SeqCo-DETR: Sequence Consistency Training for Self-Supervised Object Detection with Transformers

Guoqiang Jin, Fan Yang, Mingshan Sun, Ruyi Zhao, Yakun Liu, Wei Li, Tianpeng Bao, Liwei Wu, Xingyu Zeng, Rui Zhao

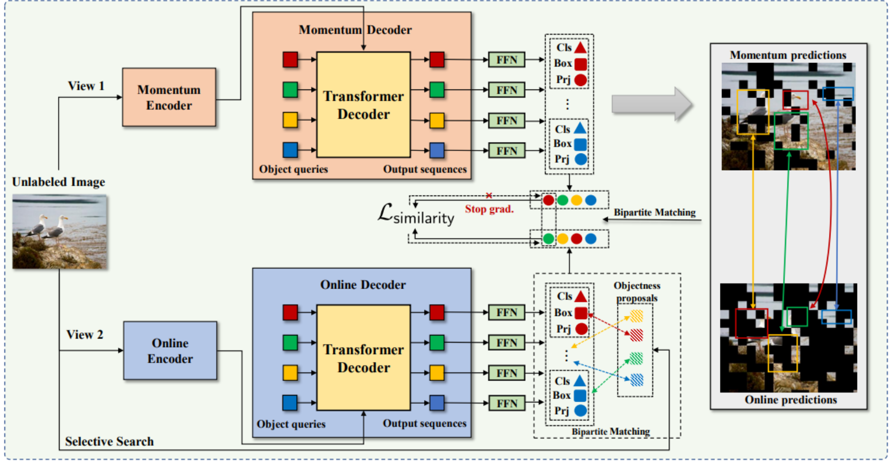
BMVC 2023

Motivation

- Transformer-based methods have introduced a novel paradigm of object detection tasks. However, prior pre-training approaches for transformer-based object detection have primarily relied on **unsupervised** methods, which would be limited by the hand-crafted pseudo labels.
- Most self-supervised methods are designed for image classification tasks and rely on **image-level** features. However, object detection requires **object-level** features.

Method

The sequence consistency strategy and the complementary mask strategy



Experiments

Comparison results. Pretrained on ImageNet, finetuned on COCO or VOC.

Model	COCO val2017			VOC test07		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Faster R-CNN [22]	42.0	62.1	45.5	56.1	82.6	62.7
Deformable DETR (Supervised CNN) [45]	43.8	62.6	47.7	59.5	82.6	65.6
Deformable DETR (SimCLR CNN)†	41.5	59.8	45.4	57.3	80.0	63.6
Deformable DETR (BYOL CNN)†	44.7	63.8	48.8	59.9	82.7	66.7
Deformable DETR (MoCo CNN)†	43.1	61.6	46.9	59.6	81.8	66.0
Deformable DETR (SwAV CNN)†	45.0	63.8	49.2	61.0	83.0	68.1
UP-DETR (Deformable DETR) ‡	44.7	63.7	48.6	61.8	83.4	69.6
JoinDet [33]	45.6	**64.3**	49.8	63.7	83.8	70.7
DETReg w/o feature embedding †	45.2	63.7	49.5	63.0	83.5	70.2
DETReg [1]	45.5	64.1	49.9	63.5	83.3	70.3
SeqCo-DETR	**45.8**	64.2	**50.0**	**64.1**	**83.8**	**71.6**

Mask strategy.

Model	Mask strategy	AP
DETReg	w/o Mask (baseline) [1]	45.4
	w/ Mask_{50} †	45.0
SeqCo-DETR	w/o Mask	45.6
	$\text{Mask}_{online@50}$	45.6
	$\text{Mask}_{online@50} + \text{Mask}_{momentum@50}$	45.4
	$\text{Mask}_{online@70} + \text{Mask}_{momentum@30}$	45.6
	$\text{Mask}_{online@70} + \text{Mask}_{\neg(online@70)}$	**45.8**

Pre-training datasets and region proposal strategy.

Method	IN100	IN100 (Rnd bbox)	COCO	COCO+	COCO GT
DETReg †	45.4	44.1	45.1	45.1	45.6
SeqCo-DETR	**45.8**	**44.3**	**45.6**	**45.6**	**45.8**

Sequence utilization methods.

Model	One-by-one matching	Bipartite matching	Multi-feature	AP
SeqCo-DETR	✓			45.6
	✓		✓	45.3
		✓	✓	45.5
		✓		**45.8**

Conclusions

We introduce SeqCo-DETR, **a novel self-supervised learning method for object detection based on transformers.**

1. We exploit the **sequential nature of transformer** networks to achieve self-supervised learning for object detection, maintaining **sequence consistency** under different image views.
2. We propose a **complementary mask strategy** incorporated with the sequence consistency strategy to extract more global context information for object detection.
3. We adopt bipartite matching to optimize **sequence-level** self-supervision.
4. Extensive experiments on both single-object and multi-object detection datasets demonstrate the effectiveness, resulting in state-of-the-art performance on MS COCO (45.8 AP) and PASCAL VOC (64.1 AP).