Cognitive Learning and the Multimodal Memory Game: Toward Human-Level Machine Learning

Byoung-Tak Zhang

***Abstract*— Machine learning has made great progress during the last decades and is being deployed in a wide range of applications. However, current machine learning techniques are far from sufficient for achieving human-level intelligence. Here we identify the properties of learners required for human-level intelligence and suggest a new direction of machine learning research, i.e. the cognitive learning approach, that takes into account the recent findings in brain and cognitive sciences. In particular, we suggest two fundamental principles to achieve human-level machine learning: continuity (forming a lifelong memory continuously) and glocality (organizing a plastic structure of localized micromodules connected globally). We then propose a multimodal memory game as a research platform to study cognitive learning architectures and algorithms, where the machine learner and two human players question and answer about the scenes and dialogues after watching the movies. Concrete experimental results are presented to illustrate the usefulness of the game and the cognitive learning framework for studying human-level learning and intelligence.**

I. Introduction

Much knowledge has accumulated with respect to the cognitive substrate for human intelligence and brain during the last several decades. However, artificial intelligence research did not pay much attention to the cognitively-plausible ways for building intelligent machines. It is only recently that people start to address more ambitious problems based on these scientific data. These include the several groups working toward human-level intelligence, i.e. creating high-performance machine intelligence based on the cognitive substrate of human intelligence [2], [4], [7], [8], [11]. However, there is not much work, especially in computational intelligence, emphasizing the role of learning and memory in building human-level machine intelligence.

In this paper we argue that without "cognitive" learning the goal of achieving human-level intelligence is far from being completed. As McGaugh nicely put, we are, after all, our memories:

- It is our memory that enables us to value everything else we possess. Lacking memory, we would have no ability to be concerned about our hearts, achievements, loved ones, and incomes. Our brain has an amazing capacity to integrate the combined effects of our past experiences together with our present experiences in creating our thought and actions. This is all possible by the memory and the memories are formed by the learning process. (from [12])

Thus, learning and memory is a fundamentally function in realizing human-level intelligence. Much, if not most, of human intelligence lies in memory mechanisms and the memories are formed by the learning process. Thus, understanding and use of the architecture and mechanism of human cognitive learning is essential for constructing synthetic intelligence systems, as detailed in Section II below.

We identify two principles of cognitive learning which are ubiquitously used by humans but not machines (Section III). One principle involves with the temporal aspect of learning: learning is a continual, incremental process lasting the whole life of an individual. We call this the continuity principle. The other principle involves with the spatial aspect of learning: learning is the formation and adaptation of global and local representations combined in a memory system. We will refer to this as the glocality principle. We argue that learning machines should be built upon these principles to achieve human-level intelligence.

To test the principles we develop an experimental platform for cognitive learning in Section IV. The platform serves as a game consisting of two humans and one machine learner (Fig. 1). The game is played in a multimedia movie theater. The goal of the learner is to imitate the players by watching the movies, reading the captions, and observing the players playing the games. There are two types of questions and both are cued memory-recall tasks [3], [6]. In one type, the players are shown a scene of the movie and they are supposed to remember the captions or describe the scene in text. In the other type, the players read the captions of the movie without images and they have to find the images corresponding to the captions.

Many variations of this basic game are possible to study the architectures and algorithms for cognitive learning as well as various tasks and strategies in the multimodal "learning by viewing" framework. In Section V we present some experimental results on image-to-text and text-to-image versions of the memory game.

The main contribution of this paper is two-fold. First, we identify two principles for cognitive learning that are a basis of any cognitive learning machines aiming at human-level intelligence. Second, we propose a concrete experimental platform to develop the architectures and algorithms for cognitive machine learning. We discuss in Section VI the long-term research issues and short-term applications of the multimodal memory game and the cognitive learning approach.

