アノテータごとのばらつきを考慮した音響イベント検出

古賀 直樹[1,2]　坂東 宜昭[2]　井本 桂右[1,2]

[1]同志社大学　[2]産業技術総合研究所

1. はじめに

本稿では，複数のアノテータが作成した，ばらつきのあるアノテーションを用いた音響イベント検出 （SED）について述べる．SED とは，音響イベントの発生時刻と種類を推定するタスクの 1 つであり，見守りシステムや警備システムなど様々な分野への応用が期待されている技術の 1 つでもある．ここで音響イベントとは，車の走行音，機械の駆動音，波の音のような音を指す．SED では，畳み込み再帰形ニューラルネットワーク (CRNN) [1] や，長期短期記憶モジュール (LSTM) [2]，Transformer [3] を用いた教師あり学習による手法が高い性能を達成していることが広く知られている．

教師あり学習に基づく SED では，教師データとして用いる音響信号に，音響イベントの種類と発生時刻を付与する必要がある．実世界の環境音では，複数の音響イベントが同時に発生したり，背景雑音により目的の音響イベントが聞き取りづらい場合があり，個々のイベントの発生時刻を判断するのは難しい [4]．その結果アノテーションがばらつき，個々の音響イベントの種類と発生時刻を一意に決定するのが困難になる．

そこで本研究では，ばらつきのある複数のアノテーションから直接ネットワークを学習できるようにする Crowd layer [5] を用いて SED に取り組む（図 1）．Crowd layer をネットワークを組み込むことによって，アノテータの特徴を捉えることが可能になり，誤差逆伝播法のみを用いて音響特徴量とアノテーションからネットワークを学習できるようになる．実験では，新たに複数人の SED 用アノテーションを実施した実録音公開データセットに対して Crowd layer を組み込んだネットワークの識別性能が改善することを確認する．

2. ばらつくアノテーションを用いた SED

本節ではばらつくアノテーションから直接学習する機構である Crowd layer をネットワークに組み込む．

2.1 Crowd layer

Crowd layer はネットワークの出力層の 1 種で，CRNN など従来の推論モデルと組み合わせて用いる．この出力層は，CRNN が抽出した特徴量時系列に対し，アノテータごとに異なる線形変換を施すことで，アノテータごとの偏りを表現する．Crowd layer によって，各アノテータのラベリングにおける偏った部分を認識・修正でき，複数のアノテータが作成したばらつきのあるアノテーションからネットワークを直接学習できるようになる．$\mathbf{x}$ と表記するネットワークの出力が与えられると，アノテータ k における Crowd layer の活性化関数は $\mathbf{a}^{(k)} = f^{(k)}(\mathbf{x})$

Sound Event Detection Handling Variations Among Annotators: *N. Koga, Y. Bando, K. Imoto*

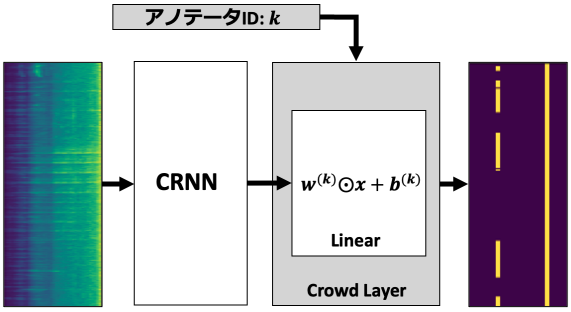

図 1: Crowd layer を適用した CRNN の概要

と表すことができる．Crowd layer の原論文 [5] では，各クラスの事後確率に対して，重みを掛ける関数や，バイアスを追加する関数など複数の活性化関数が提案されているが，本研究では重みを掛け，バイアスを足しこむ以下の関数を用いる：

$$f^{(k)} = \mathbf{w}^{(k)} \odot \mathbf{x} + \mathbf{b}^{(k)} \tag{1}$$

ここで，$\mathbf{w}^{(k)}, \mathbf{b}^{(k)}$ はアノテータ固有のベクトルであり，音響イベントのクラス数を C とすると，両パラメータのサイズは $C \times 1$ である．

