数理・情報のフロンティア
2019 年度採択研究代表者

2020 年度
年次報告書

井上　中順

東京工業大学 情報理工学院
助教

データ大統一に向けたマルチモーダル事前学習

§1. 研究成果の概要

本研究では、「タスクを横断した音と画像の事前学習」方法の確立を目指しています。具体的には、音声からの話者認識と、画像からの顔認識を中心に、複数のタスクを相補的に学習するマルチモーダル学習の研究を実施しています。

2020 年度の主な成果は、話者認識モデルの自己教師あり学習法と、顔画像認識モデルの距離学習法の2つに関するものです。前者では、音声から発話者を認識するモデルを、教師ラベル情報を用いずに学習する手法を確立しました。具体的には、画像分類で用いられている対象損失(Contrastive Loss)を音声向けに設計・改良することで、大規模な教師なし音声データ上で、話者情報を抽出するニューラルネットワークの学習が可能であることを示しました。後者では、同一の学習方式を顔画像認識に適応した上で、損失関数における距離尺度を改良しました。具体的には、学習時にミニバッチ上で画像サンプルを頂点とするグラフ構造を定義し、結合の重みを最適化することで、顔画像認識に適した画像表現を得る方法を確立しました。これらの成果により、音と画像が同一の方式で学習可能であることが明らかとなったため、現在は、音と画像の表現の共通化に取り組んでいます。