音の聴き方を自ら学ぶ AI

～自己教師あり学習による様々な音の汎用表現学習技術から、大規模言語モデルを活用した音の理解の最前線へ～

AI that learns to listen on its own
– Advancing self-supervised audio representation toward cutting-edge sound understanding with large language models –

NTT コミュニケーション科学基礎研究所 メディア情報研究部
事象モデリング研究グループ リサーチスペシャリスト
仁泉　大輔
Daisuke Niizumi

●プロフィール

1997年九州工業大学大学院情報工学研究科博士前期課程修了。メーカーにおける電子楽器や民生品のソフトウェア開発経験を経て、2020年よりリサーチスペシャリストとしてNTT コミュニケーション科学基礎研究所に勤務。深層学習を用いる音響信号を中心とした表現学習に関する基礎研究に従事。日本音響学会、人工知能学会各会員。

音や画像などメディア情報から、それらの理解に有用な特徴量の自動抽出処理を学習により獲得する表現学習は、AIによる優れたデータの理解を可能にしました。本講演では、私たちの身の周りの様々な音をAIに理解させるための音の表現学習技術を紹介します。学習したモデルは人の声や動物の鳴き声といった音の種類の認識や音楽ジャンルの識別など、幅広い音の問題に応用できます。表現学習の中でも、音データの内容を示すラベルを推論する従来の教師あり学習手法の代わりに、「音の穴埋め問題」を解くデータそのものを使った自己教師あり学習手法が、ラベルを持たない大量のデータを学習する手段として注目されており、大規模言語モデルを利用して言語の持つ意味と結びつけた音の理解へと発展しています。

データを理解するための基礎技術「表現学習」

表現学習は、音や画像などのメディア情報をコンピュータで理解するために、それらの有用な特徴量（＝表現）の自動抽出処理を学習する技術です。従来は、信号処理により得られる音の大きさや高さ、周波数成分など基本的な情報を基に人の手で用途に特化した特徴量を設計していましたが、例えば会話音声や乗り物の音、動物の鳴き声など、多様で複雑な音の理解には限界があります。そのため、多様な音のデータから互いに区別するうえで重要となる特徴量の自動抽出を表現学習で実現できれば、より優れた理解が可能になります。表現学習を実現する手法として、近年、大量のデータに内在する豊かなパターンから教師ラベルなしで表現を学習できる自己教師あり学習が注目されています。この学習手法は、教師ラベルの代わりとなる教師信号をデータ自体から生成することで、教師ラベルのコストを必要とせず、効率的な表現学習を実現しています。

音の違いを抑えた表現「BYOL for Audio」

多様な音を理解するうえで有用な特徴量をどのように学習すればよいか、その答えの一つとして画像分野の手法BYOL（Bootstrap Your Own Latent）を基に提案した自己教師あり学習手法BYOL for Audio（BYOL-A）は、音のパターンを捉えながらも細かな違いを抑えた数値の表現を学習します。例えば犬の鳴き声は音の高さや大きさ、または犬の種類によって異なりますが、「ワン」と擬音語で表されるように、大まかには「犬の鳴き声」と理解されます。BYOL-Aは、このように音に違いがあっても同じ音のパターンであれば同じ数値表現を抽出することを学習目標とすることで、有用な表現の学習を可能にしました（図1）。その結果、従来の手法と比べて発話キーワード認識で20ポイント以上精度を改善するなど、表現学習が問題解決の飛躍的な性能向上をもたらしました。

