C-002

監視カメラシステムの FPGA 実装に向けた物体検出モデルのパラメータ数削減
Reducing the Number of Parameters in Object Detection Models for FPGA Implementation of Surveillance Camera Systems

石川 翔太† 黒木 修隆† 沼 昌宏†
Shota Ishikawa Nobutaka Kuroki Masahiro Numa

1. はじめに

近年，監視カメラシステムを CNN により実現する技術が注目されている。一般的に利用されるクラウドでの監視カメラシステムでは，監視カメラの映像に対して，データセンタで物体検出や個人の識別処理が行われるが，監視カメラの死角を減らすためにカメラの台数を増やすと，通信帯域の制限による画像転送レートの低下が深刻な課題となる。問題解決の手段として，監視カメラシステムを処理速度とエネルギー効率に優れた FPGA（Field-Programmable Gate Array）に実装し，処理を FPGA 上で行うことで，クラウドに転送するデータ量を削減することが考えられる。しかし，FPGA のリソースには制限があるため，監視カメラシステムのために実装する CNN の軽量化が必要となる。また，一般的に CNN の軽量化を行うと精度低下を招く傾向があるが，可能な限り精度低下を抑制する必要がある。

そこで本稿では，従来のモデルと比較して，精度低下を抑えつつ，軽量化を実現する CNN の構造を提案する。提案する CNN では，従来の畳み込み演算と比較して軽量な Depthwise Separable Convolution を利用して，パラメータ数の削減を行う。さらに，再パラメータ化を利用して，ソフトウェア上で学習させた複雑な CNN を FPGA 実装に適した軽量な CNN に変換することで，精度の向上を図る。また，提案する CNN は層多重化回路として実装可能で，従来手法では CNN の構造上，検出が困難であった小さい物体の検出が比較的容易なネットワークである。また，監視カメラシステムで人を検出することを想定して，横長のデフォルト・ボックス（略称：Dbox）を消去し，パラメータ数を削減する。

2. 提案手法

2.1 Depthwise Separable Convolution

Depthwise Separable Convolution は，MobileNet [1] において提案された畳み込み手法である。図 1 に Depthwise Separable Convolution の構造を示す。空間方向の畳み込みとチャネル方向の畳み込みを分割しており，それぞれ Depthwise Convolution，Pointwise Convolution と呼ぶ。

従来の畳み込みにおける入力チャネル数を c_{in}，出力チャネル数を c_{out}，カーネルサイズを $k \times k$，特徴マップサイズを $p \times p$ とするとき，パラメータ数 P は，

$$P = c_{\text{in}} c_{\text{out}} k^2 \quad (1)$$

で表される。

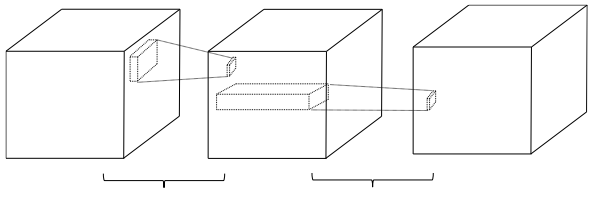

図 1 Depthwise Separable Convolution の構造

Depthwise Separable Convolution のパラメータ数 P' は，

$$P' = P_{\text{DW}} + P_{\text{PW}} = (c_{\text{out}} + k^2)\, c_{\text{in}} \quad (2)$$

で表せる。$c_{\text{out}} >> k^2$ と近似すれば，パラメータ数を $1/k^2$ に削減することが可能である。

2.2 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から別のアーキテクチャにパラメータを変換することである。再パラメータ化では，複雑なモデルでトレーニングされたパラメータを，より単純なモデルで利用できるように変換を行う。再パラメータ化により，FPGA 実装のために単純化された CNN においても高度な推論を行うことが可能になる。

2.3 提案ネットワーク

本稿では，特徴抽出部とクラス・位置推定部の 2 つの部分に対して変更を加えることで，監視カメラシステムの FPGA 実装に適した CNN を提案する。図 2 に提案手法のネットワーク構造を示す。

