ショートノート

YOLOX-Nano を構成する層を MobileNetV3 を構成する層に組み替えた効果の検証

Verification of Effect caused by Replacement of Layers in YOLOX-Nano with Layers in MobileNetV3

堀田 忠義，吉久 翔悟，秋葉 将和
Tadayoshi Horita, Shogo Yoshihisa, and Masakazu Akiba

This paper proposes a new object detection model, which is based on YOLOX-Nano, but some layers not only in backbone but also in neck and head parts of YOLOX-Nano are replaced by layers in MobileNetV3. This paper also shows the comparison between the proposed model and the original YOLOX-Nano and the other YOLO models in terms of number of parameters, inference time, and object detection accuracy.
Keywords: deep learning, object detection model, YOLOX-Nano, MobileNetV3, layer replacement

1. はじめに

物体検出モデルは，多くのコンピュータビジョンタスクで活用されている．例えば自動走行車では，車に搭載されたカメラで，歩行者や標識を認識するシステムの中核部分に，物体検出モデルが使われている．また品質検査では，製造現場における目視検査の作業を自動化するシステムの中核部分に，同様に物体検出モデルが使われている．

IoT 機器や小型ロボットなどの組み込みの分野では，システムに組み込むコンピュータの消費電力や物理サイズに関する制限がある場合がほとんどである．畳み込みニューラルネットワーク[1]を使った物体検出モデルは，多くの計算リソースを必要とする．そのため，工夫なしでは，組み込みシステムでの実装は非現実的である．

MobileNetV3[2]は，組み込みシステムに向けて開発された，軽量な画像分類モデルである．また YOLO[3]は，物体検出モデルの 1 つであり，YOLOv2[4]や YOLOX[5]はその派生モデルである．

YOLO(またはその派生モデル)のバックボーンを MobileNetV3 に組み替えた物体検出モデルは，例えば参考文献[6]など，他者研究が多く存在している．それらの物体検出モデルは，オリジナルモデルと比較して，モデルの軽量化，検出精度向上のいずれか，もしくは両方の効果が得られている．

ところで，バックボーンだけでなく，ネックやヘッドを構成する層を，MobileNetV3 を構成する層に組み替えたモデルは，オリジナルモデルと比較して，モデルの軽量化，もしくは検出精度向上の効果が得られる可能性がある．しかし，このようなモデルを扱う他者研究は見当たらない．

本研究では，YOLOX-Nano[5]のバックボーンを MobileNetV3-Small[2]に置き換え，さらにネックおよびヘッドの一部を，MobileNetV3 を構成する層に置き換えることによって構成した物体検出モデルを提案する．また本研究の目的は，オリジナルモデルと比較した提案モデルの，パラメータ数，推論時間および検出精度の変化を検証することである．

2. 提案モデルの構築

2.1. MobileNetV3 Bottleneck と CSPLayer の違い

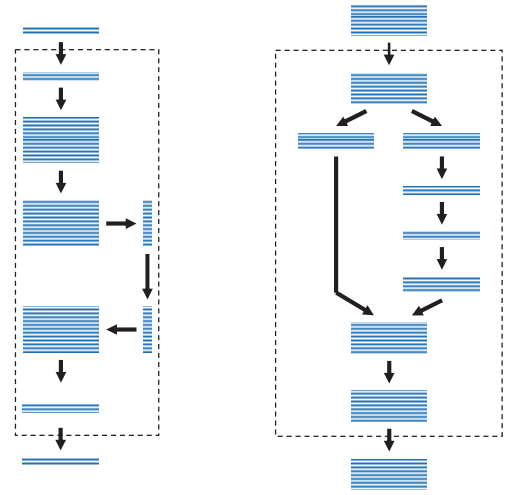
MobileNetV3 Bottleneck は，MobileNetV3 を構成する主なモジュールである．このモジュールでは，軽量化のために，Inverted residuals という手法が利用されている．この手法では，圧縮された特徴マップを入力とし，層内で展開し，深さ方向畳み込みで特徴抽出をした後，圧縮して特徴マップを出力する．特徴マップを圧縮して伝搬することにより，パラメータ数を削減している．

CSPLayer は，YOLOX のバックボーンである CSPDarkNet を構成する，主なモジュールである．このモジュールは，CSPResNet を参考に構成されており，入力された特徴マップの一部を畳み込み，それと入力されたそのままの特徴マップを結合する構造を持つ．入力された特徴マップの全体ではなく，一部を畳み込むことによって，パラメータ数を削減している．

