Knowledge extraction from Deep Belief Networks

Introduction

The aim of this project is to implement a system that helps to understand the running of a deep belief network which was conceived in [1]. This system is based on knowledge extraction and this knowledge is represented by logical rules. This implementation will then have to be tested on a specific dataset and the results will be analysed.

So the products that will be generated are a knowledge extraction library and an analysis of its application on the Visual Genome dataset.

The research question for this project is:

How can I extract knowledge from DBN using logical rules and how can I apply it to the Visual Genome Dataset?

The principal beneficiaries of this work:

* The Research Centre for Machine Learning of City, University of London for which the members published most of the papers dealing with knowledge extraction for deep learning and neural networks
* Artificial Intelligence industry which needs a better understanding of the behavior of its systems, especially in fields like Defense, Medicine, Law…
* Scientists and researchers working with Deep Learning and Neural Networks working in explanation of Artificial Intelligence (Explainable AI, The Defense Advanced Research Projects Agency …)

Critical Context

The need of Knowledge extraction

Artificial Intelligence allows to create systems that learn and act by following and imitating the behaviour of humans. Artificial Intelligence and especially Machine Learning and Deep Learning have spread out their wings and are nowadays applied in most of our daily tools while having a predominant role in many fields like Military Defence, Health Research or Finance. However, the processes followed by these machines often call out complicated mathematical models while dealing with a huge amount of information. The power and the effectiveness of these systems are then limited by their ability to make understand their reasoning and their logic to humans [3].

Therefore, Deep Learning systems and especially Artificial Neural Networks are famous for their high level of learning performance particularly in pattern and voice recognition. But the learning process is often assimilated to a black-box ([4] and [5]) unable to provide justifications for the results which can be crucially required in some areas of application. Hence, if we take the case of a neural network trained to classify pictures of animals, when the network recognizes an image as a cat, we have absolutely no idea why the system has converged to this conclusion, it could be because of the shape of the eyes, the ears or something else. The same issue can be raised with an image of tumour and a neural network which identify if the tumour is benign or malign. In this case, the doctor or the researcher would need to know the reasoning of the system. So this can be very confusing for the user who need to understand the system’s decision in order to know when it fails and when it is reliable.

However, this lack of interpretability was not an issue with the first AI systems which used to reason with logical inference, providing a feedback of their inference process which can then be understood by humans [3]. These systems were much less efficient than the current ones though. But most of the new works and papers dealing with the subject of knowledge extraction on the new techniques of Artificial Neural Network are inspired of the logical process used in these earlier systems [5].

Algorithms and methods for knowledge extractions

Decision trees

Various first attempts to extract rules from neural network require specific assumptions and specific structures for the network [6]. That is why Schmitz et al [6] has implemented an algorithm that extracts binary decision trees from a trained neural network without any assumptions made on the network. This approach is specific to feedforward neural networks dealing with supervised learning. A simple definition of the decision tree would be “it is an analysis diagram, which can help decision makers to decide which is the best option between different options, by projecting possible outcomes. The decision tree, gives the decision maker an overview of the multiple stages by that will follow each possible decision” [7]. Let’s take the example of an individual who wants to choose between going to a party or no:

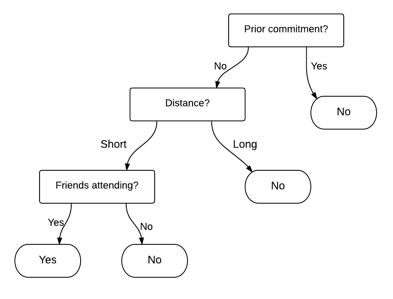


Figure 1: Binary decision tree, each node defines a test on values of property, terminal nodes are Boolean value of the goal class

The TREPAN algorithm was the first one developed to create decision trees that mimics the behavior of ANN ([8] and [5]).

