, 2009). Such machines, called *artificial scientists*, not only enhance our ability to carry out scientific research, but building them also guides our understanding how human scientists come to comprehend the world. Creating an artificial scientist that is not restricted to a specific domain, but performs scientific research *in general* can be considered a great challenge for artificial general intelligence.

To construct an artificial scientist, we need to have some idea of *what* it is that human scientists do and *how* they do this. Since the societal, fundamental or personal goals of science are not contained in its domain, I here rather define the *activity* of scientists as the development of theories that explain the past and predict the future, and are consistent with other theories. These theories result from systematic reasoning about observations, whether obtained accidentally or intentionally, for example in a controlled experiment.

A theory that not only explains how scientific progress is achieved by human beings, but also specifies how scientific investigation can be carried out with computer algorithms, is the theory of compression-driven progress (Schmidhuber, 2009). This theory considers both human and artificial scientists as computationally limited observers that try to represent observations in an efficient manner. Finding efficient representations entails identifying regularities that allow the observer to compress the original observations and predict future observations. Discovered regularities then serve as an explanation for the observed phenomena. Compression *progress* is achieved when an observer discovers previously unknown regularities that provide increased compression of observations. The theory of compression-driven progress further postulates that scientists direct their attention to interesting data, that is, data that is neither impossible to compress (i.e. truly random) nor easily compressible with existing methods, but is expected to hold previously unknown regularities that allow for further compression.

Based on this theory, it is possible to implement an artificial scientist that can operate in a variety of scientific disciplines. In this paper I explore the possibilities and challenges for the construction of a compression-driven artificial scientist.

Compression-Driven Artificial Scientists

A compression-driven artificial scientist is a machine that aims to predict future and unobserved observations by identifying the regularities underlying its sensory input. It consists of the following components: (1) A sensory input module that collects observations, such as a camera, microphone or software interface. (2) An adaptive compressor that discovers regularities in the observational data. A compressor that is particularly suitable for this task is the deep autoencoder of Hinton and Salakhutdinov (2006), which learns to convert high-dimensional input data to short codes. Of course it is possible to use another, possibly even more general algorithm, but the Hinton and Salakhutdinov autoencoder has the advantage that it can *reconstruct* and thus *predict* data from its coded representations. (3) A reinforcement learning algorithm that learns to select actions (e.g. manipulate the world, perform experiments, direct attention, move) that take the artificial scientist to interesting data. *Interestingness* is defined as the improvement of the adaptive compressor on parts of the input data, and is determined from the number of bits needed to reconstruct the original input from its coded representation. (4) Optionally, a physical implementation, such as a robot. The use of existing datasets,

however, allows for the implementation of the artificial scientist as a software program, which can significantly reduce the costs and complexity of its construction.

Representation

The compression-driven artificial scientist is not immediately useful to its human colleagues, because the regularities it discovers are not represented in understandable form (i.e. in a connectionist architecture). A related problem is that the artificial scientist has no a-priori notion of objects[1] in its raw sensory inputs (e.g. a stream of bits or numbers), while its theories should preferably be about such objects, not about bits or numbers. These two problems reflect the more general challenge of constructing symbolic representations from subsymbolic data (see e.g. Smolensky, 1988). Here I explain how both artificial and human scientists construct mental *objects* from sensory inputs using the compression principle, and how this process is the basis for communicating discovered regularities in symbolic form.

Using the basic operations of its reasoning apparatus, the artificial scientist builds methods that compress those parts of its sensory input signal that have certain structure. Note that compression does not merely apply to static objects in space, but also extends to structural relations in time. Different types of structure require different compression methods, allowing the artificial scientist to distinguish individual entities or phenomena by their *method of compression*. When compression methods are organized in a hierarchical fashion, the artificial scientist can construct more abstract concepts and find increasingly general relations between objects on different abstraction levels.

Human scientists discovered many parts of the world that are compressible and predictable to some extent, while other parts seem to resist compression and prediction. Interestingly, the inability to describe and predict certain parts of the world is mostly not because the fundamental forces of nature are unknown to science, but because the deterministic laws of nature produce chaos in some parts and order in other parts of the world (where chaos and order are equivalent to incompressible and compressible observations, respectively). That is, the most fundamental relations express only the most general aspects of the world, not all specific details relevant to our lives. Human scientists therefore try to find intermediate levels on which the world exhibits regularity, give those parts names and relate them in a systematic way to already identified entities. As a result, different *levels of organization* materialize into individual objects[1] of scientific thought.

Discovered structure in parts of the world can only be communicated in a meaningful sense through a shared language. While mathematics and logic are rather popular languages in science, the relations they express have no intrinsic meaning, but need to be related to concepts that are recognized by all communicating parties (e.g. Schmidt and Lipson (2009) used symbolic regression on variables whose human interpretation was established beforehand, not discovered independently by their algorithms). Artificial scientists therefore need to learn how to map their internal representations of discovered objects and structure onto the entities (e.g. symbols) of a shared language. Such a shared language can, in principle, be learned among different instances of artificial scientists in an *un*supervised fashion. However, this artificial language is probably not easily accessible to human scientists. Instead, an artificial scientist should learn a language that is easily understandable for human scientists. For this, the artificial scientist needs to learn from labeled data, either by augmenting the reinforcement learning algorithm with an external reward based on label prediction, or by a function (e.g. an additional neural network) that learns to map internal representations onto labels in a supervised fashion.

[1] objects in the most general sense, such as material objects like molecules and robots, but also more abstract objects like a rainbow, a supercluster (of galaxies) or musical notes

Conclusion

In this paper I explored the possibilities and challenges for the construction of a compression-driven artificial scientist. While the theory of compression-driven progress provides the basic mechanism for scientific investigation, an ongoing challenge is the human interpretation of theories constructed by artificial scientists. In the future I aim to implement the proposed architecture and demonstrate its capability to discover known and novel forms of structure in scientific data.

References

Hinton, G. E., and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507.

King, R. D.; Rowland, J.; Oliver, S. G.; Young, M.; Aubrey, W.; Byrne, E.; Liakata, M.; Markham, M.; Pir, P.; Soldatova, L. N.; Sparkes, A.; Whelan, K. E.; and Clare, A. 2009. The automation of science. *Science* 324:85–89.

Rorty, R. 1991. *Objectivity, relativism, and truth: philosophical papers*, volume 1. Cambridge, UK: Cambridge University Press.

Schmidhuber, J. 2009. *Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes.* Lecture Notes in Computer Science. Berlin / Heidelberg: Springer. 48–76.

Schmidt, M., and Lipson, H. 2009. Distilling free-form natural laws from experimental data. *Science* 324:81–85.

Smolensky, P. 1988. A proper treatment of connectionism. *Behavioural and Brain Sciences* 11:1–74.