6U-06

深層学習を用いた物体検出と認識の応用

王 相衡† 岳 学彬† 孟 林†
立命館大学・理工学部・電子情報工学科†

1. はじめに

近年，深層学習技術は邁進し，コンピュータビジョン，特に画像認識分野において，画像処理と比べると，遥かに優れている高精度を達成してきた．産業や人々の生活に大きな革新をもたらすことが期待されている．その中には，物体の検出と認識を同時に実現できる技術である YOLO[1]モデルが注目され，さまざまな場所で応用されている．本研究は，YOLO を用いて，我々が進んできたリアルタイム物体の検出と認識の応用を紹介し，その結果を報告する．その中には，ロボットと融合する下膳システムと，IoT と融合する古代文献解読システムが含まれる．認識結果から見ると，YOLO モデルは複数分野での物体検出と認識の有効性を示した．

2. 先行研究

2.1 物体認識モデル

実際には，深層学習の基本理論は 1980 頃に，福島らにより提案された[2]．深層学習に対して基盤理論を提供したが，当時のハードウェアデバイスの制限により，十分に発展できなかった．近年，ハードウェアデバイスの進化により，2012 年に，最初の物体認識深層学習モデルとして，AlexNet が提案された．その後，VGG[3]，InceptionNet，ResNet などの高精度な深層学習モデルが提案されてきた．しかし，これらのモデルは，高精度を追究することとともに，深い層を使用することで，パラメータ量が非常に多く，認識時間が長いとの問題を生じている．さらに，エッジデバイスに実装される場合は，メモリ不足の問題で，大規模で複雑なモデルを適用することが困難で，深層学習の応用範囲が制限された．そのため，研究者らは ShuffleNet，MobileNet[4]， MNasNet などの複数の軽量モデルを提案されていた．その中には，MobileNet がモデルのパフォーマンス（精度）を維持しながら，速度を上げることで，世の中に知らされてきた．

2.2 物体検出と認識を同時に実現できるモデル

現在，深層学習において，画像認識のみのモデルは，社会のニーズに十分に答えできなくなるため，物体の検出と認識を同時に実現できるモデルが提案された．その中には有名なのが SSD[5]と YOLO である．SSD は認識モデルである VGG を物体検出モデルに組み込み，物体の検出と認識を同時に実現する．YOLO シリーズは近年優秀な物体検出と認識モデルとして，精度を落とさずに速度を向上させる．さらに，YOLOv4[1]のパフォーマンスは，YOLOv3 より大幅に向上している．

3. モデル

図 1 は使用した YOLO モデルの構成を示す. Back bone は，モデルの基幹ネットワークで，特徴量の抽出で使用されている．従来，Backbone は，VGG モデル構造に似て，畳み込みにより特徴の抽出を行い，プーリングにより次元の圧縮を行う．しかし，これらの処理に大量なパラメータと実行時間を必要であるため，Backbone を軽量化することが考えられる．

図 1 に示しているように，サイズは 416 x 416 の画像を入力し，Backbone により特徴マップを抽出する．その後，特徴マップは SPP と PANet に転送する．SPP ネットワークは 3 つの異なるスケールプーリング 5x5，9x9，13x13 を使用する．これにより，より豊富な特徴が得られ，受容野を広げことができる．PANet は，主にターゲット検出におけるマルチスケールの問題を解決するために設計されている. SPP によって処理した特徴マップをアップサンプリングとダウンサンプリング処理を行う．したがって，さまざまなサイズのオブジェクトの検出能力を向上させる．

4. 実験

本研究は，ロボットと融合する下膳システムと，IoT と融合する古代文献解読システムの複数のアプリケーションを用いて，提案した YOLO での認識結果を評価する．

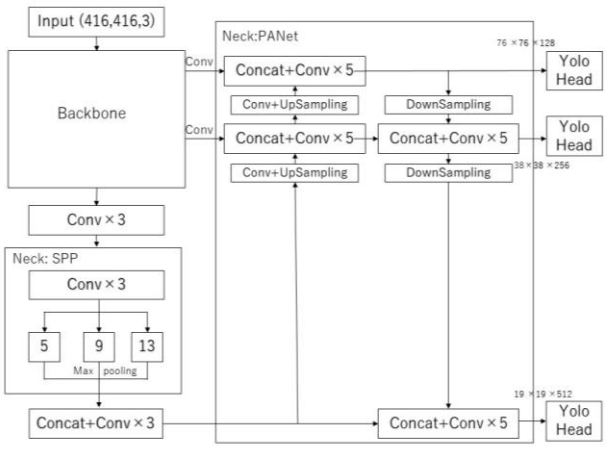

図 1．モデルの構造

Applications of deep learning based object detection and recognition
† Xiangheng Wang, Xuebin Yue and Lin Meng, Department of Electronic and Computer Engineering, College of Science and Engineering, Ritsumeikan Univ.

