Artificial Intelligence for Human-Robot Interaction
Papers from the AAAI 2015 Fall Symposium

Pororobot: A Deep Learning Robot that Plays Video Q&A Games

Kyung-Min Kim[1], Chang-Jun Nan[1], Jung-Woo Ha[2], Yu-Jung Heo[1], and Byoung-Tak Zhang[1,3]

[1]School of Computer Science and Engineering & [3]Institute for Cognitive Science, Seoul National University, Seoul 151-744, Korea
[2] NAVER LABS, NAVER Corp., Seongnam 463-867, Korea
{kmkim, cjnan, yjheo, btzhang}@bi.snu.ac.kr, jungwoo.ha@navercorp.com

Abstract

Recent progress in machine learning has lead to great advancements in robot intelligence and human-robot interaction (HRI). It is reported that robots can deeply understand visual scene information and describe the scenes in natural language using object recognition and natural language processing methods. Image-based question and answering (Q&A) systems can be used for enhancing HRI. However, despite these successful results, several key issues still remain to be discussed and improved. In particular, it is essential for an agent to act in a dynamic, uncertain, and asynchronous environment for achieving human-level robot intelligence. In this paper, we propose a prototype system for a video Q&A robot "Pororobot". The system uses the state-of-the-art machine learning methods such as a deep concept hierarchy model. In our scenario, a robot and a child plays a video Q&A game together under real world environments. Here we demonstrate preliminary results of the proposed system and discuss some directions as future works.

Introduction

Human-robot interaction (HRI) is an emerging field which makes an effort to create socially interactive robots that help humans in various aspects, including healthcare (Fasola and Mataric 2013), and education (Kory and Breazeal 2014). Now, robots can interact with children and help their educational development (Saerbeck et al. 2010; Kory et al. 2013; Fridin 2014), and personalized tutor robots have shown to notably increase the effectiveness of tutoring (Leyzberg et al. 2014). Particularly, recent advancement in machine learning enables robots to deeply understand visual scene information and describe the scene in natural language as compared to that of humans (Fang et al. 2015; Karpathy and Fei-Fei 2015; Vinyals et al. 2015). Many studies on image question & answering (Q&A) have also shown successful results (Gao et al. 2015; Malinowski et al. 2015; Ren et al. 2015) and improve the level of HRI. However, in order to interact with a human in an effective manner, a robot agent must deal with real-world environments including dynamic, uncertain, and asynchronous properties based on lifelong learning (Zhang 2013). In this paper, we propose a prototype system using state-of-the-art deep learning methods for a child-robot interaction scenario. The deep concept hierarchy (Ha et al. 2015) is used as a knowledge base and deep learning methods including deep convolutional neural networks (CNNs) (Krizhevsky et al. 2012) and recurrent neural networks (RNNs) (Mikolov et al. 2013; Gao et al. 2015) are used for representing features of the video data and generate answers from the questions. The scenario is based on a video Q&A game under the real world environment where a child and a robot asks questions on the cartoon video contents and answers these questions to each other. As a robot platform, we use the Nao Evolution V5. Nao can synthesize speech from the text and recognize the human voice. For preliminary experiments, we generate questions from the video and the results determined by a BLEU score and human evaluators show that the questions are somewhat appropriate to be raised to the child and the robot. Furthermore we discuss some directions of future work for achieving human-level robot intelligence.