Biointelligence Laboratory (http://bi.snu.ac.kr/), School of Computer Science and Engineering and Graduate Programs in Cognitive Science, Brain Science, and Bioinformatics, Seoul National University, Seoul 151-744, Korea. E-mail: btzhang@bi.snu.ac.kr

978-1-4244-1821-3/08/$25.00©2008 IEEE

II. Intelligence and Memory in Humans

What are the fundamental properties of human intelligence especially in comparison to machines? Humans are creative, compliant, attentive to change, resourceful, and able to take a variety of circumstances into account [14]. In comparison to machines, however, humans are imprecise, sloppy, distractable, emotional, and illogical.

Here we focus on the following three properties

- resourcefulness
- situatedness
- integration

and then discuss how these are related to human memory and learning.

One of the distinguishing features of human intelligence is the versatility or resourcefulness [13]. Humans can come up with many new ideas and solutions to a given problem. Machines are good at solving a specific problem that they are designed to deal with. But, they are brittle outside the scope of the problem domain. Humans are not just reactive, they are proactive and imaginative. To be resourceful it is important to make associations or finding the relationship between two or more seemingly distant concepts. Imagination requires recall memory [6].

Human intelligence is developed situated in multimodal environments. It is interesting to note that situated or embodied intellects may solve the problem easier than isolated or disembodied intellects. This seems counter-intuitive since, for machines, embodiment or more modalities may mean more noisy data and more computing power. However, intelligent behavior in a disembodied agent requires a tremendous amount of knowledge, lots of deep planning and decision making, and efficient memory storage and retrieval [14]. When the agent is tightly coupled to the world, decision making and action can take place within the context established by the sensory data from the environment, taking some of the memory and computational burden off the agent.

The human mind makes use of a variety of representations and problem-solving strategies. This is why human decision making is robust in a wide range of domains. One important issue in achieving human-level intelligence is how to integrate the multiple tasks into a coherent solution [10]. This is an important part of the binding problem [18]. Memory plays an important role in this regard. We are our memories. Our records of our personal past are essential in enabling us to act, make decisions, and survive. All of our knowledge of our world, and our skills in living in it, are based on memories of our experiences.

In the discussion above, we noted that memory forms the substrate for intelligent behavior in humans. In particular, we see that the three properties regarding intelligence, i.e. resourcefulness, situatedness, and integration, are related with the three aspects of memory and learning, i.e.

- recall
- multimodality
- binding

respectively.

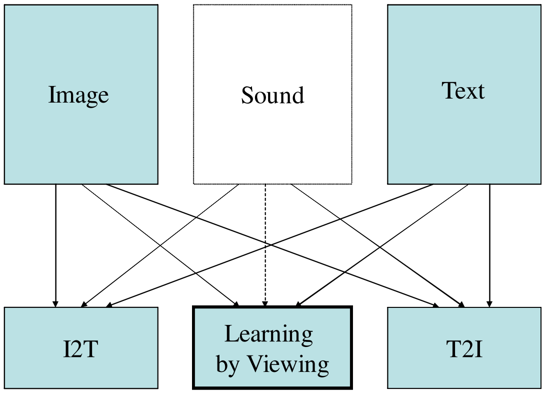

Fig. 1. The multimodal memory game environment. Two humans (I2T and T2I) and a machine learner are situated in a multimedia theater. The learner acquires visuolinguistic knowledge by watching the movies and by observing the human players play image-and-text memory games.

III. Cognitive Learning

In this section we review previous work on human learning and machine learning and then propose two principles that distinguish them in the most fundamental way.

Learning in humans can be characterized in several ways. One classical, but still effective theory says that humans use three different types of learning: accretion, tuning, and restructuring [15].

- Accretion is the accumulation of facts like in learning new vocabulary or the spelling of already known word.
- Tuning is a slow process of practice requiring prolonged, laborious effort. It takes thousands of hours to become an expert pianist or soccer player.
- Restructuring is to form the conceptual structures. It requires exploration, comparison, and integration of concepts.

