Active Representations: A Primitive for Analogical Computing in the Brain

Yoonsuck Choe (choe@tamu.edu)
Department of Computer Science, Texas A&M University
3112 TAMU
College Station, TX 77843-3112

Abstract

The static nature of representations (or symbols) makes them require an active interpretational (or computational) mechanism to render them useful or meaningful. To avoid the problems of infinite hierarchy of interpreters, more active approaches have been proposed. These are called active representations (Hofstadter 1985; Mitchell 2001) or active schemas (Narayanan 1999), and neurons can perform analogical or metaphorical tasks. Surprisingly, such active units are very much alike the neurons in our brains, and they can indeed perform analogical tasks. In this paper, a detailed neural mechanism that may be implementing such a function is proposed, and the implications of this new connection between analogy and the neural substrate in building intelligent agents and in understanding the brain function will be discussed.

Introduction

The static nature of representations (or symbols) makes them require a separate, active interpretational (or computational) mechanism to render them useful or meaningful. The difficulty that is caused in such a framework is that of an infinite regress of higher and higher interpretational units. However, if an active role is assigned to the representations, an emergent behavior native to the system of representations can arise not requiring an infinite hierarchy (Hofstadter 1985; Mitchell 2001). A related approach with distributed active schemas resulted in a similar behavior (Narayanan 1999). Central to the function of these active units is the ability to find relations across different domains, often termed *analogy* (Hofstadter 1995) or *metaphor* (Narayanan 1999). In fact, analogy and metaphor are highly inter-related in that they refer to similarities in relations and attributes, but metaphor encompasses a broader spectrum than analogy (Gentner 1989). However, in this paper, I will use the term analogy to focus more on shared *relations* than shared *attributes* (Edelman 1998; Shepard and Chipman 1970).

What is interesting is that such active representations are very much alike the neurons in our brains. It turns out that these *active* neurons as a collection can indeed perform analogical tasks if certain conditions are met. In fact, an exact circuit that may be implementing such a conditional requirement exists in the brain. Cortico-cortical connections together with the thalamo-cortical loop in the brain are ideal for implementing such a constraint. In this paper, I will focus on describing how the neurons can perform rudimentary analogical tasks and what is the role of the thalamus and cortical connections in the process.

Analogy is commonly attributed to higher cognitive faculties only, but it does not always have to be the case as Chalmers et al. (1992) suggested. If this is true, analogy may be part of a larger set of human brain function including perception and motor function, as well as cognition. With this new framework of active representations and analogical process, we can start to take a more focused approach in building intelligent agents and in understanding the nature of brain function.

Neurons as Active Representations

Adopting the active approach for representations, we can think about what kind of unit in the brain can embody such a functionality. It turns out that the neurons can implement such active representations. Instead of focusing on understanding what kind of information the neurons encode and process, we can ask what *action* is taken when they sense a certain feature in the incoming input, be that temporal or spatial. The action performed by neurons is simply *invoking* activity in other neurons (figure 1). In this way, neurons represent a certain input feature, and take immediate action by invoking other neurons once the feature is detected. Thus, in this framework, which neurons are invoked by a single neuron becomes as important as what features it encodes or is sensitive to.

The question then is what kind of general principle can such active neurons implement? Such a unit alone cannot achieve much, neither can a serial chain of such units. The true power of this simple unit is revealed when it is used in a massively parallel way. This may be an obvious line of thought because that is what our brains seem to do. However, it turns out that the collective effort of these simple units can embody a simple yet powerful functional principle of analogy.

