時系列データ対応可能なオープンボキャブラリー物体検出システム

金沢工業大学
工学部　情報工学科
中沢研究室　4EP4-53 丸山洸青

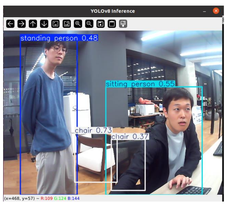
研究背景・課題点

研究背景
・物体検出は事前学習されたものによって検出できるものが制限されてしまう

・未知の物体　事前学習されていないものはほぼ検出することができない

OVD(Open-Vocabulary Object Detection)
・事前学習された物体クラスにある程度制限されるに使用者からの**テキスト入力から任意の物体を検出可能**
・従来の手法のゼロショット学習(BLCなど)と比較すると性能は**117%**上がっている

例　Yolo world[1)]



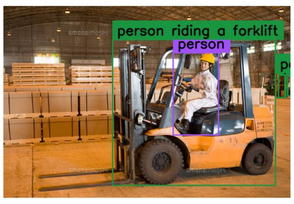


図１　実行例

model.set_classes(["standing person,person ridinng a forklift・・・etc"])

上の図のようにclassを設定することによって検出させたいクラスを検出可能、forkliftというクラスはないが検出ができている

課題点
動画・時系列データに対する検出の精度が低いことがあげられる
・フレームの前後の情報を認識できていないため扉から出てきている・歩いている人などのクラスを検出することができない

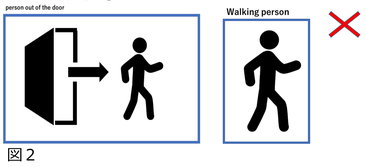


図２

提案手法

提案手法１
OVD系列のモデルで時系列に対応させるため動作の過程をすべて学習させる

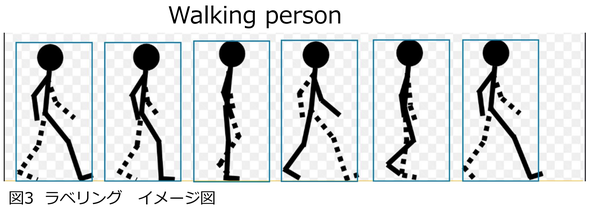


図3　ラベリング　イメージ図

上の図はWalking personを適応させるため**歩く過程の全て**（足を上げる前に出すなど）の動作を**ラベリング**し、学習させる

提案手法2

提案手法２
OVD系列のモデルを時系列に対応させるために**Timesformer**を利用する

Timesformer[2)]

動画を理解するため時空間関係を捉えられるようにしたTransformerモデルである５つの方式がある

Divided Space-Time Attention[3)]

Timesformerの中で最も**精度が高いとされる方式**

緑と青（縦）・・・時間
赤（画像）・・・空間

を意味しており、２つをアテンションを用いて動画を認識させる手法

図４Divided Space-Time Attention[3)]

OWL-Vit[4)]と**Timesformer**[2)]

OWL-Vitは**OVDを実装するための手法**の一つであり、**Transformer・Vit(Vision transformer)をbase**に作られているモデル

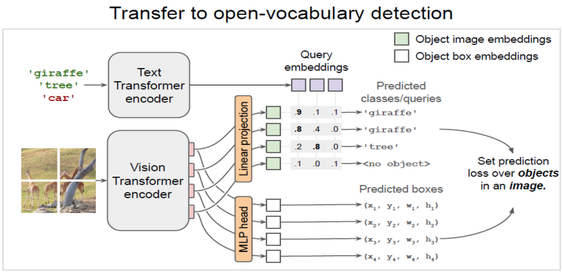


図5　OWL-Vitモデル図[4)]

Transformer・Vitそれぞれのencoderから得られた画像と物体テキストの埋め込み表現から類似度の高いクラスを出す、Vitから出されたbboxとクラスを適応させ、物体を検出する

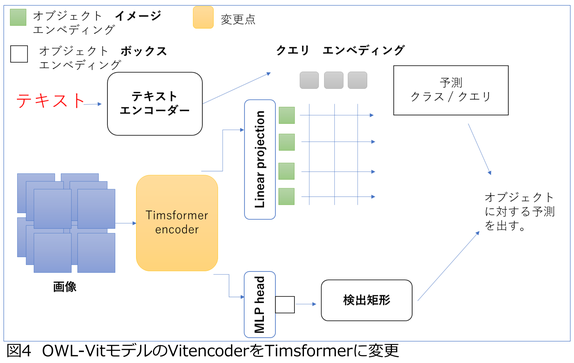


図4　OWL-VitモデルのVitencoderをTimsformerに変更

Vitのencoder部分をTimesformerに変更し、時系列に対応できるよう試みる

評価手法

・**IoUを用いて評価する**
・**IoUが0.６以上**

$$IoU = \frac{\text{二つの領域の積}}{\text{二つの領域の和}}$$

Iou 小
Iou 大

この２つの領域は
・１つ目は**予測された領域**(Predicted Bouding Box)
・２つ目は**正解の領域**(真値)(Ground-truth Bounding Box)

今後

・データセットを提供しているサイト(roboflowなど)から**のデータセット収集**

・現状できていない検出できていない行動(walking personなど) を認識できるようにするための**fine tuning**

・（モデル構造の理解・変更）

参考文献

1)Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, Ying Shan, **YOLO-World: Real-Time Open-Vocabulary Object Detection** ,https://arxiv.org/abs/2401.17270 ,2024/2/22　参照(2024/9/27)
2),3) Gedas Bertasius, Heng Wang, Lorenzo Torresani, **Is Space-Time Attention All You Need for VideUnderstanding?**, https://arxiv.org/abs/2102.05095,pp.4,2021/6/9 参照(2024/9/27)
4) Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, Neil Houlsby, **Simple Open-Vocabulary Object Detection with Vision Transformers**, https://arxiv.org/abs/2205.06230,pp.3,2022/7/22,参照(2024/9/27)