See, Hear, and Learn to Describe—Crossmodal Information Processing Opens the Way to Smarter AI

Kunio Kashino

Abstract

At NTT Communication Science Laboratories, we are researching information processing across different types of media information such as images, sounds, and text. This is known as crossmodal information processing. The point of crossmodal information processing is to create a common space, which is a place where multiple types of media data are associated. The common space enables us to realize new functions that have never existed before: new transformations between different media—such as creating images and descriptions from sound—and the acquisition of concepts contained in media information.

Keywords: crossmodal information processing, AI, concept acquisition

1. Introduction

The driving force behind the recent development of artificial intelligence (AI) is deep learning technology. Deep learning, when it is applied to object recognition, for example, involves preparing a large amount of data in which images of various objects are photographed, and the names of objects (class labels) such as *apple* and *orange* are combined (training pairs). This kind of learning is known to achieve recognition of objects in images with high accuracy.

While deep learning has been researched and used in various fields due to its excellent classification performance, we are particularly interested in its ability to find correspondence between different types of media information (for example, images and sounds). The different kinds of information such as images, sounds, and text are called modalities, and the correspondence of information across different modalities is called crossmodal information processing. In this article, we introduce the concept of crossmodal information processing and its implications.

2. New information conversion

One advantage of crossmodal information processing is that it makes it possible to transform information in a way that was not previously thought possible through a common space, which is a place where different kinds of media information are related (**Fig. 1**).

2.1 Creating an image from sounds

For example, our research team is working on the task of estimating images from sounds. We humans can imagine the visual scene from the sound around us even when we close our eyes. This may imply that it could be possible to create an image of the scene from the sound picked up by microphones. For example, several microphones can be placed in a room to record several people talking there. If you use four microphones, you will get four sound spectrograms representing the frequency components of the sound captured by each microphone, and an angular spectrum representing the directions of arrival of the sounds. The system takes these as inputs. It then processes these pieces of information using neural networks and maps them into a low-dimensional space.

Another neural network then uses this information

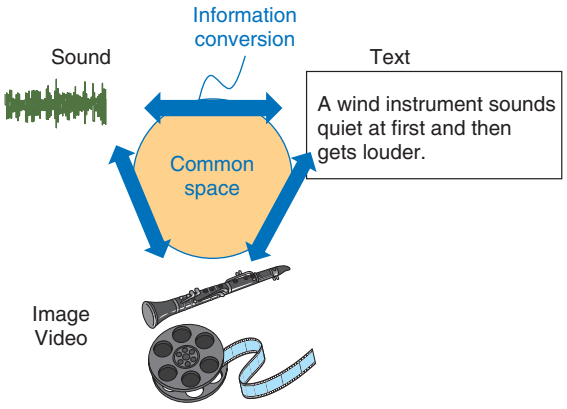


Fig. 1. Conceptual diagram of information conversion by crossmodal information processing.

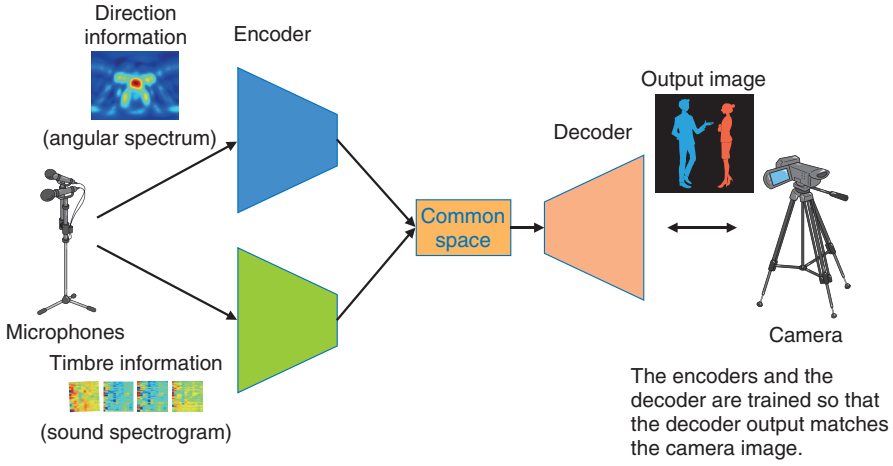


Fig. 2. Generating an image from sound.