We have to simplify matters to see how analogy can be processed by such neurons. Let us assume there are six neurons in an imaginary creature's brain inhabiting the world of fruits (figure 2). After allowing the fruit brain to experience the world of fruits, it will learn the co-occurrences between features and establish relational arrows as shown in the figure (arcs with arrows). Also suppose that the brain

Content Areas: Neural Networks, Perception, Cognitive Modeling, and Analogy ($Revision$: 1.65)

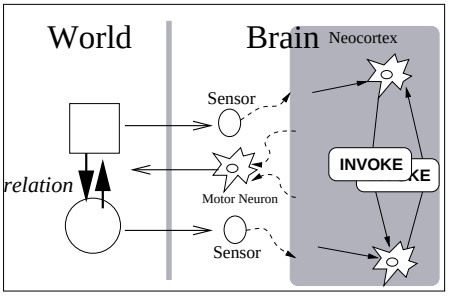

Figure 1: **Active Role of Neurons.** A simplified diagram shows how neurons represent features in the world and actively invoke other neurons. Such invocations establish a relational context among neurons, and thus represent relations between objects and events in the world.

is partitioned into several specialized map areas (or partitions), just like in the real brain. Now, suppose <apple>, <orange>, and <word-red> were presented to the creature simultaneously. If we track the activation, we can see that these detectors will turn on: apple detector, orange detector, color-red detector, color-orange detector, and finally, word-red detector. These activations are *input-driven*. Because the neurons are active, as soon as they detect what they are familiar with, they send out signals through the relational arrows horizontally across the cortex. As a result of this second order activation, the word-orange detector turns on, even without input. Now, here is the crucial moment. We can ask this question: *which neuron's firing was purely cortically-driven?*. Note that this question can be viewed as a filtering process. The result of the filtering is then <word-orange>. The significance of this observation is that this process is very similar to solving analogical problems. The input presented to the creature can be viewed as an analogical query: <apple>:<orange> = <word-red>:<?>. The filtered cortical response <word-orange> can then be the *answer* to this query.[1] Thus, active relations can perform an analogical function when the responses are filtered properly.

However, things can get complicated when combinations of objects are used as a query. Let us extend the creature's feature detectors to include concepts of small and big (not shown in the figure). Then we can allow the creature to learn the relations again. We can then present an analogical query like this: <big><apple>:<small><apple> =<big><orange>:<?>. However, in this case, if we follow the same steps as above, we come across a problem. Because the answer we expect (<small><orange>) already appeared in the query (i.e. they are input-driven), if we look for purely cortically-driven activations, the answer will be <word-red> <word-orange>. However, this problem can be overcome if we ask: *what are the most cortically-driven activities in each partition of the brain*? Because <big> and <apple> appear in the input twice but <small> and <orange> appear only once, the latter two can be selected,

[1]There is an issue of how the presence of <word-red> can affect the outcome at all. This problem will be discussed later in the discussion section.

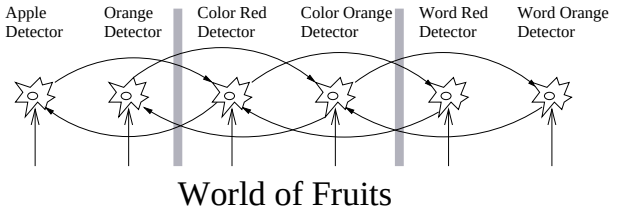

Figure 2: **World of Fruits.** A brain with fruit and color detector neurons is shown. The six neurons each respond to these input features as labeled above. At the bottom is the fruit world, and the thick vertical arrows represent afferent input. The horizontal arcs are the relational arrows that point to their most frequently co-occurring counterparts that have been learned through experience. The gray vertical bars represent the partitioning of the brain into separate map areas (from the left to right, object map, color map, and word map). Note that for simplicity, the word-orange detector connects only to the color-orange detector, but not the orange detector, i.e. it is a word-color-orange detector, not a word-object-orange detector.

as well as the purely cortically driven activities listed above. Thus, even for derived activities that are input-driven, those that are less input-driven can survive and the correct analogical response can still be found among such activities that are more cortically-driven within each partition. Note that <color-orange> also survives the filtering, but what is more important here is that a simple filtering process as described above can provide potential answers to analogical queries.

In this section, I have shown that active neurons that encode input features and relational context can collectively perform rudimentary analogical functions.[2] But does the brain function in such a way? In fact, an exact circuit that can serve such a function exists in the brain.

Neural Basis of Analogical Completion and Filtering

Two basic neural mechanisms are needed to account for the proposed analogical function: completion and filtering. Below, I will discuss how the cortico-cortical connections and thalamo-cortical loop can implement these two mechanisms.

