



属性情報と生成モデルを用いた音響イベントのゼロショット学習

上原, 康平
高島, 遼一
滝口, 哲也

(Citation)

神戸大学都市安全研究センター研究報告, 29:114-119

(Issue Date)

2025-03

(Resource Type)

departmental bulletin paper

(Version)

Version of Record

(JaLCD0I)

<https://doi.org/10.24546/0100496658>

(URL)

<https://hdl.handle.net/20.500.14094/0100496658>



属性情報と生成モデルを用いた 音響イベントのゼロショット学習

Zero-shot learning for sound event classification using an
attribute vector and a generative model

上原 康平¹⁾

Kohei Uehara

高島 遼一²⁾

Ryoichi Takashima

滝口 哲也³⁾

Tetsuya Takiguchi

概要: 本研究では、学習データが存在しない音響イベント（車の走行音や水の流れる音といった環境音）を分類するための、ゼロショット学習の方法を提案する。従来手法では、あらかじめ各音響イベントについて、音源の材質や高音／低音といった属性情報を付与しておき、音響イベントラベルの代わりに属性情報を推定することで、未学習のイベントの分類を可能にしている。しかしこの手法には、未学習の音響イベントが学習済みの音響イベントに分類されやすいという課題がある。本研究ではこの予測の偏りを低減するため、画像分野で提案された、生成モデルを用いるアプローチを音響イベント分類に適用する。提案手法では、従来手法で定義されたものと同様の属性情報を用いるが、属性情報を推定するのではなく、条件付き GAN によって未知音響イベントの潜在特徴を生成し、生成した特徴を用いてクラス分類器を学習する。実験の結果、提案手法は従来手法よりも高い分類性能を示した。

キーワード: 音響イベント分類、ゼロショット学習、生成モデル、条件付き GAN

1. はじめに

音響イベント分類は環境音の種類を分類する技術で、高齢者や乳幼児の見守り、異常監視システムなどへの応用が期待されている。しかし、一般に異常なイベントは滅多に起こらないため、学習データが入手しづらいという問題があり、学習データが存在しない音響イベントを分類するゼロショット学習の方法が研究されている。

従来手法¹⁾には、未学習の音響イベントが学習済みの音響イベントに分類されやすいという課題がある。このような学習済みのクラスへの予測の偏りは、音響イベント分類に限らず、ゼロショット学習に関する多くの研究で確認されており、ゼロショット学習における重要な研究課題となっている。本研究では、この学習済みの音響イベントへの予測の偏りを改善するために、画像認識の分野で提案された、生成モデルを用いるアプローチ²⁾を導入する。これは生成モデルによって学習データに存在しないクラスのデータを補完し、クラス分類器を学習するという方法をとるもので、生成モデルを用いた音響イベントのゼロショット学習は本研究が初めてとなる。

2. ゼロショット学習の問題設定

データセットに含まれる全クラスの集合 Y を、既知クラスの集合 Y_s と未知クラスの集合 Y_u の二つに分割する ($Y_s \cup Y_u = Y$, $Y_s \cap Y_u = \emptyset$)。モデルの学習に用いる音声データは既知クラス $s \in Y_s$ に含まれるもののみを用い、未知クラス $u \in Y_u$ に含まれる音声データはモデルの評価に用いる。また、各クラス $y \in Y$ に対して、予測の手がかりとなる属性情報 a_y を定義する。属性情報は、例えば「音源は木材か否か」といった音源の材質に関する属性や、「高い音か否か」といった音高に関する属性など、表1に示される計15種類の項目からなり、各音響イベントクラスの音に対して、15種類の属性について当てはまっていれば「1」、当てはまっていなければ「0」とする15次元のベクトルとして定義される。

表1 属性情報の内訳

音の高さ	音の長さ	材質	状況	その他
1. 高	4. 長	7. 木材	11. 落下	14. ノイズ的
2. 中	5. 中	8. 金属	12. 衝突	15. 繰り返し
3. 低	6. 短	9. 樹脂	13. 多数音源	
		10. 陶器		

モデルは二つのタスクによって評価する。一つはZSL (Zero-shot Learning) タスクで、入力と識別候補を既知クラスに限定して分類を行う。もう一つはGZSL (Generalized Zero-shot Learning) タスクで、入力と識別候補を限定せずに分類を行う。入力されたデータに対して、ZSLタスクは $|Y_u|$ 個のクラスから、GZSLタスクは $|Y_s + Y_u|$ 個のクラスから正しいクラスを予測するため、GZSLタスクはZSLタスクよりも難しい問題設定となっている。

