音を扱いやすくする汎用音響信号表現について

2022/9/14

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
仁泉 大輔

発表の前置き

NTT

😣 含まれないもの

- 高度な音響信号処理や数式。
- 音響信号の厳密な取り扱いや解析的な分析アプローチ。

😃 含まれるもの

- 深層学習を用いたブラックボックス的アプローチ。
- 画像分野の方法を取り入れた、非厳密な音の取り扱い。
- 自己教師あり学習による、ラベル不要なエンドツーエンド学習。

録ったあとの話題です。

深層学習方面からの視野をご提供できれば幸いです。

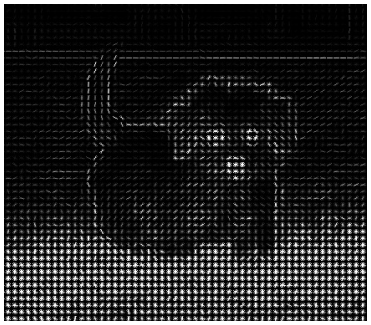
問: 犬を識別するのは簡単でしょうか?

アプローチ

1. 画像のピクセルからパターンを抽出して、プログラムでしきい値を微調整して識別。
   → プログラムで記述するのは非常に困難。
2. 画像のHand-crafted特徴量を使って統計的に学習して識別。
   → 深層学習以前の主流。
   → 2012年に深層学習によるブレークスルー。
3. 深層学習で内部表現の特徴量と識別器をデータからエンドツーエンドで獲得。

本日は一般の音に関して、
この辺りの話題をお話します。

犬画像例



HOG(勾配のヒストグラム)特徴量例

https://debuggercafe.com/image-recognition-using-histogram-of-oriented-gradients-hog-descriptor/

発表者略歴

- 所属: 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 メディア認識研究グループ
- 氏名: 仁泉 大輔 (にいずみ だいすけ)
- 製造業において電子楽器や家電製品等のソフトウェア開発に携わる経験を経て、2020年から現職にて音の説明文生成や汎用音響信号の表現学習など、深層学習を用いた音響信号の研究に従事。

発表内容

- 一般の音を取り扱う動機
  ～環境音分析の発展、大規模データセット学習済みモデル
- 画像分野の深層学習モデルを転用
  ～強力なモデルは音響信号でも性能を発揮
- **汎用音響信号表現**
  ～汎用的に適用できる強力な学習済みモデルを目指して
- むすび

一般の音を取り扱う動機～環境音分析の現状

- 人が一般に聴く主な音として、
  音声・音楽・環境音などが考えられる。
  - 音声・音楽はそれぞれ成熟した研究分野。
  - 環境音分析はDCASEにおいて近年活発な取り組み。
- 音声認識の発展に対して、身の回りで
  環境音が対象になることが少ない。
- 環境音を応用したいニーズは存在。
  - 高齢者や子供の見守り
  - メディア・コンテンツへの自動タグ付与
  - 工場における機器の自動監視
  - ライフログの自動生成

  ：

井本 桂右, 川口 洋平, 環境音分析・異常音検知の研究動向, 電子情報通信学会
基礎・境界ソサイエティ Fundamentals Review, 2021, 15 巻, 4 号, p. 268-280

一般の音を対象とする稀な応用例:
iOSでのイベント音検知

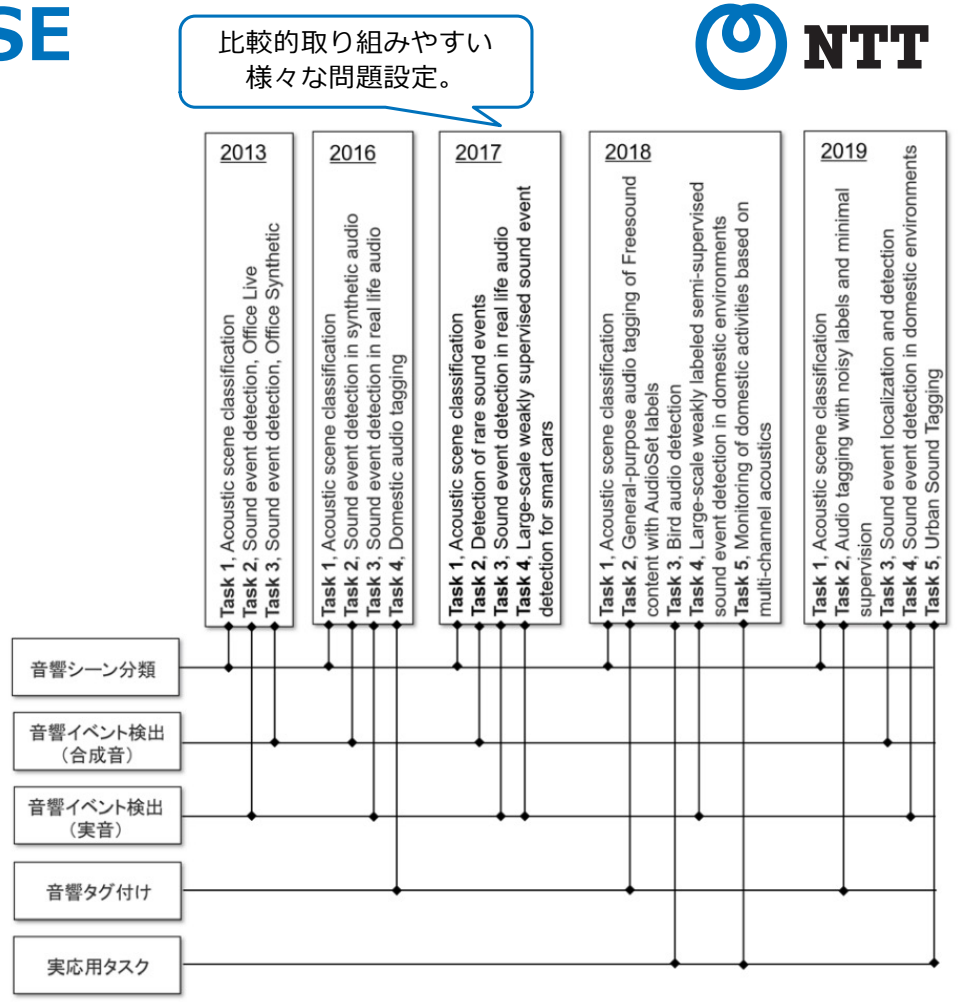
NTT技術ジャーナル「究極のプライベート音空間を実現するメディア処理技術」より
https://journal.ntt.co.jp/article/7509

環境音分析のコミュニティDCASE

Detection and **C**lassification of **A**coustic **S**cenes and **E**vents

- 環境音分析の研究に関する研究コミュニティでの活発な取り組み。
- 競争型ワークショップを毎年開催。
  - https://dcase.community/
  - 2020年からは異常音検知タスクも。

環境音分析も積極的に
取り組まれている現状。



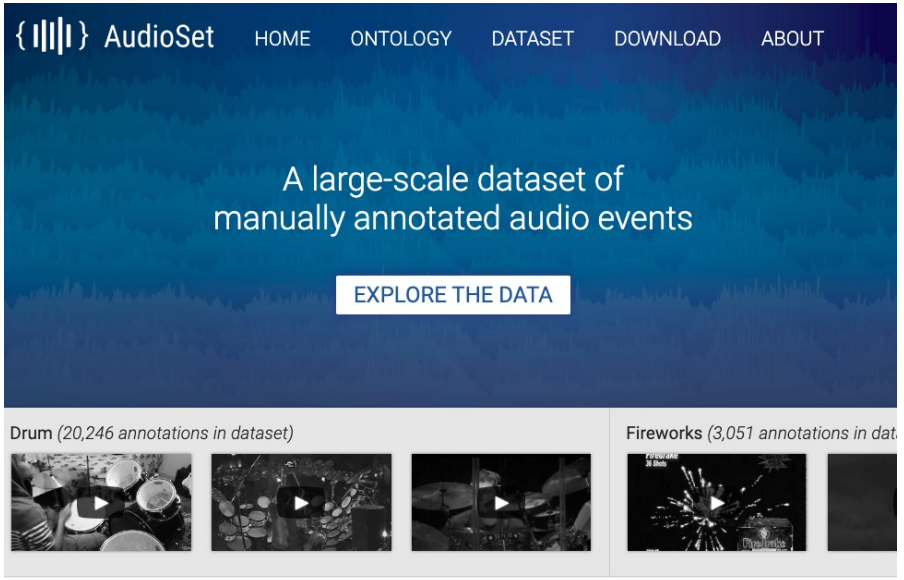
図–1　DCASE Challenge にて企画されたタスクの大分類

大石 康智, 環境音分析の研究を促進させる競争型ワークショップ, 日本音響学会誌, 2019, 75 巻, 9 号, p. 519-524

大規模データセット AudioSet

環境音を含む大規模なデータセットの存在。

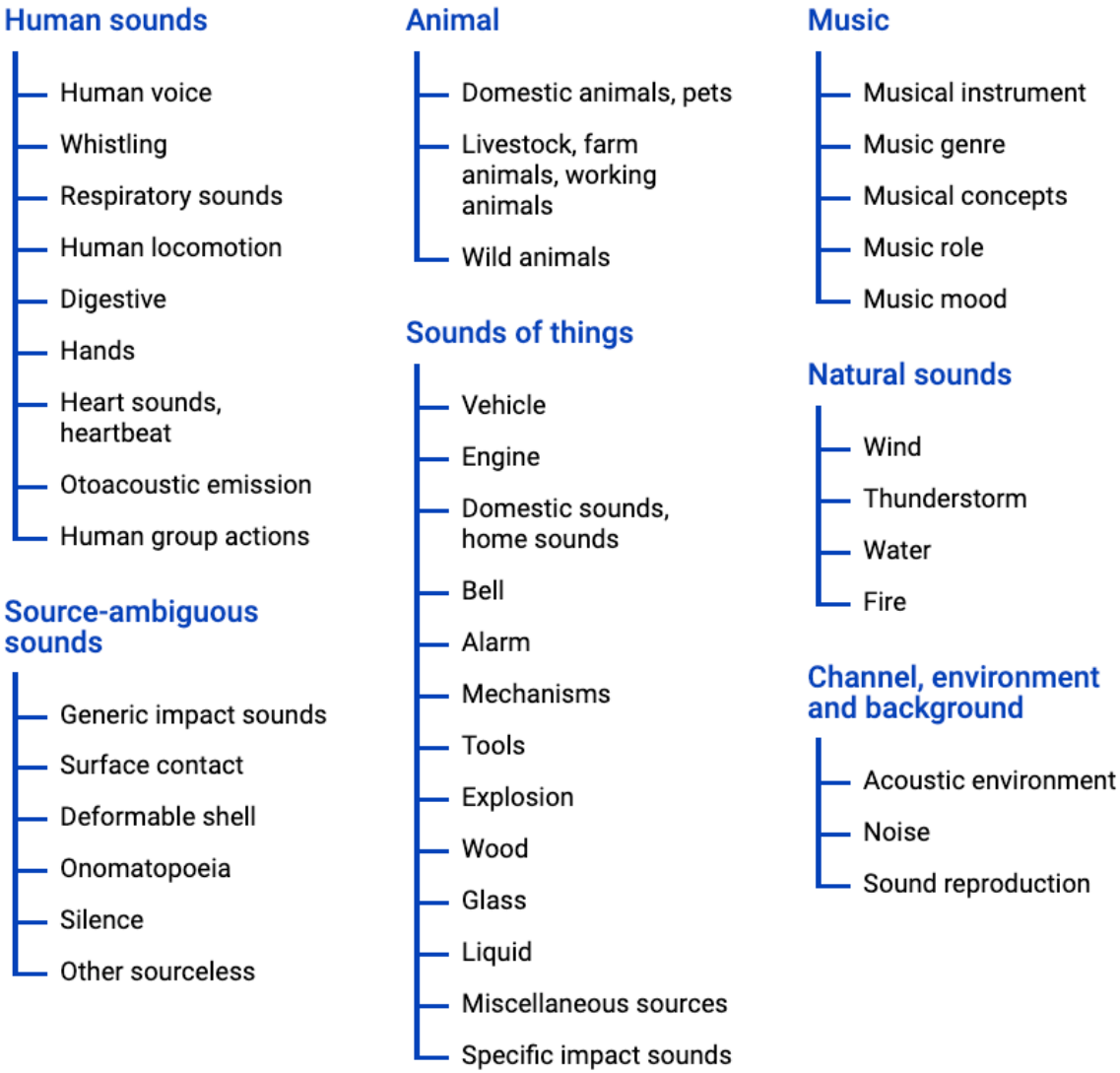
- 音のImageNetに相当する包括的な大規模データセット。
- YouTubeのビデオクリップ10秒区間に対して含まれる音のクラスを(複数)ラベル付け。
- 210万(2.1M)サンプル。 ⇔ ImageNetは約128万サンプル。
  ※現在入手可能なものは180万サンプル程度。
- **音声・音楽・環境音などから幅広い種類のクラス。**
  - Acoustic Event Recognition、または Audio Tagging と呼ばれるタスク設定。
  - 527クラス、マルチラベル。
- 画像分野のCNN(VGG)を利用したベースライン。
- **このデータセットでの性能を競って事前学習モデルが次々と提案されている。**(ImageNet同様の状況)



https://research.google.com/audioset/

環境音を含め、**一般の音を包括的に扱える**
モデル(深層表現)の提案が活発に行われている。

参考)
AudioSet Ontology



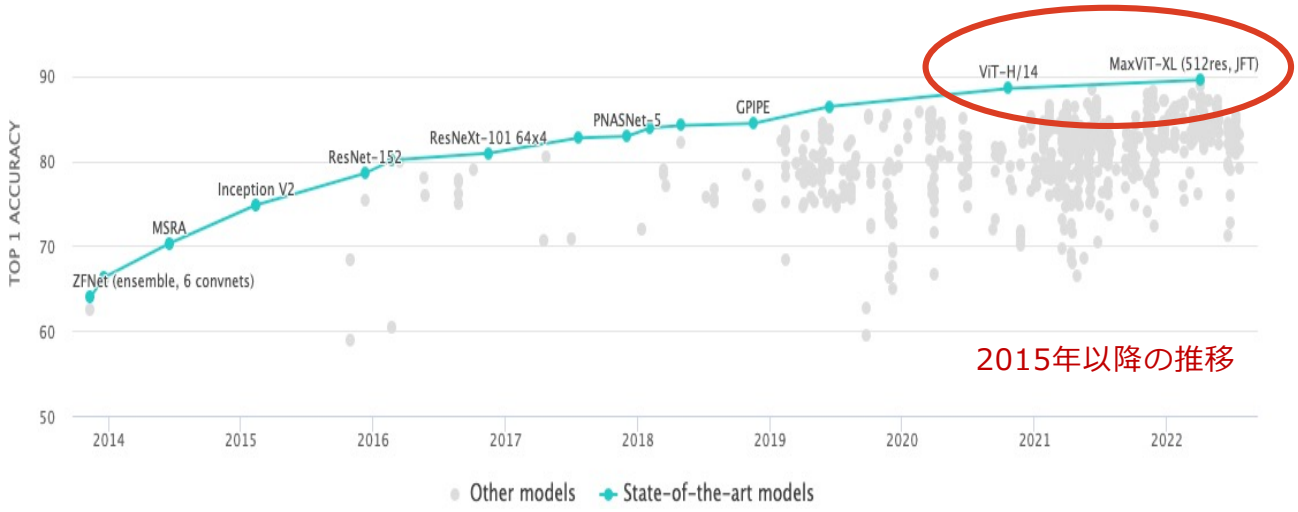
https://research.google.com/audioset/ontology/index.html

