From Synapses To Rules: The Self-Referential Perspective

BRUNO APOLLONI, GABRIELE BIELLA
Department of Information Sciences
University of Milan
Via Comelico 39/41, 20135 Milan
ITALY
apolloni@dsi.unimi.it

ANDREAS STAFYLOPATIS
Department of Electrical and Computer Engineering
National Technical University of Athens
157 73 Zographou, Athens
GREECE
andreas@cs.ntua.gr

Abstract: We consider the extraction of formal knowledge from a trained neural network in the perspective of identifying this network within our brain and the final user of this information with our brain again. We first analyze theoretical issues — mainly coming from AI, but also from neurophysiology and information theory — on relations and links between subsymbolic and symbolic knowledge in our brain. From this analysis a bipartition derives of the considered algorithms. From one side, there are direct methods for discovering Horn clauses and extensions from trained networks, a usual subject in many review papers. From the other side, we will identify symbolic knowledge with tools for efficiently managing concepts discovered in subsymbolic way in a self-referential framework where a neural network is however the user of the concepts. At first glance, this alternative perspective would just reconsider the direct methods in respect to the functionalities of the hidden_to_output nodes connections. But exactly after self referentiality, discovering formal connection should require a heavy training of the involved neural network, namely: a training capable of simulating the architectural and parametric refinement achieved by our brain along millennia. This calls for algorithms for symbolical learning that comply with neurophysiological functionality constraints, but shorten the mentioned long training phase, by using facilities now available to our brain — such as a preexisting formal knowledge and the capabilty of generating suitable examples by ourself. By definition, the output of these algorithms is exactly the goal knowledge springing from neural networks we are searching for.

Key-Words: Neural networks, rule extraction, symbolic/subsymbolic knowledge. Proc.pp.5301-5306

1 A cognitive perspective

Since the early revival of neural networks, the aim of rereading in symbolical terms a function learnt by a neural network has been pursued by many researchers. Many reasons can be listed at the basis of this objective. The most obvious reason is the fact that human beings are used to discuss and communicate about scientific topics by symbols ruled by mathematical logic, with evident benefits concerning generalization, diagnosis, inquiry and explanation of the communicated matter. In other words, although no one can refuse that subsymbolic attitudes [23], such as intuition or experience, lie at the basis of practical useful actions and that they are indispensable ingredients of common sense reasoning [24], we must at least be able to describe through symbols the core of these actions and their rationale, in order to extend these actions to similar operational contexts and/or to submit them to criticism, improvement and translation into formal rules. This delineates an inherent hierarchy where at low levels we locate actions and at higher levels we locate their formal explanation.

Answering the question of how the symbolic description of mental processes, in terms of rules and representations in the province of conventional artificial intelligence (AI), can be related to a subsymbolic description in terms of brain mechanisms is the so called *complete reduction problem* [12], as a last translation of the symbol

grounding AI problem (which discusses how symbols in a symbolic system acquire meaning). Neural networks, or the connectionist paradigm in the broad sense, play the role of a representative of brain mechanisms. In this context the aforementioned question translates in the more pragmatic question of designing architectures embodying both neural networks and finite states automata, where the latter represent the formal knowledge and deduction rules [25].

Various opinions regarding the links between the two (neural and automaton) modules can be found in the literature. We pass from:

i) the negation of a substantial difference between them [30], reducing the neural approach to a mere technique under the usual symbolic approach;

ii) a more conciliatory perspective based on the usual semeiotic distinction between concept and symbol [8], where a concept can be also achieved by a neural network, a symbol is just a label, with its own algebra, to refer to a concept, and an external agent states links between them;

iii) a utility point of view ([4], [27], and most part of neurofuzzy literature), in which *a goal oriented automatic script* [22] takes benefit from the tuning of some parameters operated by a neural network;

iv) a totally subsymbolic approach [29], where the cognitive paradigm of an agent, aimed at surviving in an hostile environment, is completely embodied by a neural network in an evolutionary framework, in which populations of networks are decimated or increased according to fitness functions and usual genetic change (chance) rules.

2 A connectionist perspective

Till here the high level, mostly philosophical discussion in the favourite style of the AI community. Seeing the matter from the connectionist point of view, we face a trained network and ask ourselves how to solve the complete reduction problem in this instance. In most connectionist literature on this matter, the transition from lower to higher levels is conceived as being operated by an external agent who resides on the top of an analogous hierarchy. We may think of this agent as an old parent who translates the confused stammering of his pupil into a clear sentence, i.e., who states the reference links between the naive concepts raised by the network and robust symbols.

