2Q-06

深層学習と画像セグメンテーション に基づいた動物種類と位置の検出

任 散陽 ， 中田　裕一 ， 田中 直樹
神戸大学大学院 海事科学研究科

1. はじめに

近年、深層学習の急速な発展に伴い、オブジェクト検出分野のいくつかのモデルが実際のプロジェクトに適用されてきた。例えば、工場の生産ラインでは、欠陥品を迅速かつ正確に検出する必要がある。自動車の自動運転では、周囲のすべての物体をリアルタイムで検出し、道路と障害物の位置を区別し、安全な運転ルートを計算する必要がある。スマートフォンの撮影アプリでは、リアルタイムで画面にフィルター効果を追加する必要がある。使用シナリオごとに計算資源が異なるため、認識速度と精度の要件も異なる。

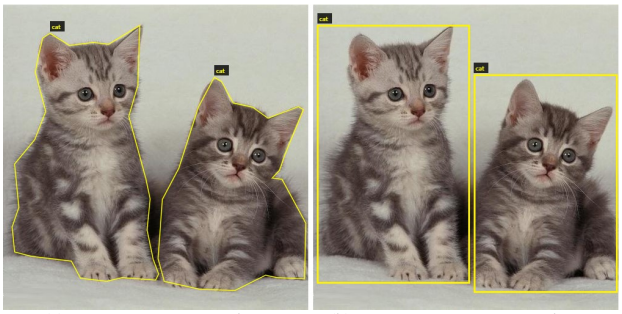
畳み込みニューラルネットワークに基づいた画像分類とオブジェクト検出の分野では、新しいモデルが毎年発表されている。これらのモデルを使用すると、画像の分類、オブジェクトの Bounding box の出力に加えて、ピクセルレベルのインスタンスセグメンテーションもできる。実際の検出では、画像ファイルを入力できるだけでなく、 PC の Web カメラとスマートフォンカメラでリアルタイム検出もできる。

これらのモデルの違い、及びそれぞれのモデルの長所と短所について、多くのモデルの中からプロジェクトに適したモデルを選択する方法について検討を行う。さらに、モデルをゼロからトレーニングする方法、オリジナルのオブジェクト検出ソフトウェアを開発する方法についても検討を行う。

2. データセット

この研究では、 Google の画像検索を使用して、40 種類の動物の著作権フリーの画像を収集する。データセットとして適切な画像を選択して、すべての画像の中心部分をカットして、500 * 500 の JPG 形式にフォーマットする。トレーニング時に、モデルの入力サイズによって、もう一度リサイズする。各動物の画像数は 150〜250 枚で、合計 9024 枚である。

オブジェクト検出モデルをトレーニングするために、これらの画像には 2 つの方法でアノテーションを行う。1 つはセマンティックセグメンテーション情報が付加された Microsoft COCO の形式で、1 つは Bounding Box 情報が付加された PASCAL VOC の形式である。アノテーションをするとき、画像のパス、オブジェクトカテゴリー、およびオブジェクト境界の座標値が XML ファイルに保存される。



(a) MS COCO Annotation　(b) PASCAL VOC Annotation

3.実験

実験段階では、我々はいくつかのモデルをトレーニングした。トレーニングしたモデルには、LeNet-5、AlexNet、ResNet、および MobileNet などの画像分類モデルと、Mask R-CNN、YOLOv3、および Tiny YOLOv3 などのオブジェクト検出モデルが含まれる。トレーニング時間、精度、モデルサイズ、認識速度などの項目を比較した。

実験結果は、 Mask R-CNN の認識精度は高く、モデルサイズが大きく、 GPU での検出速度は高

Animal detection and identification Based on Deep Learning and Image Segmentation
Ren Sanyang , Nakata Yuichi , Tanaka Naoki
Kobe University Graduate School of Maritime Sciences

速ですが、 CPU での検出速度は遅いため、 PC での使用に適している。 Tiny YOLO v3 の精度はわずかに低くなりますが、モデルサイズが小さく、高速な検出速度を持ち、携帯電話での使用に適している。モデルの構築から見ると、マトリックス演算が多い。GPU を使用して、認識速度が向上する。反対側に、CPU やスマホの計算力はより弱いので、認識速度が徐々に下がる。