画像分野の深層学習モデルを転用

画像分野での深層学習モデルの発展

ImageNetにおける2012年のブレークスルー以降、画像認識向けに提案される深層学習モデルの性能が向上を続けている。

2021年からはTransformerアーキテクチャ(ViT)がCNNに置き換わる状況に達している。



2015年以降の推移

https://paperswithcode.com/sota/image-classification-on-imagenet

認識：2012年以降のエラー率の変化

		Error
Before ディープラーニング	Imagenet 2011 winner (not CNN)	25.7%
After ディープラーニング	Imagenet 2012 winner 以降CNN	16.4% (Krizhesvky et al.)
	Imagenet 2013 winner	11.7% (Zeiler/Clarifai)
	Imagenet 2014 winner	6.7% (GoogLeNet)
	Baidu Arxiv paper:2015/1/3	6.0%
	Human: Andrej Karpathy	5.1%
	Microsoft Research Arxiv paper: 2015/2/6	4.9%
	Google Arxiv paper: 2015/3/2	4.8%
	Microsoft Research CVPR paper: 2015/12/10	3.6%

2015年2月には人間の精度を超えた

画像認識で人間の精度を超えることは数十年間、実現されていなかった 16

人工知能は人間を超えるか―ディープラーニングの先にあるもの
https://www.jsme.or.jp/iip/Japanese/Newsletter/No47/matsuo.pdf

2015年までの推移

AudioSet: 画像分野のモデルが上位を占める

NTT

ベースとなるモデル

Rank	Model	Test mAP	AUC	d-prime	Extra Training Data	Paper	Code	Result	Year	Tags
1	**MBT** (AS-500K training + Video)	0.521			×	Attention Bottlenecks for Multimodal Fusion			2021	
2	**PaSST** (Ensemble)	0.496			✓	Efficient Training of Audio Transformers with Patchout			2021	Transformer
3	**HTS-AT** (Ensemble)	0.487			✓	HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection			2022	Transformer
4	**AST** (Ensemble)	0.485			✓	AST: Audio Spectrogram Transformer			2021	Transformer
5	**PSLA** (Ensemble)	0.474	0.981	2.936	✓	PSLA: Improving Audio Tagging with Pretraining, Sampling, Labeling, and Aggregation			2021	CNN
6	**Audio-MAE** (local, AS-2M)	0.473			×	Masked Autoencoders that Listen			2022	Self-Supervised Learning
7	**HTS-AT** (Single)	0.471			✓	HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection			2022	Transformer
8	**PaSST-S** (Single)	0.471			✓	Efficient Training of Audio Transformers with Patchout			2021	Transformer
9	**MaskSpec** (AS-2M)	0.471			×					Self-Supervised Learning
10	**Audio-MAE** (global, AS-2M)	0.468			×	Masked Autoencoders that Listen			2022	Self-Supervised Learning

DeiT

Swin Transformer

DeiT

EfficientNet (CNN)

ViT

Swin Transformer

DeiT

ViT

ViT

https://paperswithcode.com/sota/audio-classification-on-audioset

2022/9時点での
AudioSetの
クラス分類性能
ランキング

音の分野でも、**画像分野で高性能を記録したモデルが次々と応用されている**

強力なモデルの普及例～DCASE主催Kaggleコンペ

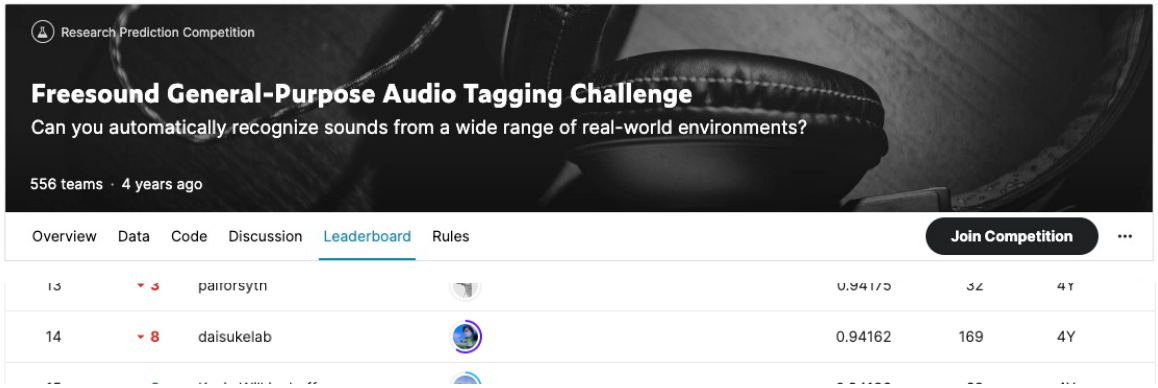
DCASE2018, 2019では、Task2 Audio tagging タスクが Kaggleで開催された。

- 画像に関するコンペが多く、音に関するものは少なかった。
- 画像コンペでは、CNNが既に優位な手法として確立。
- 音に関しても自然とCNNを使ったアプローチが注目を集め、上位を占めた。

➔ **強力なアプローチを採用しない理由がない。**

➔ 現在のAudioSetランキング。

https://dcase.community/challenge2018/task-general-purpose-audio-tagging



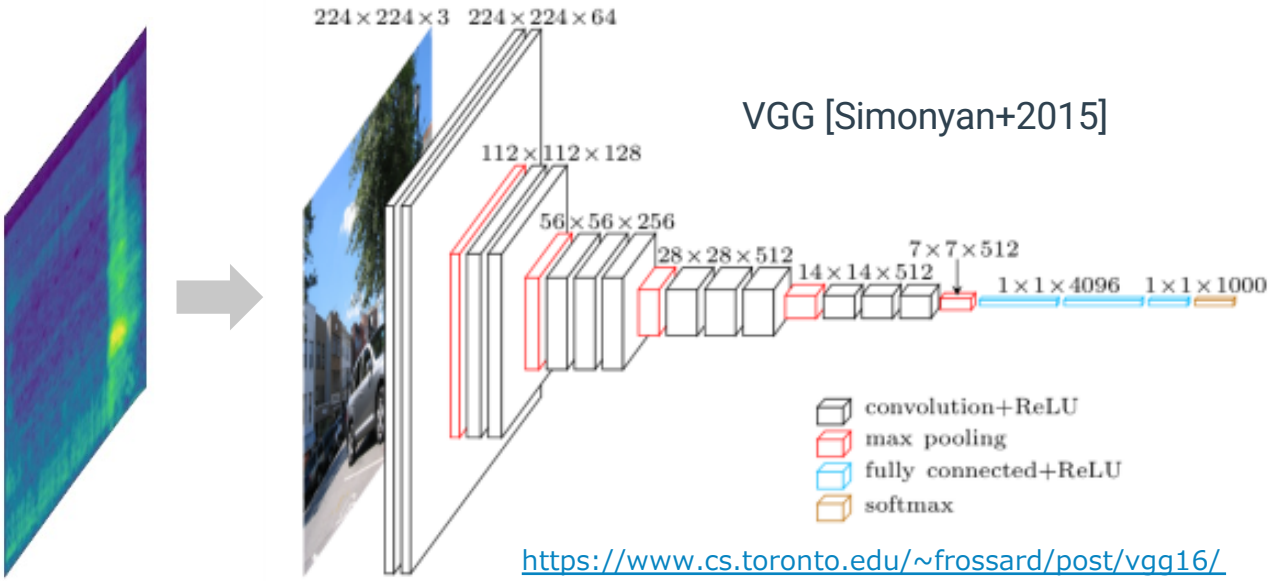
https://www.kaggle.com/c/freesound-audio-tagging

どのように画像分野のモデルを転用するか
～音響信号をスペクトログラムへと変換

時間周波数解析により音響信号をスペクトログラムに変換することで、
画像に置き換えてモデルに入力できる。 ➔ **画像分野の強力なモデルが利用できる。**

【ただし注意が必要】

- 入力データサイズ
  - 時間長固定・可変
- ノーマライズ方法
- 出力サイズ
- フィルタサイズ
- ストライド、…



VGG [Simonyan+2015]

https://www.cs.toronto.edu/~frossard/post/vgg16/

特にCNNを応用する際はネットワークパラメータと扱う
データサイズの関係を慎重に設計する必要がある

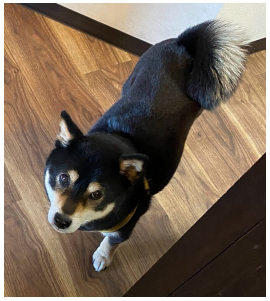
時間周波数解析: スペクトログラムへの変換

- 時間周波数解析により波形をスペクトログラムへ、画像と同等に扱える平面データに変換。
- 周波数・振幅(パワー)を対数軸で扱う、log-mel spectrogramで扱われることが多い。
- 一つのサンプルからモノラル信号のパワーのみで構成される$R^{F\times T}$データの取り扱いが多い。

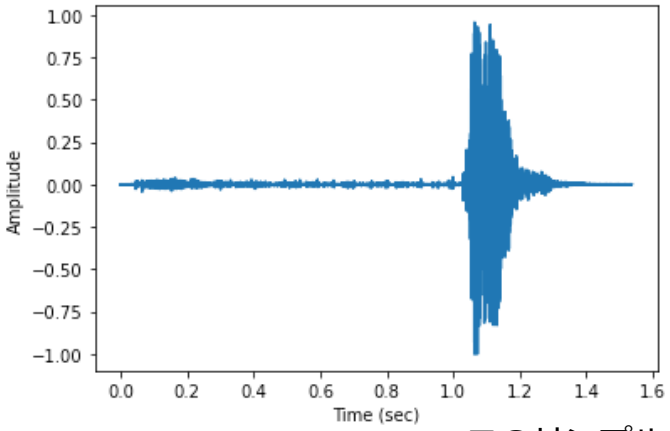
波形データ → スペクトログラム

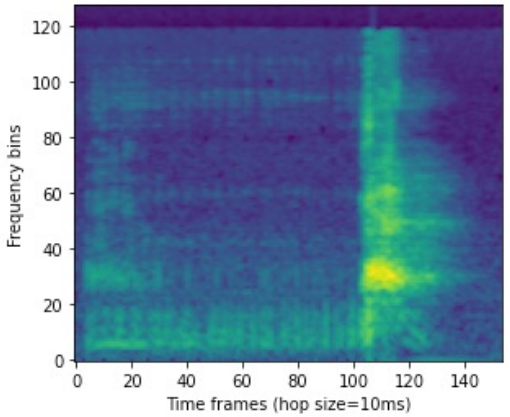
```
In [28]: wav, sr = torchaudio.load('viz/ToroBark2s.wav')
         wav = wav.mean(0)
         plt.plot(np.linspace(0.0, len(wav)/sr, len(wav)), wav)
         plt.xlabel('Time (sec)'); plt.ylabel('Amplitude');
```

```
In [34]: X = torchaudio.transforms.MelSpectrogram(sample_rate=sr, n_fft=2048,
                hop_length=sr//100)(wav).log()
         plt.imshow(X, origin='lower')
         plt.xlabel('Time frames (hop size=10ms)'); plt.ylabel('Frequency bins');
```



ウー、ワン!





このサンプル例は1.54s

この例の
要素数(FxT)
=128x154

汎用音響信号表現

～幅広く一般の音響信号に適用可能な表現を目指して～

汎用音響信号表現とは?