特徴抽出 CNN では，通常の畳み込み演算と比較して軽量な Depthwise Separable Convolution を利用して，パラメータ数を削減しており，回路の共有が可能で FPGA 上で層多重化回路を構成できるネットワークとする。また，位置・クラス推定 CNN へと接続する際に利用する特徴マップに深い層の特徴マップを利用することで，小さい物体を検出するときに利用する特徴マップと，大きい物体を検出するときに利用する特徴マップの畳み込み演算の回数の差が少なくなるように工夫した。さらに，再パラメータ化を行い，学習時と推論時もモデルを変更している。学習時は複雑なモデルで学習を行い表現力を向

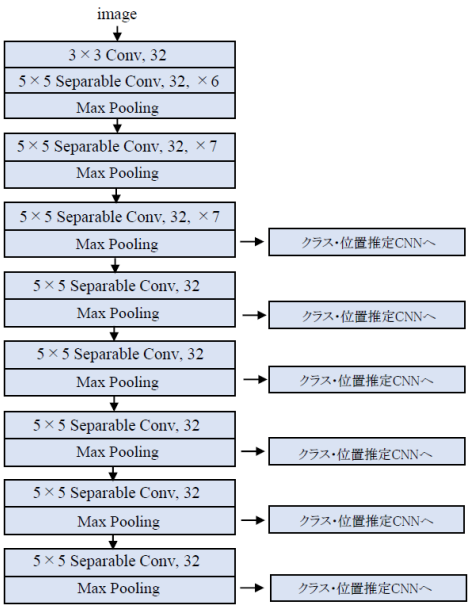

図 2 提案手法のネットワーク構造

†神戸大学，Kobe University

表 1　実験に利用したネットワークの詳細情報

ネットワーク名	Network A	Network B	Network C	Network D
パラメータ数（比）	23,745,908 (1.0000)	79,308 (0.0033)	100,116 (0.0042)	100,116 (0.0042)
特徴抽出部のネットワーク	VGG-16 +extra	Basenet2	Basenet2	Basenet3
Dbox 数（out1, out2, out3, out4, out5, out6）	(4,6,6,6,4,4)	(3,3,3,3,3,3)	(4,6,6,6,4,4)	(4,6,6,6,4,4)

表 2　監視カメラデータセットの実験結果

ネットワーク名	適合率	再現率	F 値
Network A	0.90	0.97	0.93
Network B	0.90	0.88	0.89
Network C	0.87	0.84	0.85
Network D	0.69	0.88	0.77

上させ，推論時はパラメータ数の少ない軽量なモデルで高精度な検出を行うことができる。

クラス・位置推定 CNN では，人を検出することを考慮し，横長の Dbox を削除することでパラメータ数を減少させている。

3. 評価実験と考察

3.1 実験内容

提案した監視カメラシステムの FPGA 実装に適した CNN に関して，Pascal VOC 2007 [2] で学習させたモデルに対して，監視カメラデータセットでファイン・チューニングを行った上で，認識精度を評価する。表 1 に，実験に利用したネットワークの詳細情報を示す。実験に利用したのは，Network A ~ D で，Network A は SSD の従来手法である。Network B と C は再パラメータ化を利用する CNN を利用しており，Network B はデフォルト・ボックスの削減を行っている。Network D は Network B，C の推論時のネットワークで学習を行ったネットワークとなっている。

また，特徴抽出 CNN を ZCU102 にマッピングした上で，利用リソース数を評価する。

3.2 実験結果と考察

表 2 に，Pascal VOC 2007 で学習させた重みに対して，監視カメラデータセットでファイン・チューニングを行い，再度ネットワークの重みを更新したうえで，監視カメラ画像に対してテストした結果を示す。