Current machine learning techniques do not fully exploit these types of learning. Neural networks and connectionist models focus on the tuning mode of learning. Most learning algorithms do not support the accretion mode of learning either. Learning is typically defined as a function approximation problem from a fixed set of training points [5]. The real accretion, however, in human learning is a long-lasting process of accumulation of knowledge. The worst part of existing machine learning architectures is the lack of restructuring capability. Neither the kernel machines nor the graphical models offer any effective methods for incremental assimilation of new facts [1]. Though there were some conceptual learning systems proposed decades ago, but they did not survive because of their poor performance.

What are the fundamental principles underlying human cognitive learning? Aristotle identified three laws of association, in effect, three laws of learning: similarity, contrast, and contiguity [3]. Interestingly, most of the current machine learning algorithms are based on the principles of similarity and contrast. Unsupervised learning methods are algorithms to search for similarities between objects.

Supervised learning methods, especially classification algorithms, tries to maximize the contrast between the classes. Contiguity, especially temporal contiguity, is not exploited in machine learning, except some aspects of it is considered in reinforcement learning [5].

James Mill (1773-1836) made one of the most systematic statements of the empiricist position of learning. He suggested three criteria for the strength of associations: permanence (resistance to forgetting), certainty (confidence), and spontaneity (reaction time). James Mill was also concerned with how simple ideas get combined into complex ideas. His notion was one of *mental compounding*, where the complex idea was unitary but made up of a conglomerate of simple ideas. John Stuart Mill (1806-1873), the son of James Mill's, developed the theory of mental compounding in to *mental chemistry*, a term for a complex idea that originally derived from constituent simple ideas but which was qualitatively different from the sum of the simple constituents. In 1970s Tulving proposed the encoding specificity principle which is quite similar to John Stuart Mill's in that the association of two ideas results in a unique combination that may render the constituents unrecognizable by themselves.

Importantly, the old idea of mental compounding or mental chemistry is not reflected in machine learning research so far. Connectionist models or neural networks are not suitable to model this kind of learning. Mental chemistry requires building blocks or modules that can be combined, which is not possible by weight adjustment alone. There has been some work on learning by symbolic composition in 1980's but no methods survived the 1990s. This pure symbolic approach does not assure predictive accuracy from a training set. It should also be noted that the popular learning algorithms, such as kernel machines and graphical models, are not appropriate to simulate the mental chemistry either. It will be a challenge to have a learning machine that shows high performance and can simulate "cognitive chemistry", or cognitive learning.

Based on the review above, we propose two principles that best distinguish the human learning and the machine learning. These are the principle of continuity and the principle of glocality.

- The principle of continuity. The experiences of each immediately past moment are memories that merge with current momentary experiences to create the impression of seamless continuity in our lives [12]. Memory is the consequence of learning from an experience, i.e., the consequence of acquiring new information. As noted before, Aristotle already identified this principle. Continuity is important in learning in dynamic environments [16].
- The principle of glocality. There is a long-lasting debate about the localized and distributed (or global) representations of information in the brain. The brain consists of functional modules or microcircuits [9] which are specialized to perform some specific cognitive functions, such as locomotion, olfaction, vision, and language, but a particular cognitive function, say language, involves with a large number of specialized microcircuits. Therefore, the cognitive functions of the brain make use of both global and local, i.e. glocal, representations.

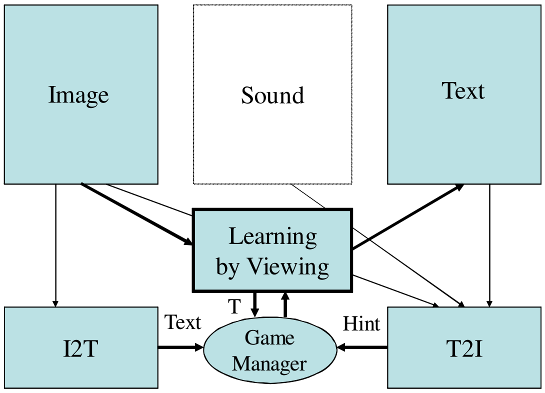

Fig. 2. Text generation game. Given an image (query), the machine learner generates a text `T`. The player I2T (human student) also generates a `Text`. In this mode of game, the player T2I plays the role of the teacher (or oracle) to give a `Hint` to the student I2T. The machine learner learns by viewing the training session `(T, Text, Hint)`. In this simplified version, the sound modality is not used for the machine learner (but it is used for the teacher).