"General-purpose Audio Representation" = 「汎用音響信号表現」
≈汎用的に有効な特徴量表現

※発表者が対応させた日本語

- 新しい研究対象で、明確な定義は存在しない様子。
- 「**幅広い問題解決に役立つ音響信号の汎用表現**」と考えられる。
  - 究極的にはファインチューニングなしに万能な表現。

学習済みモデルがそのまま使えるということ

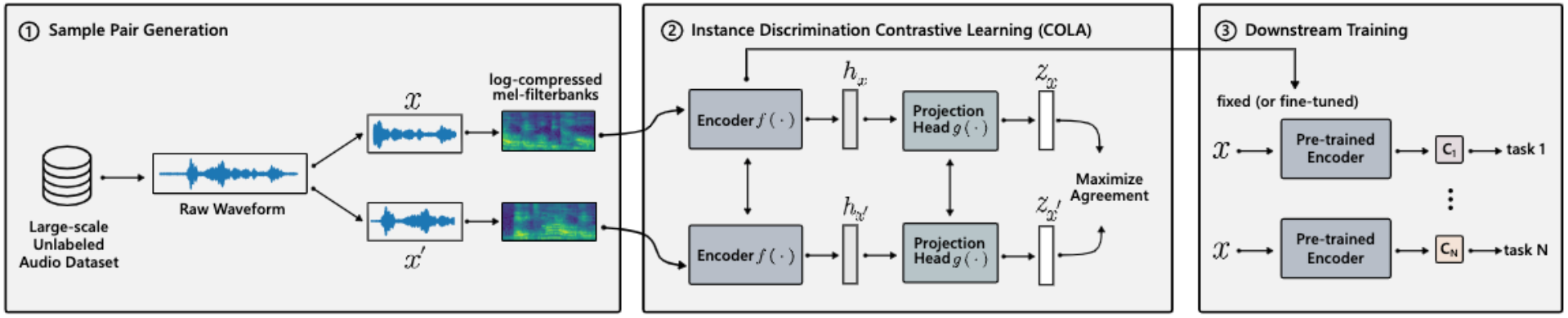
- 既存手法の特徴:
  1. 深層学習モデルを学習して得る**深層表現**。
  2. **自己教師あり**(教師ラベルなし)による学習方法が主流。
  3. **複数のモダリティ**を利用した学習も可能。

汎用音響信号表現: これまでの経緯

環境音分析の発展に加えて、画像分野の発展が強く影響。

- 画像分野のモデルの転用が一般的になっている状況。
- 画像分野で自己教師あり学習の「対照学習」が発展し注目された。

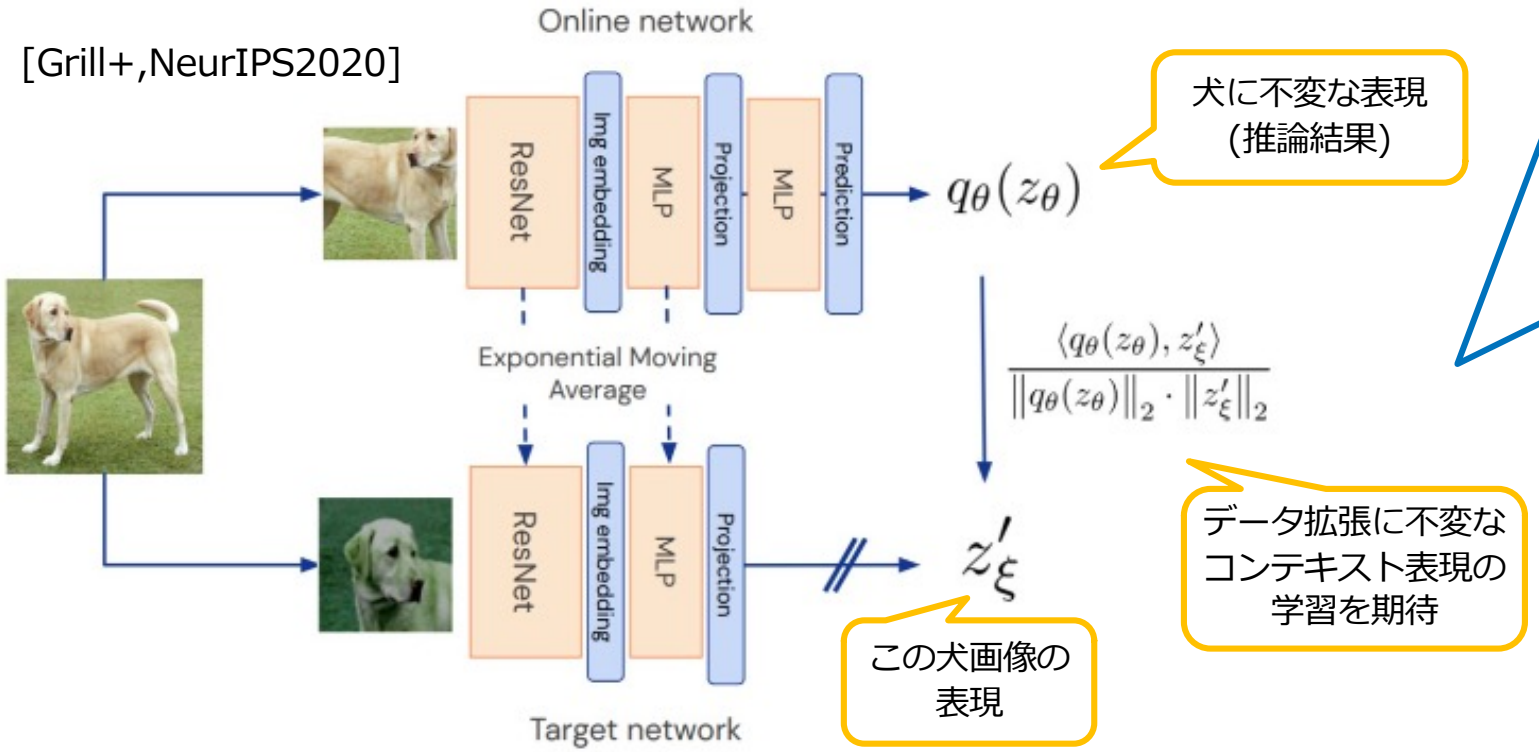
[Saeed+2021] "**Contrastive learning** of **general-purpose audio representations**."



自己教師あり学習(SSL)とは? BYOLによる例

- 入力から２つのデータ拡張結果を作り出し、オンライン側の表現からターゲット側の表現を予測する手法。ターゲットはオンラインの指数移動平均。

[Grill+,NeurIPS2020]



学習フレームワーク**自ら作り出した教師信号**で表現を学習する仕組み。

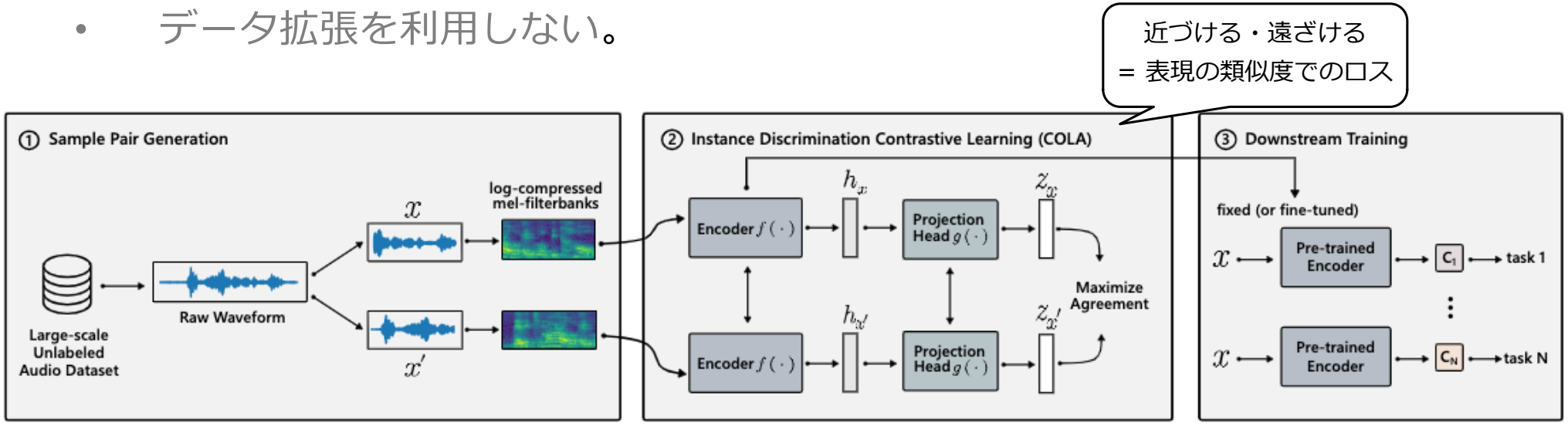
ラベル不要で、**より大規模なデータからの学習を可能**にする。

https://www.casualganpapers.com/self-supervised-contrastive-representation-learning/BYOL-explained.html

COLA: 対照学習による手法

[Saeed+,ICASSP2021]

- 音声の時系列性を利用した対照学習(Contrastive learning)。
- 切り出し位置が近いペアの表現を近づけ、遠い位置のペアを遠ざける。
  - 同じクリップ=近い vs. 違うクリップ切り出し=遠い、の関係性を利用。
  - データ拡張を利用しない。



[2] Saeed, Aaqib, David Grangier, and Neil Zeghidour. "Contrastive learning of general-purpose audio representations." *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

COLAによる性能の向上

- 性能はそれまでの手法と比較して著しく向上。

Table 2. Test accuracy (%) of a linear classifier trained on top of COLA embeddings or baseline pre-trained representations.

	CBoW [16, 25]	SG [16, 23]	TemporalGap [16, 25]	Triplet Loss [16, 25]	TRILL [13]	COLA
Speaker Id. (LBS)	99.0	**100.0**	97.0	**100.0**	-	**100.0**
Speech commands (V2)	30.0	28.0	23.0	18.0	-	**62.4**
Acoustic scenes	66.0	67.0	63.0	73.0	-	**94.0**
Birdsong detection	71.0	69.0	71.0	73.0	-	**77.0**
Music, Speech and Noise	98.0	98.0	97.0	97.0	-	**99.1**
Music instrument	33.5	34.4	35.1	25.7	-	**63.4**
Speech commands (V1)	-	-	-	-	**74.0**	71.7
Speaker Id. (Voxceleb)	-	-	-	-	17.7	**29.9**
Language Id.	-	-	-	-	**88.1**	71.3
Average (TRILL tasks)	-	-	-	-	**59.9**	57.6
Average (non-TRILL)	66.25	66.0	64.3	64.4	-	**82.5**

[2] Saeed, Aaqib, David Grangier, and Neil Zeghidour. "Contrastive learning of general-purpose audio representations." *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

汎用音響信号表現: その後の発展

NTT

[1] 2019/05 Self-supervised audio representation learning for mobile devices, Pre-Training Audio Representations with Self-Supervision

初出と考えられる。※発表者調べ

[19] 2019/12 PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition (≈ GPAR, TASLP)

2021

[2] 2020/10 (COLA) **Contrastive Learning of General-Purpose Audio Representations** (ICASSP2021)

[3] 2021/03 ※ BYOL for Audio: Self-Supervised Learning for General-Purpose Audio Representation (IJCNN2021)

[4] 2021/04 Multimodal Self-Supervised Learning of General Audio Representations

[5] 2021/09 BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition (≈ GPAR)

[6] 2021/10 SSAST: Self-Supervised Audio Spectrogram Transformer (≈ GPAR)

[7] 2021/10 Conformer-Based Self-Supervised Learning For Non-Speech Audio Tasks (ICASSP2022)

[8] 2021/10 DECAR: Deep Clustering for learning general-purpose Audio Representations

2022

[9] 2021/11 Towards Learning Universal Audio Representations (ICASSP2022)

[10] 2022/03 DeLoRes: Decorrelating Latent Spaces for Low-Resource Audio Representation Learning

[11] 2022/03 MAE-AST: Masked Autoencoding Audio Spectrogram Transformer (≈ GPAR)

[12] 2022/04 ※ Masked Spectrogram Modeling using Masked Autoencoders for Learning General-purpose Audio Representation

[13] 2022/04 Masked Spectrogram Prediction For Self-Supervised Audio Pre-Training (≈ GPAR)

[14] 2022/04 ATST: Audio Representation Learning with Teacher-Student Transformer

[15] 2022/05 Self-Supervised Learning Method Using Multiple Sampling Strategies for General-Purpose Audio Representation (ICASSP2022)

[16] 2022/05 ※ Composing General Audio Representation by Fusing Multilayer Features of a Pre-trained Model (EUSIPCO2022)

[17] 2022/06 BYOL-S: Learning Self-supervised Speech Representations by Bootstrapping