Video Q&A Game Robots

Video Q&A game robots could be a favorable technology to improve children's early education in two aspects. First, both a child and a robot can learn new concepts or knowledge in the video during the question & answering game. For example, the robot can learn an unknown fact or knowledge by asking questions to the child and vice versa. Second, the robot can leverage children's social abilities. The robot and the child have the same experiences, i.e. watching a video, and interact with each other with that experience. Here is an example scenario of a child-robot

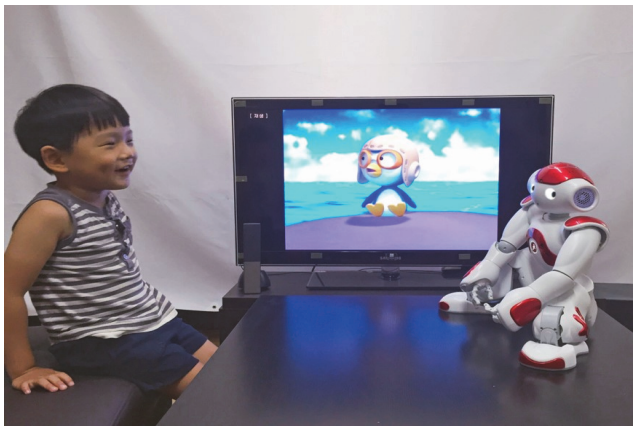

Figure 1: A Video Q&A Game Robot "Pororobot"

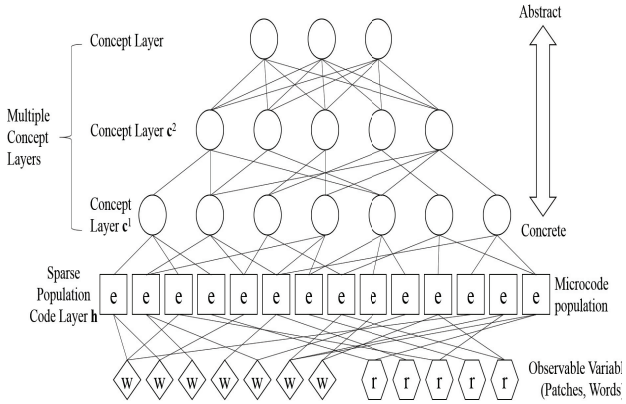

Figure 3: The Structure of Deep Concept Hierarchies

interaction whilst playing the video Q&A game. A child and a robot sit in front of a television and watch a cartoon video together for 10~15 minutes together. After watching this, the robot asks some questions to the child regarding the story (e.g. "First questions, what did they do last night?"). The robot should be able to generate questions from the observed video. If the child answers correctly, the robot agrees with the answer and gives the next question (e.g. the robot says while clapping "Correct! Second question, what is the color of the sky?"). As the game goes on, the child can learn more concepts from the video with the robot. Figure 1 shows an example of a video question & answering game played by a child and a robot "Pororobot".

Video Q&A System

The overall process of the video question & answering is described in Figure 2. The system has four parts. (i) a preprocessing part for feature generation consisting of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). (ii) deep concept hierarchies for learning visual-linguistic concepts and storing the knowledge from the video into a network structure. (iii) a vision-question conversion part, and (iv) a candidate microcode extraction component using RNNs. An RNN is used to generate the next word in the answer and the method is similar to the current image question & answering techniques (Gao et al. 2015).

The Four Parts of the System

(i) The preprocessing part converts a video into a set of visual-linguistic features. Firstly, the video is converted to scene-subtitle pairs. Whenever the subtitle appears in the video, the scene at that time is captured. Also, each scene is converted to a set of image patches using Regions with Convolutional Neural Network (R-CNN) features (Girshick et al 2014). Each patch is represented by a CNN feature, which is represented by a 4096 dimensional real-value vector using the Caffe implementation (Jia 2013). A subtitle is converted to a sequence of words and each word is represented by a real-valued vector using RNN. In this work, we use the word2vec to encode the words (Mikolov et al. 2013).