to create an image. This image shows what kind of people are speaking and where they are in the room, so we can get a rough idea of what is going on in the room (**Fig. 2**). The process of mapping (encoding) an input into a low-dimensional space and generating (decoding) high-dimensional information from it in this way is generally called the encoder-decoder model, and it can be constructed using deep learning by providing input/output pairs as training data.

Up to now, we have conducted simulation experiments and experiments using actual sound-producing objects to confirm that it is actually possible to show what is where by means of an image under certain conditions [1]. This kind of sound to image conversion is a new information processing method that has never been tried before, to the best of our knowledge. As this technology develops, we believe that it will be useful for confirming the safety of people and property in places where cameras are not allowed or in situations where the camera cannot capture images very well such as in shadows and darkness.

2.2 Explaining sounds in words

Another example of heterogeneous information conversion is from sound to text. Conventional speech recognition systems can convert spoken words into text but cannot convert sounds other than spoken words into appropriate text. However, we have developed a technology to generate onomatopoeic words to express a sound and descriptions of the sound as a full sentence from a sound picked up by a microphone [2].

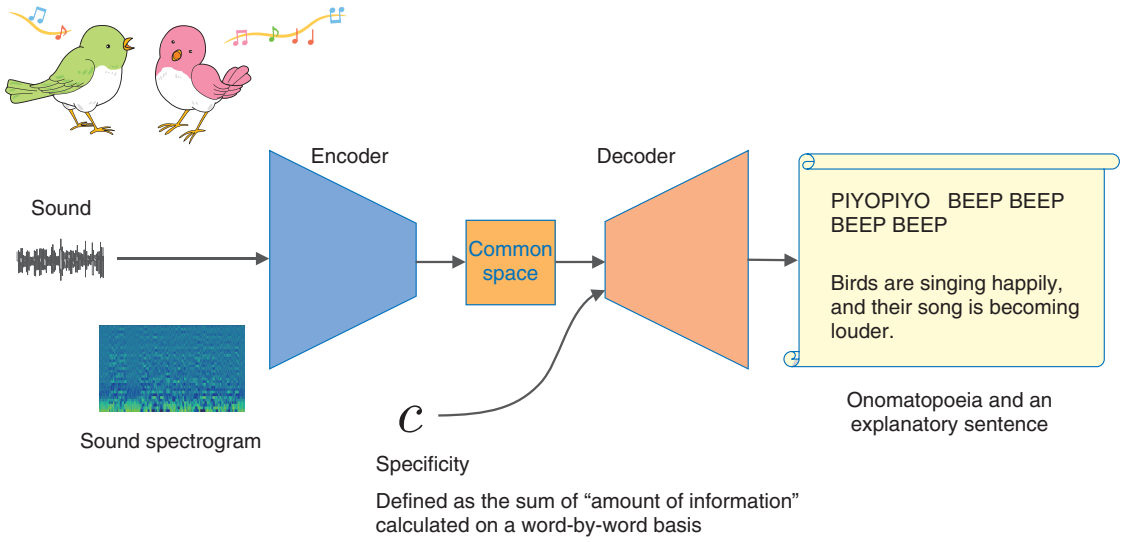


Fig. 3. Generation of explanatory statements from sounds: the conditional sequence-to-sequence caption generation method.

This method, called conditional sequence-to-sequence caption generation, or CSCG, is also based on the encoder-decoder model (**Fig. 3**). It is used here to convert one sequence to another sequence. First, the features extracted from the input acoustic signal are encoded as a time series by a recurrent neural network and mapped in the low-dimensional space. Another recurrent neural network then decodes the phoneme sequence (onomatopoeia) or the word sequence (description) from the information.