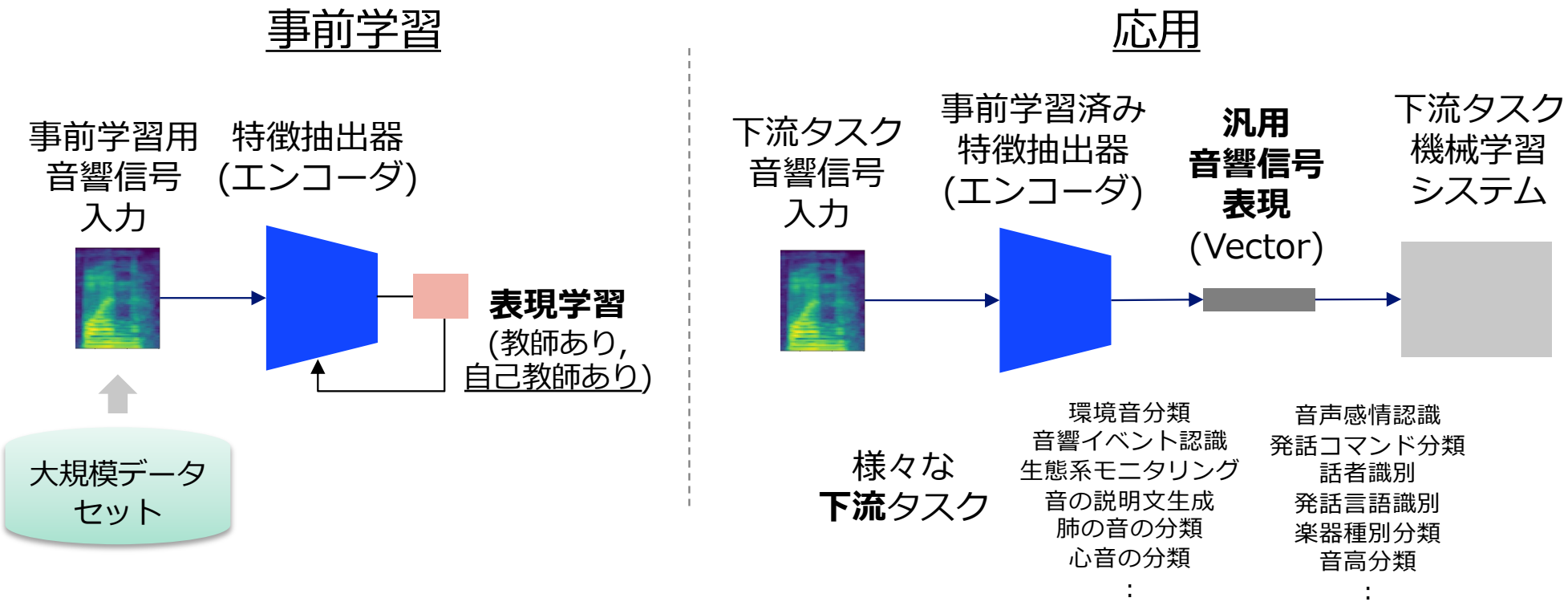
[18] 2022/07 (Audio-MAE) Masked Autoencoders that Listen

2021年後半から増加傾向。

※＝発表者の提案

汎用音響信号表現: 研究の枠組み

事前学習により得られた表現を転移学習して、様々な応用タスクの性能で評価。
＝下流タスク



汎用音響信号表現: 下流タスク例

【HEAR Benchmark】

4つの**環境音**タスク (Environmental sound)

- 環境音分類 ESC-50, FSD50K
- 録音位置分類Gunshot Triangulation
- 蜂の女王不在検知Beehive States

5つの**音声**タスク

- 言語識別VoxLingua Top 10 (Vo10)
- 擬音識別Vocal Imitations (Imit)
- 音声感情認識CREMA-D (CR-D)
- 音声コマンド分類Speech Commands v2 (SPC)
- 話者数カウントLibriCount (Cnt)

6つの**音楽**タスク (Music)

- 音楽/音声分類GTZAN Music Speech (M/S)
- 音楽ジャンル分類GTZAN (GTZ)
- 音高識別NSynth Pitch (PT)
- 打楽器音色またはストローク分類Mridingham Stroke and Tonic (MTonまたはNStk)
- 打楽器音分類Beijing Opera (Perc)

【BYOL-A, MSM-MAE利用タスク】

音事象(**環境音**)認識タスク

ESC-50: 50環境音クラス分類

UrbanSound8K: 10環境音クラス分類

非意味的発話(**音声**)認識タスク

SPCV2: 35音声コマンドクラス分類

VC1: 1,251話者クラス分類

VF: 6言語クラス分類

CRM-D: 6音声感情クラス認識

音楽タスク

GTZAN: 10音楽ジャンルクラス認識

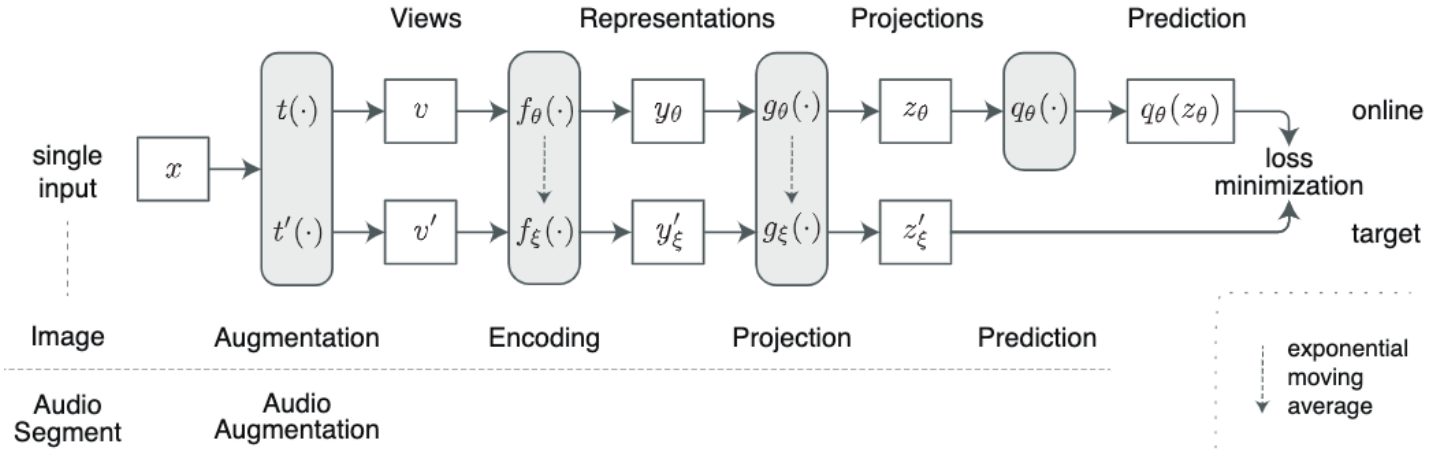
NSynth: 11楽器クラス分類

Surge: 88MIDIノートクラス分類

BYOL-A: BYOLを利用した手法

[Niizumi+(NTT),IJCNN2021]

- 対照学習と同時期に提案されたBYOL (Bootstrap Your Own Latent)を利用。
- 時系列性の前提条件を使わない。
  - 不採用「切り出し位置が近いペアの表現を近づけ、遠い位置のペアを遠ざける」
  - 採用　「**データ拡張に不変な表現を学習**」



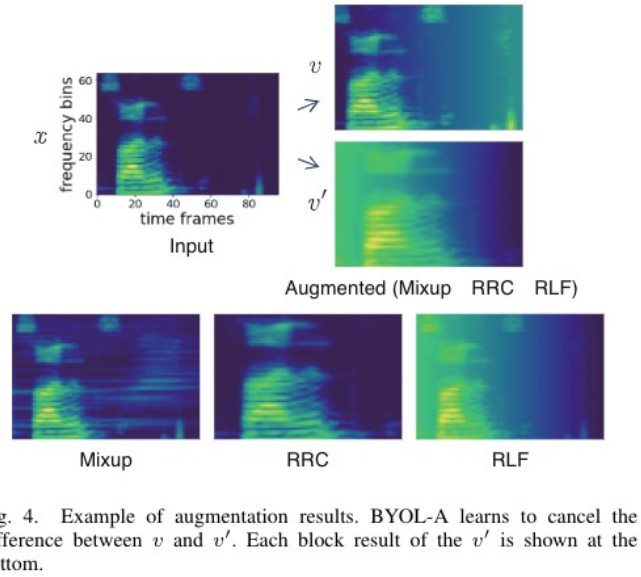


Fig. 4. Example of augmentation results. BYOL-A learns to cancel the difference between v and v'. Each block result of the v' is shown at the bottom.

[3] Niizumi, Daisuke, et al. "BYOL for audio: Self-supervised learning for general-purpose audio representation." *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021.

BYOL-Aによる性能の向上

- COLAと比較して更に性能を向上。

TABLE I
PERFORMANCE COMPARISON RESULTS FOR DOWNSTREAM TASKS

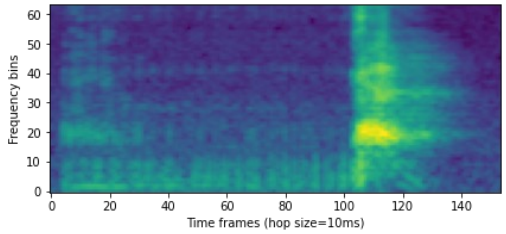
Method	Dim.	Remarks	NS	US8K	VC1	VF	SPCV2/12	SPCV2	Average
TRILL [13]		conventional	N/A	N/A	17.9%	88.1%	74.9%	N/A	N/A
COLA [14]		conventional	63.4%	N/A	29.9%	71.3%	71.7%	62.4%	N/A
OpenL3 [20] [1]		reference	N/A	78.2%	N/A	N/A	N/A	N/A	N/A
COALA [19] [2]		reference	73.1%	72.7%	N/A	N/A	N/A	N/A	N/A
COLA'	512-d	our impl.	65.4%	76.3%	25.9%	73.5%	59.1%	63.1%	60.5%
COLA'	1024-d	our impl.	69.3%	77.1%	31.2%	76.7%	71.9%	71.0%	66.2%
COLA'	2048-d	our impl.	70.2%	78.5%	30.4%	79.5%	76.7%	76.8%	68.7%
BYOL-A	512-d	**proposed**	69.1%	78.2%	33.4%	83.5%	86.5%	88.9%	73.3%
BYOL-A	1024-d	**proposed**	72.7%	78.2%	38.0%	88.5%	90.1%	91.4%	76.5%
BYOL-A	2048-d	**proposed**	**74.1%**	**79.1%**	**40.1%**	**90.2%**	**91.0%**	**92.2%**	**77.8%**

[3] Niizumi, Daisuke, et al. "BYOL for audio: Self-supervised learning for general-purpose audio representation." *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021.

そもそも表現とは具体的に?

- 典型的には実数列のベクトル。

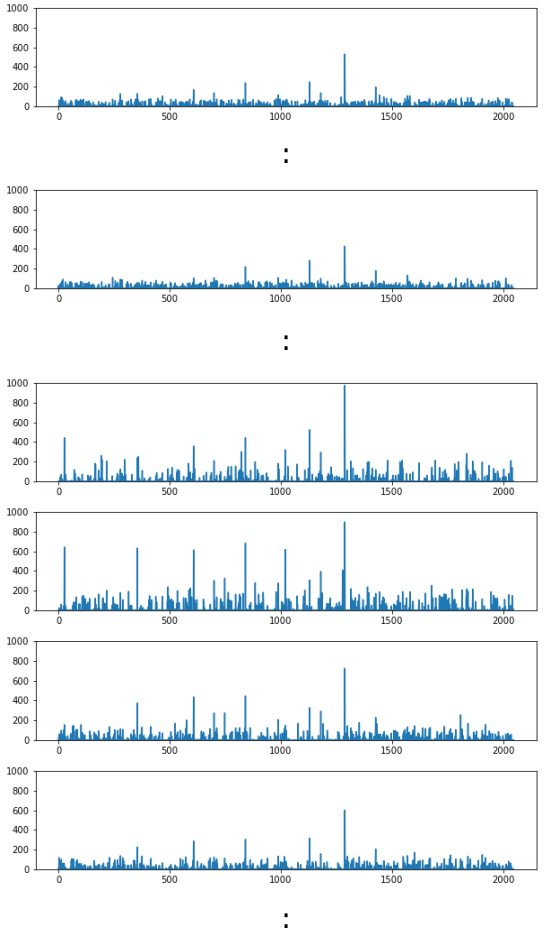
ウー、ワン!



BYOL-A

[Niizumi+(NTT), IJCNN2021]

(時間フレームごとに出力した場合)

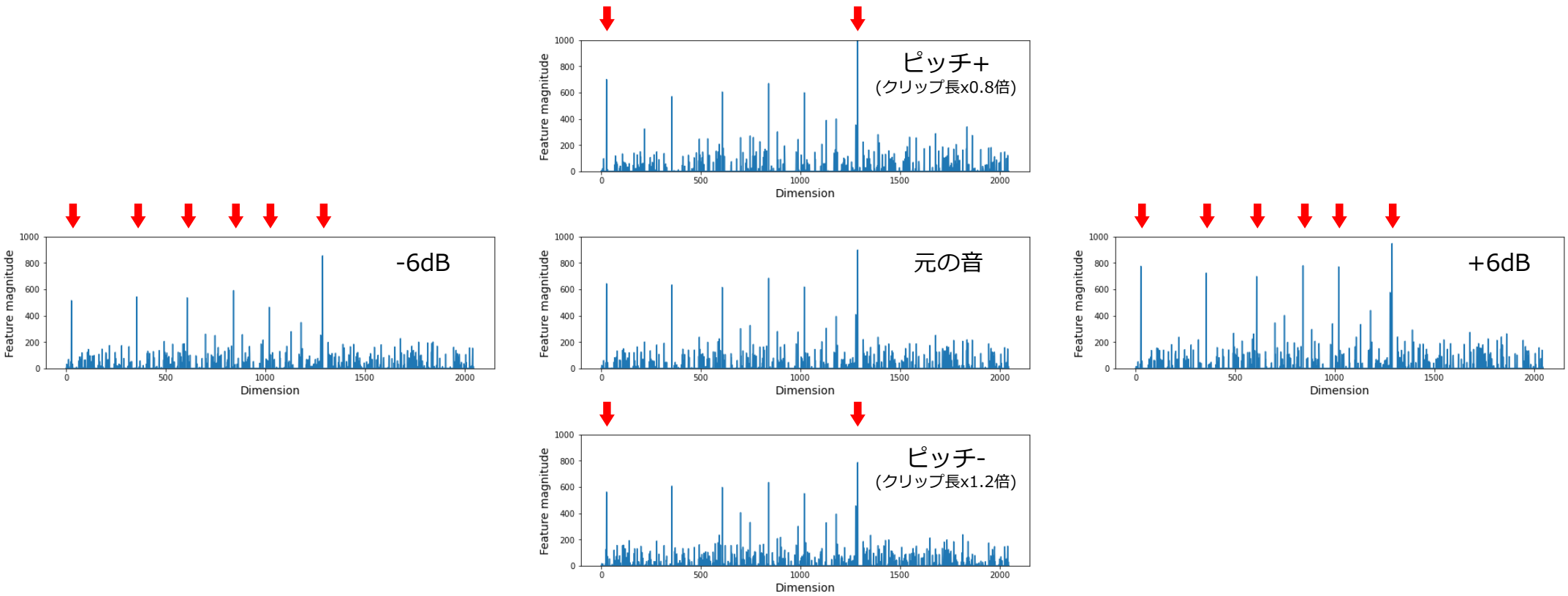


表現ベクトル例 (t=105付近, 2048次元)

[0.0, ..., 41.2, 55.1, 0.0, 975.6, 0.0, .., 0.0, 180.0, ...]

