研 究 終 了 報 告 書

「データ大統一に向けたマルチモーダル事前学習」

研究期間: 2019 年 10 月～2022 年 3 月
研 究 者: 井上　中順

1. 研究のねらい

近年、深層学習と大規模計算機の相乗的発展により、音や画像といったマルチメディアデータの高精度な認識が可能となった。産業界では、音声認識や画像認識をもとにした、様々なサービスが成功を収めている。しかし、現行の認識システムは、それぞれが特定のタスクに特化したものであり、複数の種類のタスクやデータを横断的に扱える、汎用的な認識システムは実現していない。最先端の深層学習でも、人間のように音や画像の情報を組み合わせて統合的に理解することは難しいのが現状である。

そこで本研究では、音の認識と画像の認識に共通して有効なニューラルネットワークモデルを構築し、その事前学習手法の研究を実施することで、タスクやデータの種類を横断した汎用的な学習方式の確立を目指す。具体的には、音声に関する話者認識・環境音認識と画像に関する顔認識・物体認識の４つのタスクを研究対象とし、統合的な学習フレームワークの実現を目標とする。

2. 研究成果

(1)概要

ACT-X の研究期間では、大きく分けて以下の３のテーマについて研究を実施した。

研究テーマ A「音認識モデルの自己教師あり学習」(2019-2020 年度)
研究テーマ B「画像認識モデルの自己教師あり学習」(2020-2021 年度)
研究テーマ C「音と画像を横断した学習」(2021 年度)

テーマ A は話者認識・環境音認識に関するものである。ここでは VoxCeleb2 データセットと呼ばれる大規模話者照合データセットで教師ラベルを用いずにエラー率を約 15.26%まで下げることができており、発表当時では世界最高の性能を記録している。テーマ B は顔認識・物体認識に関するものである。ここでは当初予定していた画像識別タスクへの応用を想定した自己教師あり学習に加え、画像生成タスクへの応用までを含める形で研究を実施した。テーマ C は映像データを介して画像認識モデルから音認識モデルへの知識転移を可能とする方法を提案し、その有効性を大規模実験により示した。成果は音声分野のフラッグシップ国際会議 ICASSP など、国際会議での研究発表を行った[発表文献 1-8]。以下ではそれぞれの詳細について述べる。

（2）詳細

研究テーマ A「音認識モデルの自己教師あり学習」

音の認識に関しては、話者認識と環境音認識の２つタスクを対象とするため、それらに関する自己教師あり学習手法の研究を実施した。

話者認識の目的は、音声データが与えられたもとで、そこから個人性情報を抽出し、発話者が誰であるかを認識することである。話者認識では、任意の長さ(一般的には 5～10 秒程度)の音声を、話者の個人性情報のみを有した１つの特徴ベクトルに変換するニューラルネットワークが広く用いられており、その学習方法が課題となっている。従来研究では、ほとんどの場合、教師ラベル(話者の ID)が付与された大規模な音声データセットが学習に用いられている。しかし、データ作成の人的コストは非常に高いという問題点がある。そこで、自己教師あり学習と呼ばれる、教師ラベルを用いない学習方式を実現した。詳細には、次の３点に関して成果が得られそれぞれ論文発表を行った：(1) Generalized Contrastive Loss と呼ばれる損失関数の提案とその話者照合への応用［発表文献 2], (2) Meta Learning と呼ばれる学習手法［発表文献 3]。(1)では VoxCeleb2 データセットと呼ばれる大規模話者照合データセットで教師ラベルを用いずにエラー率を約 15.26%まで下げることができており、発表当時では世界最高の性能を記録している。

環境音認識の目的は、音データが与えられたもとで、音源の種類（車、飛行機、雨、エアコンなど）を認識することである。環境音認識では、音源の種類が多い点と、録音環境への依存度が高い点が課題となっている。そこで、特定の録音環境でのデータが少ない場合でも高精度で学習を行う手法を提案した［発表文献 4]。本手法はカラーノイズと呼ばれる手続き型ノイズを用いてニューラルネットワークを事前に学習する手法である。

研究テーマ B「画像認識モデルの自己教師あり学習」

画像認識に関しては、顔認識と物体認識の２つタスクを対象とするため、それらに関する自己教師あり学習手法の研究を実施した。

顔認識の目的は、顔の画像が与えられたものとで、人物を認識することである。高画質な静止画では高精度な認識がすでに実現しているため、本研究ではそれよりも低画質な映像データを対象とした。Graph Grouping Loss と呼ばれるグラフ構造に基づいた損失関数を提案したもの[発表文献 5]が成果である。これはテーマ A における音声向けの損失関数を画像向けに改良したものである。

一方、物体認識に関しては、従来から自己教師あり学習の有効性が示されているが、

特徴量の分解(disentanglement)に関する部分は未知な部分が多く、教師あり学習と自己教師あり学習の違いが明らかとされていない。そこで、自己教師あり学習と変分オートエンコーダ（VAE)を組み合わせた学習方法に関する研究を実施した。詳細には、次の２点に関して成果が得られそれぞれ論文発表を行った：(1) カテゴリラベルに基づいた損失を用いた VAE の学習［発表文献 6]、(2) 自己教師あり学習の損失を用いた VAE の学習［発表文献 7]。