The completion may be accomplished by the long-range cortico-cortical connections (Mumford 1992). As mentioned earlier, synapses are strengthened when the presynaptic activity precede postsynaptic activity (Song et al. 2000), thus the connections can implement causal relations. Also, specific patterns of connections observed in animals (e.g. in the primary visual cortex of monkeys; Blasdel 1992) show how such patterns can implement specific completion functions. Computational models also showed how such connectivity patterns can encode feature co-occurrence and how they can affect the performance of the model (Bednar and Miikkulainen 2000; Choe 2001; Geisler et al. 2001).

In the thalamo-cortical loop, there exists a massive feedback from the cortex to the thalamus and an inhibition mechanism within the nucleus reticularis thalami (nRt) on the

[2]Analogical tasks can become much more complex than the ones shown here.

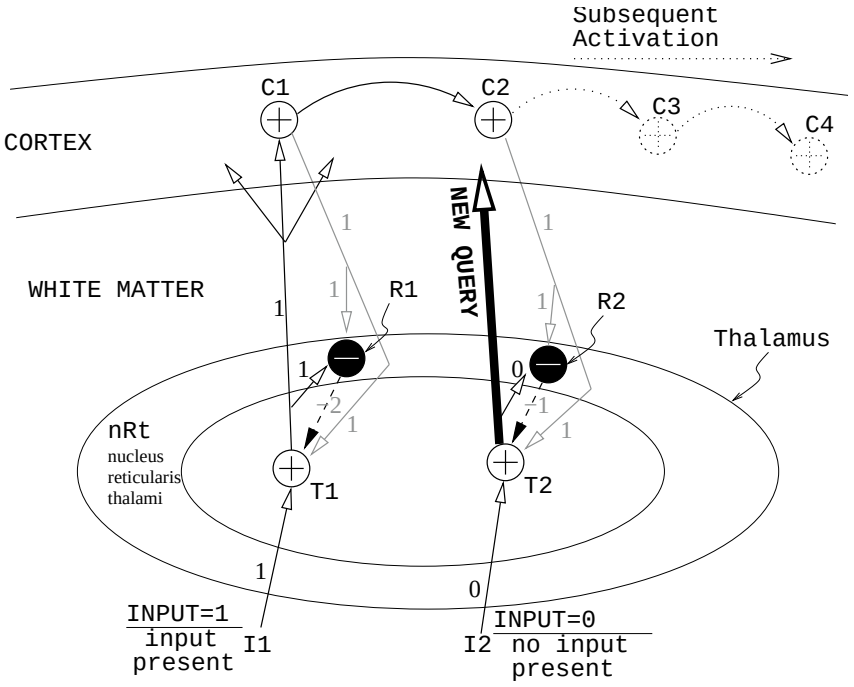

Figure 3: **Analogical Filtering in the Thalamus.** The diagram shows a simplified thalamo-cortical loop that can perform analogical completion and selection, and propagate the selection back to the cortex. I1 and I2 are input fibers, T1 and T2 are thalamic relay cells, R1 and R2 are inhibitory nRt cells, and C1, C2, C3, and C4 are cortical neurons (neurons in multiple layers of the cortex are shown as a single unit). The neurons are either excitatory (+) or inhibitory (–), and the arrows are axons (pointing in the direction of action potential propagation). The numbered labels on each arc show the activity being carried. Black solid arrows are ascending fibers to the cortex and cortico-cortical connections (relational arrows), and gray solid arrows are cortico-thalamic feedbacks. Dashed arrows are inhibitory. The diagram shows a scenario when input was presented to C1, which excites C2, and in turn generates the feedback from C2 to T2, then retransmitted to the cortex as a new query (ascending thick black arrow). The selection decision for further propagation to the cortex depends on the relative excitation and inhibition T1(T2) receive from C1(C2) and R1(R2). On the right of C2 (dotted) in the cortex is the subsequent cascade of analogical completions. Note that to avoid clutter, reciprocal connections in the cortex, as well as disinhibiting connections within the nRt layer are not shown. All connections shown are based on known anatomy of the thalamus and the cortex (Mumford 1995).

surface of the thalamus (Mumford 1995). This particular architecture has been thought to be involved in the analysis and synthesis of new memories (MacKay 1956), active blackboard (Harth et al. 1987; Mumford 1991), global workspace (Newman et al. 1997), and finally, generating attention and consciousness (Crick and Koch 1990).