Symbolic knowledge extraction

Another way to extract knowledge from the Black-Box of the network is to provide logic rules used as hypothesis [9]. In the hidden part (hidden layers) of the network, each hidden node represents a hypothesis about a specific rule. The hidden node will calculate the probability that “the rule implies a certain relation in the beliefs *b* being observed in the visible layer *V*, given the previously applied rules” [9]. For instance, a logic rule can be “The light is on and the door is not open”, which is represented by the formula: . That means that the hypothesis is true when (the light is on) is true and when (the door is open) is false. These kinds of rules are easily understandable by the user and they are based on the observations and data provided as inputs.

Rule extraction for Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are popular models mostly used for a data with a dynamic and temporal behavior like Time Series. Their particularity is that they have at least one feedback loop, which means that one of the output of one of the layers becomes an input of a previous layer [11]. To illustrate this concept, we take the example of a network with two layers each one with two neurons. One of the neuron of the second layer is connected to one of the neuron of the first layer via a feedback loop:

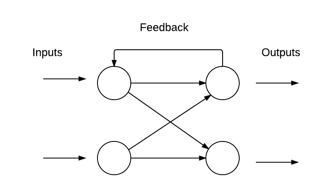


Figure 2: Illustration of recurrent neural network

A vast variety of RNN structures have been developed but only a few of them have been subject of works on rule extraction [10]. Most of them are listed by H. Jacobsson [10], and they use several kinds of logic rules (propositional logic, nonconventional logic, First-order logic and Finite state machines).

Unsupervised learning with Deep Belief Network and RBM

Unsupervised learning

RBM and Deep Belief Network

Rule extraction methods

References:

[1] S. N. Tran and A. D’Avila Garcez, “Deep Logic Network: Inserting and Extracting Knowledge from Deep Belief Networks,” in *IEEE Transactions on Neural Networks and Learning Systems,* PP(99),08 November 2016

[2] <https://www.technologyreview.com/s/603795/the-us-military-wants-its-autonomous-machines-to-explain-themselves/>

[3] D. Gunner, “Explanaible Artificial Intelligence (XAI)” in *Broad Agency Announcement,* DARPA-BAA-16-53, August 10, 2016

[4] C. Percy, A. D’Avila Garcez, S. Dragicevic, M. V. M. França, G.G. Slabaugh & T. Weyde (2016), “The Need for Knowledge Extraction: Understanding Harmful Gambling Behavior with Neural Networks,” in *Frontiers in Artificial Intelligence and Applications*, 285, pp. 974-981. doi: 10.3233/978-1-61499-672-9-974, 2016

[5] M. Rangwala and G.R. Weckman, “Extracting Rules from Artificial Neural Networks utilizing TREPAN” in *IIE Annual Conference.Proceedings,* pp. 1-6, 2006

[6] G. P. J. Schmitz, C. Aldrich, and F. S. Gouws, “ANN-DT: An Algorithm for Extraction of Decision Trees from Artificial Neural Networks”, in *IEEE Transactions on neural networks, Vol. 10, no. 6, November 1999*

[7] N. I. Khawaldah, N. A. Ishtayeh, “Binary Decision Tree”, *Al Zaraqa University*

[8] C. Percy, A. D’Avila Garcez, S. Dragičevic, M. V. M. França1, G. Slabaugh and Tillman Weyde, “The Need for Knowledge Extraction: Understanding Harmful Gambling Behavior with Neural Networks,” in *Frontiers in Artificial Intelligence and Applications*, 285, pp. 974-981. doi: 10.3233/978-1-61499-672-9-974, 2016

[9] H.L.H. (Leo) de Penning, A. d'Avila Garcez, L. C. Lamb3, J-J C. Meyer, “A Neural-Symbolic Cognitive Agent for Online Learning and Reasoning,”, in *Conference: IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, Barcelona, Catalonia, Spain, July 16-22, 2011

[10] H. Jacobsson, “Rule Extraction from Recurrent Neural Networks: A Taxonomy and Review,” in *Neural Computation* 17, 1223–1263**,** 2005

[11] S. Haykin, “Neural Networks and Learning Machines”, 3rd edition, Prentice Hall, Upper Saddle River, N.J., 2009