Cambot: A Visual Conversation Robot for Interactive Engagement

Kibeom Kim[1,3], Jin-Hwa Kim[2,3], Byoung-Tak Zhang[1,2,3]
Surromind Robotics[1]
Cognitive Science Program[2]
Cognitive Robotics and Artificial Intelligence Center[3]
Seoul National University
Seoul, Republic of Korea
Email: {kbkim,jhkim,btzhang}@bi.snu.ac.kr

***Abstract*—To achieve human-level artificial intelligence, it is crucial to develop algorithms which handle human-like visual and linguistic information. One of promising solutions is to use Multimodal Residual Networks (MRN) for the multimodal residual learning in assumption of visual question-answering tasks. It extends the idea of the deep residual learning, which learns joint representation from vision and language information effectively. While the MRN is handling with multidisciplinary problems of vision, language and integrated reasoning, a visual conversation robot can be a bridge to interact with humans. Cambot can be instantiated in any platform including robots, desktops and tablet PCs, which have a camera and microphone, engaging natural environmental situations of visual conversation for human interactions.**

I. Introduction

We have been seeking human level artificial intelligence (AI) technology. For the technology, the studies in AI have been researched narrow subject like image recognition [1], [2], [3] and language model [4], [5], [6]. But, we need to make a robot which can handle multidisciplinary problems of vision, language and integrated reasoning like a person. Like the cambot, RI-MAN [7] and Sony's robot pet, AIBO [8], are interactive robots. But these don't have visual and linguistic ability. The cambot that use a core technology as Multimodal Residual Networks (MRN) [9] that is great performance at visual question-answering tasks [10] can understand vision and language data. It is to learn multimodality of the tasks exploiting the excellence of deep residual learning [2] and Stacked Attention Networks (SAN) [11]. Fig.1. shows inference flow of the MRN. Therefore, the cambot which is made for basic human level AI robot may help to solve the real world problems.

II. Multimodal Residual Networks

MRN consists of multiple learning blocks, which are stacked for deep residual learning. Denoting an optimal mapping by $\mathcal{H}(\mathbf{q}, \mathbf{v})$, we approximate it using

$$H_1(\mathbf{q}, \mathbf{v}) = W_{q'}^{(1)} \mathbf{q} + F^{(1)}(\mathbf{q}, \mathbf{v}). \tag{1}$$

The first (linear) approximation term is $W_{q'}^{(1)} \mathbf{q}$ and the first joint residual function is given by $\mathcal{F}^{(1)}(\mathbf{q}, \mathbf{v})$. The linear mapping $W_{q'}$ is used for matching a feature dimension. It is defined the joint residual function as

$$\mathcal{F}^{(k)}(\mathbf{q}, \mathbf{v}) = \sigma(W_q^{(k)} \mathbf{q}) \odot \sigma(W_2^{(k)} \sigma(W_1^{(k)} \mathbf{v})) \tag{2}$$

where σ is $\tanh$, and $\odot$ is element-wise multiplication. The question vector and the visual feature vector directly contribute to the joint representation.

For a deeper residual learning, we replace $\mathbf{q}$ with $H_1(\mathbf{q}, \mathbf{v})$ in the next layer. In more general terms, (1) and (2) can be rewritten as

$$H_L(\mathbf{q}, \mathbf{v}) = W_{q'} \mathbf{q} + \sum_{l=1}^{L} W_{\mathcal{F}^{(l)}} \mathcal{F}^{(l)}(H_{l-1}, \mathbf{v}) \tag{3}$$

where L is the number of learning blocks, $H_0 = \mathbf{q}, W_{q'} = \Pi_{l=1}^{L} W_{q'}^{(1)}$, and $W_{\mathcal{F}^1} = \Pi_{m=l+1}^{L} W_{q'}^{(m)}$. Notice that the shortcuts for a visual part are identity mappings to transfer the input visual feature vector to each layer (dashed line). At the end of each block, it is denoted H_l as the output of the l-th learning block, and $\oplus$ is element-wise addition.

III. Cambot

Here, we suggest a cambot as a platform for visual conversation. It consists of three parts, such as a robot hardware, a web server and a visual question-answering (VQA) server. It is illustrated on fig.2. The robot hardware is equipped with

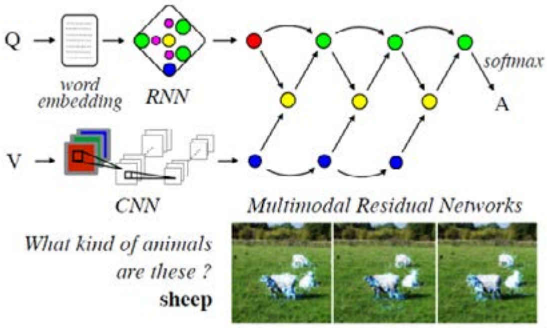
Q
word embedding
RNN
softmax
A
V
CNN
Multimodal Residual Networks
What kind of animals are these ?
sheep

Fig. 1. Inference flow of Multimodal Residual Networks [9].