MobileNetV3 Bottleneck および CSPLayer の構造をそれぞれ，図 1 (a)および(b)に示す．同図において，点線大枠

で囲われた部分はモジュールを，横線から成る四角は特徴マップを，矢印は畳み込みなどの演算を，それぞれ示す．

ImageNet データセットによる Top-1 accuracy は，CSPResNet が 65.3%であるのに対し，MobileNetV3 では 67.4%で，少し高精度である．また，パラメータ数は，CSPResNet が 2.7M であるのに対し，MobileNetV3 では 2.5M で，少し軽量である．CSPDarkNet の Top-1 accuracy およびパラメータ数は，公開されていないため，それとの比較はできない．しかし，CSPDarkNet はベースとなった CSPResNet と同等の性能があると考えられる．よって，



(a) MobileNetV3 Bottleneck　　(b) CSPLayer

図 1 各モジュールの構造

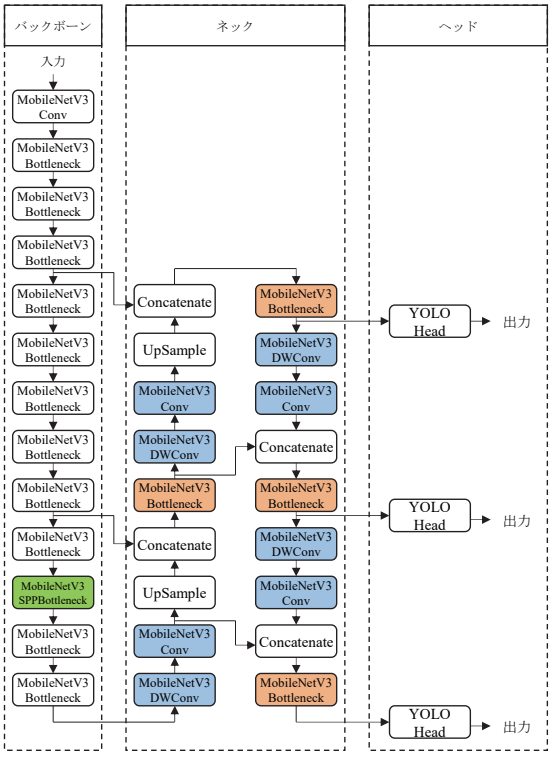


図 2 本研究で検証する物体検出モデルの全体構成

表 1 学習時設定対象ハイパーパラメータ

ハイパーパラメータ	値
最適化アルゴリズム	RMSProp
学習係数	0.001
バッチサイズ	16
エポック数	100
入力サイズ	416×416

表 2 学習時ハードウェア環境

ハードウェア	型式
CPU	Intel Core-i7 11700F
RAM	DDR4-16GB×2
GPU	NVIDIA GeForce GTX1070
VRAM	GDDR5-8GB

表 3 学習時ソフトウェア環境

ソフトウェアまたはライブラリ	バージョン
Windows 10 Pro	21H2
Anaconda	2.3.1
Python	3.7.13
TensorFlow-gpu	2.3.0
CUDAToolkit	10.1.168
cuDNN	7.6.0
Numpy	1.19.2
OpenCV	4.6.0

YOLOX のバックボーン，ネックおよびヘッドを構成する層を，MobileNetV3 を構成する層に組み替えることによって，パラメータ数や検出精度の観点から優位性を持つ可能性があると言える．

2.2. 提案モデル構成の概要

本研究で提案する物体検出モデルの全体構成図を，図 2 に示す．

YOLOX-Nano は，主に CSPLayer と BaseConv で構成されている．提案モデルは，YOLOX-Nano の CSPLayer を MobileNetV3 Bottleneck に，YOLOX-Nano の BaseConv を MobileNetV3 DWConv 層と Conv 層にそれぞれ組み替えたモデルである．これにより，検出精度向上もしくはパラメータ数削減を期待する．

バックボーンでは，全体を MobileNetV3-Small に組み替える．さらに，図 2 の緑色で示すように，MobileNetV3-SPP Bottleneck 層を追加する．この層は，CSPDarknet の SPP Bottleneck の BaseConv 層を，MobileNetV3 を構成する層に組み替えたものである．

ネックでは，図 2 のオレンジ色で示すように，YOLOX-Nano の CSPLayer を MobileNetV3 の Bottleneck 層に組み替える．さらに，図 2 の青色で示すように，YOLOX-Nano の BaseConv 層を MobileNetV3-small の DWConv 層と Conv 層に組み替える．ヘッドにおいても，同様の組み替えを行う．

3.　検出精度やパラメータ数の検証

3.1.　検証環境

学習時の設定対象ハイパーパラメータを表 1 に，学習時のハードウェア環境を表 2 に，学習時のソフトウェア環境を表 3 に，使用した損失関数を表 4 に，それぞれ示す．