Knowledge representation in artificial intelligence systems involves the use of symbols. Among several possible definitions, we will adopt the following circular definition of symbols. A symbol is a meaningful quantity that represents a body of knowledge that may need to be accessed in processing the symbol. Stated differently, a symbol represents an abstract notion as well as relevant properties of that abstract notion. Symbolic processes transform symbol structures into other symbol structures that are relevant in a specific context. Although heuristic and linguistic knowledge used by experts can be the basis on which AI systems are constructed, the mentioned symbol grounding problem remains one of the main problems of AI [12]. Roughly speaking, this is the top-down view of the problem of subsymbolic to symbolic mapping. Several methodologies (as fuzzy set theory, probabilistic reasoning etc.) have been used in order to support the conventional AI systems in order to bridge the gap between symbols and subsymbols. For example, fuzzy expert systems or decision trees are using numerical data in order to determine the symbolic knowledge.

On the other hand, subsymbolic representation is usually defined by the origin of the information (e.g. sensors or database), in a more opaque, subconceptional manner, rather than by the content of the information. A main objective of mapping subsymbolic information into a symbolic representation is to find an abstract representation of the symbol or object, which is invariant with respect to various features (e.g. invariant with respect to position and orientation). Moreover, the representation should be such that direct links are possible to properties of the object or symbol and such that the representation can easily be used in symbolic reasoning. This problem can be viewed as the bottom-up case of the subsymbolic to symbolic mapping problem. The main approach of subsymbolic processing is the connectionist approach, that is biologically inspired neural networks that try to model the human brain and artificial neural network models that try to emulate the intelligent behaviour of humans. Lately, supporting methodologies like fuzzy set theory have been used in order to improve the symbolic representation and processing capabilities of NNs.

Although each of the subsymbolic and symbolic framework can specify intelligent systems by itself, the limits of each category are not clear and become increasingly vague. The main issue in the research on intelligent systems is now how the symbolic description of mental processes, in terms

of rules and representations in the province of conventional AI, can be related to a subsymbolic description in terms of brain mechanisms (the top-down approach). The same question is how the subsymbolic processes of neurons, synapses and interconnections of a conventional artificial neural network can be related to the symbolic description of human logic and behaviour (the bottom-up approach). One of the principal reasons that neural networks have been considered a useful vehicle for such a development is that there is an existence evidence for the solution of creating such a system (and of course answering the above questions), that of the human being.

3 A physiological oriented perspective

Our perspective looks far different. Namely, we start from the perspective that the rule extractor is not an external agent but a rule located in the mentioned hierarchy, i.e. a piece of neural network, by itself. Thus, rather than an external subroutine, we are looking for an inner loop *priming* symbols from synapses.

Let us schematize the problem in this way: all that exists is a plenty of neurons variously interconnected through plastic synapses. Plasticity means that synapses are capable of answering to external stimuli by modifying their weights according to some entropic rule. Neuron plasticity is regulated by entropic rules under the common aim of preserving and improving neurons' life (cognitive target without pre-existing script). Plasticity plays the double role of modifying the synaptic weights and, as an extreme consequence, of modifying the neural network architecture. Thus, as a solution of the mind-brain dilemma [31], in place of searching for an architecture and weights that may result optimal for a special brain task, we try to solve the same search problem under the preeminent target of optimizing the implementation of the above entropic rules. Stated in other words, we look at physiological models for studying realistic models for passing from features to symbols. Namely, we start from already trained neural networks by definition, thus skipping the wide chapter of efficient and realistic learning algorithms. Rather, we focus on neural network architecture and dynamics.

Let us start from the stimulating remarks of [6]: "It is easy to recognize a histological preparation as being cortex rather than cerebellum or tectum. It is much more difficult to tell whether it is human or bovine, motor, sensory or associative cortex", and [32]: "the morphological and physiological characteristics of cortical neurons are equivalent in different species, as are the kinds of synaptic interactions involving cortical neurons. This similarity in the organization of the cerebral cortex extends even to specific detail of cortical circuitry".

Actually we have some differences concerning:

- the thickness of the different cortical layers (sensory layer is thicker in bovines),
- biochemical composition (different markers) characterize the two species,
- connection fan-in and branching diffusion are higher in humans.