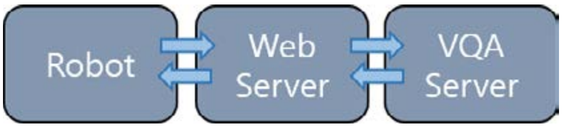

Fig. 2. Three parts of the Cambot.

microphones, cameras and speakers like a human. From the microphones and cameras, the robot gets audio signals and images from the human. If a person ask a question to the robot, the Cambot take a picture and the audio signals, and the signal is converted to Korean text. We used the Google Speech Recognition APIs for converting the signal to the text. Then, the data sends to web server that is second part. If the Cambot will speak when returned data is received from web server. Fig.3. is a scene depicting a conversation that a man ask a question and the Cambot has an answer.

The web server which is made for working on any platforms like desktops and tablet PCs connects the robot hardware and the VQA server. In this part, a sentence of input data is translated Korean into English via the Japanese using the MS Bing translator APIs, as the VQA server is learned English. And the data from the third part is translated English into Korean. If the input sentence translates Korean into English directly, the translation is not clean. But Japanese and Korean is better to translate because the sentence structure is similar. Furthermore, translation of Japanese and English are cleaner than Korean and Japanese because of the large amount of data.

The last part is the VQA server. The server which is using the MRN is received the input data that is an English sentence and image. After process, output data will return to web server as an answer. The MRN is learned the VQA dataset [10] which are collected via Amazon Mechanical Turk from human subjects, who satisfy the experimental requirement. It includes 614,163 questions and 7,984,119 answers, since ten answers are gathered for each question from unique human subjects. The images come from the MS-COCO dataset [12], 123,287 of them for training and validation, and 81,434 for test. The images are carefully collected to contain multiple objects and natural situations, which is also valid for visual question-answering tasks.

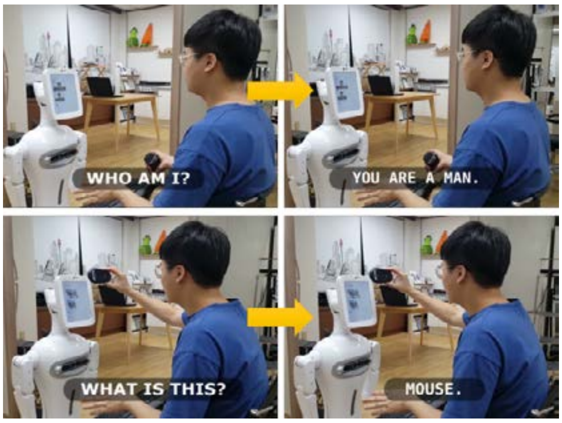

Fig. 3. Cambot with interactive communication.

IV. Conclusion

The Cambot with MRN is state-of-the-art in visual question-answering. It is an advanced technology that handles a sentence and image. And the web server which connects a robot and the VQA server translate Korean into English and English into Korean. Using the cambot, it is possible to play with children and teach a foreign language to them. Furthermore, we expect this technology may help blind people to make it easy to find something by talking with them. Although the cambot handle captured image, the cambot will process video data and other sensing like a human.

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[4] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[5] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." in *Interspeech*, vol. 2, 2010, p. 3.

[6] S. Sukhbaatar, J. Weston, R. Fergus *et al.*, "End-to-end memory networks," in *Advances in neural information processing systems*, 2015, pp. 2440–2448.

[7] M. Onishi, Z. Luo, T. Odashima, S. Hirano, K. Tahara, and T. Mukai, "Generation of human care behaviors by human-interactive robot riman," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*. IEEE, 2007, pp. 3128–3129.

[8] P. H. Kahn Jr, B. Friedman, D. R. Perez-Granados, and N. G. Freier, "Robotic pets in the lives of preschool children," in *CHI'04 extended abstracts on Human factors in computing systems*. ACM, 2004, pp. 1449–1452.

[9] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, "Multimodal residual learning for visual qa," *arXiv preprint arXiv:1606.01455*, 2016.

[10] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[11] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," *arXiv preprint arXiv:1511.02274*, 2015.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.