Pororobot: Child Tutoring Robot for English Education

Kyung-Min Kim[1], Chang-Jun Nan[1], Min-Oh Heo[1], and Byoung-Tak Zhang[1,2]
[1]School of Computer Science and Engineering / [2]Cognitive Science and Brain Science Programs
Seoul National University
Seoul, Korea
{kmkim, cjnan, moheo, btzhang} @bi.snu.ac.kr

***Abstract*—The recent success of machine learning has lead to advancements in robot intelligence and human-robot interaction. It is reported that robots can well understand visual scene information and describe the scenes in language using computer vision and natural language processing methods. Image Question-Answering (QA) systems can be used for human-robot interaction. However, to achieve human-level artificial intelligence based on lifelong learning, a model must deal with real-world environments including dynamic, uncertain, and asynchronous properties based on lifelong learning. In this paper, we propose a prototype system for a video QA robot "Pororobot". The system uses the state-of-the-art machine learning system using dual memory model for implementing robot intelligence. In our scenario, a robot and a child plays a video QA game. Here we demonstrate preliminary results of our system.**

Keywords—videoQA; tutoring robot; human-robot interaction; robot intelligence; machine learning;

I. Introduction

Solving question and answering (QA) problem has been an important theme in artificial intelligence discipline and many computational models have been proposed for few decades. However, in order to achieve human-level artificial intelligence, the models must deal with real-world environments including the dynamic, uncertain, and asynchronous properties based on lifelong learning [1]. Thus, to tackle video QA problem, it is necessary to consider that the concepts in the video dynamically change according to the story. Because of this property, it is hard to answer the questions with only generalized knowledge in a learned model. The model needs to remember not only generalized information but also precise information in short-term context like working memory in human brain.

Here we develop methods i.e. dual deep-learning architectures, for video QA based on two Deep Concept Hierarchies (DCH) [2]. DCH can effectively represent multimodal concepts and efficiently captures the conceptual changes from incrementally observed data. In the architecture, there are two memories that the one learns short-term information with a fixed structure and small memory space and the other learns long-term information with a more flexible structure and large memory space. Thus, they can learn both short-term context and long-term context in the video. As the video stories unfold, each memory independently learns the observed video and have different emergence and evolution phase.

The proposed method is evaluated on a cartoon video series 'Pororo' consisting of approximately 200 episodes with the total playing time of 1232 minutes. We evaluate the model with 100 QA pairs from the video and 20 human evaluators.

II. Video QA System Overview

In our scenario, a child and a robot see the video and they interact each other with the same experience, i.e. watching a video. This could be advantageous for children's early education [3]

To make this possible considering that the video data are continuous, dynamic, and multimodal, implementing a video QA system requires some issues as follows:

1) Multiple time resolution knowledge
2) Flexible and generalized knowledge representation using multimodal features
3) Continuous knowledge acquisition and update

For solving three issues, we implement a knowledge representation methods based on concepts for enhancing the generalization. Concepts represented with multimodal variable are incrementally learned from continuously increasing data. Especially, as we consider a dynamic environment where children ask questions while observing cartoon videos, we place the two knowledge base in our system.

Figure 1 shows an overview of our video QA system. The QA system can automatically generate the questions from the observed video and answer the questions. Also, because of the

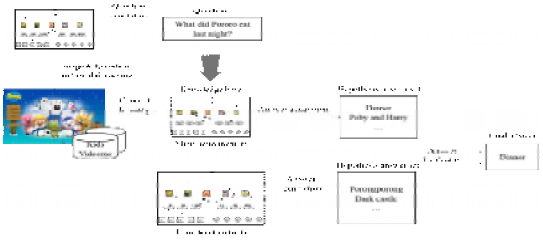

Figure 1: Overview of Video QA System

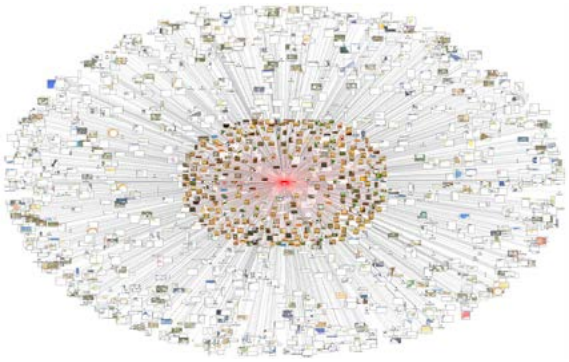

Figure 2: Multimodal Semantic Networks

dynamic, asynchronous properties of the video, the model need to handle the short-term context and long-term context of the video story. The system learns the video with concept learning methods and there are two memories in which process observed video data of different time resolution. To generate questions from the video, the system has additional QA data for the video and learns the pattern of the image-question pairs. Generating question is similar to machine translation problem. To generate answers for the questions, it is needed to search both the short-term and long-term memory.

