TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

利用課題名：生成型音声・画像 AI のための学習アルゴリズムの開拓と自動機械学習の研究
英文：Investigation of Training Algorithms and AutoML for Generative AI

利用課題責任者：坂東　宜昭和
First name Surname: Yoshiaki Bando

所属：国立研究開発法人　産業技術総合研究所
Affiliation: National Institute of Advanced Industrial Science and Technology (AIST), Japan
URL: https://www.airc.aist.go.jp/cosine

邦文抄録（300 字程度）
本利用課題では，我々の日常生活で人々とコミュニケーションをとりながら適切に仕事をこなす人工知能（AI）技術を実現するため，音声・画像 AI のための学習アルゴリズムの開拓と自動機械学習の研究を進めた．特に，映像信号と音響信号を入力として音響イベントの発生時刻と音響物体を表す領域を推定する視聴覚音源定位（AV-SSL）に取り組み，Deep Clustering に基づく視聴覚埋め込み空間を学習することで，複数音源を容易に弁別可能する特徴量抽出器を構築した．さらに，データセットのキャッシュ機構の構築や動画データの効率的な HDF5 形式への保存など，高速に研究ループを進展させるための研究ツールを整備した．

英文抄録（100 words 程度）
We investigated machine learning research for building multimodal artificial intelligence (AI) systems that can communicate humans in our daily lives. Specifically, we investigated audio-visual sound source localization (AV-SSL), which is a task to detect and localize sound events from video and audio recordings. We trained audio and visual encoders whose embedding space is designed to easily distinguish multiple sounding objects based on the deep clustering technique. In addition, we built supporting tools for training neural models on high-performance computing (HPC) systems.

Keywords: Audio-visual training, generative AI, contrastive learning, self-supervised learning, automated machine learning

背景と目的

膨大なデータに対し効率的にスケールする大規模言語モデル（LLM）や自己教師あり学習の台頭により，多くの人工知能（AI）タスクが実用に足る性能を達成している．一方，我々の日常生活で人々とコミュニケーションをとりながら適切に仕事をこなす AI 技術は未だ実現に至っていない．本研究の目的は，このような人間と協働する AI 技術を確立することである．特に，ユーザの利用環境に応じて適応的に高い性能を出すための枠組み，AI 自体の開発段階における効率的な学習や MLOps パイプライン，ユーザの要望に応じて適宜自身を変化できるメタ学習技術は，未だ発展途上の課題であり，本研究課題ではこれらの解決を目指した．

概要

本課題では，実世界で人間と協調する次世代 AI 技術を確立するため，その効率的な学習方法の開拓とともに実応用で課題となる自動機械学習（AutoML/AutoPEFT）に取り組む．具体的には，音

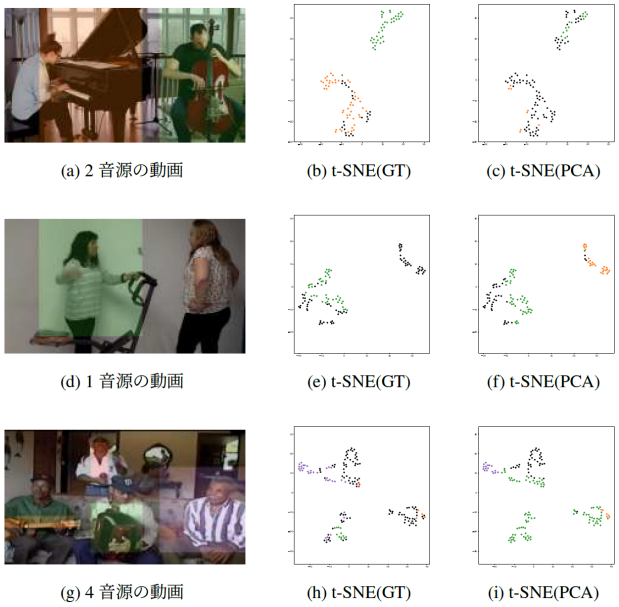

図 1 学習した埋め込み空間の可視化結果例 [1]

(a) 水平角 0°　(b) 水平角 90°　(c) 水平角 180°　(d) 水平角 270°　(e) 結果の統合

図 2 得られた埋め込みモデルを用いた実収録データでの動作検証 [1]

声・画像 AI のためのスケーラブルな学習アルゴリズムの開拓，実世界データにおける性能評価ベンチマークの確立など，実世界を理解し人々と協調するAIに必要な要素技術の開拓を進めた．

結果および考察

具体的には，映像信号と音響信号を入力として音響イベントの発生時刻と音響物体を表す領域を推定する視聴覚音源定位（AV-SSL）に取り組み，Deep Clustering に基づく視聴覚埋め込み空間を学習することで，複数の音響イベントを容易に弁別可能にする特徴量抽出器を構築した [1].

図 1 に得られた埋め込み空間の例を示す．図 1-(b), (e), および(h)から，得られた埋め込み空間は，音響物体ごとに異なる特徴（色）へ埋め込めていることが分かる．また，図 2 に示すように，対話ロボットへの搭載を目的として，実収録した全方位映像に対して構築したモデルを適用する実験を行った．構築したモデルは，実時間推論できるよう軽量なアーキテクチャを採用している．また，ロボットでは周囲の状況を理解するため全方位画像の活用が望ましいが，そのような学習データの収集にはコストを要する．そこで，幾何変換した空間での埋め込みのクラスタリングを活用することで，Web から収集した動画で学習したモデルでも，容易に全方位画像に適用できる枠組みの構築を進めた．

本研究の学習では，我々が収集した約 160 時間の視聴覚データを用いた．視聴覚データとしては比較的大規模なデータセットであり，このような動画データから効率的な学習を実現するため，Lustre ファイルシステムへのアクセスを最小限とするキャッシュ機構を構築した．torch.utils.data.Dataset のラッパーとして記述でき，既存の Dataset クラスをラップすることで，1 度アクセスした結果を共有メモリまたはローカルストレージにキャッシュする．素朴な，学習開始前にデータをローカルストレージに転送する場合と比べて，ジョブ実行から学習開始までの時間を大幅に削減でき，モデル学習のデバッグ効率を改善することができた．さらに，動画データの HDF5 ファイルへの効率的な保存方法の検討もすすめた．HDF5 ファイル内に素朴にフレーム情報を保存すると，動画に特化した(非可逆)圧縮方式に比べて圧縮効率が悪い．一方，小さな大量の動画ファイルを逐一ファイルシステムから読み出すと，Lustre ファイルシステムへの負荷が上がってしまい効率が下がる．そこで，HDF5 ファイル内に mp4 コンテナをバイナリとして保存し，Dataset クラス内で動的に展開する機構を実装・評価した．殆どの動画デコードプログラムは，ファイルからの読み出しを前提としているが，本実装ではメモリ・ファイルシステム上での展開として実装することで，オーバーヘッドに短縮することができた．

まとめ、今後の課題

本利用課題では，我々の日常生活で人々とコミュニケーションをとりながら適切に仕事をこなす人工知能（AI）技術を実現するため，音声・画像 AI のための学習アルゴリズムの開拓と自動機械学習の研究を進めた．特に，実時間で動作する AV-SSL の構築を通じて，LLM などの記号処理システムに入力するための情報抽出技術の開発を実施した．

参考文献

[1] 大規模 Web データを用いたロボットのための視聴覚音源定位システム，櫻井　舜，坂東宜昭，佐々木洋子，大西正輝，応用音響・電気音響研究会，査読無，40-44，2024