When a description is generated, the kind of description that is appropriate depends on the case, and you cannot specify a single, generic correct answer. For example, a scene may need to be expressed simply in a short sentence, such as "A car is approaching; it is dangerous," or a scene may need to be expressed in detail in terms of subtle nuances of engine sounds according to the type of vehicle and vehicle speed.

To achieve this, we controlled the function of the decoder using an auxiliary numerical input called *specificity* and made it possible to adjust the detail of expression. The specificity value is defined as the sum of the amounts of information contained in words in a sentence. Consequently, smaller specificity values produce shorter descriptions, while greater ones produce more specific and longer descriptions. Experiments under certain conditions show that our technology can generate onomatopoeic words that are more receptive than the onomatopoeic words given by humans, and can also effectively generate explanatory texts according to the designated specificity.

We believe that this technology is useful for creating subtitles for video and real environments and for searching media. Traditionally, attempts have been made to assign known class labels to sounds, such as *gunshot*, *scream*, *the sound of a piano*, and so on. However, with sound, the correspondence between the sound signal and the name of the sound source is not always obvious, and it is quite common to encounter sounds that we cannot identify. In such cases, the effectiveness of classification alone is limited.

This technology makes it possible to search for sounds based on the description by linking the sound with the description. In fact, it is possible to directly measure the distance between sounds and onomatopoeic words or descriptions in the common space, and therefore, to search for sounds using onomatopoeic words or descriptions. In such cases, one may want to specify in detail the nuances of the desired sound in the description. With this technology, it is possible to specify not only class labels such as *car* and *wind* but also the pitch, size, and temporal changes of the sound. To our knowledge, this is the first method that can generate descriptions of sounds in the form of a full sentence.

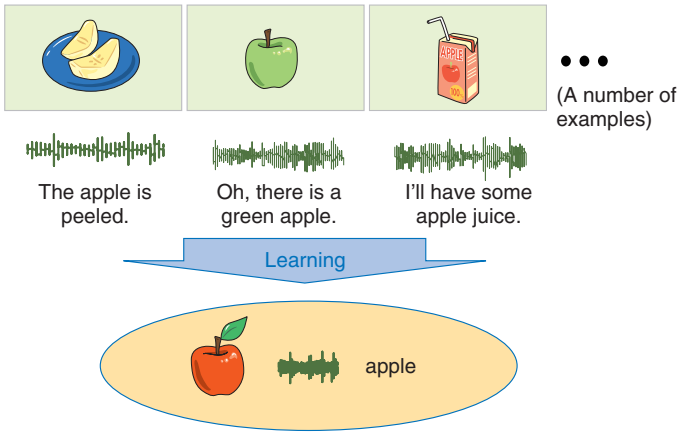


Fig. 4. Obtaining concepts from crossmodal information.

3. Learning new concepts by itself

Another advantage of crossmodal information processing is that it is possible to acquire concepts by finding the correspondence between different kinds of information in the common space. Preparing the large amount of data required for deep learning is often no simple task. It is often difficult to collect sufficient data for rare events, for example, and in many cases, it is also difficult to define classes in advance. We are therefore working on research that aims to automatically acquire a set of concepts contained in media information and use these concepts for recognition and retrieval.