3. 提案手法

(1) 概要

提案手法の概要を図1に示す。ゼロショット学習では未知クラスのデータをモデルの学習に用いることができない。そこで本研究では、条件付きGANによって未知音響イベントの属性情報からイベントの潜在特徴を生成し、生成した潜在特徴によってクラス分類器を学習する。未知クラスのデータを補完することで、既知クラスと未知クラスの間での学習データの不均衡を緩和し、既知クラスへの予測の偏りを低減する狙いがある²⁾。

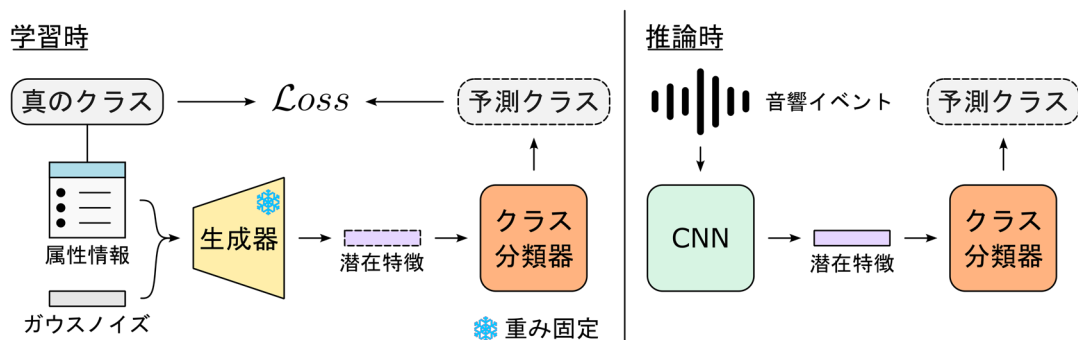


図1 提案手法の概要

(2) 条件付き GAN の学習

まず、従来手法と同様に、既知クラスの音響イベントから属性情報を推定するタスクにより、音響イベントの対数メルフィルタバンク特徴 (FBANK) x を潜在特徴 $e = E(x)$ に変換する CNN エンコーダ E 、 e から属性情報 $\tilde{a} = R(e)$ を推定する属性推定器 R を事前学習する。学習済みの二つのモデルを用いて、属性情報からイベントの潜在特徴を生成する条件付き GAN を学習する。条件付き GAN の学習プロセスを図 2 に示す。生成器 G 、識別器 D について、Wasserstein GAN の敵対的損失 $\mathcal{L}_{\text{GAN}}$ を式(1)のように計算する。

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}[D(e_s, a_s)] - \mathbb{E}[D(\tilde{e}_s, a_s)] - \lambda_1 \mathbb{E}[(\|\nabla_{\hat{e}} D(\hat{e}, z)\|_2 - 1)^2] \quad (1)$$

ここで、 $e_s = E(x_s)$ は既知クラス $s \in Y_s$ のイベント x_s の潜在特徴、 a_s は既知クラス s に付与された属性情報、 $\tilde{e}_s = G(z, a_s)$ は属性情報 a_s とガウスノイズ $z \sim N(\mathbf{0}, \mathbf{1})$ から生成された潜在特徴である。 $\hat{e} = \alpha e_s + (1 - \alpha) \tilde{e}_s$ は e_s と $\tilde{e}_s$ の間の任意の点を表し、 $\alpha \sim U(0, 1)$ である。 λ_1 はハイパーパラメータである。また、属性情報と生成特徴の整合性をとるためのサイクルー貫性損失 $\mathcal{L}_{\text{CYC}}$ を式(2)のように計算する。

$$\mathcal{L}_{\text{CYC}} = \mathbb{E}[\|R(\tilde{e}_s) - a_s\|_2^2] + \mathbb{E}[\|R(\tilde{e}_u) - a_u\|_2^2] \quad (2)$$

ここで、 a_u は未知クラス $u \in Y_u$ に付与された属性情報、 $\tilde{e}_u = G(z, a_u)$ は属性情報 a_u とガウスノイズ z から生成された潜在特徴である。最終的に、条件付き GAN は式(3)のようにして学習される。

$$\min_G \max_D \mathcal{L}_{\text{GAN}} + \lambda_2 \mathcal{L}_{\text{CYC}} \quad (3)$$