But all these are indicators of a wider or different subsymbolic activity of humans w.r.t. bovines. Looking at prefrontal or frontal cortex, the sites deputed to mental abstraction processes, we have that these regions are more wide in humans. But:

- prefrontal and frontal areas are less extensive in a 2 years child than in an adult bovine. Neverthless child is capable of sophisticate abtraction processes, bovine isn't.

Today, we have not much perplexity on computational mechanisms of a neural network. In fact, usual formulas based on a weighted mixture of the signal coming from a set of neurons seems a reasonable approximation to physiological mechanisms being based on sufficient statistics in general. Physiological connections actually change weights and pointers at different time scales, even at a rate of fractions of seconds, as it happens in our computer simulations.

Thus, we do not distrust the subsymbolical computing mechanism. But where is soul? As mentioned before we have a lot of ingenious brain architectures capable of explaining some consciousness phenomena. And a counterpart series of neurophysiological lay-outs are available as well. However they just move the problem one step back: where is the architect soul?

Far from willing to engage a religious or philosophycal discussion let us consider the brain life. It is well known in neurophysiology that the neruon population in our brain is in a sharp continuous evolution. This consists of a population increasing in the early stage, till 5 years childhood, a turn over in the subsequent ten years or so, and decreasing at a rate of some thousands of neurons per day in the rest of the life. This massive pruning however is not a symptom of decline, at least till the age of 40 years (hopefully later as well), but of brain strengthen, on the contrary, according partly to codified genetic laws, partly to actual interactions with the environment. In the

assumption that a robust veering from fantasy to rationality occurs in the same human life period [19], we could argue that the sizing and then pruning of the neural population is functional to a bias from subsymbolical to a symbolical organization of the knowledge in the brain neural network.

Thus, an analysis of the mentioned brain genetic laws could give hints on the complete reduction problem. Stated in other words, from among the huge amount of evolutionary lives we can assume for brain, we will simulate only a restricted family of evolutions that satisfy neurophysiological and neuroevolutionary constraints. From one side this family satisfies the constrained connectionist paradigm evoked by [31] as a proof of the connectionism worthiness, from the other one it acts as a very accelerated replica of a possible brain life from the starting of human life till now. If this simulation is at least partially correct, the resulting configuration should shed light on the essential connectivity maps in the brain and its functionality in building symbolic concepts from sensory data. This should look as an architectural counterpart of the Giles strategy [15] of identifying finite automaton transition rules within the evolution of the state vector of a recurrent neural network. The mentioned physiological constraints should inhibit the search space exponential explosion that undermines Giles strategy.

From a more pragmatic - some time practitioner biased - point of view, many static rules have been proposed in the literature for reshaping a neural architecture. On this concern, in [3] a whole process dynamics is proposed that is capable of dynamically sizing a neural network, some times pruning, some times adding nodes or connections to the network, under the global aim of maintaining neurons alive, young and efficient.

4 Structuring the brain

Assuming that a neural network is able to design its architecture by itself, and that this architecture at least partly schlerotizes, means to assume that the network is able to produce concepts in hardware as unmodifiable pieces of the network itself, as actually prefigured by many authors such as [29]. The crucial problem remains the building of the reference links table as the basic symbolic activity.

This table could be learned by the network as well, like in the expert mixture of Jacobs and Jordan [13] or in the various issues of boosting algorithms [9] (a last rereading in late sense of ART networks [11]). In both approaches single networks are trained on subdomains of the input space, so that they can be identified with subsymbolic concepts each. Then on a given input their responses are properly weighted — or just selected with a majority rule — to achieve the final verdict . But this mixture requires a complex training procedure operated by an external agent again, and the inferred table is just a mapping not susceptible of symbolic manipulation.

Another alternative is that the boolean logic becomes coded at highest levels of a hierarchically organized neural network such that the above pieces of network constituting concepts represent the floor of this hierarchy, and meditation chains are wired in the higher level according to a protocol which proves preserving information. It is possible to show that wirings arranged to code boolean aggregates of the lower level concepts are highly information preserving in terms of *sentry points* [5]. Thus it could be hypothesized that the entropic criterion of minimizing information waste drove brain to sclerotize its higher level layers in order to code such a kind of protocol — as a physiological counterpart of the Peirce pragmatism and an inherent (deviated in some points) implementation of his abduction methods [17]. In relation to recent models of *counter stream* [14], [28] or of bidirectional communication between different hierarchical level, such as relaxation-type refinement [20], CONSYDER strategy [24], sycronizing Contextual Fields [18], attention control [10] or control of the recognition process [1], [21], the present model supplies a way of building up and managing in hardware the symbolic knowledge to be circulated from higher to lower level brain layers (i.e. top-down from mind to brain). Actually, what is intended as an abductional process in those papers is not far different from the idea underlying LPC (linear predictive coding) [16]: use an adaptive model to foresee next signal and assume the difference between actual and predicted signal as signal information to be employed for checking and/or modifying the model itself. For instance, in the Koerner model, a top-down flow of concepts matches the bottom-up flow of the sensory data. At each level, a sufficient agreement between the two kinds of knowledge primes both information for adapting the concept to the sensory data and a new step to go deeper in this symbolic-subsymbolic comparison. This step looks as an abduction step where formal knowledge is enriched. However it remains unexplained, at least at an operational level, which mechanism completely rules the top-