モデルの学習および検出精度評価のために，PASCAL VOC[7](2007 および 2012)データセットを使用する．同データセットは，人，動物，乗り物，家具などの 20 種類のクラスがあり，それらの写真およびアノテーション情報データで構成されている．それらのデータセットについて，用途（学習時の使われ方），画像枚数および物体の数を，表 5 に示す．

3.2.　パラメータ数

提案モデルおよび YOLOX-Nano のパラメータ数を，表 6 に示す．提案モデルのパラメータ数は，YOLOX-Nano と比較して 0.22M (24%)増加した．

3.3.　推論時間

NVIDIA 社製シングルボードコンピュータである，Jetson Nano[8]での，提案モデルと YOLOX-Nano の推論時間を表 7 に示す．推論時間は，trtexec の--avgRuns オプションを使用し，100 回推論した推論時間の中央値として求めた．提案モデルの推論時間は，YOLOX-Nano と比較して，同等であることが確認できた．

表 4　損失関数

出力	損失関数
領域回帰	二乗和誤差
背景分類	バイナリークロスエントロピー誤差
クラス分類	バイナリークロスエントロピー誤差

表 5　データセットの画像枚数と物体の数

データセット	用途	画像枚数	物体の数
PASCAL VOC 2007	test	4,952	11,724
PASCAL VOC 2012	train	5,717	13,609
PASCAL VOC 2012	validation	5,823	13,841

表 6　パラメータ数の比較

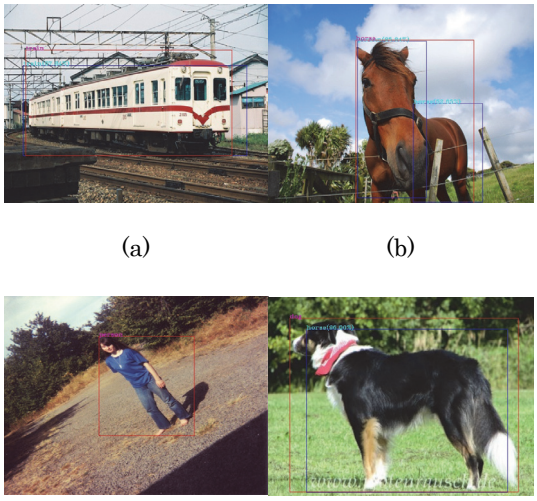
モデル名	パラメータ数
YOLOX-Nano	0.91M
本研究検証モデル	1.13M

表 7　推論時間の比較

モデル名	Enqueue Time	GPU Compute
YOLOX-Nano	11.70 ms	24.52 ms
本研究検証モデル	11.14 ms	25.45 ms

表 8　モデルの *mAP* 比較

モデル名	*mAP*
YOLO	0.634
YOLOv2	0.786
本研究検証モデル	0.229



(a)　(b)　(c)　(d)

図 3　データセットの一部の画像とその出力

3.4.　検出精度

提案モデルと，YOLO および YOLOv2 の *mAP* 値を表 8 に示す．なお，公開されている YOLOX-Nano の *mAP* 値は，COCO データセット[9]で評価されているため，比較できない．

表 8 より，他者研究モデルの *mAP* が 0.5 以上であるのに対し，提案モデルでは 0.229 と，検出精度が低い結果となった．

また，紙面の制約上詳細は省略するが，提案モデルでは，クラスごとの *AP* 値が 0.424～0.051 の範囲となった．

3.5.　推論結果

用途が test であるデータセット内の 4 つの例と，それらに対する提案モデルの推論結果を図 3 に示す．同図において，青色および赤色のマークは，推論結果および正解情報を，それぞれ示す．

同図の(a)では，推論結果と正解情報はほぼ同じ(クラス分類は正しく，かつ領域回帰の誤差は小さい)である．(b)では，クラス分類は正しいが，領域回帰での誤差が大きい．(c)では，person クラスの物体部分が背景として誤認識されている(背景分類誤り)．(d)では，領域回帰の誤差は小さいが，クラス分類を誤っている(正解は dog だが推論結果は horse)．

4.　検出精度低下等についての考察

提案モデルでの，検出精度の低下およびパラメータ数増加の理由について，以下に考察する．

第 1 に，調節可能なハイパーパラメータ探索が不十分と考えられる．本研究で設定対象としたハイパーパラメータは表 1 のとおりであるが，これ以外にも，ネックなどの中間層で扱う特徴マップのチャンネル数，ネック内の MobileNetV3 Bottleneck 層での膨張率（層間のチャンネル数の増減の度合い），および使用する活性化関数の種類，などが調節可能である．しかし，これらのパラメータの探索空間は非常に高次元であるため，容易に予測ができない．つまり，最適なパラメータを手作業による操作のみで探索するには，膨大な時間を要することは明らかである．