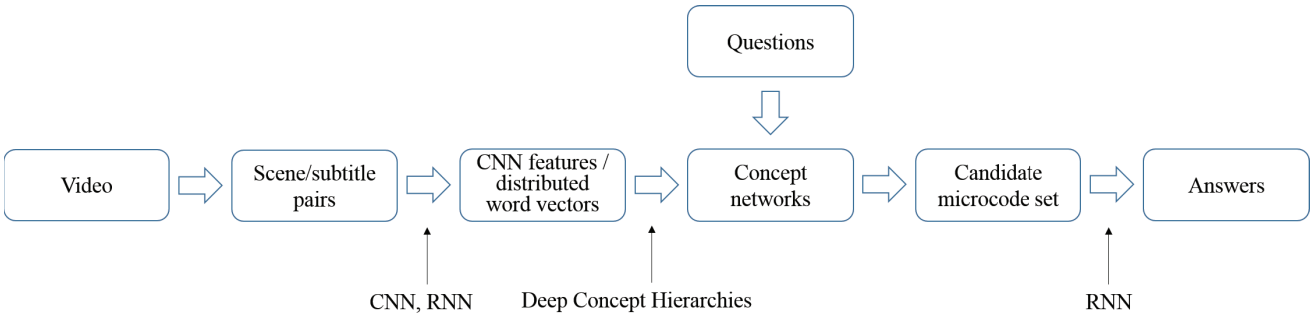

Figure 2: Diagram of the Video Question & Answering Process

(ii) The objective of the second part is to learn visual-linguistic knowledge, i.e. concepts or stories, in the video and store the knowledge into a network structure. For learning concepts, we use a recently proposed concept learning algorithm, deep concept hierarchies (DCH). The overall structure of DCH is described in Figure 3. The DCH contains a large population of microcodes in the sparse population code (SPC) layer (layer **h**) which encodes the concrete textual words and image patches in the scenes and subtitles of the video (Zhang et al 2012). A flexible property of the model structure allows the model to incrementally learn the new conceptual knowledge in the video (Ha et al 2015). A node in concept layer 1 (layer $\mathbf{c}^1$) is a cluster of microcodes in layer **h** and as the model continuously observes the video, the number of nodes in $\mathbf{c}^1$ dynamically change according to the similarities of the microcodes in the cluster. A node in concept layer 2 (layer $\mathbf{c}^2$) represents an observable character in the video. The number of $\mathbf{c}^2$ nodes matches the total number of characters and the connections between the $\mathbf{c}^1$ and $\mathbf{c}^2$ layer are constructed according to the appearance of the characters in $\mathbf{h}^m$. The weight of the connection between the node in $\mathbf{c}^1$ and the node in $\mathbf{c}^2$ is proportional to the frequency of appearance of the character in $\mathbf{h}^m$. DCH stores the learned concept into a network which becomes a knowledge base. The knowledge base is used to generate questions for the given video clip in the third part and give a hypothesis microcodes set which is input to the fourth part.

(iii) The third part converts scenes to questions using DCH. The conversion is similar to a machine translation problem. For this conversion, we make an additional question & answering dataset and convert it into scene-question pairs. Another DCH is used to learn the patterns between scenes and questions and generate questions from the images. The vision-question translation can be formulated as follows.

$$q^* = \arg\max_q P(q \mid \mathbf{r}, \theta) = \arg\max_q P(\mathbf{r} \mid \mathbf{q}, \theta) P(\mathbf{q}, \theta), \quad (1)$$

where θ is a DCH model and q^* are the best questions generated from the model.

(iv) The fourth component contains a candidate microcode extraction component and a RNN. The extraction component receives a question Q and selects a hypothesis microcode set m.

$$m = \arg\max_i s_m(Q, E_i), \quad (2)$$

E_i is the i-th microcode in the DCH and s_m is a selection function. s_m converts Q into a set of distributed semantic vectors with the same embedding vector space used in the preprocessing part and computes the cosine similarity between Q and E_i. The selected hypothesis microcode set is then fed into the RNN to generate answer.

$$a = \arg\max s_a(Q, [E_1 ... E_M], w), \quad (3)$$

	BLEU Score (with 1-gram precision)	Averaged Human Rated Score (from 1 to 5)
Generated Questions from DCH	0.3513	2.534

* A BLUE score and human evaluations are measured on 200 questions and 80 questions respectively

Table 1: Results of the Evaluation for Our Generated Questions

where M is the total number of the selected microcodes and s_a is an answer generation function. In this paper, this function will be similar to recent RNN techniques used in image question & answering problems (Gao et al. 2015].