2.2 ネットワークの学習

時間フレーム t，クラス c における正解ラベル $y_{t,c} \in [0, 1]$ と，ネットワークが出力する時間フレーム t におけるクラス c の事後確率 $\hat{y}_{t,\,c} \in [0, 1]$ に対して，以下の BCE (Binary Cross Entropy) を損失関数とし，ネットワークを学習させた．

$$L = -\sum_{t,c} y_{t,\,c} \log(\hat{y}_{t,\,c}) + (1 - y_{t,\,c}) \log(1 - \hat{y}_{t,\,c}) \tag{2}$$

また，ネットワークのテストには学習時の最後のエポックで用いたモデルを使用した．

3. 評価実験

本節では，Crowd layer を用いたモデルで SED におけるアノテータの特徴を学習できるかを確認した．

3.1 データセット

本実験では SED のためのデータセットである TUT Sound Events (SE) 2016/2017 を用いて教師あり学習を行った [6]．データセットの音響信号には，1 つあたり 10 個のアノテーションを付けた．また，音響信号の長さは全て 10 秒に揃えた．図 2 は TUT SE 2017 の信号 a131.wav に付けられた 10 個のアノテーションから抜粋した 3 個を比較したもので，アノテータによって選んだ音響イベントもその発生区間も異なるアノテーションになっていることがわかる．音響信号にアノテーションを付けた結果，イベントラベルの種類は，TUT SE 2016

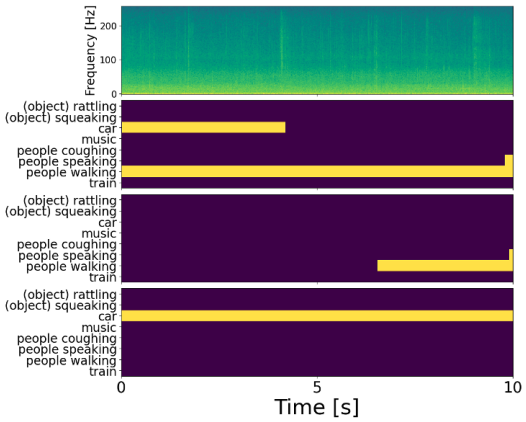

図 2: 音響信号と 3 人分のアノテーション結果

表 1: TUT Sound Events 2016 (20 アノテータ) での検出性能 (%)

モデル	データ	Rec.	Prec.	micro F1	macro F1	micro ER	macro ER
Crowd layer なし	Train	34.63	51.59	41.44	11.38	74.69	307.82
Crowd layer あり	Train	34.13	52.71	41.43	11.34	73.69	119.80
Crowd layer なし	Eval	**61.59**	77.31	68.56	35.93	46.01	78.66
Crowd layer あり	Eval	61.36	**77.95**	**68.67**	**36.20**	**45.79**	**78.51**

表 2: TUT Sound Events 2017 (10 アノテータ) での検出性能 (%)

モデル	データ	Rec.	Prec.	micro F1	macro F1	micro ER	macro ER
Crowd layer なし	Train	42.37	54.57	47.70	13.70	68.70	101.98
Crowd layer あり	Train	43.65	54.60	48.51	13.95	68.03	101.14
Crowd layer なし	Eval	65.65	67.45	66.54	34.39	51.51	82.86
Crowd layer あり	Eval	**65.76**	**67.68**	**66.71**	**35.02**	**51.11**	**82.52**

では 31 種類，TUT SE 2017 では 22 種類となった．また教師データの数は TUT SE 2016 では訓練用に 3,849 個，テスト用に 2,192 個，TUT SE 2017 では訓練用に 5,670 個，テスト用に 1,627 個となった．学習ではアノテーションと音響特徴量のセットをランダムにサンプルしてバッチを構成した．

