C-015

Separable 畳み込みを用いた環境音認識モデルの軽量化
Lightweight Environmental Sound Classification Model Using Separable Convolution

福岡 直樹† 黒木 修隆† 沼 昌宏†
Naoki Fukuoka Nobutaka Kuroki Masahiro Numa

1. はじめに

近年，環境認識技術分野において，画像などの視覚的情報を用いた環境認識だけでなく，非言語的かつ非音楽的音声である環境音を利用した環境認識が注目を集めている。環境音の利用により，視覚的情報だけでは認識できないイベントの認識が可能になる。現在，環境音認識手法の多くには，画像認識などと同様に深層学習が利用されているため，大きな計算量を必要とする。この点は，リソース数に限りがあるエッジデバイス上に環境音認識モデルを実装する上で問題となる。

エッジデバイスへの実装を想定した環境音認識モデル ACDNet [1] は Mohaimenuzzaman らによって提案され，畳み込みニューラルネットワーク（CNN：Convolutional Neural Network）を利用している。ACDNet に関して，いくつかのモデル圧縮技術を用いて MCU（Micro Controller Unit）への実装が試みられているが，FPGA（Field-Programmable Gate Array）の MCU に対する並列処理能力や消費電力などにおける優位性の点で，FPGA による実装が有利と考えられる。一方で，FPGA には実装可能な回路規模に制限があるため，実装対象とする ACDNet の軽量化が必要となる。また，一般的に CNN の軽量化による精度低下を抑制する必要がある。

そこで本稿では，従来の ACDNet と比較して精度低下を抑えつつ，軽量化を実現する環境音認識モデルを提案する。提案モデルでは， Separable 畳み込みの導入により，パラメータ数および計算量を削減する。さらに，軽量化にともなう精度低下の抑制を目的として，Multi-branch アーキテクチャと再パラメータ化を導入する。

2. 提案手法

2.1 Separable 畳み込み

Separable 畳み込みは，空間方向とチャネル方向で別個に畳み込みを行うことで，計算量を削減する畳み込み手法である。空間方向の畳み込みは Depthwise 畳み込み，チャンネル方向の畳み込みは Pointwise 畳み込みと呼ばれ，両者を連続して行うことで，通常の畳み込みと同様に扱うことができる。

通常の畳み込みの計算量N_{conv}とパラメータ数 P_{conv} は，

$$N_{conv}= K^2HWC_{in}C_{out}, \quad P_{conv}= K^2C_{in}C_{out} \tag{1}$$

で表される。ただし K はカーネルサイズを，H，W はそれぞれ特徴マップの高さと幅を，C_{in}，C_{out} はそれぞれ入力，出力特徴マップのチャンネル数を表す。

一方で，Separable 畳み込みの計算量N_{sep}とパラメータ数 P_{sep} は，

$$N_{sep} = (K^2 + C_{out})HWC_{in}, \quad P_{sep}=(K^2+C_{out})C_{in} \tag{2}$$

と表すことができる。$C_{out} >> K^2$ と近似すれば，計算量とパラメータ数を $1/K^2$ に削減できることがわかる。

2.2 Multi-branch アーキテクチャ

CNN のアーキテクチャは，分岐路をもたない Single-path と，分岐路を含む Multi-branch に分類される。Multi-branch アーキテクチャは，Single-path アーキテクチャと比べて高い精度を実現できる一方で，処理速度やメモリ効率の低下，アーキテクチャ上の制約がかかるなどデメリットももつ。

2.3 再パラメータ化

再 パラメータ化とは，あるアーキテクチャを別のアーキテクチャとして扱えるようにパラメータを変換することをいう。再パラメータ化により，Multi-branch アーキテクチャの利点である精度と，Single-path アーキテクチャの利点である処理速度の両立が可能になる。