Experiment Design and Preliminary Results

Cartoon Video and Q&A Dataset Description

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. 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. For the experiment, we use a famous cartoon video 'Pororo' of 1232 minutes and 183 episodes.

Also, we make approximately 1200 question & answer pairs from '*Pororo*' video by making five questions for every five minutes in the video. In detail, there exist two types of questions. One can be answered using the image information only (e.g. "What did *Pororo* and his friends do after eating?", "What did *Eddy* ride?") and the other type needs additional information like subtitles or the story to be answered (e.g. "Why did *Pororo* and his friends go into cave?", "How was *Loopy*'s cooking?").

Robotic Platform

For a robotic platform, the Nao Evolution V5, 58-cm tall humanoid robot is used. Nao is equipped with text-to-speech / face detection software and mainly used in an educational environment.

Video Q&A Game Environment

A child and a robot will play a video question & answering game. The game will be situated in a notebook and the game event can be streamed to the robot. Based on the video contents, the robot asks questions to the child and the child answers as described before. Some simple sound effects or animations may be included.

Question Generation Results

To evaluate the reasonability of the questions generated from DCH, we measure a BLEU score of the questions and conduct human evaluations. Table 1 summarize the results of the evaluation. A BLEU score is typically used in machine translation problems and can range from 0 to 1 (Pap-

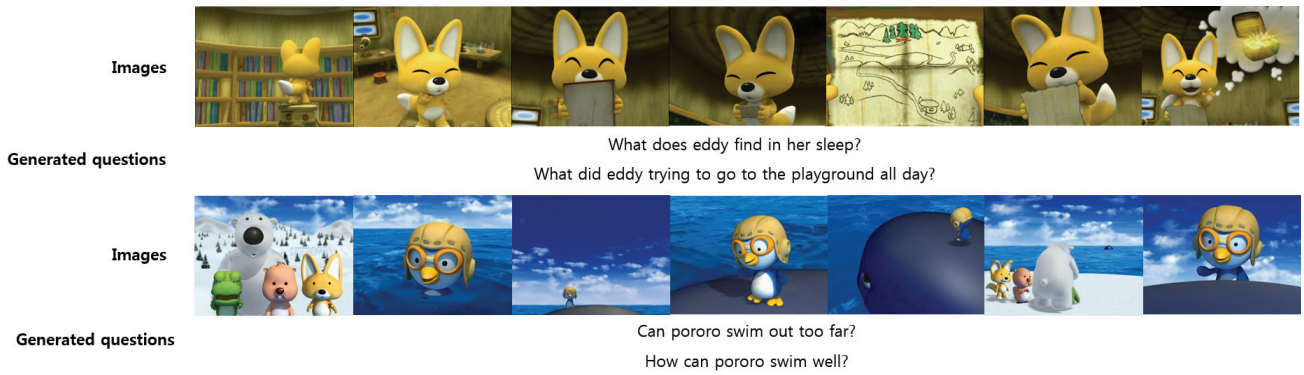

Figure 4: Example Questions Generated from DCH

ineni et al. 2012). Higher scores indicate close matches to the reference translations. In this paper, we match the generated questions to the ground truth questions and the DCH achieves a BLEU score of 0.3513 with 1-gram precision on 200 generated questions. For human evaluations, seven human judges rate the generated questions with scores from 1 to 5. The average score on 80 generated questions is 2.534. The results show that the generated questions are somewhat appropriate to be raised to the child and the robot. Figure 4 shows the examples of the generated questions. A query scene image is positioned in center (4^{th} image) in an image row and images are ordered sequentially as are in the video.