ここで、 λ_2 はハイパーパラメータである。生成器 G と識別器 D の学習時、エンコーダ E と属性推定器 R のパラメータは固定する。

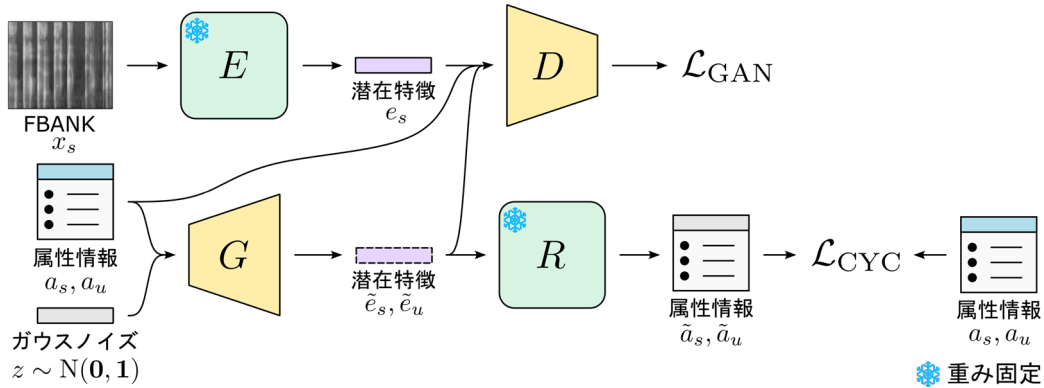


図2 条件付き GAN の学習プロセス

(3) クラス分類器の学習

学習済みの生成器 G によって各音響イベントクラスの潜在特徴を生成し、生成特徴を用いてクラス分類器 P を学習する。クラス分類器 P は一般的な多クラス分類モデルと同様に Softmax 関数によって各クラスの確率を計算するもので、式(4)で示す損失関数を最小化するように学習される。

$$\mathcal{L}_{\text{CLS}} = -\mathbb{E}_{y \sim \mathbb{P}_{Y_c}} [\log P(y|G(z, a_y))] \quad (4)$$

ここで、 $\mathbb{P}_A$ は集合 A の要素が従う分布、 Y_c は識別候補のクラスの集合を表し、ZSL タスクでは $Y_c = Y_u$ 、GZSL タスクでは $Y_c = Y_s \cup Y_u$ となる。分類時は入力される音響イベントの FBANK x を、以下の式(5)で表すようなクラス $\tilde{y}$ に分類する。

$$\tilde{y} = \underset{y \in Y_c}{\operatorname{argmax}} P(y|E(x)) \quad (5)$$

3. 評価実験

(1) 使用するデータ

「コインを落とす」、「手をたたく」のような計 105 クラスの音響イベントが収録された RWCP-SSD³⁾ データセットを用いて実験を行った。先行研究¹⁾によって、105 クラスのうち 68 クラスに対して属性情報が定義されているが、いくつかのクラスは音が似ており、属性情報が全く同じになっている。そこで、属性情報が他のクラスと重複しない 6 クラスを未知クラスに定め、残りの 62 種類のクラスのうち、属性情報が一致するクラスは 1 つのクラスとしてまとめることで、30 種類の既知クラスに再定義した。データ分割の詳細を表 2 に示す。

表 2 データ分割

	既知クラス	未知クラス
クラス数	30	6
学習データのサンプル数	4,734	N/A
検証データのサンプル数	610	N/A
評価データのサンプル数	305	541

(2) モデルの詳細

モデルに入力する FBANK は、音声データをゼロ埋めにより 1 秒 (100 フレーム) に長さを揃えた上で、フレームごとに 80 次元の対数メルフィルタバンク特徴を計算することで得られた 80×100 次元の特徴量である。CNN エンコーダに VGGish⁴⁾を採用し、最終層でグローバル平均プーリングを行うことで、512 次元のベクトルとして潜在特徴が得られる。条件付き GAN を構成する生成器、識別器はともに数層の全結合層で実装し、生成器にはガウスノイズと属性情報を、識別器には潜在特徴と属性情報をそれぞれ結合して入力する。また、生成器に入力するガウスノイズは 15 次元とした。クラス分類器の学習の際に用いる生成器には、条件付き GAN を 100 エポック学習した中で、生成した潜在特徴に対するサイクルー貫性損失 $\mathcal{L}_{\text{cyc}}$ の平均が最も小さくなるエポックのものを用いた。

