Supplementary Materials: Recurrent Glimpse-based Decoder for Detection with Transformer

Zhe Chen[1] Jing Zhang[1] Dacheng Tao[2,1]
[1] The University of Sydney, Australia [2] JD Explore Academy, China
{zhe.chen1,jing.zhang1}@sydney.edu.au; dacheng.tao@gmail.com

1. Introduction

In the supplementary materials, we first present the detailed algorithmic description of the REGO for DETR. We also include more implementation details. We then discuss how the REGO can improve DETR for free during inference. Lastly, we present some additional qualitative results about how our REGO module can help reveal the object relations that are useful for detection.

2. More Descriptions

Processing Algorithm In general, the REGO follows the algorithm described in Alg. 1 to process the visual features within each stage. We will release the code shortly.

Algorithm 1 Processing of i-th REGO Stage

Require: $H_{dec}(i-1)$, $O_{box}(i-1)$
Ensure: $H_{dec}(i)$, $O_{box}(i)$, and $O_{cls}(i)$
1. Calculate RoIs by enlarging bounding box areas of $O_{box}(i-1)$ according to a scale $\alpha(i)$
2. Extract glimpse features $V(i)$ based on enlarged RoIs;
3. Perform multi-head attention on glimpse features $V(i)$ and $H_{dec}(i-1)$ to obtain decoded features $H_g(i)$;
4. Concatenate $H_g(i)$ and $H_{dec}(i-1)$ to obtain refined attention modeling outputs $H_{dec}(i)$ of current stage;
5. Predict object bounding boxes $O_{box}(i)$ and corresponding labels $O_{cls}(i)$ using $H_{dec}(i)$;

More Implementation Details In addition to the details discussed in the paper, we would also like to mention the following aspects of implementation. In particular, regarding the extraction of multi-scale features, this can be easy for applying REGO on DETR methods like Deformable DETR [3] that already extract multi-scale features for attention modeling. When applying the REGO on DETR methods like the original DETR [1] that only extract the single-scale feature from the last convolutional stage of the backbone, we attach 1×1 convolutions on the output of different convolutional stages (stage level 2 to stage level 5) to obtain multi-scale features. In this case, the attached 1×1 convolutions reduce the channel numbers to 256. Note that we do not use FPN for extracting multi-scale features to save costs. Besides, for the 'DC5' DETR variants which use single-scale features but enlarge the scale of the last convolutional stage, we still attach 1×1 convolutions to extract features and reduce channel numbers, except that the features of the last convolutional stage is down-sampled to its normal scale (*i.e.*, 1/32 of the input image) to save costs.

In addition, in the REGO, we use different weight parameters to initialize the decoders of different stages, thus the decoder at each stage can be trained to be more sensitive to the glimpse features of the corresponding stage. At each stage, to stabilize training, we follow [2] and do not back propagate gradients into the outputs of the previous stage.

Inference Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Deformable DETR [3]	43.8	62.6	47.7	26.4	47.1	58.0
Inference w/ REGO	45.9	65.2	49.7	27.6	48.9	61.5
Inference w/o REGO	44.9	65.0	48.7	26.5	48.4	61.1

Table 1. Performance comparison of whether using REGO for inference. Deformable DETR [3] is used as baseline. The models related to the performance of inference with or without REGO all use REGO for training.

3. Improve DETR For Free

As mentioned in our paper, the proposed REGO method can improve the DETR for free during inference. This means that, after training with REGO, the obtained DETR model can still achieve improved performance by removing the REGO from the detection pipeline during inference. Table 1 shows the results evaluated on MS COCO *val* set. We use the Deformable DETR [3] as our baseline DETR method.

From these results, we can find that the complete REGO method (training and inference) improves around 2 points in AP, and the Deformable DETR trained with REGO but tested without REGO still achieves around 1 point gain in AP comparing to the baseline method. This demonstrates

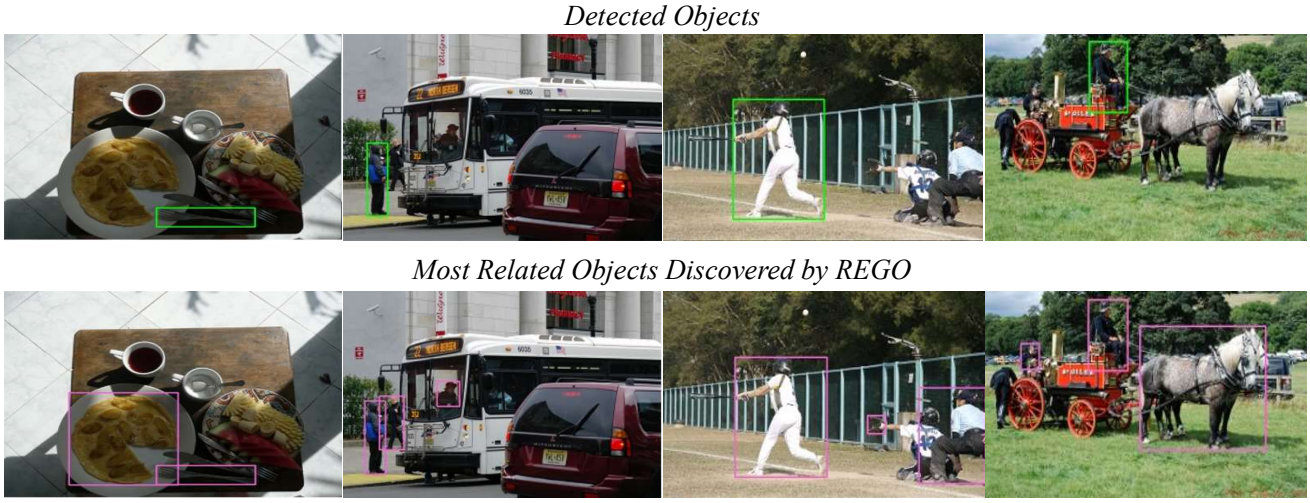

Figure 1. Illustration of using the REGO to reveal object relations. Each figure of the top row shows one of the detected objects of the previous stage, and each figure of the bottom row shows the most related detection results of the previous stage discovered by the decoder of the REGO. Results are obtained from the last stage of the REGO.

that the REGO is effective to enhance the feature of the original DETR method based on the RoI-based refinement procedure, which also suggest that the proposed REGO method can indeed improve the attention modeling in DETR during optimization. By removing the REGO, the obtained 1 point improvement does not introduce any extra complexity for inference comparing to the baseline DETR method.

4. Learned Object Relations

As described in the paper, the decoder of the REGO in each stage correlates the glimpse features extracted based on previously detected bounding boxes with the previous attention modeling outputs corresponding to the same set of detected bounding boxes. Therefore, the obtained correlation results can reveal the relations between any two detected bounding boxes of the previous stage, and the detection results with higher correlation weights can be considered as more important for refining the attention modeling outputs and detection results.

In Fig. 1, we show some examples of the objects with the highest correlation weights *w.r.t.* a detected object from the previous stage. From the presented figures, we can find that the most related objects discovered by the REGO generally have strong semantic connections. For example, in the first column, the 'fork' and 'pizza' are most correlated for refining the detection related to the 'fork' object; in the fourth column, the 'person' and the 'horses' are most correlated for refining the 'person' object who is driving the carriage. Both examples are intuitive to human as well. This illustrates that the REGO can help DETR explore the information from semantically meaningful areas without wasting attention on obviously irrelevant areas, which is beneficial for improving the training efficiency and effectiveness of attention modeling in DETR.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1

[2] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 1

[3] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1