例) ピッチや音量が変わると表現は?

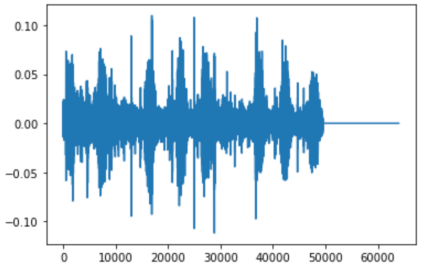
- BYOL-A を使った例
  → ベクトルに含まれるいくつかの次元の数値が変動している。



可視化例 – 各手法の表現ベクトル

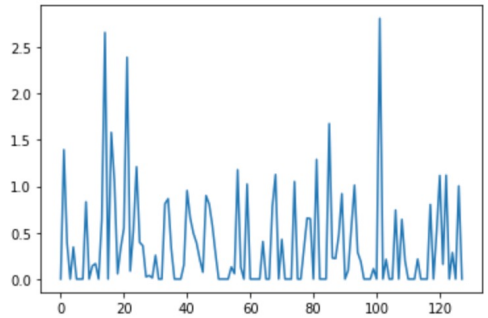
```
# Example waveform
x, y = test_loader.dataset[0]
plt.plot(x); print(test_loader.dataset.classes[y]); print(x[:8])
```

```
dog_bark
tensor([-0.0021, -0.0072, -0.0083, -0.0081, -0.0035,  0.0019,  0.0113,  0.0201])
```




```
cfg = load_yaml_config('config/vggish.yaml')
vggish = evar.ar_vggish.AR_VGGish(cfg)
feature_vggish = vggish(x.unsqueeze(0))[0].detach().numpy()
plt.plot(feature_vggish); print(feature_vggish.shape);  print(feature_vggish[0:6])
```

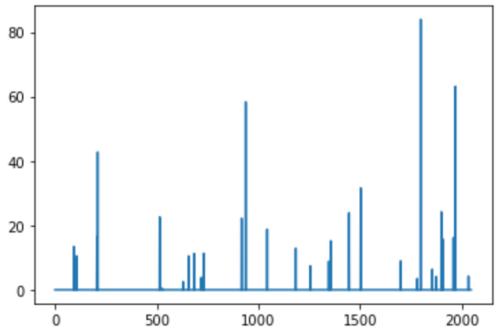
```
(128,)
[0.        1.3931446 0.39216742 0.        0.3441395 0.       ]
```



VGGish

```
feature_cnn14 = cnn14(x.unsqueeze(0))[0].detach().numpy()
plt.plot(feature_cnn14); print(feature_cnn14.shape);  print(feature_cnn14[500:520])
```

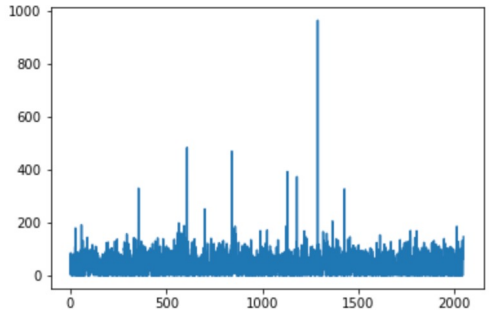
```
(2048,)
[ 0.       0.       0.       0.       0.       0.       0.
  0.       0.       0.       0.       0.       0.       0.
  0.       0.       0.       0.       0.       0.      ]
```



CNN14
スパースな表現

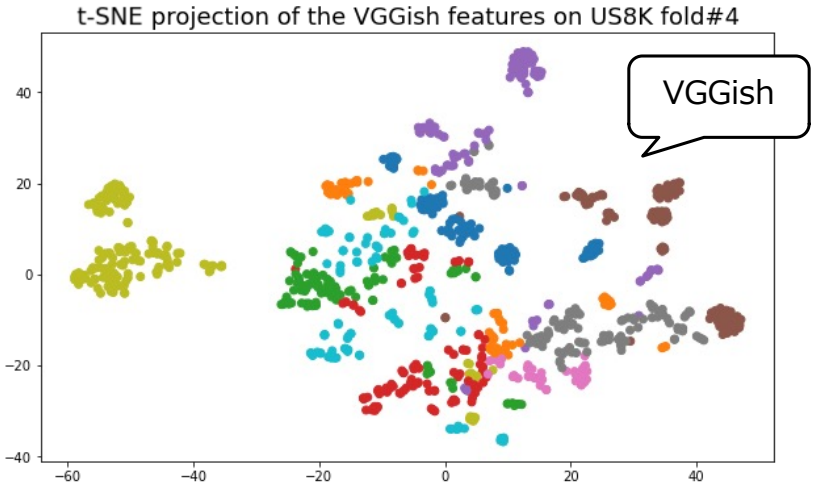
```
from evar.ar_base import calculate_norm_stats
cfg = load_yaml_config('config/byola.yaml')
byola = evar.ar_byola.AR_BYOLA(cfg)
byola.norm_stats = torch.tensor([-6.465687, 4.7879076]) # precomputed by:
feature_byola = byola(x.unsqueeze(0))[0].detach().numpy()
plt.plot(feature_byola); print(feature_byola.shape);  print(feature_vggish[0:6])
```

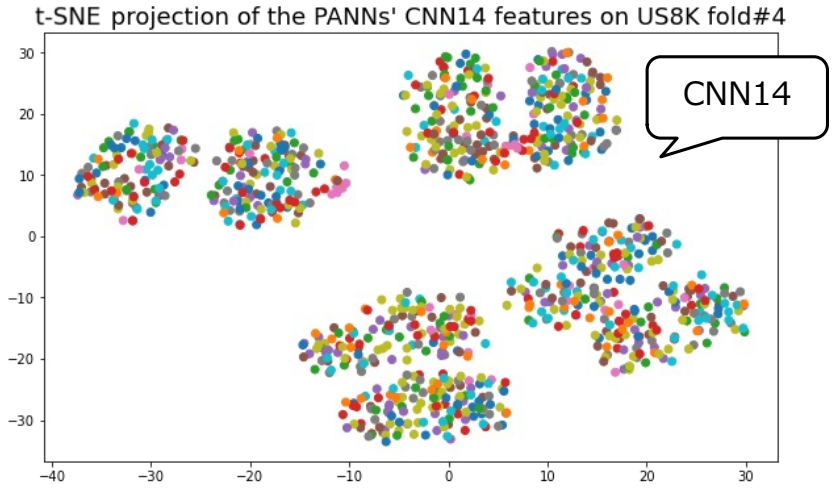
```
(2048,)
[0.        1.3931446 0.39216742 0.        0.3441395 0.       ]
```

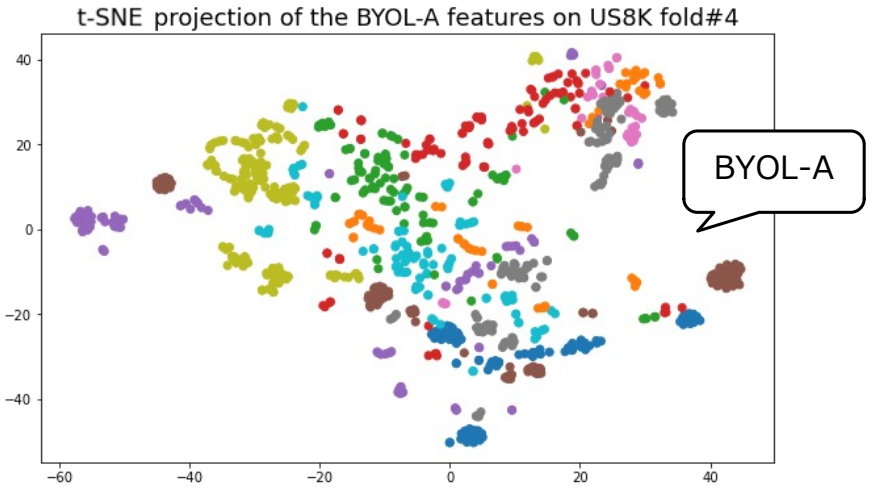


BYOL-A

可視化例 – UrbanSound8Kの各表現







Representation	SER tasks ESC-50	US8K
[S] VGGish [5]	68.2 ± 1.1	75.1 ± 0.3
[S] VGGish-4K [5]	79.5 ± 0.4	78.5 ± 0.3
[Sas] PANNs [12]	90.1 ± 0.4	82.0 ± 0.7
[Sas] ESResNeXt [14]	89.0 ± 1.2	84.3 ± 0.4
[Sas] AST [15]	**93.5** ± **0.4**	**85.5** ± **0.2**
[Ux] COALA [16]	74.7 ± 1.3	71.9 ± 1.0
[Ux] OpenL3-E [17]	81.2 ± 1.3	80.7 ± 0.4
[Ux] OpenL3-M [17]	82.2 ± 0.8	80.4 ± 0.3
[U] TRILL [19]	75.4 ± 0.7	75.2 ± 1.3
[U] Wav2Vec2-F [40]	65.6 ± 1.7	67.8 ± 0.3
[U] Wav2Vec2-C [40]	57.6 ± 0.8	66.9 ± 0.4
[U] BYOL-A	83.2 ± 0.6	79.7 ± 0.5

性能はどの手法も
それほど変わらない。

汎用音響信号表現: 学習方法

これまでご紹介してきた内容では、

- **自己教師あり学習(Self-supervised learning; SSL)**の手法
  →**明示的な汎用音響信号表現。**

しかしながら、

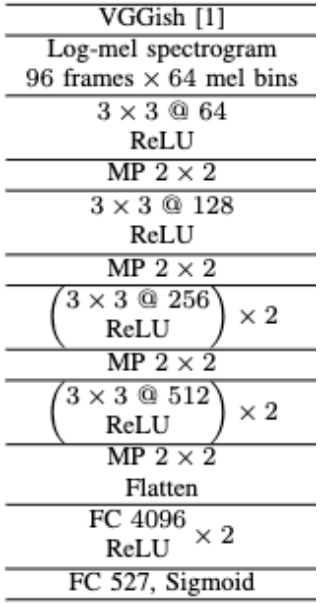
- AudioSetなどを**教師あり学習**する手法
  → **非明示的な汎用表現**として実質的に利用されている。

教師あり学習による手法

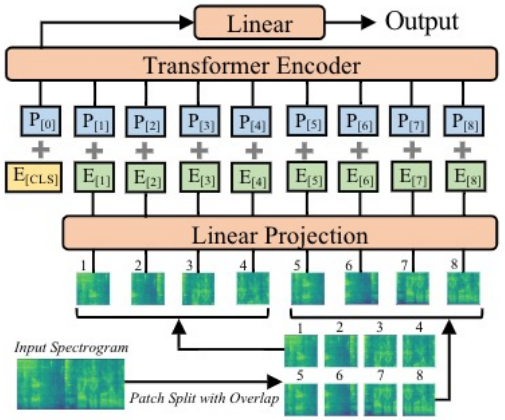
～画像分野のImageNet学習済みモデルと似た発展

教師あり学習～暗黙の汎用表現学習

- **大規模データセット**によるラベルを利用した**教師あり学習**。
  - VGGish (2017): VGGをベースにしたモデルでYouTube-8Mを学習。
  - AST (2021): ViTベースのモデルでImageNet事前学習を初期値にAudioSetを学習。
- **評価タスクが限られる**。(明示的な汎用手法ではない)
  - VGGishはイベント認識タスクで評価。
  - ASTはESC-50 (シーン分類), Speech commands(発話コマンド分類)で評価。
- **実際の応用研究における利用例**が見られる。



[VGGish]



[AST]

[VGGish] Hershey, Shawn, et al. "CNN architectures for large-scale audio classification." *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017.
[AST] Gong, Yuan, Yu-An Chung, and James Glass. "AST: Audio spectrogram transformer." arXiv preprint arXiv:2104.01778 (2021).

教師あり学習～応用例

- 特に**VGGish**の利用例が散見される。
- しかし**汎用的な性能が確認された上での応用ではない。**

音の説明文生成への応用例

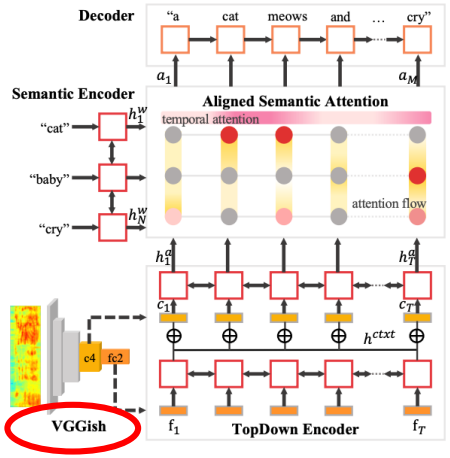


Figure 4: The audio captioning model with *top-down multi-scale encoder* and *aligned semantic attention*.

