

End-to-End Object Detection with Adaptive Clustering Transformer

BMVC 2021 Submission # 709

A Supplementary Experiments

DETR with the ResNet-101 backbone can achieve higher performance while the FLOPs will increase accordingly. We add experiments using the ResNet-101 backbone and adding a dilation to the last stage of the backbone. Other network structures and hyper-parameters remain unchanged. In Figure 1, we compare the AP of ACT with DETR-DC5 and the K-mean clustering. We refer to hash rounds in our model as L and refer to the number of clusters in K-means as C . In Table 1, we also compare the AP, AP_L , AP_M , and AP_S of our model with DETR-DC5 in detail.

We can draw conclusions similar to using resnet50 as the backbone network. We reduce the FLOPs of DETR while the loss in AP is only 0.5%. And comparing with k-means clustering, our adaptive clustering can better solve this task, and we have obtained higher accuracy with lower FLOPs. Most of the loss in AP comes from small and medium-sized objects. We guess this is due to the shift of the bounding box predicted by ACT, which is discussed further in Section C.

This experiment further proves the generalization of our method and switching the backbone network will not make ACT invalid. Therefore, we are confident to believe that applying our method to other attention models in computer vision can also achieve ideal results.

B Visualisation of Adaptive Clustering

To analyze which queries are represented by the same prototype, we visualize some representative clusters in Figure 2. We can easily find that the three clusters displayed are the features of the cow, the sky, and the field. This indicates that our clustering is related to semantics and location. Those queries with similar semantics and similar locations will easily be grouped.

C Comparison of the Prediction

In order to analyze the impact of our adaptive clustering attention on the final results, we compare the predictions of ACT and DETR. We found that our method has very little impact on the label prediction, but it will cause the bounding boxes to shift, disappear, or repeat.

Model	GFLOPs	AP	AP _L	AP _M	AP _S
Backbone (ResNet-101)	192.3				
DETR-DC5	+58.1	44.9	62.3	49.5	23.7
ACT (L=32)	+43.1	44.4	62.1	49.2	23.0
ACT (L=24)	+37.8	43.4	62.1	48.3	21.0
ACT (L=20)	+34.1	42.4	61.7	47.4	19.9
ACT (L=16)	+29.6	40.0	60.8	45.3	16.5

Table 1: We compare the AP of our model with DETR-DC5. All the models use dilated ResNet-101 as the backbone. We refer to the bbox AP of large, medium and large size instance as AP_L, AP_M and AP_S respectively.

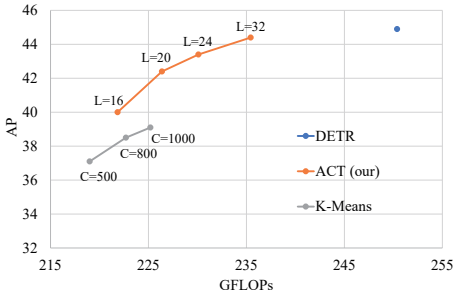


Figure 1: We compare the AP of ACT with DETR-DC5 and the K-mean clustering. All the models use dilated ResNet-101 as the backbone. We refer to hash rounds in our model as L and refer to the number of clusters in K-means as C.

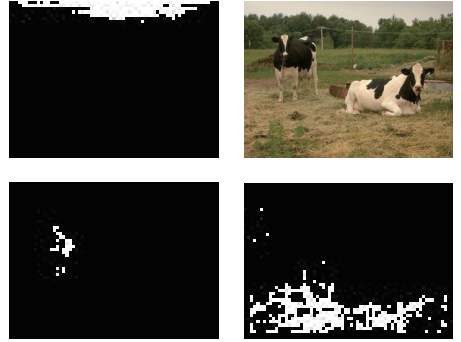


Figure 2: We visualize some representative clusters in the encoder. The queries where the white pixel is located belong to the same cluster.

The most common impact is the shift of the bounding box, which is observed in most of the predictions. However, the shift of the large size bounding box has little effect on the final prediction result, which explains why the AP loss of ACT mainly occurs on small and medium-sized objects. The red arrow in Figure 4 marks some shifted bounding box. The shift of the bounding box in Figure 4(a) is much more serious than others. But due to its large size, most of the shifted bounding box still overlaps the target object. However, the small size bounding box of the tie (in Figure 4(b)) and the person (in Figure 4(c)) has deviated from the target object. Our MTKD transfer the knowledge of the bounding box regression branch, so it can obtain a significant improvement in AP.

DETR sometimes produces multiple overlapping bounding boxes for some objects, so does the ACT. This is because of the wrong prediction of “no objects”. ACT can sometimes correct the misjudgment of DETR, while sometimes it will produce new misjudgments. For example, in Figure 3(a) there are multiple bounding boxes for the toilet both in ACT and DETR. In Figure 3(b), ACT eliminates the redundant bounding box of the person in DETR, while in Figure 3(c), ACT creates a new redundant bounding box for the train.