3.2 実験条件

本実験では，ベースラインモデルとして双方向ゲート付き再帰ユニット (BiGRU) [7] と CRNN を組み合わせた，CNN-BiGRU を用いる．CNN-BiGRU は，CNN と時系列データの長期伝達が可能な GRU を組み合わせたモデルで，SED で高い性能を達成することが報告されている [8, 9]．CNN-BiGRU の最終層に Crowd layer を追加したモデルを「Crowd layer あり」，CNN-BiGRU を「Crowd layer なし」として，両者を比較した．

評価指標には，200ms 単位で集計した，適合率 (Rec.)，精度 (Prec.)，micro/macro F1 および micro/macro エラー率 (ER) を用いた．音響信号には，窓長を 1,764 サンプル，ホップ長を 882 サンプルとした短時間フーリエ変換を適用し，その後，メルビン数を 40 としたメルスペクトログラムに変換した．また，全ての音響特徴量とアノテーションを 500 フレームに整えた．学習に関して，学習率が 1.0×10^{-4}，荷重減衰が 1.0×10^{-5} の AdamW を用い，エポック数は 300，バッチサイズは 128 とした．50%の確率で Dropout をパラメータに適用した．

3.3 実験結果

TUT SE 2016 の場合の両モデルの評価指標の数値を表 1 に，TUT SE 2017 の場合の結果を表 2 に示す．2 つの表より，TUT SE 2016 と TUT SE 2017 の両データセットにおいて，「Crowd layer あり」の方が「Crowd layer なし」よりもほぼ全ての指標でわずかに性能が向上したことが確認できたが，その差は全て 1%以内に収まる結果となった．また，TUT SE 2016 の精度 (Prec.) だけは「Crowd layer あり」よりも「Crowd layer なし」の方が数値が向上した．

実験の結果，Crowd layer をネットワークに組み込むことによる性能向上はわずかであった一方，ほぼ全ての指標で数値が向上したことはアノテータの特徴を Crowd layer によって把握できた裏付けとなると考えられる．Crowd layer の原論文 [5] では，我々の実験よりもデータ数の多い，25,000 枚の画像を用いた分類問題に取り組んでいたため，データ数を増やしてネットワークを学習させて実験することが今後の課題として挙げられる．

4. おわりに

本研究では，ばらつきのあるアノテーションに対して Crowd layer を導入して SED に取り組んだ．実験の結果，各アノテータの特徴を把握しつつ音響イベントの種類と発生区間を学習できたことを確認した．今後は Crowd layer を用いて異なる条件下での実験を行いながら，学習性能をより向上させられるような，Crowd layer を取り入れた新しいモデルの開発に取り組みたい．

謝辞: 本研究の一部は，NEDO および JSPS 科研費 22H03639, 23K16908，ROIS-DS-JOINT (006RP2023) の支援を受けた。

参考文献

[1] E. Çakır *et al.* Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1291–1303, 2017.

[2] S. Jung *et al.* Polyphonic sound event detection using convolutional bidirectional lstm and synthetic data-based transfer learning. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 885–889, 2019.

[3] K. Miyazaki *et al.* Convolution-augmented transformer for semi-supervised sound event detection. In *Proc. workshop detection classification Acoust. Scenes events (DCASE)*, 100–104, 2020.

[4] A. Mesaros *et al.* Sound event detection: A tutorial. *IEEE Signal Processing Magazine*, 38(5):67–83, sep 2021.

[5] F. Rodrigues *et al.* Deep learning from crowds. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[6] A. Mesaros *et al.* Tut database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*, 1128–1132, 2016.

[7] Kyunghyun C. *et al.* Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[8] R. Lu *et al.* Bidirectional (gru) for sound event detection. Technical report, DCASE2017 Challenge, September 2017.

[9] S. Adavanne *et al.* A report on sound event detection with different binaural features. *arXiv preprint arXiv:1710.02997*, 2017.