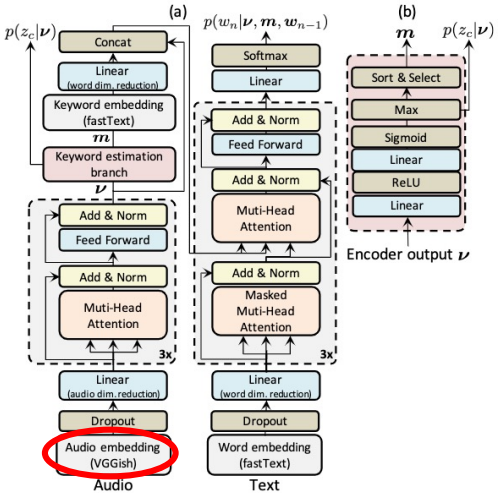


Figure 2: *(a) Architecture of TRACKE and (b) details of keyword-estimation branch $\mathcal{M}$.*

AudioCaps: Generating Captions for Audios in The Wild (Kim et al., NAACL 2019)

Y. Koizumi et al., "A Transformer-Based Audio Captioning Model with Keyword Estimation." (Interspeech 2020)

肺の音の分類への応用例

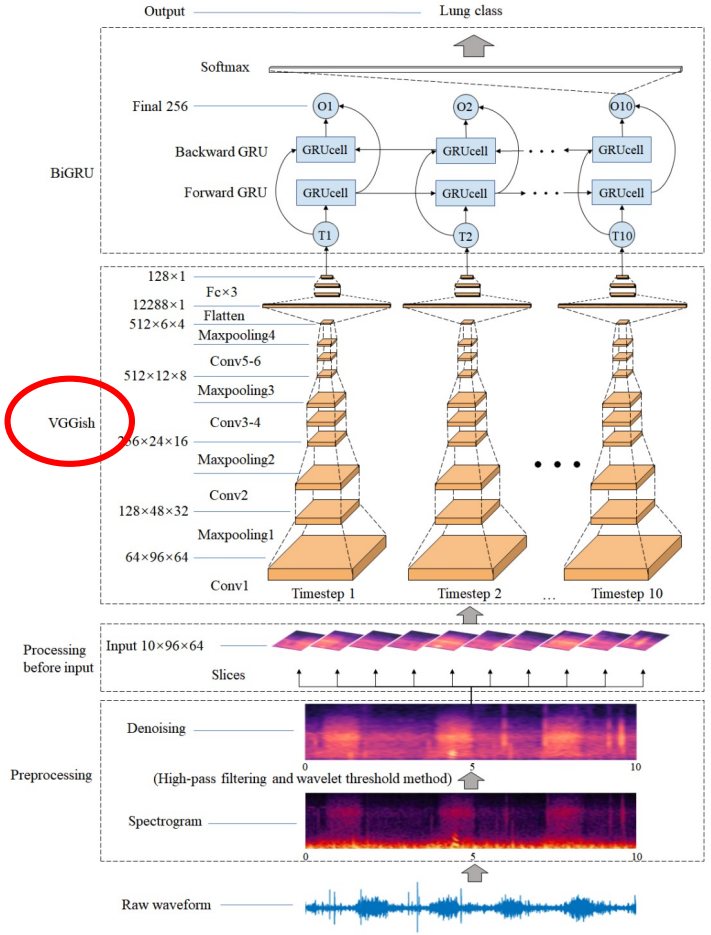


FIGURE 1. The lung sound recognition model based on VGGish-BiGRU.

L. Shi et al., "Lung Sound Recognition Algorithm Based on VGGish-BiGRU," in IEEE Access, vol. 7, pp. 139438-139449, 2019, doi: 10.1109/ACCESS.2019.2943492.

教師あり学習～汎用性能は十分か?

- 教師あり学習モデルは十分汎用に耐えるか?
  - 一般に利用される最終層付近は性能が劣ることも。
  - 層ごとに得意なタスクが異なる。

そのまま最終層を活用したのでは性能が出ない
タスクがある
(性能はタスクに依存)

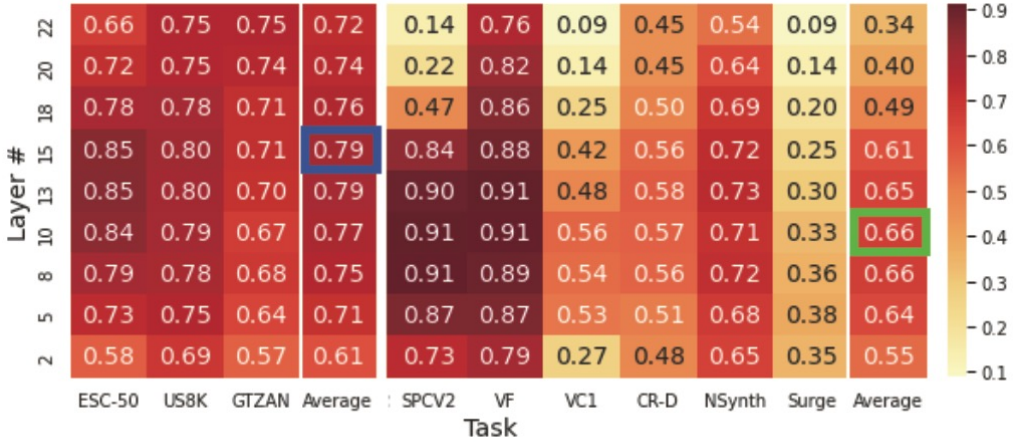


図 2: VGGish における層別の精度 (%)

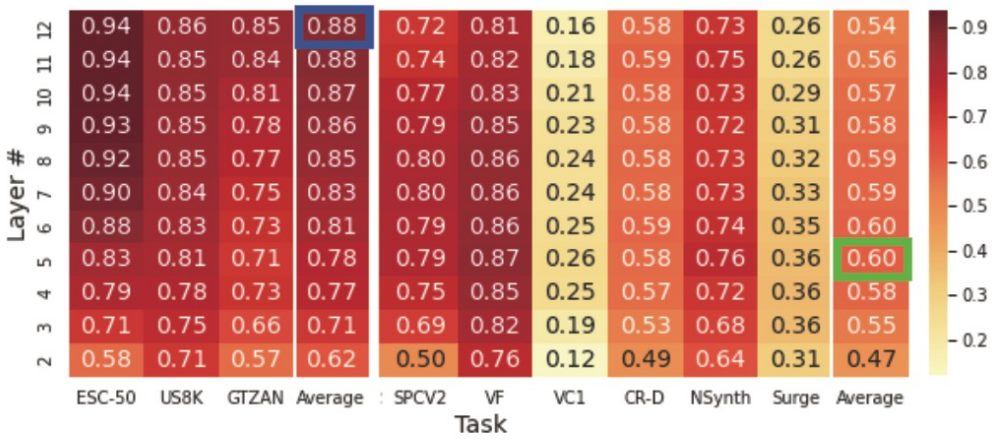


図 4: AST における層別の精度 (%)

- 仁泉+, “事前学習モデルの複数層特徴量の融合を用いた汎用音響信号表現", 信学技報 (EA2022-9) (2022).
- Niizumi, Daisuke et al. “Composing General Audio Representation by Fusing Multilayer Features of a Pre-trained Model.” ArXiv abs/2205.08138 (2022) & EUSIPCO2022.

教師あり学習～汎用性能を向上させることも可能

NTT

性能改善手法 [Niizumi+(NTT),EUSIPCO2022/信学技法2022]

中間・後半層の特徴量出力を融合 ➔ 汎用性能が向上する

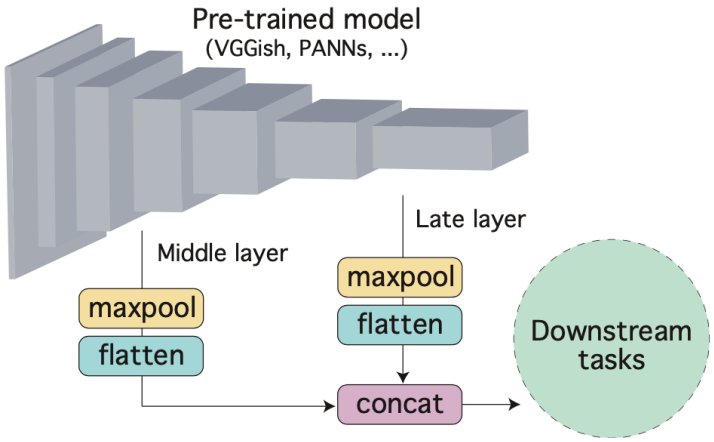


表 3: 提案手法による事前学習モデルの精度改善 (%)

	SER tasks		NOSS tasks				Music tasks			
Representation	ESC-50	US8K	SPCV2	VC1	VF	CRM-D	GTZAN	NSynth	Surge	Avg.
VGGish	68.2	75.1	14.3	9.0	75.7	44.4	75.3	53.9	8.8	47.2
VGGish-Fusion#10#15	86.5	80.9	91.4	54.5	91.6	59.3	70.8	73.6	33.3	71.3
difference	+18.2	+5.8	+77.1	+45.5	+16.0	+15.0	-4.5	+19.7	+24.5	+24.1
CNN14	90.1	82.0	51.4	8.0	75.0	50.7	79.7	66.0	10.4	57.0
CNN14-Fusion#3#6	93.0	85.8	91.3	50.6	90.5	59.0	77.4	73.8	32.4	72.6
difference	+2.9	+3.8	+39.9	+42.7	+15.5	+8.3	-2.3	+7.8	+22.0	+15.6
AST	93.5	85.5	71.8	16.5	81.2	57.9	84.3	73.2	25.8	65.5
AST-Fusion#5#12	94.2	85.5	80.4	24.9	87.6	60.7	82.9	77.6	34.6	69.8
difference	+0.6	+0.0	+8.6	+8.4	+6.4	+2.8	-1.4	+4.5	+8.9	+4.3

- 仁泉+, “事前学習モデルの複数層特徴量の融合を用いた汎用音響信号表現", 信学技報 (EA2022-9) (2022).
- Niizumi, Daisuke et al. “Composing General Audio Representation by Fusing Multilayer Features of a Pre-trained Model.” ArXiv abs/2205.08138 (2022) & EUSIPCO2022.

教師あり学習モデル: まとめ

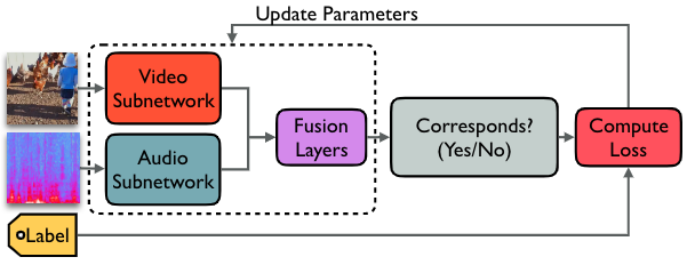
- ImageNet同様に、多様なラベルで事前学習することで汎用性が期待されている。
- 応用例が散見されるが、そのまま最終層を使うと汎用性能に劣ることが報告されている。
- 複数層の出力を組み合わせることで、汎用的に役立つものにカスタマイズできることも報告されている。

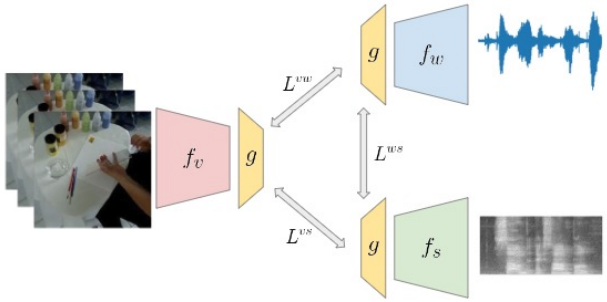
マルチモダリティを利用した手法

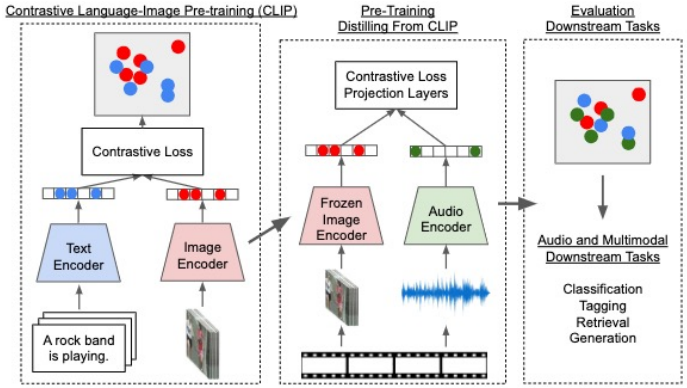
～マルチモーダルな自己教師あり学習

マルチモダリティSSL

- OpenL3
  - 音と画像の対応関係を利用した学習
- Wang et al.
  - Raw, Spectrogram, 画像の対応関係から学習
  - 汎用音響信号表現として提案
- Wav2CLIP
  - CLIP事前学習済みモデルを音響信号に蒸留して音の表現を学習。







[OpenL3] Cramer, Jason, et al. "Look, listen, and learn more: Design choices for deep audio embeddings." ICASSP 2019.
[Wang et al.] Wang, Luyu, et al. "Multimodal self-supervised learning of general audio representations." *arXiv preprint arXiv:2104.12807* (2021).
[Wav2CLIP] H. -H. Wu, P. Seetharaman, K. Kumar and J. P. Bello, "Wav2CLIP: Learning Robust Audio Representations from Clip," *ICASSP 2022*.
[CLIP] Radford, Alec et al. “Learning Transferable Visual Models From Natural Language Supervision.” *ICML* (2021).