一方で，MobileNetV3 を提案している文献[2]では，これらのパラメータをネットワーク検索(Neural Architecture Search : NAS)を利用し最適化しているが，この手法を使うには大規模な計算機リソースを必要とする．つまり，ネットワーク検索が動作可能な計算機リソースを使用可能な状況であったならば，提案モデルについても，より高精度な結果が期待できる．

第 2 に，ネックおよびヘッドの一部を，MobileNetV3 を構成する層に組み替える際に，本研究では単純な置き換えを行ったことも，理由の 1 つとして考えられる．つまり，MobileNetV3 のモジュールをネックおよびヘッドに組み込む際には，組み込み方の検討が必要であることが判明した．

5. まとめ

本研究では，YOLOX-Nano のバックボーンを MobileNetV3-Small に置き換え，ネックおよびヘッドの一部を，MobileNetV3 を構成する層に置き換えることによって作成したモデルを提案し，組み換え前後でのパラメータ数，推論時間および検出精度の変化を明らかにした．

パラメータ数は，少し増加する結果となった．Jetson Nano での推論時間は同等であった．検出精度については，他者研究モデルに比較して，提案モデルの値はかなり悪い値となった．

この検証結果に対し，パラメータ最適化およびモジュール組み込み方法の観点から，考察を述べた．

今後の課題として，ハイパーパラメータの探索，およびモジュール組み込み手法の検討，が挙げられる．

参考文献

[1] Yann LeCun, Léon Bottou, Yoshua Bengio and Patrick Haffner: “Gradient-based learning applied to document recognition”, Proceedings of the IEEE, Vol.86, No.11, pp.2278-2324(1998).

[2] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le and Hartwig Adam: “Searching for MobileNetV3”, arXiv:1905.02244, pp.1-11(2019).

[3] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, “You Only Look Once: Unified, Real-Time Object Detection”, arXiv:1506.02640, pp.1-10(2016)

[4] Joseph Redmon and Ali Farhadi: “YOLO9000: better, faster, stronger”, IEEE Conference on Computer Vision and Pattern Recognition, pp.6517-6525(2017).

[5] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li and Jian Sun: “YOLOX: exceeding YOLO series in 2021”, arXiv:2107.08430, pp.1-7(2021).

[6] Lili Wang, Qinghang Ni, Chen Chen and Hailu Yang: “Lightweight target detection algorithm based on improved YOLOv4”, IET Image Processing, Vol.16, No.14, pp.3805-3813(2022).

[7] Mark Everingham, S.M.Ali Eslami, Luc Van Gool, Christopher K.I.Williams, John Winn and Andrew Zisserman: “The PASCAL visual object classes challenge: A retrospective”, International Journal of Computer Vision, Vol.111, pp.98-136(2015).

[8] NVIDIA: “Jetson Nano 開発者キット”, NVIDIA, https://www.nvidia.com/ja-jp/autonomous-machines/embedded-systems/jetson-nano-developer-kit/(2023 年 2 月 7 日閲覧).

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick and Piotr Dollár: “Microsoft COCO: Common objects in context”, arXiv:1405.0312, pp.1-15(2015).

（原稿受付 2024/3/4，受理 2024/8/28）

*堀田 忠義，博士（工学）
職業能力開発総合大学校，能力開発院，〒187-0035　東京都小平市小川西町 2-32-1
Tadayoshi Horita, Faculty of Human Resources Development, Polytechnic University of Japan, 2-32-1 Ogawa-Nishi-Machi, Kodaira, Tokyo 187-0035.
Email: horita@uitec.ac.jp

*吉久 翔悟，修士（生産工学）
群馬職業能力開発促進センター，〒370-1213 群馬県高崎市山名町 918
Shogo Yoshihisa, Gunma Polytechnic Center, 918 Yamana-Cho, Takasaki, Gunma 370-1213
Email: Yoshihisa.Shogo@jeed.go.jp

*秋葉 将和，博士（工学）
職業能力開発総合大学校，能力開発院，〒187-0035　東京都小平市小川西町 2-32-1
Masakazu Akiba, Faculty of Human Resources Development, Polytechnic University of Japan, 2-32-1 Ogawa-Nishi-Machi, Kodaira, Tokyo 187-0035.
Email: akiba@uitec.ac.jp