(3) 実験結果

従来手法との比較を行う。表 3 にそれぞれの ZSL タスクおよび GZSL タスクの分類結果を示す。GZSL タスクにおける U と S はそれぞれ未知クラス、既知クラスに対する分類結果を表し、H は両者の調和平均を表している。また、分類性能の目安となる、クラスをランダムに予測する場合の正解率の期待値を“ランダム予測”として、提案手法の性能の上限値である、クラス分割を行わずに (未知クラスのデータも用いて) モデルを学習した場合の結果を“提案手法(理想)”として、それぞれ掲載している。

表 3 分類正解率 [%]

	ZSL	GZSL		
		U	S	H
ランダム予測	16.7 (=1/6)	2.8 (=1/36)	2.8 (=1/36)	2.8
従来手法	70.2	5.2	99.7	9.9
提案手法	95.6	22.9	86.6	36.2
提案手法(理想)	100.0	99.3	98.0	98.6

まず ZSL タスクについて、提案手法は従来手法より高い分類精度を示した。両者の最終的なクラス
の予測方法の違いが考えられる。従来手法では、イベントの属性情報を推定し、各クラスの属性情報
とのユークリッド距離の比較によって分類を行う。このため、ユークリッド距離に近い属性情報を持
つクラス同士の分類は難しくなると考えられる。提案手法ではクラス分類器によって潜在空間上の
クラス間の非線形な決定境界を学習するため、属性情報空間上での物理的な距離は分類に直接影響
しにくい。このような利点から、ZSL タスクの性能が向上した可能性がある。

次に GZSL タスクについて、提案手法は既知クラス音響イベントに対する分類正解率 S が従来手法
より劣るものの、 S と未知クラス音響イベントに対する分類正解率 U のバランスが改善され、その調
和平均である H が向上した。どれだけの未知クラスのイベントが既知クラスに分類されたかを議論
するため、未知クラスと既知クラスに関する混同行列を表 4 に示す。従来手法は未知クラスのイベ
ントのほとんどを誤って既知クラスに分類してしまっていたところを、提案手法では改善することに
成功している。一方で、従来手法では既知クラスのイベントはすべて既知クラスに分類されていたの
に対して、提案手法では既知クラスのイベントのいくつかは未知クラスに分類されてしまうなど、 U
と S はトレードオフの関係にあることがわかる。

表 4 混同行列 [%]

		予測クラス (従来手法)		予測クラス (提案手法)	
		既知クラス	未知クラス	既知クラス	未知クラス
実際のクラス	既知クラス	100.0	0.0	97.4	2.6
	未知クラス	94.8	5.2	77.1	22.9

4. まとめ

本研究では、音響イベントのゼロショット学習について、生成モデルを用いるアプローチを提案し
た。提案手法は ZSL タスク、GZSL タスクの双方で従来手法より高い分類性能を示した。しかし、改
善はされたものの、GZSL タスクにおいて未知クラスのイベントが既知クラスに誤分類される傾向は
依然として強く、さらなる精度向上のため、モデルの学習に有利な潜在特徴を得るためのエンコーダ
の改良や、属性情報の見直し等を行う予定である。

参考文献

- 1) Lin, Y., Takashima, R., and Takiguchi, T., Zero-shot sound event classification using a sound attribute vector with global and local learning, *ICASSP*, 1-5, 2023.
- 2) Xian, Y., Lorenz, T., Schiele, B., and Akata, Z., Feature generating networks for zero-shot learning, *CVPR*, 5542-5551, 2018.
- 3) Nakamura, S., Hiyane, K., Asano, F., Nishiura, T., Yamada, T., Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition, *LREC*, 965-968, 2000.
- 4) Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K., CNN architectures for large-scale audio classification, *ICASSP*, 131-135, 2017.

筆者：1) 工学部情報知能工学科、学生；2) 都市安全研究センター、准教授；3) 都市安全研究センター、教授

Zero-shot learning for sound event classification using an attribute vector and a generative model

Kohei Uehara
Ryoichi Takashima
Tetsuya Takiguchi

Abstract

This paper presents a zero-shot learning method to classify unseen sound events without training data. A previous study proposed a method to classify unseen events by estimating attribute information instead of sound event labels, where each sound event is associated with attribute information such as sound source material and the pitch (high or low) of the sound. However, this method often leads to the misclassification of unseen sound events as seen events. In this paper, we propose a method using a generative model to reduce the bias of prediction toward seen sound events, adapting an approach from the field of computer vision. The proposed method generates latent features of unseen sound events from their attribute information using a conditional GAN, and a classifier is trained using the generated latent features. Experimental results demonstrated that the proposed method achieved higher classification accuracies than the previous method based on attribute information.

©2025 Research Center for Urban Safety and Security, Kobe University, All rights reserved.