表 1 のパラメータ数に関して，従来の Network A に対して提案した Network B〜D では 99% 以上削減されている。

表 2 の結果から，Network A が適合率，再現率，F 値で最も高い精度を示した。Network B，C に関しては，Network A より精度は劣るものの，大きな低下は見られなかった。これは，監視カメラシステムを想定すると，撮影した画像の背景が変わらず，また画像の手前から奥に人間が移動したとしても大きくバウンディング・ボックスの形状が変わらないという特性があるため，検出難度が低いためだと考えられる。

Network B と Network C を比較すると，Dbox の数を削減した Network B の F 値が Network C と比べて 0.04 pt 高くなった。また，Network B は Network C と比べてパラメータ数が 20.8% 削減されたことから，Dbox 削減の効果が確認できた。

また，Network C と Network D は推論時のパラメータ数が等しいが，再パラメータ化を行った Network C の F 値について，0.08 pt の精度向上効果が確認できた。このことから，再パラメータ化を行うことで，シンプルなネットワークで精度の高い推論を行えることが確認できた。

図 3 に Network B を利用した物体検出例を示す。人の位置と形を正確に検出できていることがわかる。

表 3　マッピング結果

リソース	利用数（個）	利用可能数（個）	利用率（%）
LUT	258,274	274,080	94.23
LUTRAM	6,308	144,000	4.38
FF	235,588	548,160	42.98
BRAM	612.50	912	67.16
DSP	2,398	2,520	95.16

表 3 にマッピング結果を示す。各リソースの利用率について，LUT 94.23%，LUTRAM 4.38%，FF 42.98%，BRAM 67.16%, DSP 95.16% となった。畳み込み演算には多数の積和演算が含まれるため，DSP の利用率が高くなった。また設計した回路では畳み込み後の値をすべて保存する必要があるため，多くの BRAM が消費されている。26 層で構成される特徴抽出 CNN を，FPGA が搭載するリソースの範囲内で実装することができた。

4. まとめ

本稿では，監視カメラシステムの FPGA 実装を目的として，精度低下を抑制しつつ，パラメータ数を削減した物体検出モデルの提案を行った。具体的には，従来の畳み込み演算と比較して軽量な Depthwise Separable Convolution を利用して，パラメータ数の削減を行った。さらに，再パラメータ化を利用することで，ソフトウェア上で学習させた複雑な CNN を，FPGA 実装に適した軽量な CNN に変換を行い，精度の向上を図った。また，層多重化回路として実装可能で，小さい物体の検出が比較的容易なネットワークとした。また，監視カメラシステムで人を検出することを想定して，横長のデフォルト・ボックスを消去し，パラメータ数を削減した。

評価実験を行った結果，従来のネットワークと比較して，F 値の精度低下を 0.04 pt に抑えつつ，パラメータ数を 99% 以上削減可能であることを確認した。また，FPGA 実装時のリソース数に関しては，提案する CNN を構成する全 26 層の特徴抽出 CNN を，FPGA が搭載するリソースの範囲内で実装可能であることを確認した。

今後の課題として，提案した CNN のネットワークを利用した監視カメラシステムの処理を FPGA 上で行うことが挙げられる。

参考文献

[1] A. G. Howard et al, "MobileNets: Efficient convolution neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861v1, 17 Apr. 2017.

[2] The PASCAL Visual Object Classes Challenge 2007, http://host.robots.ox.ac.uk/pascal/VOC/voc2007/

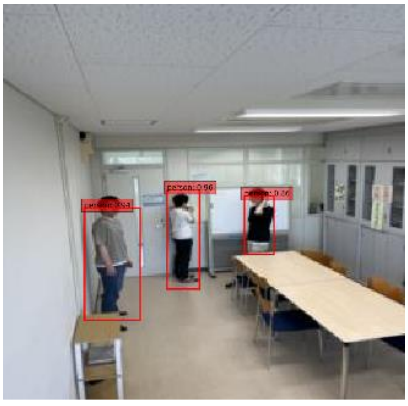

図 3　物体検出を行った出力画像