H-051

Transformer-based モデルによる運転者視界に対する物体検出性能の検討

Studying Object Detection Performance for Driver's Field of View by Transformer-based Models

小林 陽[1)]　猿田 和樹[2)]　陳 国躍[2)]　寺田 裕樹[2)]
Hinata Kobayashi　Kazuki Saruta　Guoyue Chen　Yuki Terata

1. 序論

1.1 研究背景

安全運転教育を目的として，物体検出と視線計測の融合によって，運転者が注視した物体を明らかにする研究が行われている [1, 2]．これらの研究における「注視物体判定システム」では，まず，運転者に眼鏡型視線計測器を装着させ，運転中の視界映像と視線座標を取得する．そして，視界映像に対して物体検出後，視線座標を重畳することで注視物体判定を実現している．ここで，この注視物体判定は，運転者自身の眼を通して見た視界中の**全物体**について，その領域と種類が**完全に正しく**検出されている前提の元に成り立つものである．例えば，運転者がある物体を注視していたとしても，その物体が検出されていない，あるいは検出領域が適切でなければ，視線座標上には物体が存在しないことになり，「注視していない」と判定されてしまう．このように，物体検出パートに課せられているタスクは非常に難易度の高いものである．

また，「運転者自身の眼を通して見た視界」として扱う視界映像は，自然環境を可視光カメラで撮影した映像である．晴天の昼間のような好条件時を基準に考えると，カメラに起因するショットノイズやぶれなどに加えて，自然環境に起因する明度変化や天候変化など，頻繁に発生する「ノイズ」により，映像の品質は容易に劣化する．加えて，非常に多種類のノイズが存在し，訓練時に網羅することは現実的ではない．こうしたデータ品質劣化が生じやすく，かつ，運用時に初見のノイズに直面する可能性の高い中で，前述の厳しい前提を達成する必要がある注視物体判定システム内の物体検出モデルの性能向上は重要な課題である．

文献 [1] の池田らは Faster R-CNN [3] を，文献 [2] の和田らは YOLOX [4] を物体検出モデルとして用いている．しかし，表 1 に示すように，どちらも注視物体判定システムの物体検出パートに求められる厳しい前提の保証には十分な値を示していない．また，両者は CNN-based 物体検出モデルであり，近年主流の高性能な Transformer [5] を基盤とする物体検出モデルの有効性は評価されていない．さらに，前述のように運転者視界の撮影時に生じる様々な種類のノイズに対する推論時のロバスト性は議論されていない．

1.2 目的

こうした背景から，本研究では，注視物体判定システムの前提下において，Transformer-based モデルの検出性能を多角的に検証し，注視物体判定システムの物体検出パートでの有効性を明らかにする．

1） 秋田県立大学大学院　Graduate School of Akita Prefectural University
2） 秋田県立大学　Akita Prefectural University

2. 物体検出への Transformer の応用

Transformer は，自然言語処理分野で提案された Encoder-Decoder モデルである．Self-Attention により，系列データの全トークン間の依存関係の濃薄を強調する大局的な特徴抽出を実現できる．以下では，物体検出モデルの Backbone に用いられる画像認識モデルへの応用と物体検出モデルへの応用について説明する．

2.1 画像分類モデル

Transformer を応用した画像分類モデルとしては Vision Transformer [6] が代表的である．画像をパッチ状に分割して系列データとみなすことで，Transformer の Encoder をほぼそのまま用いている．シンプルな構造だが，大きなパッチサイズの固定解像度のまま処理されたり，計算量が二次関数的であるといった課題があるため，Backbone として用いることは現実的ではない．

そこで，派生手法の Swin Transformer [7] では，Self-Attention の計算方法の工夫により，CNN のような階層的な特徴マップを形成しつつ，線形な計算量も実現している．このモデルを Backbone として用いた物体検出モデルの多くは高い性能を発揮している．

2.2 物体検出モデル

Transformer を応用した物体検出モデルとしては DETR [8] が代表的である．各 Ground-Truth に複数の予測が紐づく一対多対応の既存モデルは重複検出を排除する後処理が必要だが，Transformer により全予測間の関係性をモデル化することで，後処理無しに各 Ground-Truth に 1 つだけ予測が紐づく一対一対応の物体検出を実現している．一方で，一対一対応の設計上，Decoder のトークン（Object Queries）のうち，物体が存在すると予測された Positive Queries が物体が存在しないと予測された Negative Queries よりも圧倒的に少なくなり，極端な不均衡が生じる問題がある．これにより，Encoder が画像全体に均等に注意を向ける状態となり，この曖昧な特徴マップが Decoder の訓練に用いられる悪循環が生じてしまう．

この課題に対する派生手法が，Faster R-CNN などの一対多で予測する複数モデルの検出 Heads（補助 Heads）を，訓練時にだけ DETR 系モデルの補助に用いる Co-DETR [9] という訓練手法である（図 1）．この手法では，まず Encoder からの特徴マップを用いて Decoder が訓練される．加えて，Encoder の特徴マップを補助 Head に入力して得られた予測のうち，一対多マッチングによって Positive とされた予測の位置に対応する特徴マップ上の

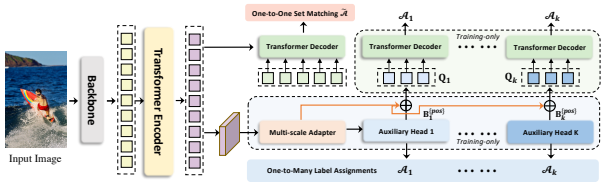

図 1　Co-DETR の概略 [9]

特徴を Object Queries として Decoder の訓練に用いる．これにより，実質的に Positive Queries が増えることで Decoder の訓練効率を向上できるほか，Encoder の特徴抽出能力も向上できる．この手法を DETR 系の DINO [10] に適用した Co-DINO が現時点で COCO の SOTA モデルである．