down concept flow and who filled the concept reservoir.

5 Two different strategies for explain networks

The assumption of these kinds of hierarchical arcitecture is not irrelevant to the way of explaining neural networks — pieces of our brain in particular.

We may exploit passing from synapses to symbols in two ways. A direct one tries to "open" the trained network in order to recognize Horn clauses or some extension of them [2], [26]. We aknowledge this is an efficient way only in elementary cases, where the information content of the connection weights is very low. Actually, these case must occur when the brain tries to organize in a very compact and unambiguous form the detailed knowledge experienced at subsymbolical level. We can expect that in this case the network wires simple circuits, such as boolean aggregates, by itself still as a result of optimization of some entropic rules, but our way of discovering these circuits must be different for sake of computational efficiency.

Concepts arise from sclerotization of trained synapses, and their management is committed to very simpe networks handling variants of Horn clauses. We can discover these boolean formulas, for instance analysing connections between hidden and output nodes, but this will result successful provided we trained the network for a very long time on a very large training set according to an appropriately devised error function and a suitable learning procedure.

Otherwise, we can shorten this subsymbolical process by symbolically learning the high level network through a process that is not neurophysiological-like, but agrees with the neurophysiological learning process hypothesized to occur in the brain during the millennia. One of the most relevant facilities that distinguish optimization from learning, available to us after this second strategy — available to higher level of neural network anyhow —, is the capability of generating suitable examples by ourselves (in a sort of unconstrained active learning [7], in place of hopefully waiting for them from the environment. The suitability of the examples, in its turn, derives from a correct use of the whole knowledge repertoire formalized by human beings in the millennia. By defenition, the output of these algorithms is exactly the formal knowledge we were searching for.

6 Conclusions and perspectives

From synapses to rules is a transition recently pursued by many researchers urged by the need of giving a new perspective to intelligent systems and, anyway, catching reliable solutions to highly complex problems. In this paper, we stretched the neuromorphological strategy of mimicing human brain activities till the extreme constraint that results coming from this strategy must be useful to a neuromorphic system as well. This forced us to consider, together with the usual technical aspects of how to convert network configurations into sets of rules, the sunbsequent problem of how to use these rules in an efficient way within complex structures. This implies to prefigure both efficient learning algorithms that enjoy both preexisting formal knowledge and random solicitations coming from Gibbs-like dynamics, and self-organizing higher level structure that manages fixed results from connectionist modules as symbols.

References:

[1] J. Ando, 3D Object recognition using hierarchical modular networks, In S.Z. Li, D.P. Mital, E.K. Troh, II. Wang (Eds.), *Recent Developments in Computer Vision*, Springer, 1996, pp. 467-478.

[2] R. Andrews, J. Diederich and A.B. Tickle, Survey and critique of techniques for extracting rules from trained artificial neural networks, *Knowledge-based Systems*, Vol. 8, No. 6, 1995, pp. 373-389.

[3] B. Apolloni and G. Ronchini, Dynamic sizing of multilayer perceptrons, *Biological Cybernetics*, Vol. 71, 1994, pp. 49-63.

[4] B. Apolloni., A. Piccolboni and E. Sozio, Hybrid symbolic subsymbolic system for controlling a single link flexible arm, *J. of System Engineering*, Vol. 6, 1996, pp. 208-222.

[5] B. Apolloni and S. Chiaravalli, PAC Learning of concept classes through the boundaries of their items, *J. Theoretical Computer Science*, Vol. 172, 1997, pp. 91-120.

[6] V. Braitenberg, Cortical architectonics: General and real, In M.A. Brazier and H. Petsch (Eds), *Architectonics of the cerebral cortex*, Raven, 1978.

