End-to-End Object Detection with Transformers

Nicolas Carion*, Francisco Massa*, Gabriel Synnaeve,
Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko

ECCV 2020

Presented by Xinyu Liu, Liujie Zheng, and Tarik Reza Toha

Arguments

Significant Research Impact

- DETR has a simple pipeline having a ResNet and a transformer with just a few hundred lines
  - Hence, many people from academia and industry are using this DETR in practice.

TITLE	CITED BY	YEAR
End-to-end object detection with transformers N Carion, F Massa, G Synnaeve, N Usunier, A Kirillov, S Zagoruyko European conference on computer vision, 213-229	9754	2020
Pix2seq: A language modeling framework for object detection T Chen, S Saxena, L Li, DJ Fleet, G Hinton International Conference on Learning Representations (ICLR)	219	2022

Technical Novelties

- DETR solves the object detection problem by introducing a new set-based bipartite matching loss function
  - In contrast, Pix2Seq solves the same problem by using some existing techniques
- DETR can be easily extended to produce panoptic segmentation. The results are shown in their paper.
  - In contrast, the authors of Pix2Seq claimed to have some sorts of extensibility. However, they did not provide any empirical evidences of such extensions.

Empirical Evidences

- Pix2Seq does not significantly outperform the DETR in terms of average precision, i.e, it has only 0.2% improvement (after two years)
- In addition, the efficiency of Pix2Seq is not evaluated against DETR in terms of a tangible metrics such as frames per second

Method	Backbone	#params	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DETR	R101-DC5	60M	**44.9**	64.7	47.7	23.7	49.5	62.3
Pix2seq (Ours)	R101-DC5	57M	**45.0**	63.2	48.6	28.2	48.9	60.4

Pix2Seq:

A LANGUAGE MODELING FRAMEWORK FOR OBJECT DETECTION

Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, Geoffrey Hinton

Presented by: Dohhyun, Nathan, Rodrigo

Argument #1: Technical Contributions

- Proposes a novel idea to reformulate CV tasks to seq2seq
- Proves the potential of language models in CV
- Introduces new techniques to combat overfitting
  - Sequence augmentation,
- All-purpose model for CV tasks

Argument #2: Training Advantage

Backbone	# params	Image size during finetuning		
		640×640	1024×1024	1333×1333
R50	37M	39.1	41.7	42.6
R50-C4	85M	44.7	46.9	47.3
ViT-B	115M	44.2	46.5	47.1
ViT-L	341M	47.6	49.0	50.0

- Outperforms DETR/state of the art models on smaller objects
- Remains competitive in medium and larger objects
- Meanwhile, in the DETR paper:
  - “This modification increases the resolution by a factor of two, thus improving performance for small objects, at the cost of a 16x higher cost in the self-attentions of the encoder, leading to an overall 2x increase in computational cost”
- Dynamic allocation and computing
- Even more room for improvement with pre-training and fine tuning

Image Size	Components	GFLOPs(Pix2Seq)	GFLOPs(DETR)
640x640	ResNet	36.12	36.9
1333x1333	ResNet	166.2	167.8

Argument #2: Figures

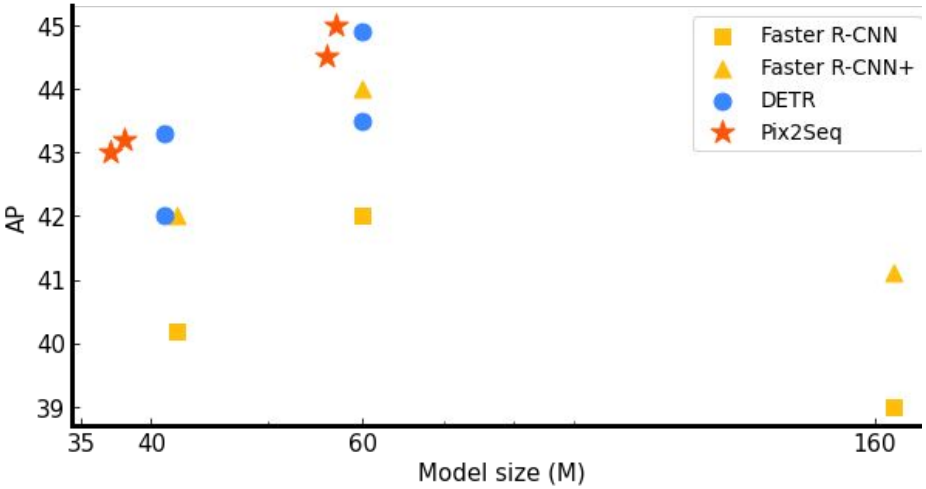

Table 1: Comparison of average precision, over multiple thresholds and object sizes, on COCO validation set. Each section compares different methods of the similar ResNet "backbone". Our models achieve competitive results to both Faster R-CNN and DETR baselines.

Method	Backbone	#params	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN	R50-FPN	42M	40.2	61.0	43.8	24.2	43.5	52.0
Faster R-CNN+	R50-FPN	42M	42.0	62.1	45.5	26.6	45.4	53.4
DETR	R50	41M	42.0	62.4	44.2	20.5	45.8	61.1
Pix2seq (Ours)	R50	37M	**43.0**	61.0	45.6	25.1	46.9	59.4
Faster R-CNN	R101-FPN	60M	42.0	62.5	45.9	25.2	45.6	54.6
Faster R-CNN+	R101-FPN	60M	44.0	63.9	47.8	27.2	48.1	56.0
DETR	R101	60M	43.5	63.8	46.4	21.9	48.0	61.8
Pix2seq (Ours)	R101	56M	**44.5**	62.8	47.5	26.0	48.2	60.3
Faster R-CNN	R50-DC5	166M	39.0	60.5	42.3	21.4	43.5	52.5
Faster R-CNN+	R50-DC5	166M	41.1	61.4	44.3	22.9	45.9	55.0
DETR	R50-DC5	41M	**43.3**	63.1	45.9	22.5	47.3	61.1
Pix2seq (Ours)	R50-DC5	38M	**43.2**	61.0	46.1	26.6	47.0	58.6
DETR	R101-DC5	60M	**44.9**	64.7	47.7	23.7	49.5	62.3
Pix2seq (Ours)	R101-DC5	57M	**45.0**	63.2	48.6	28.2	48.9	60.4

"Pix2Seq achieves competitive AP results compared to existing systems that require specialization during model design, while being significantly simpler."
(Google Research Blog, Pix2Seq: A New Language Interface for Object Detection)

Argument #2: Example

(a) Failure case with overlapping objects. PanopticFPN misses one plane entirely, while DETR fails to accurately segment 3 of them.

(b) `Things` masks are predicted at full resolution, which allows sharper boundaries than PanopticFPN

Pix2Seq outperforms DETR at labeling in complex and densely populated scenes with overlaps

Figure 14: Examples of the model's predictions (at the score threshold of 0.5). Original images accessed by clicking the images in supported PDF readers.

Argument #3: Simplicity

Pix2Seq

- Task agnostic
- Lighter model
- All purpose model for CV different tasks.

VS

DETR

- Highly customized for task
- Extra-long training schedule
- Additional loss functions
- Designed for object detection only

Compared to the DETR model, the Pix2Seq framework is:

- easier to reproduce
- more versatile
- more accessible