	LeNet5	AlexNet	ResNet18	MoblieNet v2
Training Time	00:09:08	01:40:27	01:40:11	03:09:28
Accuracy Top1	0.241	0.584	0.805	0.839
Accuracy Top5	0.541	0.845	0.961	0.956
Model Size	815.2KB	467.6MB	134.7MB	39.4MB
FPS-GPU(a)	555.6	217.4	128.2	138.9
FPS-CPU(b)	416.7	26.2	33.8	30.1
FPS-Phone(c)	79.7	23.1	21.4	24.9

Table 1. Image Classification Model

	Mask R-CNN	YOLO v3	Tiny YOLO v3
Training Time	02:39:54	03:00:32	03:17:07s
mAP@IoU=50	0.82	0.87	0.61
Model Size	256.1MB	247.0MB	34.0MB
FPS-GPU(a)	4.31	17.1	187.4
FPS-CPU(b)	0.48	1.3	66.1
FPS-Phone(c)	0.23	0.73	24.7

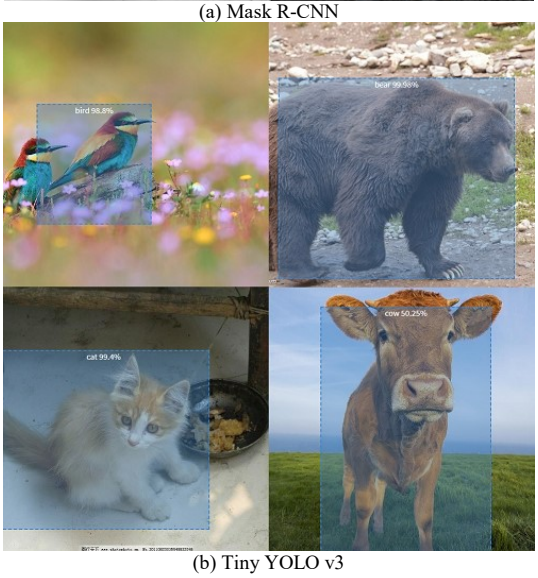
Table 2. Object Detection Model

(a) NVIDIA GeForce RTX 2080Ti/PCIe/SSE2.
(b) Intel(R) Core(TM) i7-7800X CPU @ 3.50GHz * 12.
(c) iPad Mini 4 with iOS 13.1.3.

4.結論

結論として、実際のソフトウェア開発では、使用シナリオを前提として適切なモデルを選択する必要がある。PC の場合、 GPU と CPU のパフォーマンスは強力である。構造が複雑なモデルを使用すると、画像分類を高速で実行できるだけでなく、オブジェクトの位置、さらには境界をより正確に検出できる。携帯電話の場合、GPU はなく、 CPU のみで計算する。リアルタイムのセグメンテーション検出を実行する場合、計算速度が遅いため、効果が良くない。構造が単純なモデルを使用すると、正確な分類を維持しながら、オブジェクトの Bounding Box をすばやく検出できる。

(a) Mask R-CNN



(b) Tiny YOLO v3

5. 参考文献

[1] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner (1998). Gradient-Based Learning Applied to Document Recognition.
[2] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton (2012). ImageNet Classification with Deep Convolutional Neural Networks.
[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun (2015). Deep Residual Learning for Image Recognition.
[4] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik (2014). Rich feature hierarchies for accurate object detection and semantic segmentation Tech report (v5).
[5] Ross Girshick (2015). Fast R-CNN.
[6] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun (2016). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.
[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick (2018). Mask R-CNN.
[8] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi (2016). You Only Look Once: Unified, Real-Time Object Detection.
[9] Joseph Redmon, Ali Farhadi (2016). YOLO9000: Better, Faster, Stronger.
[10] Joseph Redmon, Ali Farhadi (2018). YOLOv3: An Incremental Improvement.
[11] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.
[12] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen (2019). MobileNetV2: Inverted Residuals and Linear Bottlenecks.