3. 物体検出実験

3.1 データセットの概要

本実験では，主に昼間の晴天・曇天時に眼鏡型視線計測器で撮影された運転者視界映像のフレームで構成される独自データセットを用いる．画像データ例を図 2 に示す．クラスは 10 クラスあり，訓練用データが 2,045 枚，検証用データが 395 枚ある．正解バウンディングボックスは，物体の大小を問わず，目視や物体位置の関係性によってクラスを判別できる全物体に付与されている．

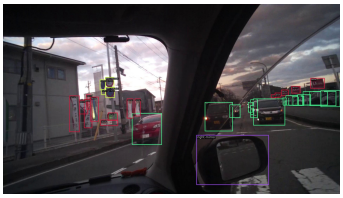

図 2 画像データ例

3.2 実験概要

本実験では，Transformer-based モデルとして 2.2 で述べた SOTA モデルの Co-DINO を用い，先行研究の CNN-based モデルと比較する．手順として，事前学習済みモデルを前述の訓練用データによりファインチューニング後，「検証用データに対する検出性能 mAP」と「データ品質を劣化させた検証用データに対する検出性能」を測る．後者については，文献 [11] を参考に，1.1 で述べたようなノイズを人工的に 14 種類再現し，mPC（各ノイズ下の検証用データでの mAP の平均）と rPC（基の検証用データでの mAP に対する mPC の比）を測る．

3.3 実験結果

表 1 に実験結果を示す．Transformer-based な Co-DINO (Swin-L) の各 mAP は，CNN-based モデルと比較すると，明確に大きな差が見られる．Small サイズの物体を除く全ての mAP が 50%を超え，特に Large サイズの物体については 78.5%という高い値を示している．データ品質劣化時の性能を見ると，Co-DINO (Swin-L) の mPC は各 CNN-based モデルの「mAP」に迫る値を示し，rPC も 77.1%と特に高い．このことから，Transformer-based モデルが推論時に初見のノイズに直面しても本来の性能から劣化しにくいことがわかる．

同じ CNN-based Backbone (ResNet-50) を用いた場合でも同様に，検証データの品質劣化の有無を問わず，Transformer-based アーキテクチャが性能向上に寄与していることがわかる．

表 1 実験結果

Method	Backbone	mAP	mAP_S	mAP_M	mAP_L	mPC	rPC
Faster R-CNN	ResNet-50	49.6	13.8	45.5	63.7	29.0	58.5
YOLOX-S	CSPNet	42.6	8.3	26.9	57.7	26.4	62.0
YOLOX-L	CSPNet	52.1	20.5	41.9	67.3	31.9	61.2
Co-DINO	ResNet-50	57.3	20.9	**53.2**	72.3	38.3	66.8
Co-DINO	Swin-L	**60.7**	**26.9**	52.7	**78.5**	**46.8**	**77.1**

4. 考察

3. では，Transformer-based モデルは，CNN-based モデルと比較して，検出性能が一回り高く，また，訓練では扱われない初見のノイズがかかった品質劣化データに対しても同様で，大幅な性能劣化が生じにくいことを示した．運転者視界に対する物体検出では，しばしば物体間の位置関係などのようなコンテキスト情報を必要とする場面があり，ノイズがあっても同様の判断によって物体検出が可能なことが多い．この点において，画像全体を考慮した大局的な特徴抽出が可能で，全検出物体間の関係性を捉えられる Self-Attention の効果により，高い検出性能とロバスト性を発揮できると考えられる．以上より，運用時のデータ品質が揺らぐ中で厳しい前提を達成する必要のある注視物体判定システムの物体検出パートにおいて，Transformer-based モデルの高い有効性を明らかにできた．

5. 結論

本研究では，注視物体判定システムの物体検出パートに課せられるタスクに則った独自データセット上で，物体検出モデルの性能を視界映像中に見られる各ノイズへのロバスト性も含めて検証した．結果として，Transformer-based モデルは先行研究の CNN-based モデルより遥かに性能が高く，当該の物体検出パートにおいて非常に有効であることを明らかにした．今後の課題として，有効性を示した Transformer-based モデルが注視物体判定システムの物体検出パートに適格かを検討するために，実運用の中でシステム全体としての性能を評価する必要がある．

参考文献

[1] 池田光汰ほか, ”Deep Learning を用いたドライバーの注視対象物の評価”, 情報処理学会第 81 回全国大会講演論文集, pp.217–218 (2019).

[2] 和田健太郎ほか, ”360 度カメラ映像を用いた注視物体の判定と評価”, 日本交通科学学会誌, Vol.23 (Suppl.), p.88 (2023).

[3] S. Ren, et al., ”Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, *Advances in Neural Information Processing Systems 28*, pp.91–99 (2015).

[4] Z. Ge, et al., ”YOLOX: Exceeding YOLO Series in 2021”, *arXiv:2107.08430* (2021).

[5] A. Vaswani, et al., ”Attention is All you Need”, *Advances in Neural Information Processing Systems 30*, pp.5999–6009 (2017).

[6] A. Dosovitskiy, et al., ”An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale”, *ICLR* (2021).

[7] Ze Liu, et al., ”Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows”, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.10012–10022 (2021).

[8] N. Carion, et al., ”End-to-End Object Detection with Transformers”, *Computer Vision - ECCV 2020*, pp.213–229 (2020).

[9] Z. Zong, et al., ”DETRs with Collaborative Hybrid Assignments Training”, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.6748–6758 (2023).

[10] H. Zhang, et al., ”DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection”, *ICLR* (2023).

[11] C. Michaelis, et al., ”Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming”, *Machine Learning for Autonomous Driving Workshop* (2019).