INTERVIEW WITH RESEARCH TEAM OF HUMAN MULTIMODAL LEARNING IN KAIST, SOUTH KOREA

Professor Yong Man Ro
School of Electrical Engineering,
Korea Advanced Institute of Science
and Technology (KAIST), Korea

Yong Man Ro is a Professor at School of Electrical Engineering, KAIST (Korea Advanced Institute in Science and Technology), South Korea. Prof. Ro established Image and video systems (IVY) Laboratory and the center for applied research in AI (CARAI) at KAIST in 1997 and 2019, respectively. Among the years, his research team has been conducting research in a wide spectrum of image and computer vision research topics. Recent research interests are deep learning, machine learning in computer vision and image processing, multimodal learning, explainable AI and Robust AI. He and his team has produced over 140 SCI indexed papers and 329 international conference papers.

Human Multimodal Research Team: (Left to right) Minsu Kim (PhD candidate), Jeung Hun Yeo (PhD candidate), Joanna Hong (PhD candidate), Prof. Yong Man Ro, Jeongsoo Choi (PhD candidate), Chae Won Kim (MS candidate), Hyun Jun Kim (PhD candidate), Bella Godiva(MS candidate), Se Jin Park (PhD candidate)

What are your team's main objectives and research topics?

In my recent research team in KAIST, main objectives are to design and develop creative and efficient learning methods and neural architectures to provide useful deep representations of multi-modal data for diverse applications. In particular, we are conducting research to develop advanced neural learning methods such as self-supervised and unsupervised, to put the technologies into more practical situations beyond the supervised learning which requires large human resources. Also, we try to provide a creative method to utilize rich multi-modal information even if some modalities are not available during inference, in order to give the consumers more powerful performances.

The main research topics of my team are developing human interpretable deep learning models, human-oriented multi-modal applications such as audio-visual speech recognition, talking face generation, speech synthesis, and multimodal dialogue systems.

In your opinion, why are those research topics interesting for you compared to other AI topics?

Research topics covered by my team are conducted with the hope of producing useful impacts on the community. My team works along with industry and government to keep being updated with the current demand of the real world. For me, research would be more interesting if it is not only kept on a paper but also be able to be applied in the community and contribute to the improvement of people's life quality. We believe that AI would become more meaningful when it can work along with humans. Therefore, we need to equip AI with senses that human has like audio, vision, and language to enable them to communicate with humans. This goal is hoped to be achieved through human-oriented multimodal research.

What are the biggest obstacles you're facing with your current research topics, and how do you deal with them?

Over the years, machine learning technology has evolved, and the size of the model has become wider and deeper. In the multimodal field, in particular, the amount of data required is even greater because the model is more complex and numerous than the single model. At the same time, there is a condition that data of various domains must be well-aligned for multimodal learning. These data are hard to obtain naturally in the real world, and there are many difficulties in collecting them.
The simplest and obvious way is to collect and label more real-world data. It is very time-consuming and expensive work, but it is something that all researchers must do consistently. Therefore, we need to cooperate with the various institution and research teams to do it efficiently. Another way is data generation. It is difficult to generate data, but once set, it is much cheaper and easier to collect data in massive quantities. In addition, we can generate data that can never be obtained naturally, so it is more helpful for various data augmentation. Finally, we adopt few-shot learning techniques to overcome data sparsity. Nowadays, many learning techniques for few-shot learning such as meta-learning and transfer learning have developed a lot, so if these methods are properly applied, effective learning is possible even on limited data.

Due to the heterogeneous nature of data, combining modalities in multimodal learning is problematic. Do you have any strategy for combining them effectively?

Multimodal learning is difficult due to the heterogeneous nature of data. Reducing heterogeneity gap, induced by inconsistent distribution of different modalities, is still considered as a challenging problem in various task (e.g. metric learning and knowledge distillation in multimodal).
To avoid heterogeneity gap issue in multimodal learning, we utilize multi-modality associative bridging through memory. The memory network maps each multimodal feature to each latent space. To use multimodal data without gap, the features are saved in each space. And, we can load the other modal data using one of the multimodal data. Therefore, learning one-to-one mapping through memory, can combine multimodal data effectively.

Human multimodal learning is very useful for many applications. For practical, a high-performance and efficient model is highly recommended. Could you tell us a bit about a robust multimodal learning system?

Human-oriented robust multimodal system contains human-AI dialog system which utilizes multimodal information: text, video, sound, etc. Using multimodal information, various robust techniques have been developed: speech enhancement and recognition system, speech reconstruction system, facial video generation system, automatic human-AI dialog system, and so on.
A robust multimodal learning system may require training with huge and diverse datasets. Since multimodal dataset has complex characteristics due to its diverse domain condition, combining all the modalities in the same representation space would provide the model's robustness. To do so, we also need multi-task and transfer learning among various dataset with various modalities.

Currently, Covid-19 has a huge impact on human life change. If you have an unlimited research budget from any source, what kinds of AI applications or systems will you develop to help people stay safe from Covid-19?

Because of Covid-19, Many people are reducing human contact and increasing their time at home. We will develop an audiovisual analysis system to help facilitate communication with others remotely within the home. The system contains audiovisual speech recognition in the wild using deep learning and this can help communication in the way of reducing the influence of noisy environment.
Also, demand for video conferencing has grown rapidly since advent of the advent of Covid-19. We will be able to develop technologies about generating talking face from speech information in real time using deep learning for both keeping privacy of participants and providing video conferencing service stably by reducing the amount of information transmitted.

In the last 10 years, technology has rapidly advanced, including Artificial Intelligence (AI). From your perspective, what will AI and deep learning be in the next 10 years?

AI evolution is approaching faster than we think. At the moment, AI is on the front lines of the industry. It's becoming rare to find workplaces that do not apply AI in their

work. Various lines of AI research are pouring out every year, and the industry is rapidly turning the research into useful technologies. I'm very excited about what would come in the next 10 years and how much more helpful it will turn out to be. It will become a part of not just an organization, but our daily lives in a wider scope and at a lower cost easily accessible for everyone.
As it integrates into most aspects of our lives and becomes pervasive, all of us should prepare for such transformations so that no one is marginalized from the educated. For AI applications to become reality, a user-friendly interface will have to be developed along with the AI research. Also, we should utilize AI as an assistant, more like a tool to augment human skills. Thus, we would have to make choices on restrictions of the use that benefit us so that it does not yield privacy and cyber-related crimes.

Do you have any encouragement messages for young scientists that are interested in your research field?

The field of AI has shown remarkable progress and results in single modality realms. However, research not only remains in single modalities, but is expanding to using various modalities, combining language, vision and audio. In parallel to the development of multimodal learning, datasets are becoming more large-scale and high resolution to support multiple modalities. By learning joint representations and aligning these modalities, we are able to capture more complex and meaningful information compared to using single modalities. Nevertheless, the real-world contains even more data beyond textual, audio and visual senses, and multimodal learning will become an important stepping stone to representing the real-world in the future. Researching multimodal learning can become burdensome and tedious, especially if you are in academia and have limited means, as training requires increasingly more resources. However, if you have the passion and patience to discover and solve problems, you will be naturally guided during the process to make an impact on the real-world. It is important to keep up with the rapidly changing field of research, but also be patient on working with projects. Take it slow, but pursue with passion!