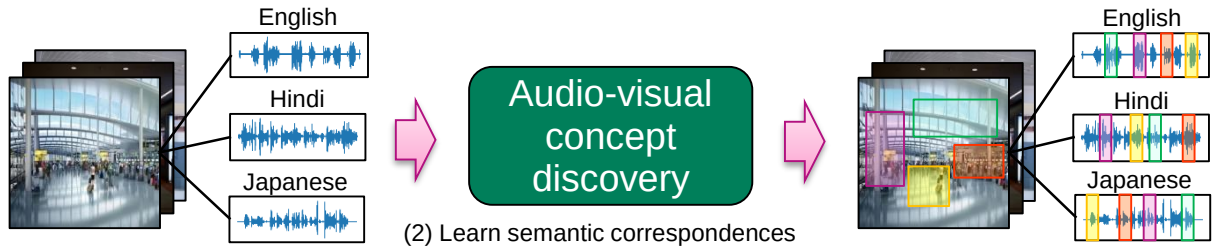
20 Learning unknown objects from speech and vision

- Crossmodal audio-visual concept discovery -

Abstract

In order for AI to visually perceive the world around it and to use language to communicate, it needs a dictionary that associates the visual objects in the world with the spoken words that refers to them. We explore a neural network models that learn semantic correspondences between the objects and the words given images and multilingual speech audio captions describing that images. We show that training a trilingual model simultaneously on English, Hindi, and newly recorded Japanese audio caption data offers improved retrieval performance over the monolingual models. Further, we demonstrate the trilingual model implicitly learns meaningful word-level translations based on images. We aim for a future in which AI discovers concepts autonomously while finding the audio-visual co-occurrences by simply providing media data that exists in the world such as TV broadcasting. We also consider the application to large-scale archive retrieval and automatic annotation that involves interactions between different sensory modalities such as vision, audio, and language.



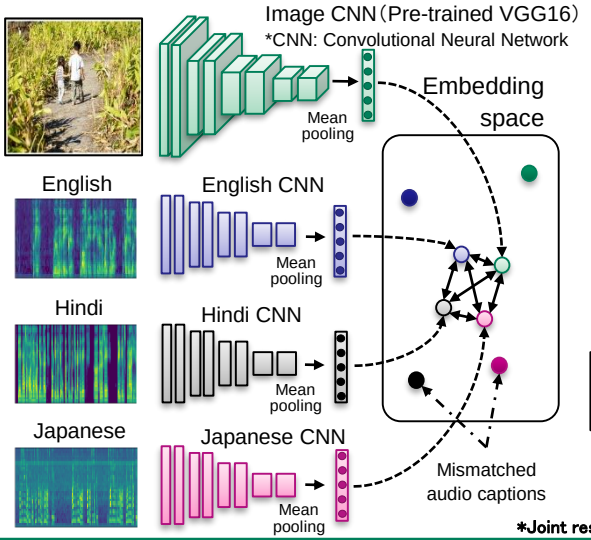
(1) Collect images and speech waveforms describing those images.

(2) Learn semantic correspondences between visual objects and spoken words.

(3) Highlight relevant regions of image as they are being described.

Learning neural network embeddings for images and spoken audio captions

Paired image and audio captions are more similar in embedding space than mismatched pairs



Evaluation of embedding space learned from image and audio captions

(1) Audio-visual retrieval performance

Recall scores for the top 10 hits (1,000 image/caption pairs)

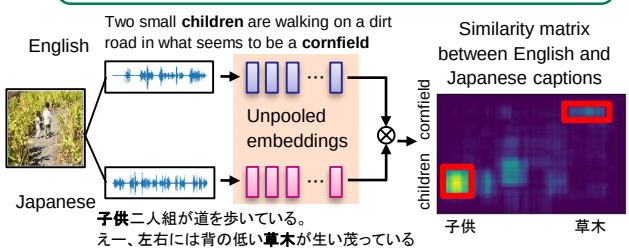
Monolingual model: 0.45 ⇨ Multilingual model: 0.50

(2) Cross-lingual audio2audio retrieval performance

Recall scores for the top 10 hits (1,000 cross-lingual caption pairs)

w/o using images: 0.01 ⇨ w/ using images: 0.50

Exploring visually grounded speech-to-speech translation



*Joint research results with MIT Computer Science and Artificial Intelligence Laboratory

References

[1] Y. Ohishi, A. Kimura, T. Kawanishi, K. Kashino, D. Harwath, and J. Glass, “Crossmodal Search using Visually Grounded Multilingual Speech Signal,” *IEICE Technical report on Pattern Recognition and Media Understanding (to appear)*

[2] D. Harwath, G. Chuang, and J. Glass, “Vision as an Interlingua: Learning Multilingual Semantic Embeddings of Untranscribed Speech,” In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2018)*, April 2018.

Contact

Yasunori Ohishi Email: cs-liaison-ml at hco.ntt.co.jp

Media Recognition Group, Media Information Laboratory

Innovative R&D by NTT

Open House 2019