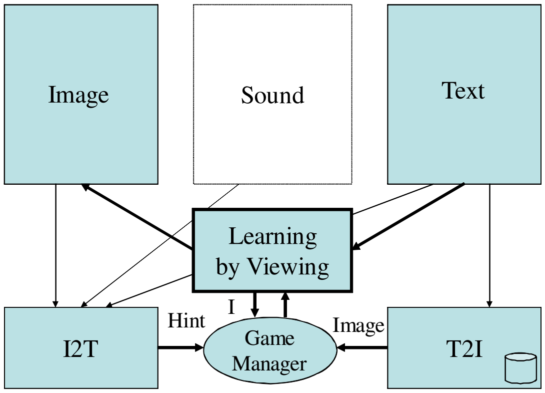

Fig. 3. Image generation game. Given a text (query), the machine learner generates an image `I`. The player T2I (human student) also generates an `Image`. In this mode of game, the player I2T plays the role of the teacher (or oracle) to give a `Hint` to the student I2T. The machine learner learns by viewing the training session `(T, Text, Hint)`. In this simplified version, the sound modality is not used for the machine learner (but it is used for the teacher).

From the machine learning point of view, it is important to maintain local and distributed representations in a single model. Localized modules are essential to add or accumulate new facts. These are especially necessary if the learning system is situated in a changing environment and to track the change as a function of time. Global or distributed representation is necessary to form a generalized representation in the memory. Existing machine learning algorithms are based on a single representation, i.e. each of them is either a localist or a globalist representation. Exceptions are hierarchical and ensemble approaches. To be useful as a cognitive learning model it is essential to have both of these properties. The diversity of this representation was also emphasized by many AI experts, including Minsky [13].

How do we develop learning algorithms based on these principles? As discussed in Section II it is important for the agent to be situated and get multimodal sensory inputs from the environment to evolve into human-level intelligence. Thus, a cognitively-plausible learning agent should be situated in a multimodal environment. In the following section we design a memory game as a research platform for studying cognitive learning that combines situatedness with contiguity and glocality.

IV. The Multimodal Memory Game

Knowing the principles is one thing and building a machine based on the principles is another. Therefore, we have designed a research platform in which the principles for cognitive learning can be explored by varying the architectures and algorithms as well as tasks and strategies. The game consists of two humans and a machine learner (Fig. 1) in a digital cinema. First, the learner learns by watching the movie. The two humans also watch the movie. After watching, the two humans play the game by questioning-and-answering about the movie scenes and dialogues. The task of one player, called I2T (for image-to-text), is to generate a text given a movie cut (image). The other player, named T2I (for text-to-image), is to generate an image given a text from the movie captions. While one player is asking, the other is answering. The two players alternate their roles. When the player is in the questioning mode, he receives all the multimodal inputs. When the player is in an answering mode, he receives the multimodal inputs except the modality in which he has to give an answer. The goal of the learner in this "multimodal memory game" is to imitate the human players by watching the movies, reading the captions, listening the sounds, and observing the players enjoying the games over the shoulder.

This "learning by viewing" is a real challenge for current machine learning technology. The basic framework can be adapted to make the task easier for the cognitive learner without loosing the spirit of the original challenge. The platform can also be used to develop short-term applications of practical interest. For example, if the game is to generate a text from an image (a picture description task), the I2T (answerer or student) is given only the image while the T2I (questioner or teacher) is given the image with the text (Fig. 2). If the game is to generate an image from a text (the image search mode), the roles of the two players are reversed (Fig. 3). The players get points if the answerer gives the right answer which is evaluated by the questioner based on his full multimodal inputs. In case the answerer gives a wrong answer the questioner can provide a hint. The points are discounted if the answer is given with hints. Note that the two human players are playing games to generate teaching signals for the machine learner. The game can be played as long as the human players have fun with watching the movies and scoring the points.