It turns out that these feedforward and feedback connections from nRt to the cortex together with the nRt inhibitions can filter feedbacks from the cortex to promote the most cortically-driven feedback, i.e. the analogical completions. Let us first see how the purely cortically-driven activities are selected (figure 3). In the thalamus, ascending fibers (T1 to C1) branch out and excite the inhibitory nRt neuron R1 (T1 to R1). When the feedback from C1 to T1 comes back, it branches and stimulates R1. As a result, if the descending feedback had a matching ascending signal, the inhibition T1 receives is twice as high as other neurons in the thalamus that are activated by purely cortically-driven feedback that came around the first time (T2). If the synaptic weights are appropriate (i.e. $w_{\mathrm{TC}} = 2$ and $w_{\mathrm{TR}} = 1$)[3], at T1 the feedback will be canceled out, but at T2 the feedback will survive the inhibition and be retransmitted to the cortex (the *new query* arrow). Such a surviving cortical feedback, together with the input stimulus at the next moment form a new analogical query to the cortex, and the same process is repeated. That is, C2 elicits activities in C3, and in turn C4 through the thalamo-cortical loop (note that they can be quite far away). For the selection of the *most cortically driven* feedback, the mutual inhibitions in the nRt layer (e.g. between R1 and R2) can be used to disinhibit (inhibiting an inhibitory neuron results in less net inhibition) each other and allow the more cortically driven feedback to go back to the cortex, even when there are paired inputs to all current cortical activity.

[3]Here, w_{YX} is the synaptic connection strength from neuron X to neuron Y.

Discussion

The neural mechanisms described in this paper can only account for simple kinds of analogies, and in some case it can even seem as simple pattern completion. For example, <orange> = ? will result in the same answer <word-orange> as in the *Active Neurons:...* section. How can the term <word-red> in the original query affect the outcome at all? For this, I believe that among many possible completions, the general map area (i.e. the partitions in figure 2)

that are activated by input gets higher preference. In this example, the fruit-map, word-map and color-map will turn on, thus purely cortical activations in other general maps (say odor-map, etc.) will not be as salient as that of <word-orange>. Thus, in this way, the presence of <word-red> can indeed affect the outcome of the analogical query. A more precise neural mechanism for this kind of selection needs to be investigated further.

Researchers regard analogical capacities as the crux of high-level cognition (see Gentner et al. 2001 for a collection of current work on analogy). However, analogy does not need to be limited to high-level cognition. Recent results suggest that analogy may be needed in perception as well (Davis and Goel 2001; Morrison 1998), and may even be a crucial requirement for cognitive development (Chalmers et al. 1992). Then it is not unthinkable that the motor functions also obey the general principle of analogy in a similar manner, thus we can then start to understand perception, cognition, and motor functions under the unifying framework of analogy instead of trying to understand those as embodying separate functional principles.

How can such a diverse functionality be integrated under the general principle of analogical processing? Massive connections exist within and across different functional areas in the brain, and the sensory/motor maps are topologically organized, i.e. nearby neurons are responsive to nearby features in the input space (Kohonen 1982; von der Malsburg 1973). Within each map, the feature detectors and cortico-cortical connections learn to encode the relations (Choe 2001; Sirosh et al. 1996). It is possible that cognitive maps also have a topological organization where nearby areas learn to encode similar concepts, such as semantic maps or episodic memory maps (Miikkulainen 1993), or even temporal sequences (James and Miikkulainen 1995). When the sensory, cognitive, and motor maps are connected in an orderly way preserving their local topology, analogies within and (more importantly) *across* different domains can be drawn.