[7] D. Cohn, L. Atlas and R. Ladner, Improving Generalization with Active Learning, *Machine learning*, Vol. 15, 1994, pp. 201-221.

[8] G. Dorfner, E. Prem and H. Trost, Words, symbols and symbol grounding, Austrian Research Institute for Artificial Intelligence, TR 93-303, Vienna, 1993.

[9] Y. Freund, Boosting a weak learning algorithm by majority, *Information and Computation*, Vol. 121, No. 2, 1995, pp. 256-285.

[10] K. Fukushima, A neural network model for selective attention in visual pattern recognition, *Biological Cybernetics*, Vol. 55, No. 5, 1985.

[11] S. Grossberg, *The adaptive brain*, Elsevier, Amsterdam, 1987.

[12] J. Haugeland, The nature of plausibility of cognitivism, *Behaviour and Brain Science*, Vol. 2, 1978, pp. 215-260.

[13] M.I. Jordan and R.A. Jacobs, Hierarchical mixtures of experts and the EM algorithm, *Neural Computation*, Vol. 6, 1994, pp. 181-214.

[14] E. Koerner, H. Tsujino and T. Masutani, A cortical-type Modular Neural Network for Hypothetical reasoning, *Neural networks*, Vol. 10, No. 5, 1997, pp. 791-814.

[15] C.W. Omlin and C. Lee Giles, Constructing Deterministic Finite-State Automata in Recurrent Neural Networks, *Journal of the ACM*, Vol. 43, No. 6, 1996, pp. 937-972.

[16] T.W. Parson, *Voice and speech processing*, McGraw-Hill, 1986.

[17] C.S. Peirce, How to make our ideas clear, *The Popular Science Monthly*, 1878.

[18] W.A. Philips and W. Singer, In search of common foundations of cortical computations, *Behavioural and Brain Sciences*, Vol. 20, 1997, pp. 657-722.

[19] J. Piaget, *Biology and Knowledge*, University of Chicago Press, 1971.

[20] T. Poggio, V. Tone and C. Koch, Computational vision and regularization theory, *Nature*, Vol. 317, 1985, pp. 314-319.

[21] R.O.N. Rao and D.H. Ballard, Dynamic model of visual memory predicts neural response properties in teh visual cortex, TR 95, 4, Dept. of Computer Science, Univ. of Rochetster, 1995.

[22] J.R. Searle, The intentionality of intention and action, *Cognitive Science*, Vol. 4, 1990, pp. 47-70.

[23] P. Smolensky, Information processing in dynamical systems: foundations of harmony theory, In D.E. Rumelhart, *Parallel Distributed Processing*, Vol. 1, MIT Press, Cambridge, MA, 1986, pp. 194-218.

[24] R. Sun, *Integrating rules and connectionism for robust commonsense reasoning*, Wiley, New York, 1994.

[25] R. Sun and F. Alexandre (Eds), *Connectionist-Symbolic Integration: From Unified to Hybrid Approaches*, Lawrence Erlbaum, 1997.

[26] A.B. Tickle., R. Andrews, M. Golea and J. Diederich, The Truth will Come to Light: Directions and Challenges in Extracting the Knowledge Embedded within Trained Artificial Neural Neworks, *IEEE Trans. on Neural Networks*, Vol. 9, 1998, pp. 1057-1068.

[27] G.G. Towell, J.W. Shavlic and M.O. Noordewier, Refinement of approximately correct domain theories by knowledge-based neural networks, In Proc. of the Eighth National Conference on Artificial Intelligence, San Jose, CA, AAAI Press, 1990, pp. 177 - 182.

[28] S. Ullman, Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex, *Cerebral Cortex*, Vol. 5, 1995, pp. 1-11.

[29] J. Vaario and S. Ohsuga, An emergent construction of adaptive neural architectures, *Heuristics - The Journal of Knowledge Engineering*, Vol. 5, No. 2, 1992, pp. 1-12.

[30] P.F.M.J. Verschure, Formal minds and biological brains: AI and Edelman's extended theory of neuron group selection, *IEEE Expert: Intelligent Systems and their Applications*, Vol. 8, No. 5, 1993, pp. 66-75.

[31] P.F.M.J. Verschure, Connectionist explanation: taking position in the mind-brain dilemma, In G. Dorffner (ed.), *Neural Networks and a New Artificial Intelligence*, International Thomson Computer Press, London, 1997, pp. 133-188.

[32] E.L. White, *Cortical circits: Synaptic organization of the cerebral cortex: Structure, function and theory*, Birkhauser, 1989.