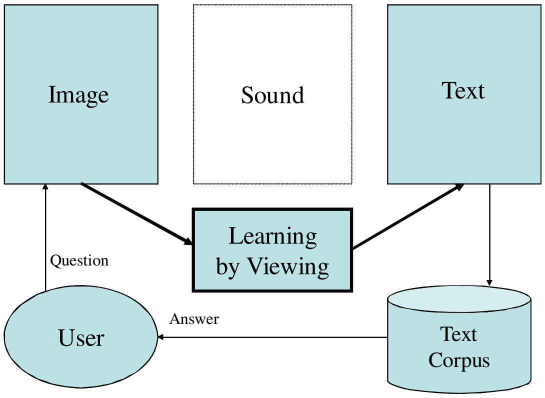

Fig. 4. Text description application. Given an image query, the learner is to generate an answer in text that describes the image. A text corpus might be used to edit the results of the linguistic recall memory of the learner. In effect, the learner produces a text description of the given image.

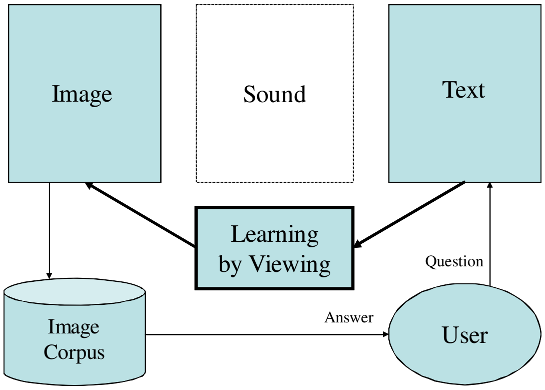

Fig. 5. Image search application. Given a text query, the learner is to find an image that best matches with the text. An image corpus might be used to filter the results of the visual recall memory of the learner. In effect, the learner retrieves an image corresponding to the text query.

Why is the game platform appropriate to study cognitive machine learning toward human-level intelligence? The multimodal memory game meets the two important criteria for human-level machine learning. First, the movies can be played as long as time allows (evaluation of continuity). In addition, the lifespan or difficulty of the continuity can be controlled by increasing or decreasing the movie length and scenarios. Second, the game involves the vision and language problems, the hardest core of human-level intelligence. Any solution to these problems will involve the representation problems, especially the global and local representations and their combinations (evaluation of glocality).

The learned models can be used for recognition and recall memory tasks involved with multimedia data. Thus, for recognition, the model can be used for text retrieval by image (Fig. 2) or, also, for image retrieval by text (Fig. 3). For cued recall tasks, the model can be used to generate natural language from images and to generate images from linguistic descriptions. The system can be used to study the linguistic memory and the visual memory.

The basic modes of the game can be modified to make the task harder or easier and to scale the problem size up or down.

- New modality. Additional modalities can be added. For example, if a haptic sensor is available it might be incorporated in to the memory model.
- More players. More players can participate in the game. This will increase the diversity and reliability of the answers. This may make the learner learn faster and more consistent.
- New media. The media contents can be extended. We can use the web UCC documents instead of the DVD videos. In the domain of education, the documents can be educational material for the school children.
- Gaming strategy. More variability can be added to increase the fun or duty of the participants to the games. For example, the gamers are allowed to watch the next scene only if they have passed the tests for the previous scenes.

By modifying the parameters above, the platform can be tuned for long-term human-level performance studies as well as for short-term applications of practical interest.

V. Experimental Results

We used the multimodal game platform to evaluate the "learning by viewing" paradigm for cognitive learning. We collected DVD videos for the TV dramas. These include "Friends", "Prison Break", "House", and "24". From a corpus we extract images and the corresponding captions in text. The image data are preprocessed to generate a visual vocabulary V_I. The text data are converted to a linguistic vocabulary V_T. The words in V_I and V_T can be considered as primitive concepts or features for building higher-level concepts or cognitive models of the multimodal sensory information.

