Neural-Symbolic Cognitive Agents: Architecture, Theory and Application

(Extended Abstract)

Leo de Penning
TNO Earth, Life and Social Sciences
Soesterberg, The Netherlands
leo.depenning@tno.nl

Artur S. d'Avila Garcez
Department of Computing, City University
London, UK
aag@soi.city.ac.uk

Luis C. Lamb
Instituto de Informatica, UFRGS
Porto Alegre, Brazil
lamb@inf.ufrgs.br

John-Jules C. Meyer
Department of Information and Computing Sciences, Utrecht University
Utrecht, The Netherlands
jj@cs.uu.nl

ABSTRACT

In real-world applications, the effective integration of learning and reasoning in a cognitive agent model is a difficult task. However, such integration may lead to a better understanding, use and construction of more realistic multiagent models. Existing models are either oversimplified or require too much processing time, which is unsuitable for online learning and reasoning. In particular, higher-order concepts and cognitive abilities have many unknown temporal relations with the data, making it impossible to represent such relationships by hand. In this paper, we develop and apply a Neural-Symbolic Cognitive Agent (NSCA) model for online learning and reasoning that seeks to effectively represent, learn and reason in complex real-world applications.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Learning—*Concept learning, Connectionism and neural nets, Knowledge acquisition*

General Terms

Theory, Design, Algorithms

Keywords

Neural-Symbolic Learning and Reasoning, Restricted Boltzmann Machines (RBM), Temporal Logic

1. ARCHITECTURE

The effective integration of online learning and robust reasoning in cognitive agents is a difficult task [8]. High-level human cognitive behaviour is difficult to model, elicit and represent in an automated system. There are many temporal relations between lower and higher-order aspects of human cognition, which are often non-deterministic and subjective (i.e. biased by personal experience and other factors like stress or fatigue). What is known about these relations is often limited to explicit behaviour (i.e. explainable behaviour), and frequently described too vaguely by domain experts. And in real-world applications, reasoning and learning with data observed in real-time, containing errors, missing values and inconsistencies, the task is made ever harder.

In this abstract, we present the architecture and theory of a Neural-Symbolic Cognitive Agent (NSCA) that is capable of meta-level learning and reasoning, and can be used for the modelling of complex and supportive cognitive agents that interact with humans in dynamic and non-stationary environments [2, 4]. This is achieved by taking advantage of neural-symbolic integration [1] to capture the many temporal relations related to human behaviour by learning from observation, using a Recurrent Temporal Restricted Boltzmann Machine (RTRBM) [7]. This approach enables the NSCA to encode prior knowledge, reason with this knowledge probabilistically (deduction), infer beliefs about dynamic observations (abduction), learn new knowledge from observations (induction) and extract this knowledge in symbolic rules in the form of temporal logic rules. For example, the NSCA can encode a temporal logic rule like $H_1 \leftrightarrow B_1 \wedge \bullet H_2$ (where $\bullet$ means 'at time $t-1$'; everything else happens at time t) in a RTRBM by mapping the hypotheses H to the hidden units, and beliefs B to the visible units. The temporal relations $\bullet H$ are mapped as recurrent connections between the hidden units.

When the NSCA observes B_1 (i.e. activating the related visible unit with some probability or real value), it can calculate the probability of H_1 by forward propagation of the probability of B_1 and $\bullet H_2$ to the hidden unit representing H_1 in the RTRBM. We refer to this as **deduction** in the NSCA. Deduction is similar to Bayesian inference, where for all hypotheses H the conditional probability $P(H|B, \bullet H)$ is calculated using the weights in the RTRBM.

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

From the posterior probability distribution, the NSCA can then assume that hypothesis H_1 is true, and infer the probabilities of all beliefs and temporal relations (i.e. $\bullet H$). We refer to this process as **abduction** in the NSCA. Via abduction the NSCA can infer the most likely beliefs B based on hypothesis H from the conditional probability $\mathbf{b} = P(B|H)$. The differences between the observed and inferred beliefs are then used by the NSCA to determine the intentions and actions of the agent.

Finally, the NSCA can learn new relations from observed beliefs B and $\bullet H$. We refer to this as **induction** in NSCA. Induction can be achieved by using the differences between observed and inferred beliefs to strengthen or weaken the correlation between activated hypotheses H and the observed beliefs B. NSCA does so by updating the weights in the RTRBM using contrastive divergence and backpropagation through time as done in [7].

