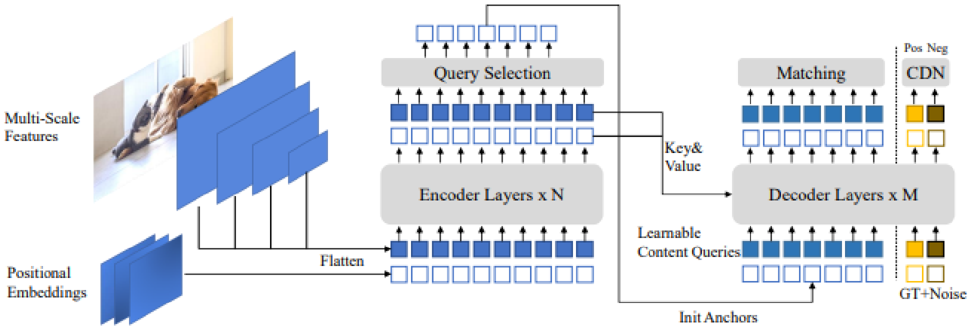
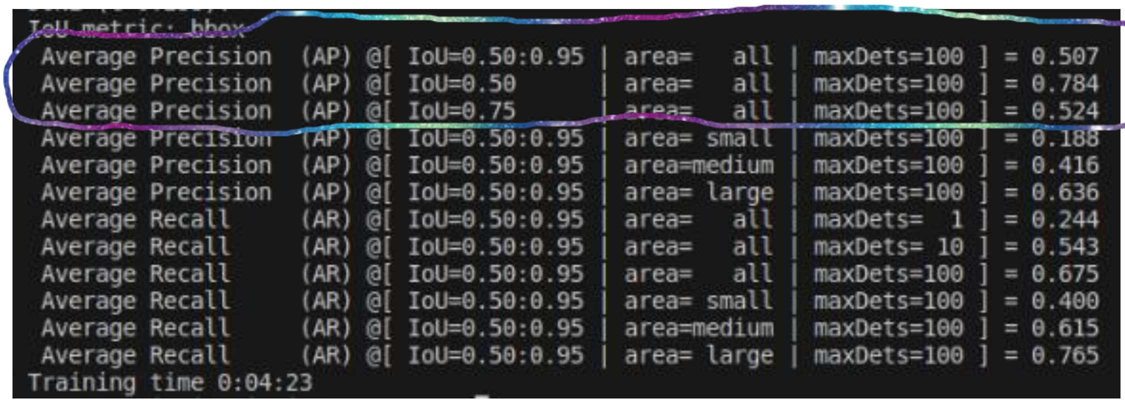
DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection

1. Architecture



首先給定一張圖像，使用 backebone 為例如 ResNet 或 Swin Transformer，從圖像中提取多尺度的特徵。這些多尺度特徵經過提取後，被丟入一個 Transformer 編碼器，同時伴隨相對應的位置嵌入。特徵通過編碼器層進行增強處理，作者提出了一種新的" mixed query selection strategy"，用來初始化解碼器的位置查詢，這些位置查詢也被稱為 "anchors"，有了初始化的位置查詢和可學習的內容查詢，作者使用"deformable attention"來結合編碼器輸出的特徵，並逐層更新查詢，最終輸出由經過改進的 anchor boxes 和內容特徵預測的分類結果組成。

2. AP



```
IoU metric: bbox
 Average Precision  (AP) @[ IoU=0.50:0.95 | area=   all | maxDets=100 ] = 0.507
 Average Precision  (AP) @[ IoU=0.50      | area=   all | maxDets=100 ] = 0.784
 Average Precision  (AP) @[ IoU=0.75      | area=   all | maxDets=100 ] = 0.524
 Average Precision  (AP) @[ IoU=0.50:0.95 | area= small | maxDets=100 ] = 0.188
 Average Precision  (AP) @[ IoU=0.50:0.95 | area=medium | maxDets=100 ] = 0.416
 Average Precision  (AP) @[ IoU=0.50:0.95 | area= large | maxDets=100 ] = 0.636
 Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets=  1 ] = 0.244
 Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets= 10 ] = 0.543
 Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets=100 ] = 0.675
 Average Recall     (AR) @[ IoU=0.50:0.95 | area= small | maxDets=100 ] = 0.400
 Average Recall     (AR) @[ IoU=0.50:0.95 | area=medium | maxDets=100 ] = 0.615
 Average Recall     (AR) @[ IoU=0.50:0.95 | area= large | maxDets=100 ] = 0.765
Training time 0:04:23
```

3. Code

num_work=2

num_class=8

epoch=12

將 models/dino.py 中的 717 行修改為

```
match_unstable_error=args.match_unstable_error

dn_labelbook_size = args.dn_labelbook_size

if dn_labelbook_size < num_classes:

    dn_labelbook_size = num_classes
```

一個視覺化的 testvisual. py

一個產生 output. json 的 testwrite. py

Visualization