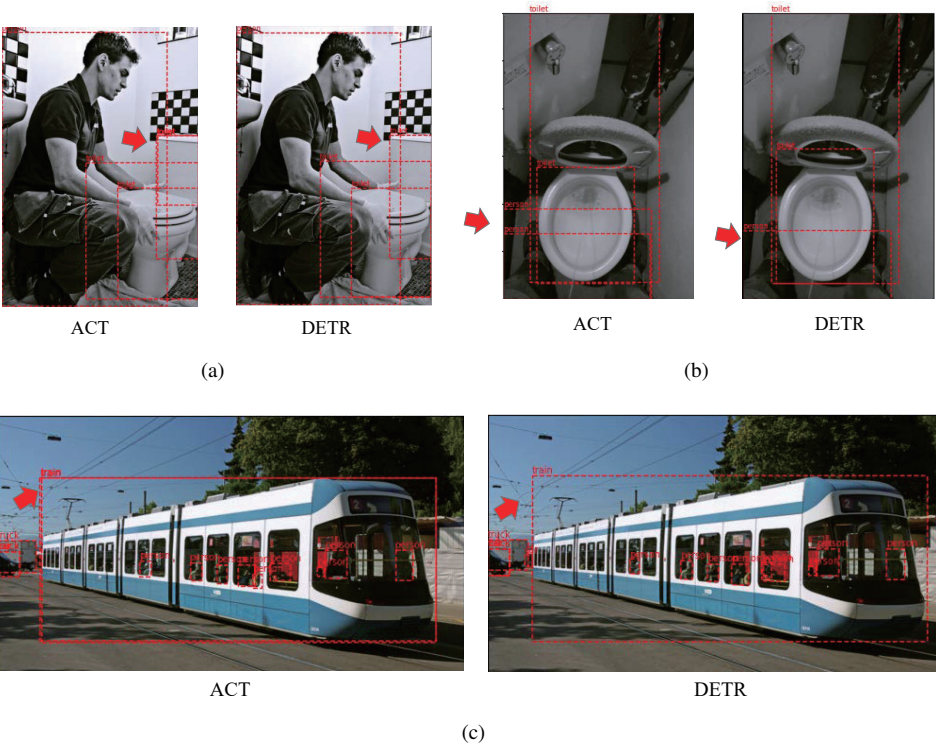
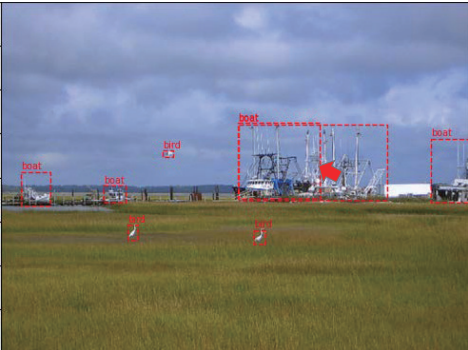
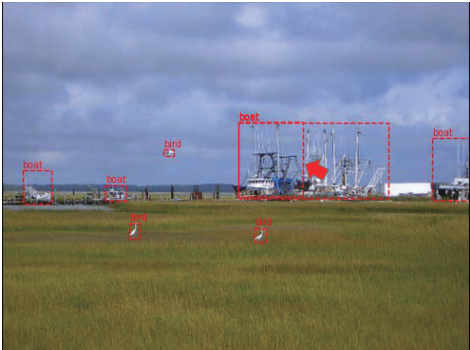


Figure 3: DETR and ACT sometimes produce multiple overlapping bounding boxes for some objects. The red arrow marks the redundant bounding boxes.



ACT



DETR

(a)

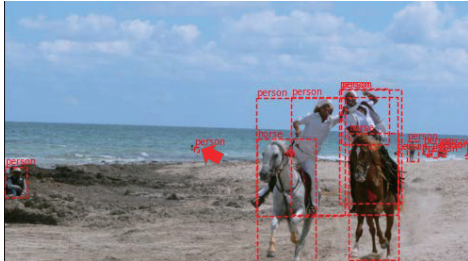


ACT

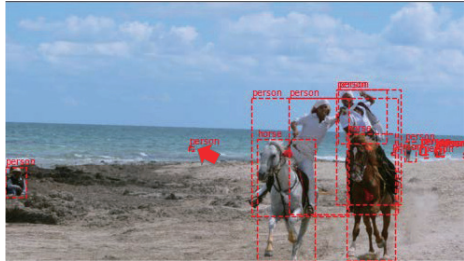


DETR

(b)



ACT



DETR

(c)

Figure 4: ACT sometimes causes the bounding box to shift. The red arrow marks the shifted bounding boxes.