音の穴埋め問題で優れた表現を学習する「Masked Modeling Duo」

BYOL-Aでは表現されない細かな違いも含め、音を構成する本質に迫る特徴を抽出するため提案した新たな自己教師あり学習手法Masked Modeling Duo（M2D）は、「音の穴埋め問題」である音の隠された部分の予測を通じて表現を学習します（図2上）。マスクによって隠された部分や未来のデータ予測を通じた学習は大規模言語モデルなど主に自然言語の分野で大きな成功を収め、優れた表現を学習できることが知られています。この学習の枠組みでは、見えている部分の入力データを表現に変換し、その表現を使って隠された部分の表現を予測します。この学習において予測の正確さを向上するよう最適化されるため、BYOL-Aとは異なり、細かな違いも表現するよう促されます。例えば、メロディの一部が欠けた音楽データから欠損部分を予測する場合、曲の構造やリズムに関する特徴が表現に反映される必要があります。同様に、犬の鳴き声、演奏される音楽、さざ波の音など、多様な音を学習することで、それぞれの音の背後にある構造的なパターンを示唆する特徴が効果的に学習されます。

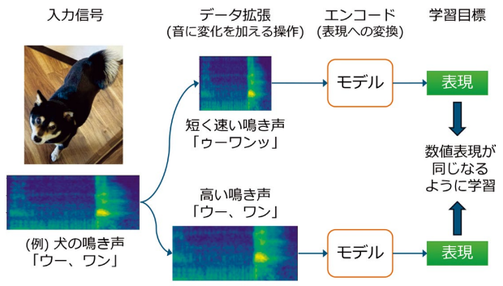

図1：表現 BYOL for Audio の学習

音とその説明文を対応付ける「M2D-CLAP」

M2D-CLAPは音の数値表現に加えて、音をその内容を表す説明文に対応付けられる表現を兼ね備えるため、M2Dによる音の穴埋め学習と、CLAP（Contrastive Language-Audio Pre-training）による音と言語の対応関係の学習を統合した手法です。このアプローチにより、音響信号と言語の意味を結びつけ、内容の近さを数値で表せる表現を獲得できます（図2下）。例えば、「雷雨の音」という説明文とその実際の音が近い数値表現を持ち、異なる音とは異なる表現になるように学習されます。これにより、「犬の鳴き声」といったテキストに対応する音の検索や、事前に学習していない「電気自動車」「バイク」などの新しく定義されたラベルへの分類問題を解くゼロショット学習が可能になります。更に、大規模言語モデルの膨大な知識に基づく言語の意味を音の表現に結びつけて学習することで、様々な音の種類についての理解がより高度化し、音の検索や説明文生成の性能が向上しました。

まとめ

これらの研究により、多様な音を理解するための表現は、音の特徴を効果的に表し、更に言語の持つ意味との結びつきを持つ汎用表現へと進化しました。しかし、音は画像や言語と比べて非常にデータが少ないなど、実用化に向けた様々な課題があります。将来的には、日常の様々なシーンの音を学習し、私たちの身の周りの音を理解するAIが、音を通じて日々の健康状態をモニタリングするサービスへの応用など様々な問題解決に役立つことが期待されます。

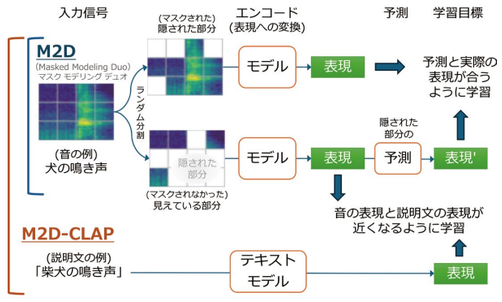

図2：音の穴埋め問題で学習する M2D および、テキストモデルを統合し拡張した M2D-CLAP

●参考文献

[1] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, K. Kashino, "BYOL for Audio: Exploring Pre-Trained General-Purpose Audio Representations," IEEE/ACM Trans. Audio, Speech, Language Process., Vol. 31, pp. 137-151, 2023.
[2] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, K. Kashino, "Masked Modeling Duo: Towards a Universal Audio Pre-Training Framework," IEEE/ACM Trans. Audio, Speech, Language Process., Vol. 32, pp. 2391-2406, 2024.
[3] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, M. Yasuda, S. Tsubaki, K. Imoto, "M2D-CLAP: Masked Modeling Duo Meets CLAP for Learning General-purpose Audio-Language Representation," in Proc. Interspeech, pp. 57-61, 2024.