Modeling Implicit Learning : Extracting Implicit Rules from Sequences using LSTM

Humans acquire different kinds of knowledge employing different types of memory systems. Implicit knowledge is a non-expressible knowledge of which the individual is not aware of and that is acquired through implicit learning. The main characteristics of implicit learning are [1]: a) encoded rules can not be categorized explicitly, b) it impacts the subsequent reasoning process when new rules are encoded, c) there is no notion of positive or negative example learned through the implicit learning ability in the case of humans, d) the knowledge, i.e the rules, is hidden in the temporal expression of behaviour and more specifically in sequences of behaviourally significant events.

In this work, we study the process of extracting structured knowledge from data corresponding to sequences of behaviour. We argue that this structured knowledge reflects the expression of skills acquired by implicit learning. In a connectionist approach, we explore the question as whether a recurrent neural network (RNN) that is trained to acquire some skills from sequences of behaviour can be subsequently analyzed in order to extract underlying knowledge and express it in a structure such as a graph or an automaton. Many attempts have been made in the field of neural network interpretability for extracting knowledge using basic RNN models [2]. In the present work, we propose to extend the methodolody to more complex and more powerful RNNs, namely Long Short Term Memory (LSTM) networks. We primarily focus on the construction of the representation of rules inside the latent space of the LSTMs after learning non-binary sequences of variable sizes and with strong sequential dependencies. The grammars that are chosen for generating the corpus of sequences, the Reber grammar and its variations, were used for cognitive psychology experiments about the study of implicit learning ability in humans [1].

The first phase of our work proves that an RNN-LSTM, a RNN with a the hidden layer composed of LSTMs, knows how to recognizes sequences that respect the rules it has implicitly encoded during its training. The second and central part of our work focuses on the extraction of these rules which had been implicit, and their representation in the form of graphs (automata), a format that can be used and understood by a human operator. We propose an adaptation of [3] that allow us, for each grammar, to extract a representation in the form of a graph, with three different systems of notation, each carrying information on the internal functioning of the RNN-LSTM : the configuration of states and transitions between them, the temporal arrangement of patterns between the different detected states, and a final notation system that offers the possibility to get a contextual explanation regarding the management of patterns by the RNN-LSTM.

Lastly, the control phase validates that the extracted automata verified the same language as the original grammar. Over 10 consecutive simulations, the percentage of recognition of valid sequences exceeded 80% for Reber's grammar and a variation of it, that generates ambiguous sequences. Finally, we show that this performance is not a limit of our algorithm itself, but a compromise to be made between the degree of precision desired during extraction and the computing power allocated. We argue that our work addresses the question of the modeling of implicit learning in the field of computational cognition and the question of the interpretability of neural networks.

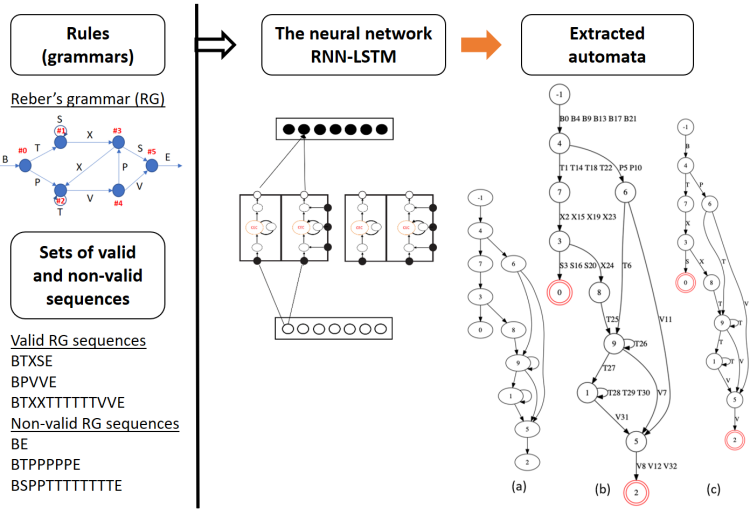

Figure : The global experimental approach for knowledge extraction from the neural network RNN-LSTM after learning valid-sequences generated from an artificial grammar.

[1] Arthur S Reber. Implicit learning of artificial grammars. *Journal of verbal learning and verbal behavior*, 6(6):855–863, 1967.

[2] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2018.

[3] Christian W Omlin and C Lee Giles. Extraction of rules from discrete-time recurrent neural networks. *Neural networks*, 9(1):41–52, 1996.