Crossmodal audio-visual concept discovery

20 声と画像から知らないモノを学びとるAI

～音声と画像によるクロスモーダル概念獲得～

どんな研究

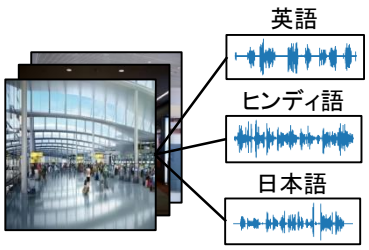
AIがモノを認識するためには、モノの見え方とその言語表現（名前）を紐付けた辞書が必要です。この展示では、画像を説明する音声データだけから、モノが映る画像領域と声による言語表現を教師ラベルなしで対応付けて、AIが知らないモノを自ら学び取る（辞書を自動作成する）技術を紹介します。

どこが凄い

日本語音声データセットを新たに構築し、既存の英語／ヒンディ語音声と合わせて、画像との潜在空間を深層学習することで、従来よりも精度良く、画像領域と言語表現が紐付けられることを確認しました。この手法により、画像を通じて異なる言語の単語翻訳知識が学習されることも大きな特徴です。

めざす未来

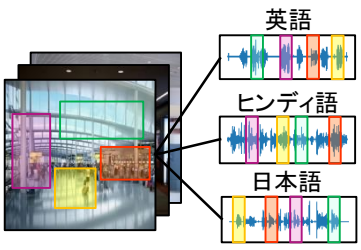
TV放送のような世の中に多く存在するメディアデータを与えるだけで、AIが音と映像の共起（時空間的な偏り）を見つけながら自律的にモノやコトの概念を学び、賢くなる未来を目指しています。音や映像、言語を自在に横断する超大規模アーカイブ検索や自動アノテーションなどへの応用を検討しています。



①画像とその内容を説明する音声のペアデータを収集

クロスモーダル概念獲得

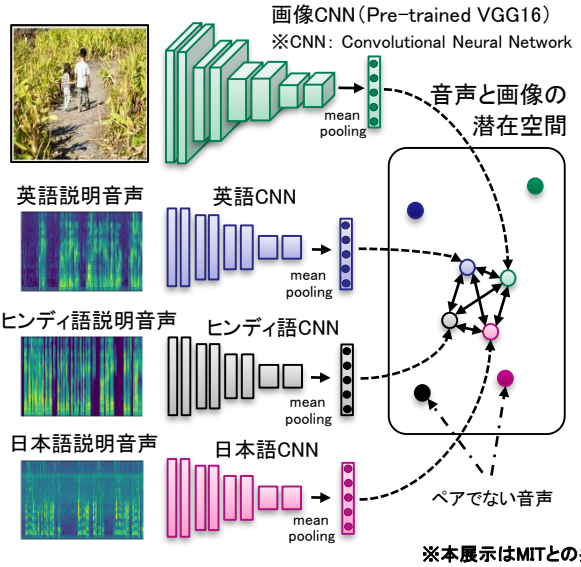
②画像領域と音声言語の対応関係を学習



③画像領域と音声言語の対応関係を辞書化

画像と多言語音声を対応付ける深層学習モデルを提案

潜在空間で画像と音声ペアが近くなるように学習



クロスモーダル探索の観点で音声と画像の潜在空間を評価

(1) 多言語音声を学習に利用することの有効性

※画像と音声1,000ペアの相互の探索精度をRecall@10で評価

単一言語音声による潜在空間学習：0.45 ⇒ 多言語音声による潜在空間学習：0.50

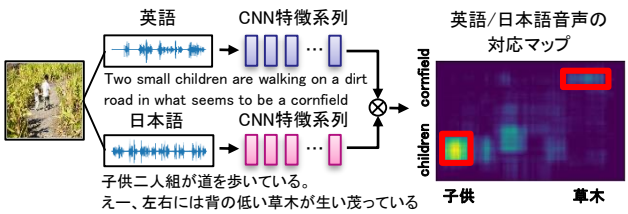
(2) 画像を中間言語として学習することの有効性

※異言語の音声1,000ペアの相互の探索精度をRecall@10で評価

画像を利用しない潜在空間学習：0.01 ⇒ 画像を利用した潜在空間学習：0.50

（Recall@10：10位以内にペアとなる音声 or 画像が見つかる割合。高い値の方が良い。）

画像（中間言語）を通じて異言語の翻訳知識が獲得されることを発見



※本展示はMITとの共同研究成果です（MIT Computer Science and Artificial Intelligence Laboratory）

関連文献

[1] 大石康智, 木村昭悟, 川西隆仁, 柏野邦夫, D. Harwath, J. Glass, “画像を説明する多言語音声データを利用したクロスモーダル探索,” *電子情報通信学会技術報告, パターン認識・メディア理解(PRMU)研究会, (to appear)*

[2] D. Harwath, G. Chuang, 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)*, 2018.

連 絡 先

大石 康智 (Yasunori Ohishi)　メディア情報研究部　メディア認識研究グループ

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

Innovative R&D by NTT

オープンハウス 2019