In this illustrative experiment, we used 112 video cuts and the same number of caption sentences. The vocabulary sizes for cognitive memory were $|V_I| = 80 \times 60$ and $|V_T| = 1000$. The preprocessed image and text data are used to train the learner. We used the hypernetwork model [17] as the memory structure of the cognitive learner. Mathematically, the hypernetwork represents the joint probability of image and text, i.e.

$$P(I,T) = P(i_1, ..., i_n, t_1, ..., t_m). \tag{1}$$

Then the goal of the learner is to approximate this probability to perform the two tasks: generating a sentence (text) T given an image I and generating an image I given a sentence T. To do this the learner computes the conditional probability

$$P(I|T) = \frac{P(I,T)}{P(T)} = \frac{P(i_1, ..., i_n, t_1, ..., t_m)}{P(t_1, ..., t_m)} \tag{2}$$

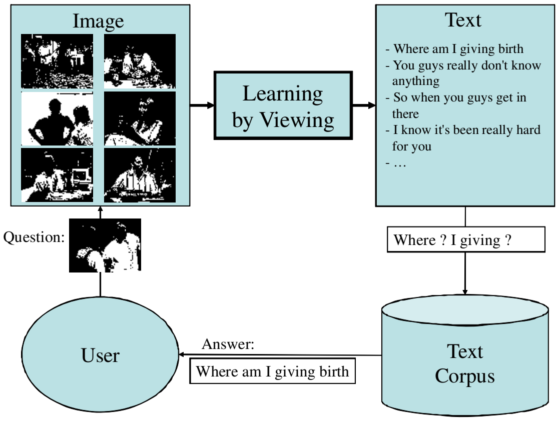
Image
Learning by Viewing
Text
- Where am I giving birth
- You guys really don't know anything
- So when you guys get in there
- I know it's been really hard for you
- …
Question:
Where ? I giving ?
User
Answer:
Where am I giving birth
Text Corpus

Fig. 6. Generating a text from an image.

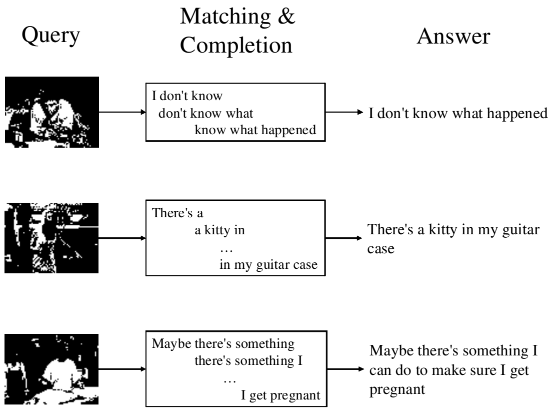

Fig. 7. From image to text: results

for solving the text-to-image (T2I) problem or the conditional probability

$$P(T|I) = \frac{P(I,T)}{P(I)} = \frac{P(i_1, ..., i_n, t_1, ..., t_m)}{P(i_1, ..., i_n)} \tag{3}$$

for solving the image-to-text (I2T) problem.

For a given data set $D = \{\mathbf{x}^{(n)} | n = 1, ..., N\} = \{(i_1, ..., i_n, t_1, ..., t_m)^{(n)} | n = 1, ..., N\}$ consisting of the text and image pairs, the hypernetwork approximates the probability

$$\begin{aligned} P(D|W) &= \prod_{n=1}^{N} P(\mathbf{x}^{(n)}|W) \\ &= Z(W)^{-1} \prod_{n=1}^{N} \exp\{-E(n)\}, \end{aligned} \tag{4}$$

where $E(n) = \sum_{k=1}^{K} \frac{1}{k!} \sum_{i_1, i_2, ..., i_k} w^{(k)}_{i_1 i_2 ... i_k} x^{(n)}_{i_1} x^{(n)}_{i_2} ... x^{(n)}_{i_k}$ and $Z(W)$ is the normalization term. The parameters W for the hypernetwork model consist of a collection of hyperedges

