Transparent Neural Networks

Integrating Concept Formation and Reasoning

Claes Strannegård[1], Olle Häggström[2],
Johan Wessberg[3], and Christian Balkenius[4]

[1] Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Sweden and Department of Applied Information Technology, Chalmers University of Technology, Sweden
claes.strannegard@gu.se
[2] Department of Mathematical Sciences, Chalmers University of Technology, Sweden
olle.haggstrom@chalmers.se
[3] Institute of Neuroscience and Physiology, University of Gothenburg, Sweden
johan.wessberg@gu.se
[4] Department of Philosophy, Lund University, Sweden
christian.balkenius@lucs.lu.se

Abstract. We present the *transparent neural networks*, a graph-based computational model that was designed with the aim of facilitating human understanding. We also give an algorithm for developing such networks automatically by interacting with the environment. This is done by adding and removing structures for spatial and temporal memory. Thus we automatically obtain a monolithic computational model which integrates concept formation with deductive, inductive, and abductive reasoning.

Keywords: transparent neural networks, developmental robotics, concept formation, deductive reasoning, inductive reasoning.

1 Introduction

Artificial General Intelligence (AGI) aims for computer systems with human-like general intelligence [1]. Thus, just like humans, AGI systems should be able to reason and learn from experience by interacting with the environment. This leads to desiderata on AGI systems that concern developmental processes and automated reasoning. It has been suggested that to build intelligent machines, it is necessary to use developmental methods where a system develops autonomously from its interaction with the environment [2][3]. This lead to the research area of developmental (or epigenetic) robotics, where models based on biological principles are used either to describe human cognitive development or to come up with novel principles for AI. The explicit goal of the area is to design cognitive

J. Bach, B. Goertzel, and M. Iklé (Eds.): AGI 2012, LNAI 7716, pp. 302–311, 2012.

architectures that can autonomously develop higher cognitive abilities. However, most research so far has focused on sensory-motor development and social interaction and has mainly ignored higher cognitive functions such as reasoning. Reasoning is commonly analyzed as in the following quote from [4]:

> Three notable hallmarks of intelligent cognition are the ability to draw rational conclusions, the ability to make plausible assumptions and the ability to generalise from experience. In a logical setting, these abilities correspond to the processes of deduction, abduction, and induction, respectively.

These problems have been studied thoroughly in traditional AI with symbolic methods such as automatic theorem proving [5], sub-symbolic methods such as artificial neural networks (ANNs) [6], probabilistic methods such as Bayesian networks [7], and many others [8]. These approaches typically focus on a proper subset of the above-mentioned types of reasoning. For instance, the symbolic approach is mainly concerned with deductive reasoning and the sub-symbolic approach with inductive reasoning. This might suggest a hybrid approach, which integrates symbolic and sub-symbolic methods such as ACT-R [9], conceptual spaces [10], or neural-symbolic systems [4]. Hybrid approaches, however, tend to be limited by the difficulty of designing interfaces for complex interaction between the different subsystems.

Human reasoning processes seem to be tightly integrated with concept formation: new concepts are created continuously and become integrated with previous knowledge and involved in new reasoning processes. Looking at developmental psychology, evidence is accumulating that infants and children use similarity-based measures to categorize objects and form new concepts [11].

For these reasons, AGI could potentially benefit from a developmental system which integrates concept formation, deduction, induction, and abduction. This is the goal of the transparent neural networks (TNNs), which are introduced in sections 2 and 3. Section 4 contains an analysis of the TNN model from the perspective of concept formation and automated reasoning and Section 5 is a conclusion.

2 Transparent Neural Networks

The TNN model arose out of an attempt to create a computational model that simultaneously accommodates two previously developed models of human reasoning: one for deductive reasoning about propositional logic [12], the other for inductive reasoning about number sequence problems [13]. In fact, both of these models are based on term-rewriting systems. In this section we define the TNNs together with their computation rules and show how they can be used for handling spatial and temporal memory.

Traditional ANNs tend to be intransparent in the sense that it is virtually impossible for a human to understand how they work and predict their computations and input-output behavior. This holds already for feed-forward networks

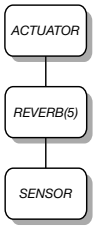

Fig. 1. Network modeling the tentacle of a sea anemone, which keeps retracting for 5 time units after being touched

and still more for recurrent networks. Therefore, they are generally not suitable for deductive reasoning and applications that are safety-critical.

TNN is a restricted type of ANN, designed with the aim of facilitating human understanding. The TNN model was heavily inspired by neuroscience, but since our only concern here is AGI, we feel free to deviate as much as we want from any existing biological or computational model.

