B-2-5

大規模鶏舎における FasterViT を用いた雛鶏検出

丹波 文冶（萩原研究室）

1　はじめに

近年，畜産業では自動化技術の発展が求められており，特に養鶏業界では雛鳥の健康状態のモニタリングが重要な課題となっている．従来は人間による目視確認が主流であったが，この方法は時間と労力を要するだけでなく，精度にも限界がある．この課題を解決するために，ディープラーニングを活用した物体検出技術が注目されているが，餌を摂取する多数の雛鳥が密集している環境では，雛鳥の分布が複雑であるため，物体検出には高度な特徴抽出が必要である．

従来のバックボーンでは，CNN[1] や Vision Transformer(ViT) [2] が広く用いられている．CNN は，局所的な特徴抽出に優れているが，大域的な特徴抽出が苦手である．これに対し ViT は画像を小さなパッチに分割し，それらの関係性を学習することで広範な特徴を捉えることが可能である．ただし，ViT は計算負荷が高く，リアルタイム処理には不向きである．

本研究では，CNN の局所的な特徴抽出能力と ViT の大域的な特徴学習能力を融合し，計算効率と精度の両方を向上させた FasterViT [3] をバックボーンとして用いた DETR with Improved deNoising anchOr boxes (DINO) [4] を活用した雛鳥の物体検出を行い，複雑な環境下でも高精度な検出を実現することを目指す．

2　システム概要

本システムは，FasterViT を用いて雛鳥の特徴量を抽出し，DINO を用いることで複雑な物体検出タスクでの安定した学習を実現させる．

2.1　FasterViT

FasterViT は，前半では CNN を用いて，局所的な特徴を効率的に抽出し，後半では，ViT を用いて，広範囲の特徴を捉えるアーキテクチャである．従来の ViT モデルでは，計算コストが高くなるという課題があった．FasterViT は画像を複数の小さな領域に分割し，各領域ごとに特徴を抽出した後，それらの情報を要約したトークンを使用して，離れた領域間の関連性を取り入れることにより，各領域の特徴を全体に取り入れる．これにより，計算コストを抑えつつ，離れたパッチ同士の関係性を正確に捉えることが可能となる．また，高解像度画像に対する処理効率が大幅に向上し，従来のモデルと比較して高速かつ高精度な推論を実現している．

2.2　DINO

DINO は，従来の DEtection TRansformer(DETR)[5] モデルを進化させた物体検出モデルであり，効率的なトレーニングを行うことで高精度な検出を実現している．特に小さなオブジェクトの検出に優れており，エンコーダの特徴マップから動的にクエリを初期化する「混合クエリ選択」を導入し，クエリの位置情報の精度を向上させている．トレーニング時にノイズ付きのバウンディングボックスを生成することで誤検出を抑え，学習の安定性を高めている．

2.3　データセットの作成

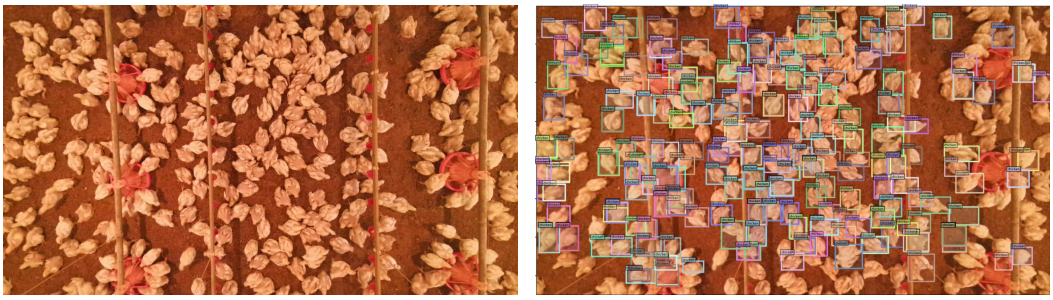
本研究では，雛鳥の物体検出を行うために，専用のデータセットを作成した．まず，鶏舎内で雛鳥が餌を摂取している様子を撮影した画像を収集する．次に，収集した画像中の雛鳥に対してアノテーションを実施する．アノテーション実施の際には，物体検出タスクで広く使用されている COCO[6] データセットにて用いられる形式を用いた．作成したデータセットは，学習用画像 5 枚，評価用画像 2 枚から構成されている．また，対応するアノテーションファイルをそれぞれ含む．