Within this huge number of maps specializing in different tasks, a cascade of multiple analogical completions can be going on in parallel, synchronized at each moment by the 40Hz rhythm to hold an instantaneously coherent state (Mumford 1995). Such state can then pose as another analogical query, and that process can repeat. When that cascade reaches a motor area, behavior will be generated. Memory content can also enter the analogical cascade, and this quasi-static contribution can prevent the continuously changing input stream from causing random cascades, thereby maintaining a more goal-directed and stable behavior. Specific mechanisms of how the memory content enter the thalamo-cortical loop, and how completed analogies are archived in memory through the interactions with subcortical centers such as the hippocampus should be studied further.

Such an integrative view of perception, cognition, and motor function under the general principle of analogy can become a powerful tool in building intelligent agents. In fact, the virtual agent developed by Morrison (1998) is a concrete example of such an idea. Morrison showed that a virtual agent with integrated perceptual, cognitive, and motor functions was able to embody a rich structure in the virtual environment, using analogical primitives. What is more interesting is that he showed that the representations in the agent can become more complex if the environment is made more complex. It is possible that if such an agent is physically implemented, it could learn more complex relations in the environment and show much more complex behavior. Thus, a promising direction for the future is to employ the analogical framework in embodied robotics research. Instead of direct mapping from sensor to motor, we can put an intermediate stage (cognition) and design the basic functions according to the analogical framework.

Neuroscience research has revealed a lot about perception and motor abilities in the brain, but understanding the cognitive faculty still remains elusive. Investigation into cognitive functions can be done under the analogical framework, where we can infer the functionality of the higher areas by backtracking the connections to the perceptual and motor areas and study their topology and analogical links. Specific predictions regarding the layout of the higher centers can be made based on the topology of the lower centers and connection structure between the two, and experiments can then focus on verifying these predictions. For example, there are orientation maps with smoothly changing orientation preference in V1 (primary visual cortex; Blasdel 1992), and there are object maps in TE (temporal area E; Tanaka 1996) that also change smoothly (for example, rotation of a head). My theory predicts that there will be a mapping from V1 to TE that preserve such local topology across different representation spaces.[4] Similar mappings may exist between sensory and cognitive areas, and if such a mapping is found, we can start to understand the abstract cognitive functions based on concrete perceptual architecture. The advantage of this theory is that it enables us to relate perceptual, cognitive, and motor functions in a unified framework. By encompassing all aspects of brain function under an analogical framework, more focused experiments can be designed to reveal specific analogical capacities, and this will help us better understand the brain function and the cognitive process. In turn, these new understandings can be utilized in building intelligent agents.

Conclusion

In this paper, I adopted the view of active representations, and observed that our neurons are no different, and they should be understood in a more active context. It turned out that collectively they can perform an *analogical function.* Specific circuits in the brain was found to be suitable for implementing such a function, thus providing further support for analogy as a general computational principle in the brain. In this new framework, the specific targets the neurons excite become as important as how they interpret incoming input. This new framework can help us take a more focused approach in building intelligent agents and in understanding

[4] Although the pathway from V1 to TE is not direct, involving V2, V4, and TEO areas, but successive mappings within this path can reveal how V1 and TE are topologically mapped.

the nature of brain function.

Acknowledgments

I am greatly indebted to Sang Kyu Shin, James Bednar, Risto Miikkulainen, Un Yong Nahm, Marshall Mayberry, Bruce H. McCormick, and Jyh-Charn Liu for their feedback and encouragement. This research was supported in part by Texas A&M University, and the Texas Higher Education Coordinating Board ARP/ATP program under grant #000512-0217-2001.

References

Anderson, J. A., and Rosenfeld, E., editors (1988). *Neurocomputing: Foundations of Research*. Cambridge, MA: MIT Press.

Bednar, J. A., and Miikkulainen, R. (2000). Tilt aftereffects in a self-organizing model of the primary visual cortex. *Neural Computation*, 12(7):1721–1740.

Blasdel, G. G. (1992). Orientation selectivity, preference, and continuity in monkey striate cortex. *Journal of Neuroscience*, 12:3139–3161.

Chalmers, D. J., French, R. M., and Hofstadter, D. R. (1992). High-level perception, representation, and analogy. *Journal of Experimental and Theoretical Artificial Intelligence*, 4:185–211.