2.1 Definition

Definition 1 (TNN). *A* TNN *consists of the following parts:*

- *A set D of labeled* nodes*. The labels are SENSOR, ACTUATOR, MIN, MAX, AVERAGE, SPACE(μ, σ), DELAY(n), and REVERB(n). Here μ and σ are real numbers and n is a natural number.*
- *A cycle-free relation $R \subset D^2$, whose elements are called connections.*

Restriction: The labels ACTUATOR, SPACE, DELAY, and REVERB are only allowed on nodes with exactly one predecessor.

In this paper we use the graphical convention that connections point upward in all figures. Examples of TNNs are given in Figures 1 and 2.

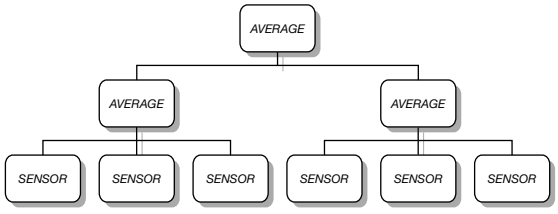

Fig. 2. Network modeling a gustatory organ for sweetness. Information on the local level is summarized and passed on to higher levels. A similar network with MAX nodes instead of AVERAGE nodes could model a sensory organ for pain.

2.2 Environments

Definition 2 (Frame). *Let V be the set of real numbers in the interval [0,1]. A* frame *for a TNN with sensor set S is a function $f : S \to V$.*

Definition 3 (Environment). *Let T be the set of natural numbers (modeling time). An* environment *for a TNN with sensor set S is a function $e : S \times T \to V$.*

Frames model momentary stimuli and environments model streams of stimuli generated by the surrounding world (which might include the TNN itself). For instance, an environment could represent the taste and smell of an apple, followed by the sound sequence [æpl], followed by the visual sequence "$6 \cdot 8 = 48$".

2.3 Activity

In contrast to the standard ANN model, our model has two types of activity. This enables us to model perception and imagination separately. For instance, it enables us to distinguish between the perceived and the imagined taste of an apple. It also enables us to model the perception of the sequence $2, 5, 8, 11$ and the imagination of the next number 14.

The inspiration behind the two types of activity comes from the distinction between (i) distal and proximal dendritic signal processing and (ii) inner and outer senses [14].

Let $N_{(\mu,\sigma)}(x) = \exp\{-(x-\mu)^2/\sigma^2\}$. This is the Gaussian density function with mean μ, standard deviation σ, and max value 1. Let A be a TNN with sensor set S and let $e : S \times T \to V$ be an environment. Then the *real* activity $r : D \times T \to V$ and *imaginary* activity $i : D \times T \to V$ are defined as follows.

Definition 4 (Real activity). *Let $r(a, 0) = 0$ and let $r(a, t+1) =$*

- *$e(a, t)$ if a is labeled SENSOR*
- *$min\{r(a', t) : (a', a) \in R\}$ if a is labeled MIN*
- *$max\{r(a', t) : (a', a) \in R\}$ if a is labeled MAX*
- *$average\{r(a', t) : (a', a) \in R\}$ if a is labeled AVERAGE*
- *$r(a', t)$ if a is labeled ACTUATOR and $(a', a) \in R$*
- *$N_{(\mu,\sigma)}(r(a', t))$ if a is labeled SPACE(μ, σ) and $(a', a) \in R$*
- *$r(a', t - n)$ if a is labeled DELAY(n) and $(a', a) \in R$*
- *$max\ \{r(a', t') : t - n \leq t' \leq t\}$ if a is labeled REVERB(n) and $(a', a) \in R$.*

SPACE nodes are used for modeling spatial memory. They output the value 1 if and only if the input is identical to a certain stored value μ. DELAY nodes and REVERB nodes are used for modeling temporal memory. DELAY nodes delay the signals before releasing them, whereas REVERB nodes make the signals linger on (reverberate). The REVERB label was inspired by reverberation among neuronal pools [14].

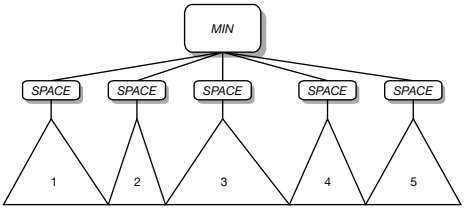

Fig. 3. Snapshot modeling apple taste. The subgraphs 1-5 model sensory organs for the five basic tastes sweetness, sourness, bitterness, saltiness and umami. For instance, subgraph 1 could be the graph of Figure 2. The top node represents a combination of memories of the basic tastes. This network can be used for detecting apple taste.