III. Experimental Results

A. Cartoon Video and QA Dataset

Cartoon videos are a popular material for early language learning for children. They have a succinct and explicit story, which is represented with very simple images and easy words [3]. These properties allow the cartoon videos to be a test bed material suitable for a video question & answering played by a child and a robot [3]. For the experiment, we use a cartoon video 'Pororo' with 1200 question & answer pairs from '*Pororo*' video. To preprocess the video, we convert the video into scene-subtitle pairs. Whenever the subtitle appears in the video, the scene at that time is captured. Each scene is converted to a set of image patches using Regions with Convolutional Neural Networks (R-CNN) [4], and the patch is represented by a 4096-dimensional fc7 feature of CNN using the Caffe implementation. In this work, we use Word2vec to encode the words

B. Network Construction

As the video stories unfold, the short-term memory and the long-term memory independently learn the observed video and thus show different emergence and evolution phases in the entire networks. Figure 2 shows that there are two area in the networks. The center area with red edges and nodes indicates the knowledge learned from the short-term memory and the outer area with gray edges and nodes indicates the knowledge learned from the long-term memory. The content of nodes related with the short-term memory continuously change during the learning process within the fixed area size while the content of nodes related with long-term memory rarely do not change within the expanded area size. We fixed the size of the short-term memory to 300 microcodes in this work.

Dataset	Video Turing Test		
	Pass	Fail	Pass Rate(%)
Pororo	313	647	32.60
MaisyABC	127	231	35.47

Table 1: The Video Turing Test Results of the Retrieved Answers

C. Evaluation

To evaluate our model, we use our model to retrieve 80 answers from 80 questions. Then, twelve human evaluators judge the results whether the answers are given by a machine or a human. Each answer is presented with the question and a sequence of seven images related to the question in order to give information about the video story to the evaluators. Additionally, to test the expandability of our model, we apply the model to another cartoon video set '*MaisyABC*' and evaluate the results. The results are shown in Table 1. It shows that 32.60% and 35.47% of the answers for each video are treated as answers from a human.

IV. Conclusion and Discussion

We proposed a deep learning- architecture of video question & answering for human-robot interaction. We demonstrate that the proposed architecture have both short-term and long-term contexts in the video and can generate questions and retrieve answers appropriately. We evaluate our methods using the real Turing Test. In future work, we plan to use much larger datasets including TV dramas and movies with more complex stories. Also, the system should be expanded to be a "purposive" or "intentional" agent. The system should be able to decide which observed data needs to be more focused on and which questions should be generated. These creative properties are necessary for lifelong learning environments [5].

References

[1] Zhang, B.-T. 2013. Information-Theoretic Objective Functions for Lifelong Learning. *AAAI 2013 Spring Symposium on Lifelong Machine Learning*. 62-69.

[2] Ha, J.-W., Kim, K.-M., and Zhang, B.-T. 2015. Automated Visual-Linguistic Knowledge Construction via Concept Learning from Cartoon Videos. *In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015)*. 522-528.

[3] Kim, K.-M., Nan, C.-J., Ha, J.-W., Heo, Y.-J., and Zhang, B.-T. 2015. Pororobot: A deep learning robot that plays video Q&A games, *AAAI 2015 Fall Symposium on AI for Human-Robot Interaction*

[4] Girshick, R., Donahue, J., Darrell, T., Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. pages 580-587, Columbus, Ohio, USA, June 2014.

[5] Zhang, B.-T. 2014. Ontogenesis of agency in machines: A multidisciplinary review. *AAAI 2014 Fall Symposium on The Nature of Humans and Machines: A Multidisciplinary Discourse.*