Discussions and Conclusions

We proposed a deep learning-based system for a video question & answering game robot "Pororobot". The system consists of four parts using state-of-the art machine learning methods. The first part is the preprocessing part consisting of the CNNs and RNNs for feature construction. The knowledge on an observed video story is learned in the second part where DCH is used to learn visual-linguistic concepts in the video. The third part converts the observed video scenes to questions and the last part uses the RNN which encodes the words in the answer and generates next answer words. We demonstrate that the preliminary results of the proposed system can allow for making a socially interactive robot which can play with a child and thus build a stepping stone for achieving human-level robot intelligence.

For further research, the robot should be able to generate questions and give a reaction differently according to the child's states like emotions. The state determined by the sensors of the robot may be an important factor to decide which questions should be generated during the child-robot interaction. To this end, the system we described in this paper should be expanded to be a "purposive" or "intentional" agent. These creative modes is prerequisite for life-long learning environments (Zhang 2014). Also, the model should be end-to-end equipped with a large-scale GPU cluster to deal with real-time, dynamic, heterogeneous nature of the real-world data in everyday life.

Acknowledgements

This work was supported by ICT R&D program of MSIP/IITP (R0126-15-1072), In addition, this work was supported in part by ICT R&D program of MSIP/IITP (10044009), and the US Air Force Research Laboratory (AFOSR 124087), and NAVER LABS.

References

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J. C., Zitnick, C. L., Zweig, G. 2015. From Captions to Visual Concepts and Back, *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*. 1473-1482

Fasola, J., and Mataric, M. 2013. A socially assistive robot exercise coach for the elderly. *Journal of Human-Robot Interaction*. 2(2):3-32.

Fridin, M. 2014. Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education. *Computers & Education*. 70(0):53–64.

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W. 2015. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. *arXiv preprint arXiv:1505.05612*.

Girshick, R., Donahue, J., Darrell, T., Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*. 580-587.

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.

Jia, Y. 2013. Caffe: An open source convolutional architecture for fast feature embedding. *http://caffe.berkeleyvision.org/*.

Karpathy, A., Fei-Fei, L. 2015. Deep Visual-Semantic Alignments for Generating Image Description. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*. 3128-3137

Kory, J. M., and Breazeal, C. L. 2014. Storytelling with Robots: Learning Companions for Preschool Children's Language Development. *In Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*.

Kory, J. M., Jeong, S., and Breazeal, C. L. 2013. Robotic learning companions for early language development. *In Proceedings of the 15th ACM on International conference on multimodal interaction*. 71–72.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *In Proceedings of Advances in Neural Information Processing Systems (NIPS 2012)*. 1097–1105.

Leyzberg, D., Spaulding, S., and Scassellati, B. 2014. Personalizing robot tutors to individuals' learning differences. *In Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 423–430.

Malinowski, M., Rohrbach, M., and Fritz, M. 2015. Ask your neurons: A neural-based approach to answering questions about images. *arXiv preprint arXiv:1505.01121*.

Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. 2013. Distributed Representation of Words and Phrases and Their Compositionality. *In Proceedings of Advances in Neural Information Processing Systems (NIPS 2013)*. 3111-3119

Papineni, K., Roukos, S., Ward, T., and Zhu, W. 2002. Bleu: A method for automatic evaluation of machine translation. *In Proceedings of ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. 311-318

Ren, M., Kiros, R., and Zemel, R. 2015. Image question answering: A visual semantic embedding model and a new dataset. *ICML 2015 Deep Learning Workshop*.

Saerbeck, M., Schut, T., Bartneck, C.; and Janse, M. D. 2010. Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1613–1622.

Vinyals, O., Toshev, A., Bengio, S., Erhan, D. 2015. Show and Tell: A Neural Image Caption Generator. *arXiv preprint arXiv:1411.4555*.

Zhang, B.-T., Ha, J.-W., and Kang, M. 2012. Sparse Population Code Models of Word Learning in Concept Drift. *In Proceedings of the 34th Annual Conference of Cognitive Science Society (Cogsci 2012)*. 1221-1226.

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

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*.