2.4 提案モデルの構造

提案モデルの全体構成を図 1 に示す。提案モデルでは，ACDNet の TFEB （Temporal Feature Extraction Block）の畳み込み層を，通常の畳み込み構造から Separable 畳み込みを用いた構造に置換している。また，置換する畳み込み層は，以下の提案手法 1 ～ 4 により実装を行う。また Multi-branch アーキテクチャを採用した提案手法 2 ～ 4 では，再パラメータ化により，推論時には提案手法 1 と同一のアーキテクチャをとる。

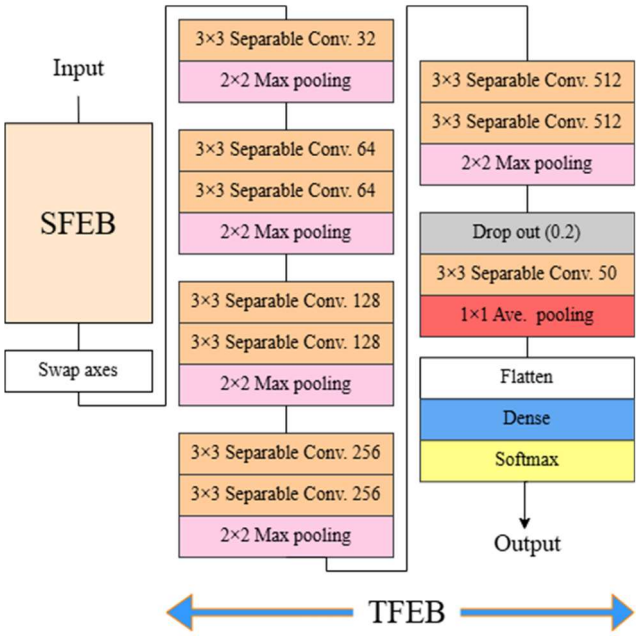

図 1 提案モデルの全体構成

† 神戸大学， Kobe University

2.4.1 提案手法 1

図 2 に提案手法 1 の畳み込み層の構造を示す。ACDNet の TFEB の畳み込み層を，そのまま Single-path アーキテクチャの 3×3 の Separable 畳み込みに置換して学習を行う。

2.4.2 提案手法 2

図 3 に提案手法 2 の畳み込み層の構造を示す。2 つの 3×3 の Depthwise 畳み込みで並行に畳み込み演算を行い，それらの結果を加算した後，バッチ正規化を行う。その後，同様の過程を Pointwise 畳み込みでも行う。

2.4.3 提案手法 3

提案手法 2 とほぼ同様の構造を取るが，Depthwise 畳み込みと Pointwise 畳み込みをそれぞれ 3 つ並行に処理する。

2.4.4 提案手法 4

図 4 に提案手法 4 の畳み込み層の構造を示す。MobileOne [2] の構成要素である MobileOne Block と同様の Multi-branch アーキテクチャによって学習を行う。MobileOne は Separable 畳み込みを利用し，かつ Multi-branch アーキテクチャと再パラメータ化の適用も行っている画像認識モデルである。

3. 評価実験と考察

提案手法 1〜4 および既存手法である ACDNet の実装および評価実験を行い，これら 5 手法および，ACDNet 以前に SOTA を達成したモデルである EnvNet-v2 [3]について比較した。また，データセットとして ESC-50 [4] を用いた。

表 1 に実験の評価結果を示す。提案手法 2 〜 4 に関しては，矢印の左側に再パラメータ化前の値を，右側に再パラメータ後の値を記載している。提案手法 1 では，既存の ACDNet と比較して，パラメータ数については約 88%，計算量に関しては約 84% の大幅な削減を達成できた。認識精度に関しても，4.5 pt 低下する一方で，ESC-50 の人間の耳による認識精度 [5] である 81.30% 以上の精度を達成していることから，このパラメータ数でのモデルの有用性が確認できたといえる。ただし FPGA への実装を実現するには，さらなるモデル圧縮による軽量化が必要であり，その結果さらなる認識精度の低下が予想される。そのため FPGA 実装を想定すると，提案手法 1 では圧縮前のモデルとしては認識精度が不十分な水準であると考えられる。