Definition 5 (Imaginary activity). *First we define an auxiliary function p, which will be used for keeping track of probabilities. Let* $p : D \times T \to V$ *be defined by* $p(b, 0) = 0$ *and*

$$p(b, t+1) = p(b, t) + \frac{r(b, t+1) - p(b, t)}{t+1}.$$

Let $i(a, 0) = 0$ *and* $i(a, t+1) =$

- $min(1, \sum\{r(b, t) \cdot p(b, t) : (a, b) \in R\})$ *if b is labeled MIN, MAX, or AVERAGE.*
- $i(a, t)$ *otherwise.*

Note that imaginary activity is defined in terms of real activity in the past and at present. Also note that imaginary and real activity propagate in opposite directions. The real activity is"mirrored" back in the form of imaginary activity. The definition of imaginary activity was inspired by (i) mirror neurons and (ii)"two-way streets" in cortex [14].

2.4 Memory Structures

Now let us show how spatial and temporal memory can be modeled.

Definition 6 (Memory). *A* memory *of a node a is a node b which is labeled SPACE and satisfies* $R(a, b)$.

Memories can be used for recording and recalling previously perceived values. For instance, the sweetness of a collection of apples can be recorded by a certain memory node and represented by a normal distribution.

Definition 7 (Snapshot). *Let* Ω *be a set of nodes. A* snapshot *of* Ω *is a structure consisting of (i) a memory* a' *of a, for each* $a \in \Omega$*, (ii) a node b labeled MIN, (iii) connections* $R(a', b)$*, for each* $a \in \Omega$.

An example of a snapshot is given in Figure 3.

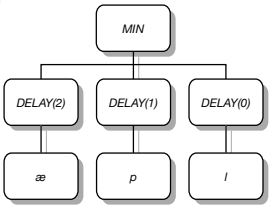

Fig. 4. Episode modeling the spoken word [æpl]. The bottom nodes may either be specialized sensors for the indicated phonemes or structures representing previously learned phonemes. A similar structure could represent the written word "APPLE" or the fact "$6 \cdot 8 = 48$".

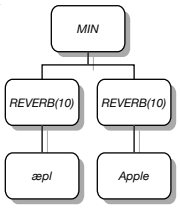

Fig. 5. Network modeling the co-occurrence of the spoken work [æpl] and apple taste. Here the node marked *æpl* could be the top node of Figure 4 and the node marked *Apple* the top node of Figure 3. The MIN node is activated if *æpl* and *Apple* are activated simultaneously modulo 10 time units. Note that real activity in *[æpl]* causes imaginary activity in *Apple* and vice versa.

Definition 8 (Episode). *Let* $a_0, \ldots, a_n$ *be snapshots. An* episode *joining* $a_0, \ldots, a_n$ *is a structure consisting of (i) nodes* $b_0, \ldots, b_n$ *labeled* $DELAY(n), \ldots, DELAY(0)$*, respectively, (ii) a node* c *labeled MIN, (iii) connections* $R(a_i, b_i)$*, for all* $0 \leq i \leq n$ *(iv) connections* $R(b_i, c)$*, for all* $0 \leq i \leq n$*.*

An example of an episode is given in Figure 4. REVERB nodes can be used when the temporal conditions relate to time intervals, as in Figure 5. Note that REVERB nodes can be modeled as a MAX of DELAY nodes.

3 Organisms

Now we shall define the notion of organism and give a basic algorithm for generating organisms.

3.1 Definition

Definition 9 (Organism). *An* organism *is a sequence of TNNs* $(A_t)_{t \in T}$ *such that*

- A_0 *contains no nodes labeled SPACE, DELAY or REVERB.*
- A_0 *is a substructure of* A_i*, for all* i*,*
- *if* $a \in A_i$ *is labeled SENSOR, then* $a \in A_0$*, for all* $i > 0$*.*

A_0 *is called the* genotype *and the* A_i *are called* phenotypes *for* $i > 0$*.*

Let A_0 be any TNN which does not contain any SPACE, DELAY or REVERB nodes. Let A_{t+1} be obtained from A_t as follows.