3　実験

3.1　実験概要

本実験では，FasterViT をバックボーンとする DINO モデルの有効性を検証するため，雛鳥の物体検出を行った．餌を摂取する雛鳥の画像を収集し，7 枚の画像を学習用と評価用に分割したものを対象のデータとする．

本実験では，学習の際に用いる雛鳥の画像データが少数である．よって，少量のデータでも高精度な学習を可能な事前学習済みモデルを用いる．本実験では，Pascal VOC[7] で学習を行ったモデルを利用する．Pascal VOC



(a) (b)

図 1: 物体検出出力結果の一例 (a) モデルへの入力画像，(b) 出力結果

は 20 クラスのオブジェクトが含まれたデータセットである．

モデルの性能を評価するために，mean Average Precision(mAP) を用いて検出精度を測定した．また，検出結果をバウンディングボックスで描画し，結果を可視化した．

3.2 実験結果

図 1 に物体検出の結果を示す．

図 1(b) から，一部の個体が検出されていないものの，多くの雛鳥の個体を適切に検出していることが確認できる．また，雛鳥が密集する環境下においても誤検出が少ないことが示された．本モデルの性能を評価した結果，mAP は 16.7%を達成しており，少ないデータセットにもかかわらず一定の精度を実現している．さらに，事前学習済みモデルを活用することでモデルの性能が大幅に向上することが明らかとなった．この結果から，FasterViT の広範囲な特徴抽出能力が，雛鳥の密集環境下における物体検出に有効であることが示唆された．

画像の外縁部では中心部と比較して検出精度が低下している傾向が見られた．カメラの画角による影響や，学習用データセット内の画像数不足が原因であると考える．また，mAP が 16.7%という結果は一定の精度を示しているものの，更なる改善の余地がある．COCO のように Pascal VOC よりも汎用性の高いデータセットを事前学習に利用することで，モデルの検出精度が向上する可能性が期待される．

4 まとめ

本研究では，FasterViT をバックボーンとする DINO モデルを用いて雛鳥の物体検出を行い，その有効性を検証した．実験結果では，学習に少数の画像データを用いた場合でも，事前学習済みモデルを用いた場合には一定の検出結果を得たことが確認できた．よって，本モデルが小規模データセットにおいても有効である可能性が示された．今後の課題は，学習に用いる画像の枚数を増やすことや，COCO などの大規模データセットを活用したさらなる追加学習を行うことで，モデルの精度向上を図る．また，他のモデルとの比較を通じて，本研究のモデルの優位性をより詳細に検証したいと考える．

参考文献

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol.86, no.11, pp.2278–2324, 1998.

[2] A. Dosovitskiy, “An image is worth 16x16 words: Transformers for image recognition at scale,” ICLR, 2021.

[3] A. Hatamizadeh, G. Heinrich, H. Yin, A. Tao, J.M. Alvarez, J. Kautz, and P. Molchanov, “Fastervit: Fast vision transformers with hierarchical attention,” ICLR, 2024.

[4] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L.M. Ni, and H.Y. Shum, “Dino: Detr with improved denoising anchor boxes for end-to-end object detection,” CoRR, abs/2203.03605, 2022.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” ECCV, pp.213–229, 2020.

[6] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, “Microsoft coco: Common objects in context,” ECCV, pp.740–755, 2014.

[7] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” IJCV, vol.88, pp.303–338, 2010.