同様に，表 1 に示された提案手法 2 〜 4 に関する結果から，モデルの Multi-branch アーキテクチャの導入および再パラメータ化により，同じ推論モデルにおいても，1.5 〜 3 pt 程度の精度向上が見込まれることがわかる。特に，提案手法 2 では 3 pt の精度向上を達成し，ACDNet 以前のモデルである EnvNet-v2 と比べ，遜色ない認識精度を達成しているといえる。

表 1　パラメータ数，計算量，認識精度に関する比較

種類	パラメータ数 [M]	計算量 [M]	認識精度 [%]
EnvNet-v2 [3]	101.25	1,620	84.70
ACDNet	4.74	544	86.00
提案手法 1	0.57	85	81.50
提案手法 2	1.14 → 0.57	146 → 85	84.50
提案手法 3	1.70 → 0.57	207 → 85	83.00
提案手法 4	0.58 → 0.57	142 → 85	84.00

4. おわりに

本稿では，FPGA 実装に適した環境音認識モデルの構築を目的として， ACDNet への Separable 畳み込みの利用に加えて，Multi-branch アーキテクチャと再パラメータ化の適用によって，モデルの軽量化と精度の両立を図る手法を提案した。

提案手法では，従来の畳み込みと比較して軽量な Separable 畳み込みを利用して，パラメータ数および計算量の削減を行った。また，Multi-branch アーキテクチャと再パラメータ化の適用により，Separable 畳み込みの軽量化効果を維持しつつ，精度低下を抑制する効果を得た。

評価実験の結果，従来の ACDNet と比較して，精度低下を最低で 1.5 pt 程度に抑えつつ，パラメータ数を約 88%，計算量を約 84% 削減可能であることを確認した。

今後の課題として，提案した環境音認識モデルはパラメータ数および計算量が依然として多く，モデル全体の FPGA 実装は困難であると考えられる。そのため，精度向上とさらなるモデルの軽量化を目指し，枝刈りの適用や量子化などによるモデルの圧縮の検討を行う予定である。

参考文献

[1] M. Mohaimenuzzaman, C. Bergmeir, I.West, and B. Meyer, “Environmental sound classification on the edge: A pipeline for deep acoustic networks on extremely resource-constrained devices,” Pattern recognition, vol. 133, no. C, 2023.

[2] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, “MobileOne: An improved one millisecond mobile backbone”, arXiv preprint arXiv; 2206.04040v2, 2023.

[3] Y. Tokozume, Y. Ushiku, and T. Harada, “Learning from Between-class Examples for deep sound recognition,” arXiv preprint arXiv: 1711.10282v2, 2017.

[4] K. J. Piczak, “ESC: Dataset for environmental sound classification,” in Proceedings of the ACM International Conference on Multimedia, pp. 1015-1018, 2015.

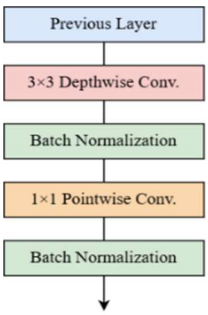

図 2　提案手法 1 の構造

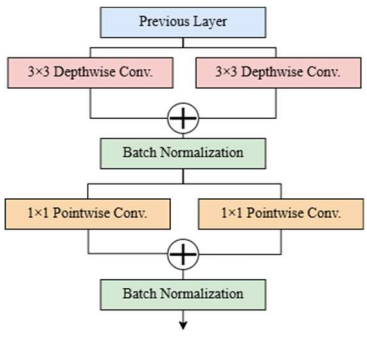

図 3　提案手法 2 の構造

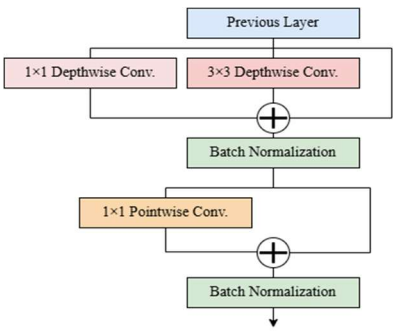

図 4　提案手法 4 の構造