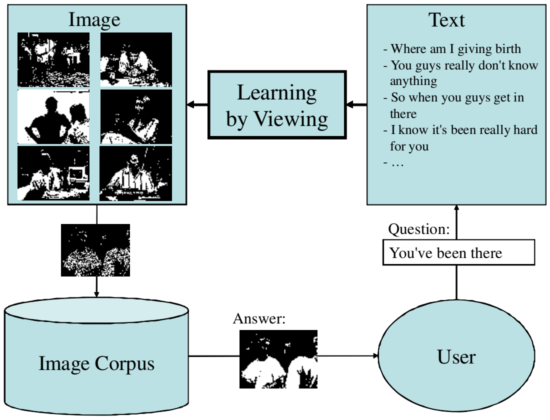
Image
Text
- Where am I giving birth
- You guys really don't know anything
- So when you guys get in there
- I know it's been really hard for you
- …
Learning by Viewing
Question:
You've been there
Answer:
Image Corpus
User

Fig. 8. Generating an image from a text.

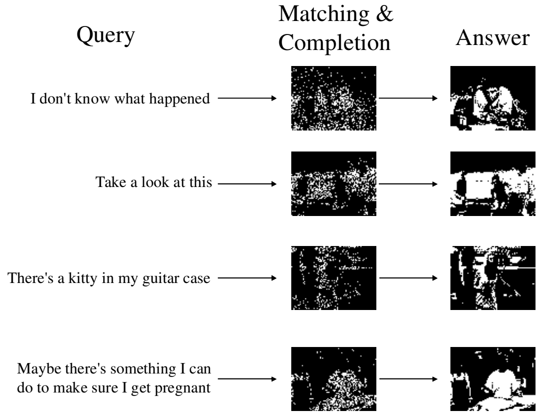
Query
Matching & Completion
Answer
I don't know what happened
Take a look at this
There's a kitty in my guitar case
Maybe there's something I can do to make sure I get pregnant

Fig. 9. From text to image: results

or higher-order features. For a fuller treatment of the hypernetwork model and its learning algorithm can be found in [17].

The hyperedges in the hypernetwork model for the multimodal memory game are conjunctive compositions of the visual and linguistic words. Smaller composition of words represent global hyperfeatures while larger compositions represent more specialized, local hyperfeatures. Combining both types realizes the principle of glocality. The hypernetwork learning process starts with a large number of random compositions of the hyperedges and "evolves" the best-fitting compositions by randomly sampling and selecting them from the existing and new observations (text-image pairs) in sequence. This incremental restructuring or "self-organizing" process reflects the continuity principle since the new or frequently-observed data strengthen the hypernetwork memory, while the trace of the old data fades away as time goes on (unless they are observed again). More discussion on the importance of the shorter-term adaptability and the longer-term persistency in intelligent systems can be found in [16].

The first task is, given an image query, for the learner to generate a text sentence using the hypernetwork memory (Fig. 6). Fig. 7 shows the results of the I2T experiments. The images in the middle column shows the sentence fragments recalled by the hypernet. The final column shows the sentence reconstructed by the fragment sentences. 98 % of the original sentences were reproduced from the image and the hypernet for this particular set of training samples.

The second task is, given a text query, for the learner to generate an image using the hypernetwork memory (Fig. 8). The results are shown in Fig. 9. The images in the middle column shows the images reconstructed from the hypernet. The final column shows the images retrieved using the reconstructed image as a query to the image database. It can be observed that the reconstructed images resemble very much the original video cuts. The learning-by-viewing approach achieved 97 % of correct retrieval. It will be interesting to see how well (or how bad) the system performs (degrades) as we increase the size of the video corpus.