研究テーマ C「音と画像を横断した学習」

最後に、音と画像に共通して有効な学習手法の研究を実施した。主な成果は、顔画像認識と話者認識に関するもの[発表文献 1]で、顔画像データを用いて音から話者を認識するモデルを学習する手法を提案した。具体的には、ラベル付き顔画像データおよびラベルなし映像データを組み合わせて利用することで、画像から音声への知識転移を実現し、画像に関するラベルのみで、音から発話者を認識するモデルが得られることを示している。VoxCeleb2 データセットにおける評価では、エラー率を 3.44%まで下げることができている。

物体認識モデルとの統合に関しては、現在(報告書執筆時点)で取り組み中であり、テーマ A,B のタスクを統合した学習方法の実装を進めている。

３．　今後の展開

ACT-X 研究を通じて、データの種類を横断した統一的な深層学習フレームワークの基礎となる損失関数や学習アルゴリズムの設計を行うことができた。個人型研究の規模では実施できない大規模実験の必要性も明らかとなったが、規模の見積もりが可能となったことには価値がある。

今後の展開としては、国内外で大きなチームを作る必要がある。画像・音・言語に関する人工知能の研究は、ここ 10 年間で巨大な IT 企業が牽引する形を取るようになり、国内は人材不足のみならず研究教育体制の見直しが追いついていない状況にある。汎用的な人工知能の実現に向けては、画像・音・言語を横断した新しい研究分野の創出が必須であるため、ACT-X 数理・情報のフロンティアのように、様々な分野の研究者・エンジニアが協力関係にある形でのプロジェクトを今後も進めて行きたい。

個々の研究テーマに関しては、学習の安定性、学習データ効率、省エネルギー化などに課題が残っている。これらについては、中規模なプロジェクトを立てて問題の解決に望みたい。

4. 自己評価

研究成果として高く自己評価できる内容は以下の２つである。

(1) 顔認識モデルから話者認識モデルへの知識蒸留という形で、画像と音を横断した学習フレームワークを実現したこと（研究テーマ B, C）

(2) 話者照合モデルの教師なし表現学習において，論文発表時点で State-of-The-Art Performance（世界最高水準の性能）を達成したこと（研究テーマ A, C）

ここで、(1)は当初の目標であった統合的な学習フレームワークの実現に関するものであり、(2)はその性能に関するものであり、どちらも期待通りの成果が得られている。一方で、物体認識モデルまでの統合を試みていたが、この部分は論文採択までに至らなかった。

研究の進め方（研究実施体制及び研究費執行状況）については、新型コロナウイルス感染症への対策から、オンラインでの研究実施が中心となり、旅費の支出を取り止め、オンラインでの研究実施のための必要最低限な物品購入へ切り替えるなどの対応をおこなった。それら以外は、計画通りに研究の実施と研究費の執行ができている。

5. 主な研究成果リスト

（1）代表的な論文発表

研究期間累積件数：8件

[1] N. Inoue, “Teacher-Assisted Mini-Batch Sampling for Blind Distillation using Metric Learning,” Proc. International Conference on Acoustics, Speech, & Signal Processing (ICASSP), 2021.

本論文は顔画像データを利用して，話者認識モデルを学習するフレームワークを提案したものである．具体的には，1) ラベル付き顔画像データ，2) ラベルなし映像データの2つを利用して，画像から音声への知識転移を実現し，音から発話者を認識するモデルが得られることを示している．評価は大規模データセット VoxCeleb2 により行い，論文発表時点で音声に対する教師ラベルを用いない手法として最高性能を達成したものである。

（2）特許出願

研究期間全出願件数：0件（特許公開前のものも含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

[2] N. Inoue and K. Goto, “Semi-Supervised Contrastive Learning with Generalized Contrastive Loss and Its Application to Speaker Recognition,” Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2020.

[3] N. Inoue and K. Goto, “Optimizing Speaker Embeddings using Meta-Training Sets,” Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2020.

[4] K. Goto and N. Inoue, “Quasi-Newton Adversarial Attacks on Speaker Verification

Systems," Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2020.

[5] N. Inoue, "Graph Grouping Loss for Metric Learning of Face Image Representations," Proc. Visual Communications and Image Processing Conference (VCIP), 2020.

[6] K. Goto and N. Inoue, "Learning VAE with Categorical Labels for Generating Conditional Handwritten Characters, " Proc. International Conference on Machine Vision Applications (MVA) 2021.

[7] N. Inoue, R. Yamada, R. Kawakami, and I. Sato, "Disentangling Groups of Factors," Proc. International Conference on Image Processing (ICIP), 2021.

[8] N. Inoue and K. Goto, "Closed-Form Pre-Training for Small-Sample Environmental Sound Recognition," Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2020.