1. (Update probabilities) Let $p'(a,t)$ be like $p(a,t)$, with the difference that the observations starting at the time when a was created. Then update this function as in Definition 5, mutatis mutandis.
2. (Make deletions) Let $a_1, \ldots, a_k$ be the nodes of $A_t - A_0$ that satisfy $p'(a_i,t) < c$. Here $c \in V$ is a fixed threshold value. Then delete each a_i along with all of its connections.
3. (Make additions) Proceed as follows:
   (a) Case: No complete snapshot is active at t.
      i. Subcase. All maximal nodes have memory nodes that are active at t. (Add snapshot) Then add a complete snapshot by connecting all active SPACE nodes to a MIN-node.
      ii. Subcase (otherwise). Some maximal nodes lack memory nodes that are active at t. Let $a_1, \ldots, a_k$ be all such nodes. (Add memories) Then add memories $b_1, \ldots, b_k$ to $a_1, \ldots, a_k$, respectively. Let the label of b_i be SPACE$(r(a_i,t), 0.25)$.
   (b) Case (otherwise): A complete snapshot is active at t. Then do both of the following.
      i. (Update snapshots) Let $b_1, \ldots, b_k$ be the SPACE nodes that are active at t and let a_i be the unique node satisfying $R(a_i, b_i)$. Suppose the label of b_i is SPACE$(\mu_{i,t}, \sigma_{i,t})$. Then compute the updated parameters $\mu_{i,t+1}$ and $\sigma_{i,t+1}$) by updating the old parameters with respect to the new data points $r(a_i,t)$ by means of Hansen's formula.
      ii. (Add episodes) Suppose there is an episode, which is active at t and joins the complete snapshots $a_1, \ldots, a_n$ (where $n \leq 9$). Then, unless it already exists, add an episode joining $a_1, \ldots, a_n, a$.

Fig. 6. Algorithm for developing organisms automatically. The algorithm uses Hansen's formula [15] for computing the mean and standard deviation incrementally.

Organisms model biological neural networks that develop over time by adding and deleting memory structures (learning and forgetting). Because of the two last conditions of Definition 9, each organism has a fixed set of sensors. Therefore the notion of environment extends to organisms in a straightforward manner.

3.2 Construction Algorithm

Next we shall give an algorithm for developing organisms automatically in a given environment. First we need to introduce some auxiliary concepts: (i) A node a is *active* at t if $r(a,t) \geq 0.95$; (ii) a node $a \in A_0$ is *maximal* if there is no node $b \in A_0$ such that $(a,b) \in A_0$; (iii) a snapshot is *complete* if it joins all maximal nodes (of A_0). The algorithm is given in Figure 3.2. Here are some remarks on the algorithm.

1. The genotype A_0 can be constructed, e.g., by modeling an existing biological or artificial network. It is a *tabula rasa*: a neural structure that has not yet formed any memories.
2. The step *Make deletions* serves the (productive) purpose of preserving memory structures that represent recurrent phenomena, while eliminating those that represent non-repeating coincidences. It was inspired by the forgetting mechanism of natural networks ("use them or lose them"), c.f. the decay theory of synapses [16].
3. The steps *Add snapshot* and *Add episode* were inspired by the Hebbian learning rule ("neurons that fire together wire together") [14].
4. The numerical values appearing in the algorithm can be changed freely. In particular this holds for the start value of standard deviation when only one data-point is available (0.25) and the maximal length of episodes (10).

4 Concept Formation and Reasoning

In this section we analyze the TNN framework from the perspective of concept formation and reasoning.

Concept Formation. The algorithm learns concepts from examples. This holds for sub-symbolic concepts, such as the taste of an apple (cf. Figure 3), and for symbolic concepts, such as the spoken word [æpl] (cf. Figure 4). Snapshots are formed on the basis of one example by means of *Add snapshot* and updated on the basis of similar examples by means of *Snapshot update*.

Concept Deletion. Concepts that are not active frequently enough are deleted by the algorithm. Conversely, repetition will make the concepts stay longer.

Classification. Snapshots and episodes serve as classifiers. In principle, once a pattern of stimuli has occurred and the corresponding snapshot or episode has been formed, it will be recognized every time it is encountered in the future. The robustness of these classifiers is determined by the values of σ of the SPACE nodes, which are in turn determined by experience. Another factor that contributes to robustness is the insensitivity of the functions MIN, MAX and AVERAGE to permutations of the inputs.

Deductive Reasoning. A simple example of deductive reasoning is given in Figure 7, where $6 \cdot 8$ is rewritten to 48. A similar rewrite step occurs when $6 \cdot 8$ appears as a subsequence of a complex expression (since real activity arises in the node $6 \cdot 8 = 48$ whenever the subsequence $6 \cdot 8$ becomes active). In general, deductive reasoning, e.g. arithmetic computations and theorem-proving, can be carried out in the TNN framework as a parallel rewrite process, based on rewrite rules that have been learned previously.