Choe, Y. (2001). *Perceptual Grouping in a Self-Organizing Map of Spiking Neurons*. PhD thesis, Department of Computer Sciences, The University of Texas at Austin, Austin, TX. Technical Report AI01-292.

Crick, F., and Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in The Neurosciences*, 2:263–275.

Davis, J., and Goel, A. K. (2001). Visual analogy in problem solving. In *Proceedings of the International Joint Conference on Artificial Intelligence*, to appear.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*, 21:449–498.

Geisler, W. S., Perry, J. S., Super, B. J., and Gallogly, D. P. (2001). Edge Co-occurrence in natural images predicts contour grouping performance. *Vision Research*. 711–724.

Gentner, D. (1989). The mechanisms of analogical learning. In Vosniadou, S., and Ortony, A., editors, *Similarity and Analogical Reasoning*, 199–241. New York, NY: Academic Press.

Gentner, D., Holyoak, K. J., and Kokinov, B. N., editors (2001). *The Analogial Mind: Perspectives from Cognitive Science*. Cambridge, MA: The MIT Press.

Harth, E., Unnikrishnan, K. P., and Pandaya, A. S. (1987). The inversion of sensory processing by feedback pathways: A model of visual cognitive functions. *Science*, 237:184–187.

Hofstadter, D. (1985). Waking up from the boolean dream, or, subcognition as computation. In *Metamagical Themas*, chapter 26. New York, NY: Basic Books.

Hofstadter, D. G. (1995). *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. Basic Books (a division of Harper Colins).

James, D. L., and Miikkulainen, R. (1995). SARDNET: A self-organizing feature map for sequences. In Tesauro, G., Touretzky, D. S., and Leen, T. K., editors, *Advances in Neural Information Processing Systems 7*, 577–584. Cambridge, MA: MIT Press.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

MacKay, D. (1956). The epistemological problem for automata. In Shannon, C. E., and McCarthy, J., editors, *Automata Studies*, 235–251. Princeton, NJ: Princeton University Press.

Miikkulainen, R. (1993). *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory*. Cambridge, MA: MIT Press.

Mitchell, M. (2001). Analogy-making as a complex adaptive system. In Segal, L., and Cohen, A., editors, *Design Principles for the Immune System and Other Distributed Autonomous Systems*, to appear.

Morrison, C. T. (1998). *Situated Representation: Solving the Handcoding Problem with Emergent Structured Representation*. PhD thesis, Bringhamton University; State University of New York.

Mumford, D. (1991). On the computational architecture of the neocortex, pt. I, the role of the thalamo-cortical loop. *Biological Cybernetics*, 65:135–145.

Mumford, D. (1992). On the computational architecture of the neocortex, pt. II, the role of the cortico-cortical loop. *Biological Cybernetics*, 65:241–251.

Mumford, D. (1995). Thalamus. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, 153–157. Cambridge, MA: MIT Press.

Narayanan, S. (1999). Moving right along: A computational model of metaphoric reasoning about events. In *Proceedings of the National Conference on Artificial Intelligence (AAAI '99, Orlando, FL)*, 121–128. AAAI Press.

Newman, J., Baars, B. J., and Cho, S.-B. (1997). A neural global workspace model for conscious attention. *Neural Networks*, 10:1195–1206.

Shepard, R. N., and Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, 1:1–17.

Sirosh, J., Miikkulainen, R., and Bednar, J. A. (1996). Self-organization of orientation maps, lateral connections, and dynamic receptive fields in the primary visual cortex. In Sirosh, J., Miikkulainen, R., and Choe, Y., editors, *Lateral Interactions in the Cortex: Structure and Function*. Austin, TX: The UTCS Neural Networks Research Group. Electronic book, ISBN 0-9647060-0-8, http://www.cs.utexas.edu/users/nn/web-pubs/htmlbook96.

Song, S., Miller, K. D., and Abbott, L. F. (2000). Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience*, 3:919–926.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19:109–139.

von der Malsburg, C. (1973). Self-organization of orientation-sensitive cells in the striate cortex. *Kybernetik*, 15:85–100. Reprinted in Anderson and Rosenfeld 1988.