2. THEORY

Based on the theory of penalty logic [6] the NSCA can extract temporal knowledge from a trained RTRBM in the form of symbolic rules R obtained directly from the network's weights W. Where each rule r is related to a hidden unit for which we infer the probabilities of the associated beliefs B and $\bullet H$ from the RTRBM, i.e. $b_r = P(B|H_r)$ and $h_r^{t-1} = P(\bullet H|H_r)$. If we do this each hidden unit, we can construct a temporal logic program Ψ using the following equations (where k is the number of beliefs, m the number of hypotheses and w_{ir} is the weight of the symmetric connection between the related visible unit v_i and hidden unit h_r, and w'_{lr} is the weight of the recurrent connection between the previous hidden unit activation h_l^{t-1} and hidden unit h_r in the RTRBM).

$$\Psi = \{\langle c_r : H_r \leftrightarrow \bigwedge_{i=1}^{k} \theta_r^{(i)} \bigwedge_{l=1}^{m} \rho_r^{(l)} \rangle, \forall r \in R\} \tag{1}$$

$$\theta_r^{(i)} = \begin{cases} B_i & \text{if } w_{ir} > 0 \wedge b_r(i) > 0.5 \\ \neg B_i & \text{if } w_{ir} < 0 \wedge b_r(i) > 0.5 \\ \emptyset & \text{otherwise} \end{cases} \tag{2}$$

$$\rho_r^{(l)} = \begin{cases} \bullet H_l & \text{if } w'_{lr} > 0 \wedge h_r^{t-1}(l) > 0.5 \\ \bullet\neg H_l & \text{if } w'_{lr} < 0 \wedge h_r^{t-1}(l) > 0.5 \\ \emptyset & \text{otherwise} \end{cases} \tag{3}$$

$$c_r = P(H_r|B = \mathbf{b_r}, \bullet H = \mathbf{h_r^{t-1}}) \tag{4}$$

The literals in $\theta_r^{(i)}$ depend on the weight w_{ir}, where a negative weight w_{ir} increases the probability of H_r when the probability of B_i decreases and thus the probability of $\neg B_i$ increases. The inverse applies to a positive weight. When w_{ir} is 0 or $b(i) < 0.5$, belief B_i has no significant influence on the hypothesis and can be left out. A similar approach is used to extract the literals in $\rho_r^{(l)}$ for $\bullet H_r$. Finally, Eq. 4 shows how we calculate the confidence value c_r of rule r, denoting the strength or 'penalty' of the equivalence relation, as done in [6]. This confidence value is based on the notion of Bayesian credibility described in [5] and is calculated in a similar way.

Based on this approach the NSCA can also encode prior knowledge in the RTRBM. This is effectively done by optimizing the joint probability distribution $P(H_r = c_r, B = \mathbf{v}, \bullet H = \mathbf{h^{t-1}})$ using the contrastive divergence algorithm with a high learning rate (assuming non-conflicting clauses).

3. APPLICATIONS

The NSCA has already been applied in various real-world environments. For example, in automated driver assessment by learning from observation of driving instructors and real-time dynamic simulation data (e.g. position and orientation of vehicles, gear, steering wheel angle, etc.) [2]. And the recognition and description of human behaviour in video (e.g. chase, exchange, jump, etc.), where the NSCA also provided meaningful temporal logic-based descriptions to explain these behaviours in terms of low-level features of detected objects [3].

4. CONCLUSIONS AND FUTURE WORK

The neural-symbolic cognitive agent model and architecture presented in this abstract offer a unified model capable of online learning, reasoning, and dynamic adaptation in complex real-world applications. The approach allows agents to learn rules about observed data in complex, data-intensive real-world scenarios and extract this knowledge for validation, reporting, maintenance, evolution and feedback. The approach also allows prior knowledge to be encoded in the model and deals with uncertainty in real-world data.

In summary, the NSCA provides an integrated model for learning, knowledge representation and reasoning capable of producing a realistic computational cognitive agent model. The NSCA seeks to address the challenge put forward in [8, 9], and contributes to the development of algorithms and tools for multiagent learning and adaptation in dynamic and non-stationary environments.

5. REFERENCES

[1] A. S. d'Avila Garcez, L. C. Lamb, and D. M. Gabbay. *Neural-Symbolic Cognitive Reasoning.* Springer-Verlag New York Inc, 2009.

[2] L. de Penning. Visual Intelligence using Neural-Symbolic Learning and Reasoning. In *Proc. of the Neural Symbolic Learning and Reasoning workshop at IJCAI*, Barcelona, Spain, 2011. IJCAI.

[3] L. de Penning, R. J. den Hollander, H. Bouma, G. J. Burghouts, and A. S. d'Avila Garcez. A Neural-Symbolic Cognitive Agent with a Mind's Eye. In *Workshop on Neural-Symbolic Learning and Reasoning at AAAI*, 2012.

[4] L. C. Lamb, R. Borges, and A. S. d'Avila Garcez. A connectionist cognitive model for temporal synchronisation and learning. In *Proc. of the AAAI Conference on Artificial Intelligence*, pages 827–832. AAAI Press, 2007.

[5] P. Lee. *Bayesian Statistics: An Introduction.* Arnold Publication, third ed edition, 1997.

[6] G. Pinkas. Artificial Intelligence Reasoning , nonmonotonicity and learning in connectionist networks that capture propositional knowledge. *Artificial Intelligence*, 77:203–247, 1995.

[7] I. Sutskever. The recurrent temporal restricted boltzmann machine. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[8] L. G. Valiant. Three problems in computer science. *Journal of the ACM (JACM)*, 50(1):96–99, 2003.

[9] M. Wooldridge. *An introduction to multiagent systems.* Wiley, 2nd edition, 2008.