4.1 ロボットと融合する下膳システム

高齢少子化が著しく進んでいる現在社会には，人手不足が深刻である．我々はこのニーズに応じて，ロボットと深層学習を融合し，人手不足を緩和することを目指す．本研究は，ロボットアームにカメラを取り付け，深層学習により食事後の食器を認識し，把持ポイントをロボットアームに渡す．それにより，下膳回収の無人化を実現する．

実験において，15 種類の食器に対して，合計 501 枚写真を撮影した．その中に，訓練データは 506 枚，検証データ 46 枚，テストデータ 51 枚とする．本研究は下記の実験条件で，OS: Ubuntu 21.04，Memory: DDR4 16*2，CPU: Intel Core i9，GPU: NVIDIA Geforce RTX 3080Ti，約 1 週間で学習を行った．学習済みモデルはテストセットに検証し，Precision と mAP (mean Average Precision)を記録する．

パラメータの量を減らし，推論時間を短縮するために、我々は YOLOv4 の CSPDarknet53 と呼ばれる Backbone を軽量認識モデルに替える．修正後モデルを学習し，モデルのサイズから見ると，YOLOv4 は 144.4MB で，軽量認識モデルを用いて書き換えたモデルは 46.9MB までに減少した．図 2 は食器認識の結果を示す．表 1 は食器の実験結果を示し，CSPDarknet53 が従来の YOLO モデルで，軽量モデルは軽量化された YOLO である．Precision から見ると，軽量モデルは CSPDarknet53 により僅かな改善しか満たさないが，mAP を多いく改善したことを確認できた．また，軽量化することにより，軽量化モデルは，CSPDarknet53 により 1.75 倍の速度向上を達成した．従って，本研究は YOLO モデルの軽量化を達成し，YOLO を用いて，産業に貢献できる可能性を示した．

表 1　食器の実験結果

Table 1　Experimental results of dish clearing.

	Precision	mAP	処理時間 s/枚
CSPDarknet53	96.19	78.82	0.0126
軽量モデル	98.64	95.26	0.0072

4.2 IoT と融合する古代文献解読システム

産業のみならず，深層学習の物体検出と認識を用いて様々な分野で応用ができると考える．また，深層学習は，古代文献解読の有効性が既に証明できた[6]．今までの我々は YOLO の研究ノウハウを用いて，IoT と融合し，古代文献の解読を実現し，文化遺産の保護と継承に貢献することを目指す．

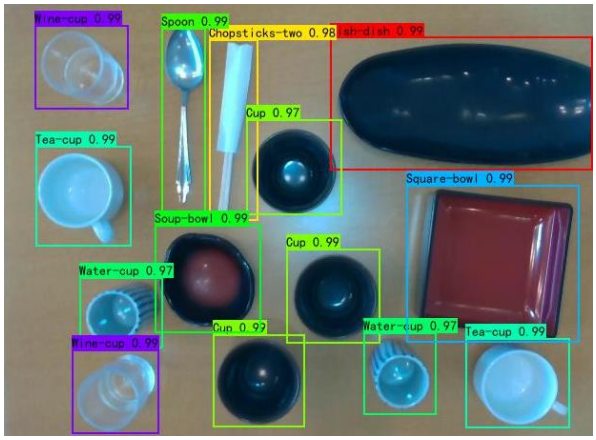

図 2．食器認識結果図

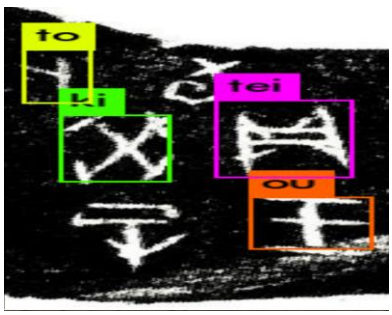

図 3．古代文献解読結果図

IoT と融合した YOLO は，ロボットと融合した下膳システと異なり，YOLO モデルをエッジデバイスに組込みではなく，クラウドサーバーに組み込み，インタネットを通じて，YOLO により物体の抽出と認識を目指す．応用において，古代文献の解読システムでは，ユーザがインタネットを通じて，認識対象とした古代文献の画像をサーバーに転送し，サーバーからの認識結果をユーザに返却する．それにより，どこでも、いつでも古代文献を解読できる．図 3 は，IoT と融合する古代文献解読システムを使用した古代文献である甲骨字の認識結果を示す．検出結果はバンディングボックスされ，バンディングボックスの上に認識された結果が表示されている．結果から見ると，複数の文字の検出と認識ができたため，モデルの有効性を示した．

5. おわりに

本稿では YOLO という物体の抽出と認識を同時に実現できる深層学習モデルを用いて，ロボットと融合する下膳システムと，IoT と融合する古代文献解読システムの複数のアプリケーションでの応用を確認し，その有効性を示した．従って，AI 技術は産業のみならず，文化遺産の保護と文化の継承などに貢献できると考える．今後，実験により確認された知見を用いて，更なる精度の向上とアプリケーションの実用化を目指す．

参考文献

[1] B. Alexey, et al.,: YOLOv4: Optimal Speed and Accuracy of Object Detection, CoRR, abs/2004.10934, 2020.

[2] K.Fukushima,: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, 36(4): 193–202,Biological Cybernetics, 1980.

[3] S. Liu, et al.: Very deep convolutional neural network based image classification using small training sample size, ACPR2015, 2015.

[4] Howard, et al.: Searching for MobileNetV3, Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV), 2019.

[5] W. Liu,et al.,: SSD: Single shot multibox detector, In ECCV, 2016.

[6] L. Meng, et al.,: Ocrale Bone Inscription Detector Based on SSD, LNCS, volume 11808, ICIAP 2019.