マルチモダリティSSL: OpenL3

- 音と画像の対応関係を利用。
- 対応しているペア、対応していないペアを作成しすることで、ラベルを作り出す。
- 学習後には、Audio, Videoそれぞれの表現を出力する学習済みモデルを得る。

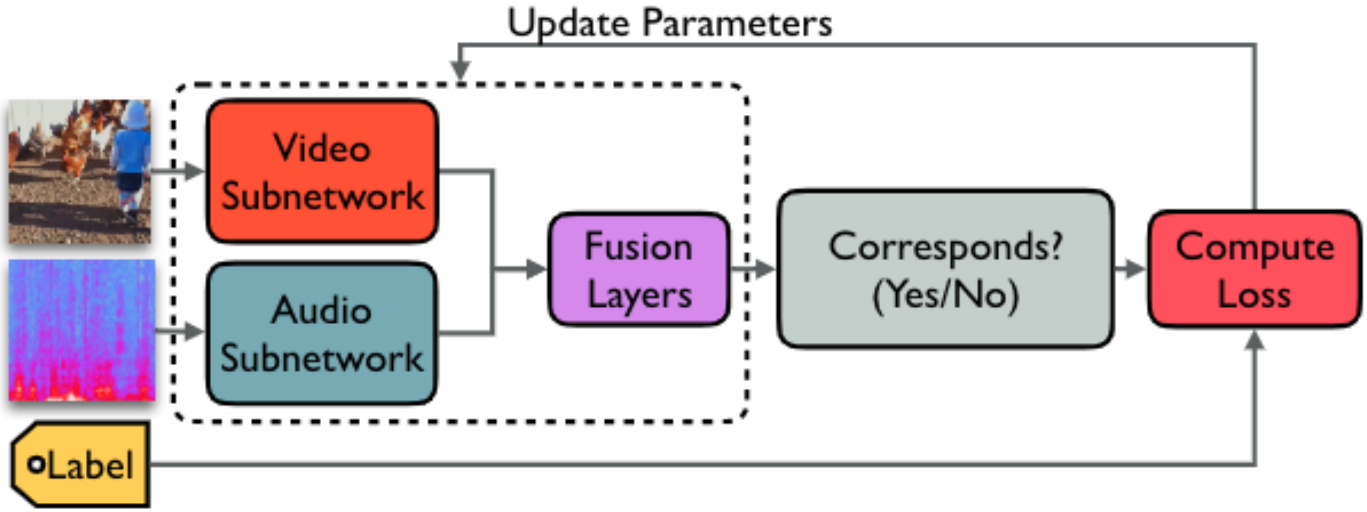


Fig. 1. High-level architecture of L^3-Net.

[OpenL3] Cramer, Jason, et al. "Look, listen, and learn more: Design choices for deep audio embeddings." ICASSP 2019.

マルチモダリティSSL: Wang et al.

時間波形・スペクトログラム・画像(映像)の対応関係を利用。

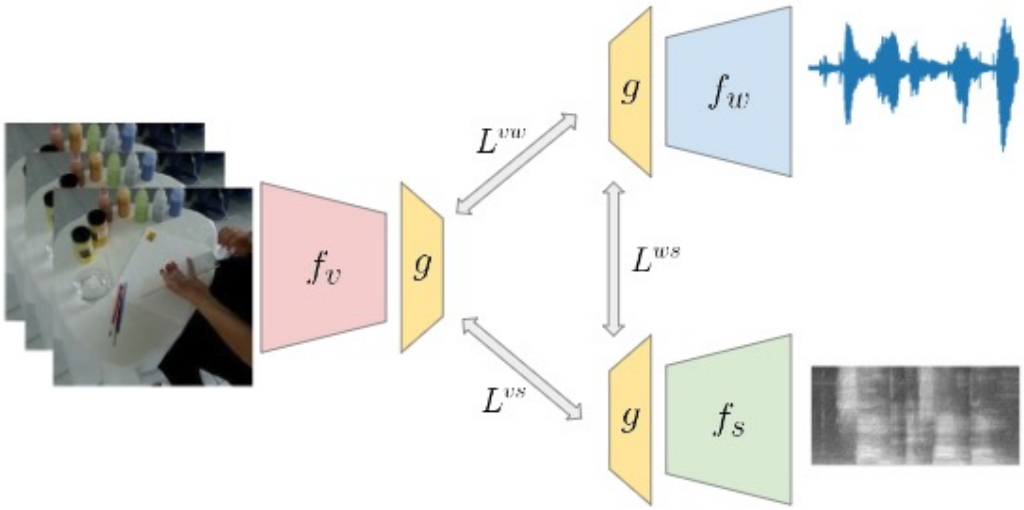


Table 5: ***Generalization to other tasks***. *We show test accuracy (%) on different downstream tasks trained with a linear classifier on top of the frozen features outputed from our pre-trained network, comparing to supervsied and unsupervsied baselines. All methods are based on EfficientNet-B0.*

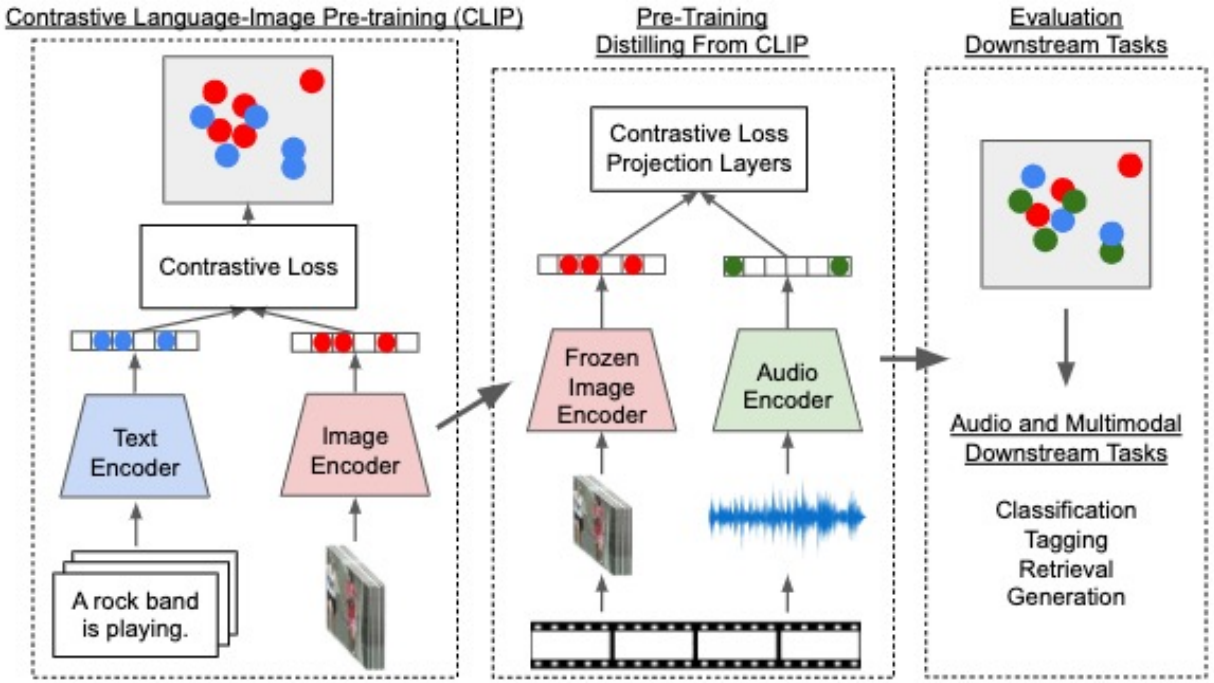
Task	COLA [44]	Ours	Sup. [30]
Speaker Id. (Librispeech)	**100.0**	99.6	-
Speech commands (V1)	71.7	**80.5**	93.4
Speech commands (V2)	62.4	**82.2**	-
Acoustic scenes	**94.0**	90.4	99.1
Speaker Id. (VoxCeleb)	29.9	**38.2**	33.1
Birdsong detection	77.0	**80.0**	81.4
Music, speech & noise	99.1	**99.6**	-
Language Id.	71.3	**79.0**	86.0
Music instrument	63.4	**68.3**	72.0
Average	74.3	**79.8**	-

[Wang et al.] Wang, Luyu, et al. "Multimodal self-supervised learning of general audio representations." *arXiv preprint arXiv:2104.12807* (2021).

マルチモダリティSSL: Wav2CLIP

- CLIP事前学習済みモデルを音響信号に蒸留して音の表現を学習。
- Zero-shot分類、などに道を拓く。
- 類似の手法が複数提案されている。

★CLIPはテキストと画像の対応を取るため幅広く利用される重要な手法として注目されている。



トレンド: 異なるモダリティのペアデータをContrastive learningで学習すると共通空間の表現が学習できる。

[Wav2CLIP] H. -H. Wu, P. Seetharaman, K. Kumar and J. P. Bello, "Wav2CLIP: Learning Robust Audio Representations from Clip," *ICASSP 2022*.
[CLIP] Radford, Alec et al. “Learning Transferable Visual Models From Natural Language Supervision.” *ICML* (2021).

最新の汎用音響信号表現手法

～画像分野 Masked Image Modeling(MIM) の応用

汎用音響信号表現: MIMベースの手法

[1] 2019/05 Self-supervised audio representation learning for mobile devices,
Pre-Training Audio Representations with Self-Supervision

[19] 2019/12 PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition (≈ GPAR, TASLP)

[2] 2020/10 (COLA) Contrastive Learning of General-Purpose Audio Representations (ICASSP2021)

2021

[3] 2021/03 ※ BYOL for Audio: Self-Supervised Learning for General-Purpose Audio Representation (IJCNN2021)

[4] 2021/04 Multimodal Self-Supervised Learning of General Audio Representations

[5] 2021/09 BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition (≈ GPAR)

[6] 2021/10 **SSAST: Self-Supervised Audio Spectrogram Transformer** (≈ GPAR)

[7] 2021/10 Conformer-Based Self-Supervised Learning For Non-Speech Audio Tasks (ICASSP2022)

[8] 2021/10 DECAR: Deep Clustering for learning general-purpose Audio Representations

2022

[9] 2021/11 Towards Learning Universal Audio Representations (ICASSP2022)

[10] 2022/03 DeLoRes: Decorrelating Latent Spaces for Low-Resource Audio Representation Learning

[11] 2022/03 **MAE-AST: Masked Autoencoding Audio Spectrogram Transformer** (≈ GPAR)

[12] 2022/04 ※ **Masked Spectrogram Modeling using Masked Autoencoders for Learning General-purpose Audio Representation**

[13] 2022/04 **Masked Spectrogram Prediction For Self-Supervised Audio Pre-Training** (≈ GPAR)

[14] 2022/04 ATST: Audio Representation Learning with Teacher-Student Transformer

[15] 2022/05 Self-Supervised Learning Method Using Multiple Sampling Strategies for General-Purpose Audio Representation (ICASSP2022)

[16] 2022/05 ※ Composing General Audio Representation by Fusing Multilayer Features of a Pre-trained Model (EUSIPCO2022)

[17] 2022/06 BYOL-S: Learning Self-supervised Speech Representations by Bootstrapping

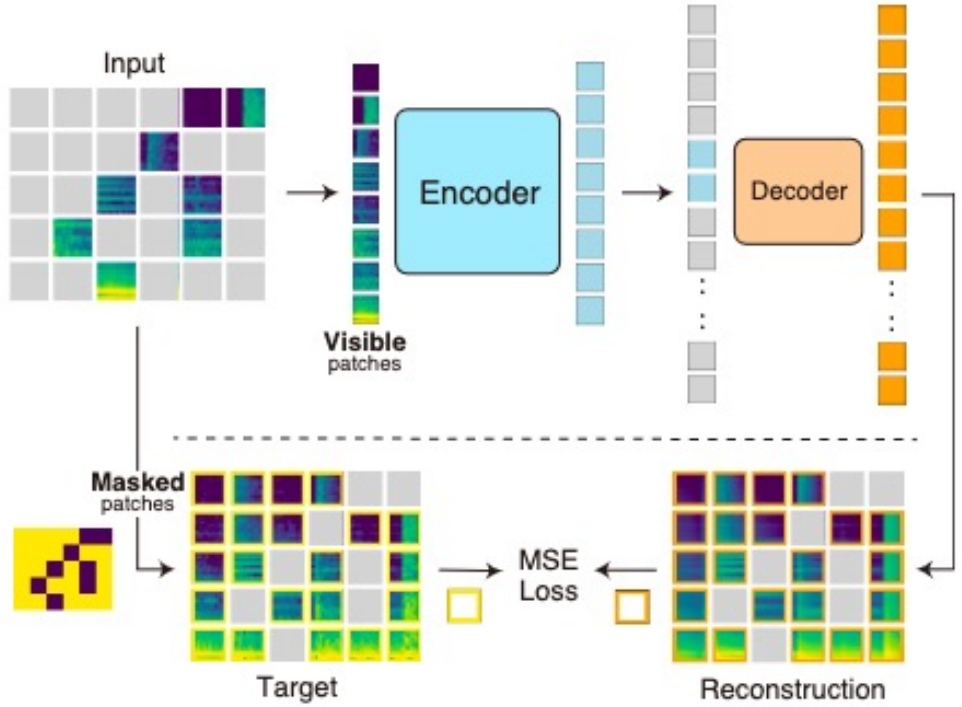
[18] 2022/07 **(Audio-MAE) Masked Autoencoders that Listen**

これまで、関連する5つの手法が提案され優位性を示している。

最新手法: Masked Image Modelingの応用

- 他の分野で発達したベストプラクティスが利用可能。
  - 高性能なVision Transformer (ViT)。
  - 自然言語処理、画像処理分野で性能を発揮している学習原理。
    例) Masked Autoencoders の学習フロー