VI. Conclusion

We have proposed a new direction of machine learning research for achieving human-level artificial intelligence. Human-level intelligence requires a system to reproduce the resourcefulness or versatility of the brain and mind. Human-level AI systems should be built on an architecture that is situated in an environment and which can integrate multimodal data. Based on the findings in cognitive neuroscience, we identify contiguity and glocality as the fundamental principles and characteristics of learning and memory underlying human intelligence. To test these principles we developed an experimental platform for cognitive learning as a multimodal memory game. The game has been used in our experiments to develop and explore the algorithms and architectures for building the text-to-image search and the image-to-text generation systems by learning-by-viewing the movies. The game platform can be adapted to deal with many cognitive learning problems of practical interest, including HCI, multimedia analysis, UCC mining, lifelog management, and cognitive robotics. Cognitive learning algorithms deal with dynamic environments with lifelong multimodal interaction and, thus, are contrasted with the conventional machine learning algorithms that assume learning as a static function approximation problem based on a fixed data set.

Acknowledgements

The author would like to thank Chan-Hoon Park, Ha-Young Jang, Minoh Heo, and Young-Jin Park for simulations. This research was supported by MOST (NRL 2002-2007), MICE (MEC 2000-2009), and BK21-IT.

References

[1] Bishop, C. M., *Pattern Recognition and Machine Learning*, Springer, 2006.

[2] Cassimatis, N. S. "A cognitive substrate for achieving human-level intelligence," *AI Magazine*, pp. 45-56, Summer 2006.

[3] Crowder, R. G. *Principles of Learning and Memory*, Lawrence Erlbaum, 1976.

[4] Duch, W., Oentaryo, R.J., and Pasquier, M., "Cognitive architectures: where do we go from here?," *First Conf. on Artificial General Intelligence*, University of Memphis, March 1-3, 2008.

[5] Duda, R. Hart, P., and Stork, D., *Pattern Classification*, Wiley, 2000.

[6] Eichenbaum, H. *The Cognitive Neuroscience of Memory*, Oxford University Press, 2002.

[7] Forbus, K. D. and Hinrichs, T. R. "Companion cognitive systems: A step tward human-level AI," *AI Magazine*, pp. 83-95, Summer 2006.

[8] Goertzel, B. and Pennachin, C. (Eds.) *Artificial General Intelligence (Cognitive Technologies)*, Springer-Verlag, 2005.

[9] Grillner, S. and Graybiel, A. M. (Eds.), *Microcircuits: The Interface between Neurons and Global Brain Function*, The MIT Press, 2006.

[10] Jones, G. F. Integrated intelligent knowledge management. In *Achieving Human-Level Intelligence through Integrated Systems and Research: Papers from the AAAI 2004 Fall Symposium*, 2004.

[11] Langley, P. "Cognitive architectures and general intelligent systems," *AI Magazine*, pp. 33-44, Summer 2006.

[12] McGaugh, J. L. *Memory & Emotion: The Making of Lasting Memories*, Columbia University Press, 2003.

[13] Minsky, M. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*, Simon & Schuster, 2006.

[14] Norman, D. *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*, Perseus, 1993.

[15] Rumelhart, D. E. and Norman, D. A. "Accretion, tuning and restructuring: Three modes of learning, In J. W. Cotton and R. Klatzky (Eds.), *Semantic Factors in Cognition*, Lawrence Erlbaum, 1978.

[16] Zhang, B.-T. "Random hypergraph models of learning and memory in biomolecular networks: shorter-term adaptability vs. longer-term persistency," *The First IEEE Symposium on Foundations of Computational Intelligence* (FOCI '07), pp. 344-349, 2007.

[17] Zhang, B.-T. and Kim, J.-K., "DNA hypernetworks for information storage and retrieval," *Proc. 2006 Int. Annual Meeting on DNA Computing* (DNA12), LNCS 4287:298-307, 2006.

[18] Zimmer, H. Z., Mecklinger, A., and Lindenberger, U. "Levels of binding: types, mechanisms, and functions of binding in memory, In Zimmer, H. Z., Mecklinger, A., and Lindenberger, U. (Eds.), *Binding and Memory*. Oxford, UK: Oxford University Press.