物体検出におけるシミュレーション画像を用いた事前学習の有効性の検証

EP18097 法華津伸一　　指導教授：山下隆義

1. はじめに

高精度な物体検出モデルを短い学習時間で実現するには，事前学習モデルの利用が効果的である．物体検出タスクの事前学習モデルでは，画像分類データセットが用いられている．このような異なるタスクのデータを用いた事前学習モデルが物体検出の精度向上にどれだけ有効であるかは明らかではない．そこで本研究では，物体検出の事前学習に適した大規模な物体検出用データセットの作成し，事前学習の有効性を調査する．

2. 物体検出

物体検出は，画像内の物体の位置情報とクラスを推論するタスクであり，自動運転に欠かせない技術である．

2.1. Your Only Look Once v4(YOLOv4)

YOLOv4[1] は，1-Stage 型の物体検出モデルである．クラス予測とバウンディングボックスの推定を同時に行い，高速に推論する．ラベルスムージングや CIoU Loss などの工夫を加えることで，YOLOv3 の計算コストを抑えつつ物体検出精度の向上を可能にした．

2.2. Pyramid Vision Transformer v2(PVTv2)

PVTv2[2] は，Vision Transformer にピラミッド構造を適用することでマルチスケールな特徴の獲得を可能にしたモデルである．計算コストを削減するために Transformer の Self-Attention に用いる Key と Value に対して空間縮小を施す Spatial Reduction Attention(SRA) を提案している．また，パッチウィンドウの拡大による局所的に連続する特徴の抽出，FFN に畳み込み層を導入して位置情報を考慮，SRA の空間縮小処理を Average Pooling に変更という 3 つの工夫を行い，PVTv1 の計算コストを抑えつつ高精度化を実現した．

3. シミュレータを使用したデータセット作成

物体検出では，明るさや天候といった環境変化の影響を受けずに物体を検出する必要がある．そこで，物体やシーンの多様性を学習するために，大規模なデータセットを用いた事前学習が必要とされている．しかし，物体検出に特化した事前学習用データセットは存在しない．そこで本研究では，シミュレータを用いて物体検出における事前学習用のデータセットの作成を行う．データセットの作成には，自動運転シミュレータである CARLA を用いる．データの収集は多様性を考慮し，時間と天候の変動させる．データは，車載カメラから 20 秒毎に画像を取得する．収集したデータは，画像サイズが 2,048 × 1,024 pixel，クラスは乗用車，トラック，バイク，自転車，歩行者，標識，信号機，その他の自動車の計 8 クラスである．収集した 200,000 枚で構成されるデータセットを CARLA V2 とする．シミュレータによるデータの生成例を図 1 に示す．

図 1: 画像データの例

4. 評価実験

作成した CARLA V2 の有効性を確認するため，ImageNet の事前学習モデルと CARLA V2 の事前学習モデルを用いて BDD100K の学習を行い，精度比較を行う．学習には，YOLOv4 と PVTv2 を用いる．

4.1. 実験結果

定量的評価結果を表 2 に示す．YOLOv4 は事前学習なしの場合に比べ，事前学習に ImageNet-1k を用いることで 8.8pt，CARLA V2 を用いることで最大 2.0pt 精度が向上した．PVTv2 では，事前学習に ImageNet-1k を用いることで 2.7pt 精度向上したが，CARLA V2 を用いた場合には 0.3pt 精度が低下することを確認した．このことから，CNN ベースの YOLOv4 には物体検出に特化した事前学習データは有効であるが，Transformer ベースモデルには効果がないといえる．

次に，CARLA V2 を用いて学習した Backbone の重みのみをファインチューニングに用いた場合と，Neck までの場合を比較すると，YOLOv4 では 0.3pt，PVTv2 では 12.4pt 精度が向上した．以上より，物体検出の事前学習には Backbone だけでなく Neck や Head を含めた事前学習が有効であることを確認した．

表 1: 定量的評価 [%]

手法	事前学習	Backbone	Neck	Head	$AP_{50:95}$
YOLOv4	なし				7.8
	ImageNet-1k	✓			**16.6**
	CARLA V2	✓			9.5
		✓	✓		9.8
PVTv2	なし				31.0
	ImageNet-1k	✓			**33.7**
	CARLA V2	✓			9.9
		✓	✓		22.3
		✓	✓	✓	30.7

定性的評価を図 2 に示す．CARLA V2 の事前学習モデルを用いることで，YOLOv4 では歩行者の後方に位置する自動車の検出，PVTv2 では自動車の手前に位置する歩行者の検出ができるようになり，オクルージョンに対する検出精度の向上を確認した．しかし，YOLOv4 では歩行者，PVTv2 ではオクルージョンが発生しているバスや自動車を検出できなくなることを確認した．YOLOv4 で歩行者が検出できない原因として，輪郭部の鮮明さがドメインにより異なるためだと考えられる．PVTv2 でオクルージョンが発生した自動車やバスに対して検出ができない原因として，渋滞したデータの不足やバスのクラスが CARLA データセットには含まれていないためだと考えられる．

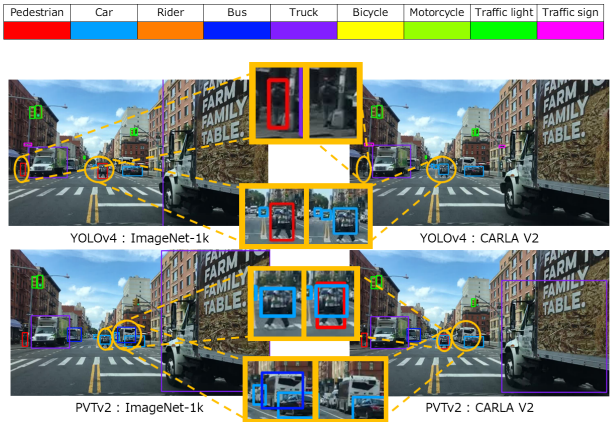

図 2: 定性的評価

5. おわりに

本研究では，シミュレータを用いたデータセットの作成，物体検出で事前学習を行った場合の傾向調査を行った．その結果，CARLA で定義されていないクラスとオクルージョンが発生した自動車の検出率の低下を確認した．そのため，高精度化を実現するにはクラス定義の統一やデータ枚数を増加させることが必要であると考えられる．

参考文献

[1] A. Bochkovskiy, *et al.*, “YOLOv4: Optimal Speed and Accuracy of Object Detection”, arXiv, 2020.

[2] W. Wang, *et al.*, “PVTv2: Improved Baselines with Pyramid Vision Transformer”, arXiv, 2021.