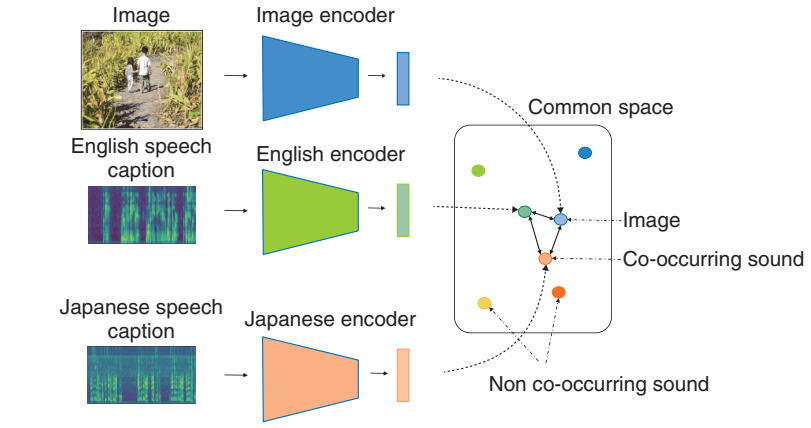
The co-occurrence of different types of media information, that is, the appearance of different types of media information originating from the same thing in the real world with specific spatiotemporal relationships instead of random relationships, makes it possible to pair media data through a common space without manually pairing media data.

For example, in daily life, it is not unusual to hear the spoken word "apple" when we see an image of an apple somewhere around us. This phenomenon means that it is no longer necessary to provide a pair of apple images and class labels in advance. Just seeing and hearing images and sounds makes the system learn their association (**Fig. 4**). In addition, the system will learn how to feel and behave, according to its circumstances. That is, if people always call an apple *ringo* (Japanese word for apple), then the system will learn it as ringo. This can be compared to the process in which we learn various things in our daily lives as we grow up.

In fact, we have confirmed that it is possible to associate words in multiple languages with objects in a photograph. As artificial co-occurrences, we prepared a set of 100,000 photographic images and their descriptions in spoken words in English and Japanese. We used them to show that the system can automatically obtain knowledge for translation between the languages regarding the objects that frequently appear in the images [3] (**Fig. 5**).

4. Future development

This article introduced crossmodal information processing. A common objective in our studies is to separate the appearance-level representation of various media information such as sound, images, and text, from the underlying common space, that is, the intrinsic information that does not depend on any specific modalities, to fully utilize both of them. If progress continues to be made in such research, it may be possible to develop AI that lives with human beings and learns by itself while sharing how to feel and behave with us humans. Such an AI could be a friendlier partner with us.



Encoder learning is performed so that co-occurring information sets are arranged close to each other in a common space, while non co-occurring information is not. Doing this for many image-speech caption pairs enables the system to extract the relationships between certain parts of the sounds and the images.

Fig. 5. Building common space by multiple encoders for concept acquisition.

References

[1] G. Irie, M. Ostrek, H. Wang, H. Kameoka, A. Kimura, T. Kawanishi, and K. Kashino, "Seeing through Sounds: Predicting Visual Semantic Segmentation Results from Multichannel Audio Signals," Proc. of the 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brighton, UK, May 2019.

[2] S. Ikawa and K. Kashino, "Generating Sound Words from Audio Signals of Acoustic Events with Sequence-to-sequence Model," Proc. of ICASSP 2018, Calgary, Canada, Apr. 2018.

[3] Y. Ohishi, A. Kimura, T. Kawanishi, K. Kashino, D. Harwath, and J. Glass, "Crossmodal Search Using Visually Grounded Multilingual Speech Signal," IEICE Tech. Rep., Vol. 119, No. 64, PRMU2019-11, pp. 283–288, 2019.

Kunio Kashino
Senior Distinguished Researcher, NTT Communication Science Laboratories.

He received a B.S. in 1990 and a Ph.D. in 1995, both from the University of Tokyo. Since joining NTT in 1995, he has been leading research projects on robust multimedia search and recognition. He is also an adjunct professor at the Graduate School of Information Science and Technology, the University of Tokyo, and a visiting professor at the National Institute of Informatics. He served as head of the Media Information Laboratory at NTT Communication Science Laboratories from 2014 to 2019.

He received the Commendation for Science and Technology from the Minister of Education, Culture, Sports, Science and Technology of Japan in 2007 and 2019. He is a senior member of the Institute of Electrical and Electronics Engineers and a Fellow of the Institute of Electronics, Information and Communication Engineers, Japan.