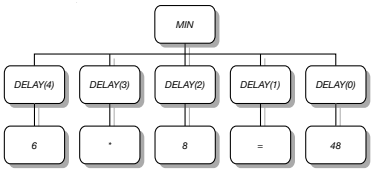

Fig. 7. What is $8 \cdot 6$? Real activity 1.0 in the four leftmost bottom nodes leads to real activity 0.8 in the top node. Imaginary activity then propagates back to all the bottom nodes, including 48.

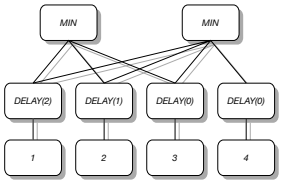

Fig. 8. What comes after 1,2? Suppose the organism has experienced both 1,2,3 and 1,2,4 before, the former more often than the latter. Then the sequence 1,2 will lead to imaginary activity in 3 and to a lesser extent in 4.

Inductive Reasoning. A simple example of inductive reasoning is given in Figure 8, where the sequence 1,2 is being extrapolated. The same mechanism can be used for interpolation, e.g. when reconstructing missing letters in words, both unambiguously as in ZEB?A and ambiguously as in H?T.

Abductive Reasoning. Abductive reasoning (for inferring possible causes) is performed via the interplay between real and imaginary activity. An example is given in Figure 9.

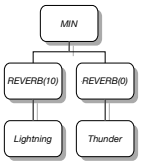

Fig. 9. Thunder appears within 10 time units after Lightning. Real activity in Thunder causes imaginary activity in Lightning and vice versa.

5 Conclusion

We presented a developmental model, which integrates concept formation and basic deduction, induction, and abduction. A version of this model was implemented in the context of an MSc project [17].

References

1. Schmidhuber, J., Thórisson, K.R., Looks, M. (eds.): AGI 2011. LNCS (LNAI), vol. 6830. Springer, Heidelberg (2011)
2. Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., Thelen, E.: Artificial intelligence. Autonomous mental development by robots and animals. Science (291), 599–600 (2001)
3. Zlatev, J., Balkenius, C.: Introduction: Why epigenetic robotics? In: Balkenius, C., Zlatev, J., Kozima, H., Dautenhahn, K., Breazeal, C. (eds.) Proceedings of the First International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems. Lund University Cognitive Studies, vol. 85, pp. 1–4 (2001)
4. d'Avila Garcez, A.S., Lamb, L.C.: Cognitive algorithms and systems: Reasoning and knowledge representation. In: Cutsuridis, V., Hussain, A., Taylor, J.G. (eds.) Perception-Action Cycle. Springer Series in Cognitive and Neural Systems, pp. 573–600. Springer, New York (2011)
5. Harrison, J.: Handbook of practical logic and automated reasoning. Cambridge University Press (2009)
6. Rumelhart, D., McClelland, J.: Parallel distributed processing: Psychological and biological models, vol. 2. The MIT Press (1986)
7. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann (1988)
8. Russell, S., Norvig, P.: Artificial intelligence: a modern approach. Prentice-Hall (2010)
9. Kurup, U., Lebiere, C., Stentz, A.: Integrating Perception and Cognition for AGI. In: Schmidhuber, J., Thórisson, K.R., Looks, M. (eds.) AGI 2011. LNCS (LNAI), vol. 6830, pp. 102–111. Springer, Heidelberg (2011)
10. Gärdenfors, P.: Conceptual spaces. MIT Press (2000)
11. Sloutsky, V., Fisher, A.: Induction and categorization in young children: A similarity-based model. Journal of Experimental Psychology 133(2), 166–188 (2004)
12. Strannegård, C., Ulfsbäcker, S., Hedqvist, D., Gärling, T.: Reasoning Processes in Propositional Logic. Journal of Logic, Language and Information 19(3), 283–314 (2010)
13. Strannegård, C., Amirghasemi, M., Ulfsbäcker, S.: An anthropomorphic method for number sequence problems. Cognitive Systems Research (2012)
14. Baars, B., Gage, N.: Cognition, brain, and consciousness: Introduction to cognitive neuroscience. Academic Press (2010)
15. West, D.H.D.: Updating mean and variance estimates: an improved method. Commun. ACM 22(9), 532–535 (1979)
16. Wixted, J.: The psychology and neuroscience of forgetting. Annu. Rev. Psychol. 55, 235–269 (2004)
17. Olier, J.S.: Transparent neural networks, an implementation. Master's thesis, Chalmers University of Technology (2012)