1-Q-17 ポスター
発表済み



1. 入力スペクトログラムをパッチ化。(例: 周波数ビンx時間フレーム = 80x208 入力に対してパッチサイズ 16x16 で分割したときのパッチ数は 5x13。80x96 入力の場合，パッチ数は 5x6。)
2. パッチをマスク: 75%をマスク(Masked)，25%を非マスク(Visible)へと分割。
3. エンコーダーにより非マスク(Visible)パッチを表現に変換。
4. デコーダーでマスクパッチを復元。入力には非マスクパッチ表現に加えてマスク部分にマスクトークンを付与しておく。復元結果は入力全体となるが，マスク部分のみをロス計算に利用。
5. 入力信号と復元結果のMSE lossにより損失を計算して逆伝播により学習。

- Niizumi et al. "**Masked Spectrogram Modeling** using Masked Autoencoders for Learning General-purpose Audio Representation." *ArXiv* abs/2204.12260 (2022)
- 本研究会 1-Q-17 **マスクスペクトログラムモデル**による汎用音響信号表現の学習

手を動かしてみたい方へ

- チュートリアルを用意しました。
  - https://github.com/nttcslab/composing-general-audio-repr/tree/main/tutorial

CPUで動きました

1.2 Convert to spectrogram

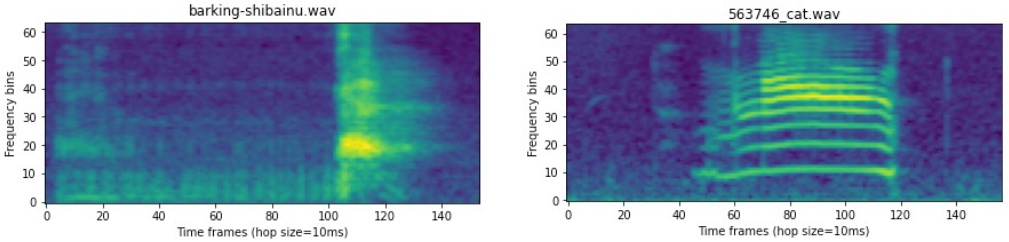
The following converts raw audio into a log-mel spectrogram (LMS) and forms a batch of the two audios.

```
def wav_to_lms(wav, sr, display=True, subplot_idx=None):
    # convert to a log-mel spectrogram.
    X = torchaudio.transforms.MelSpectrogram(sample_rate=sr, n_fft=2048, f_max=7800, f_min=60, n_mels=64,
            hop_length=sr//100)(wav).log() # 10ms

    if display:
        if subplot_idx:
            plt.subplot(subplot_idx)
        plt.imshow(X, origin='lower')
        plt.xlabel('Time frames (hop size=10ms)'); plt.ylabel('Frequency bins')
    return X

plt.figure(figsize=(15, 3))
X = []
for i in range(len(wavs)):
    x = wav_to_lms(wavs[i], sr, subplot_idx=121 + i)
    plt.title(files[i])
    X.append(x)
```

スペクトログラムに変換するレベルからの解説



4. Using GeneralPurposeCnn14

By default, this model fuses features from the layers [3, 6]. The layer 3 features have a higher magnitude, and the layer 6 features are lower range.

```
model = GeneralPurposeCnn14()
model.eval()
model.load_state_dict(weights, strict=False)
```

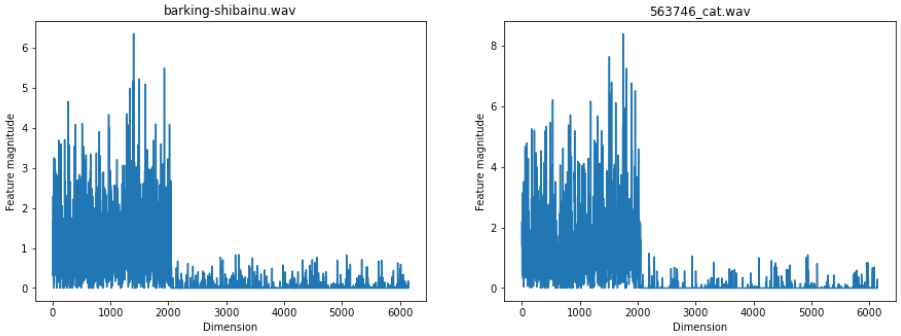
```
_IncompatibleKeys(missing_keys=[], unexpected_keys=['spectrogram_extractor.stft.conv_real.weight', 'spectrogram_extractor.stft.conv_imag.weigh
r.melW', 'fc_audioset.weight', 'fc_audioset.bias'])
```

```
# encode
features = model(X).detach()
print('Shape of features:', features.shape)

plt.figure(figsize=(15, 5))
for i in range(len(features)):
    plot_feature(features[i], subplot_idx=121 + i)
    plt.title(files[i])
```

表現を得て可視化している様子

```
Shape of features: torch.Size([2, 6144])
```



むすび

- 音声・音楽以外にも環境音分析の発展に見られるように、多様な一般の音に関するタスクへのニーズが広がっている。
- 画像分野の深層学習モデルの転用を中心として、AudioSetの事前学習済みモデルが数多く提案されている。
  - しかし、汎用性は必ずしも高くないため、応用に際しては注意が必要。
- **汎用音響信号表現**として多様な方法で学習された手法が提案され、性能を伸ばしている。
  - 学習手法、学習対象のモデル、データセット等、今後も発展が期待される。
  - 音の課題解決の特徴量抽出ツールとして、また専用のモジュールへ特化も含め、これから研究素材としての応用が期待される技術。

予備

深層表現とは?

- 表現学習の必要性

麻生 英樹, 多層ニューラルネットワークによる深層表現の学習 (<連載解説>Deep Learning(深層学習)〔第2回〕), 人工知能, 2013, 28 巻, 4 号, p. 649-659

『「内部表現の発見・獲得・学習」，すなわち，さまざまな情報が混在し，雑音で汚れている実世界の観測情報から，本質的な情報やある課題（群）に必要な情報を抽出し，処理しやすいように表現することは，人工知能やパターン認識を始めとする知的情報処理における古くからの研究課題の一つである．』

- 深層学習モデルを学習することで得られる内部表現

  → **深層表現**。

TABLE II: An overview of the recent audio self-supervised learning methods. The "speech" column distinguishes whether a method addresses speech tasks or for general purpose audio representations. The "framework" type refers to Figure I.

Model	Speech	Input format	Framework	Encoder	Loss	Inspired by
LIM [36]	✔	raw waveform	(d)	SincNet	BCE, MINE or NCE loss	SimCLR
COLA [36]	✗	log mel-filterbanks	(d)	EfficientNet	InfoNCE loss	SimCLR
CLAR [33] (semi)	✗	raw waveform log mel-spectrogram	(d)	1D ResNet-18 ResNet-18	NT-Xent + cross-entropy	SimCLR
Fonseca et al. [36]	✗	log mel-spectrogram	(d)	ResNet, VGG, CRNN	NT-Xent loss	SimCLR
Wang et al. [88]	✗	raw waveform + log mel-filterbanks	(d)	CNN ResNet	NT-Xent loss + cross-entropy	SimCLR
BYOL-A [89]	✗	log mel-filterbanks	(b)	CNN	MSE loss	BYOL
Speech2Vec [48]	✔	mel-spectrogram	(a)	RNN	MSE loss	Word2Vec
Audio2Vec [91]	✔✗	MFCCs	(a)	CNN	MSE loss	Word2Vec
Carr [67]	✔	MFCCs	(a)	Context-free network	Fenchel-Young loss	-
Ryan [68]	✗	constant-Q transform spectrogram	(a)	AlexNet	Triplet loss	- -
Mockingjay [92]	✔	mel-spectrogram	(a)	Transformer	L1 loss	BERT
TERA [93]	✔	log mel-spectrogram	(a)	Transformer	L1 loss	BERT
Audio ALBERT [94]	✔	log mel-spectrogram	(a)	Transformer	L1 loss	BERT
DAPC [95]	✔	spectrogram	(a)	Transformer	Modified MSE loss + orthogonality penalty	BERT
PASE [96]	✔	raw waveform	(a)	SincNet + CNN	L1, BCE loss	BERT
PASE+ [97]	✔	raw waveform	(a)	SincNet + CNN + QRNN	MSE, BCE loss	BERT
CPC [40]	✔	raw waveform	(a)	ResNet + GRU	InfoNCE loss	-
CPC v2 [59]	✔	raw waveform	(a)	ResNet + Masked CNN	InfoNCE loss	-
CPC2 [98]	✔	raw waveform	(a)	ResNet + LSTM	InfoNCE loss	-
Wav2Vec [84]	✔	raw waveform	(a)	1D CNN	Contrastive loss	-
VQ-Wav2Vec [85]	✔	raw waveform	(a)	1D CNN + BERT	Contrastive loss	BERT
Wav2Vec 2.0 [81]	✔	raw waveform	(a)	1D CNN + Transformer	Contrastive loss	BERT
HuBERT [99]	✔	raw waveform	(c)	1D CNN + Transformer	Contrastive loss	BERT

汎用音響信号表現

音声認識モデルを含めた手法一覧例

- "✗"=汎用音響信号表現
- 汎用音響信号表現はSpectrogram入力が多い。
- 音声認識は Inspired by BERT 手法が隆盛。

Liu, Shuo, et al. "Audio self-supervised learning: A survey." arXiv preprint arXiv:2203.01205 (2022).

音以外の分野で進む汎用志向

Towards General Purpose Vision Systems:
An End-to-End Task-Agnostic Vision-Language Architecture

Tanmay Gupta[1] Amita Kamath[1] Aniruddha Kembhavi[1] Derek Hoiem[2]
[1]PRIOR @ Allen Institute for AI [2]University of Illinois at Urbana-Champaign
https://prior.allenai.org/projects/gpv

Abstract

Computer vision systems today are primarily N-purpose systems, designed and trained for a predefined set of tasks. Adapting such systems to new tasks is challenging and often requires non-trivial modifications to the network architecture (e.g. adding new output heads) or training process (e.g. adding new losses). To reduce the time and expertise required to develop new applications, we would like to create general purpose vision systems that can learn and perform a range of tasks without any modification to the architecture or learning process. In this paper, we propose GPV-1*, a task-agnostic vision-language architecture that can learn and perform tasks that involve receiving an image and producing text and/or bounding boxes, including classification, localization, visual question answering, captioning, and more. We also propose evaluations of generality of architecture, skill-concept*[1] *transfer, and learning efficiency that may inform future work on general purpose vision. Our experiments indicate* GPV-1 *is effective at multiple tasks, reuses some concept knowledge across tasks, can perform the Referring Expressions task zero-shot, and further im-*

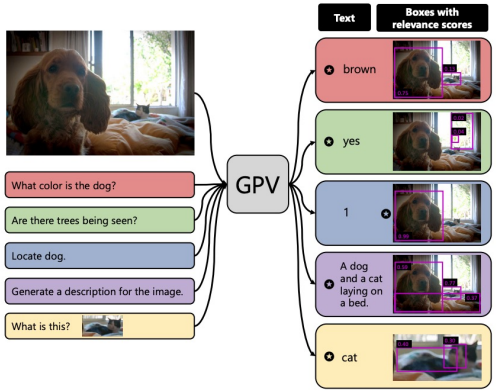


Figure 1. **A task-agnostic vision-language architecture.** GPV-1 takes an image and a natural language task description and outputs bounding boxes, confidences and text. GPV-1 can be trained end-to-end on any task that requires a box or text output, without any architecture modifications such as adding a new task-head. Results correspond to a model trained to perform VQA, localization, captioning, and classification tasks. Star indicates the output modality supervised during training for each task.

Gupta, Tanmay et al. “Towards General Purpose Vision Systems.” ArXiv abs/2104.00743 (2021)

汎用音響信号表現: 初出論文

- 初出は2019年の論文と考えられる。(発表者調べ)

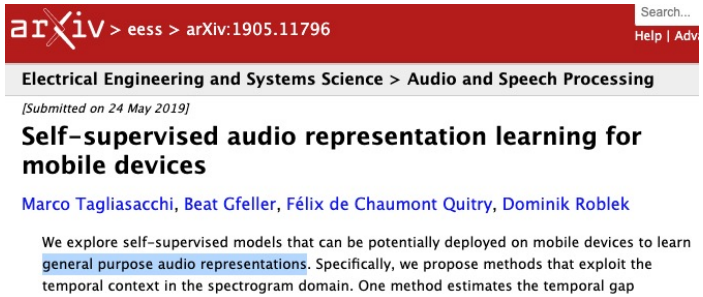


Table 2: Accuracy on downstream tasks (and fraction of accuracy recovered wrt. baselines). Downstream tasks: SPC: (*Speech Commands*), LSP: (*LibriSpeech*), TUT: *TUT Urban Acoustic Scenes 2018*, MUS: *MUSAN*, BSD: *Bird Audio Detection*, LID: *Spoken Language Identification*. In **bold** the highest accuracy attained by self-supervised models for each task.

model	SPC	LID	LSP	MUS	TUT	BSD
Spectrogram	0.16 ± .01 (+0%)	0.28 ± .04 (+0%)	0.97 ± .01 (+0%)	0.74 ± .01 (+0%)	0.36 ± .03 (+0%)	0.65 ± .02 (+0%)
Untrained	0.16 ± .01 (-1%)	0.48 ± .04 (+33%)	0.54 ± .02 (-1338%)	0.93 ± .00 (+77%)	0.57 ± .03 (+35%)	0.70 ± .02 (+31%)
AutoEncoder	0.28 ± .01 (+21%)	**0.64** ± .04 (+56%)	0.99 ± .00 (+55%)	0.94 ± .00 (+81%)	0.59 ± .03 (+38%)	0.69 ± .02 (+27%)
A2V (CRoW)	**0.30** ± .01	0.57 ± .04	0.99 ± .00	**0.98** ± .00	0.66 ± .03	0.71 ± .01

Tagliasacchi, Marco, et al. "Self-supervised audio representation learning for mobile devices." *arXiv preprint arXiv:1905.11796* (2019).