Overview

Modern object detectors like Faster RCNN require many post-processing steps like
1.NMS
2.Anchor Generation
Disadvantages of these methods are

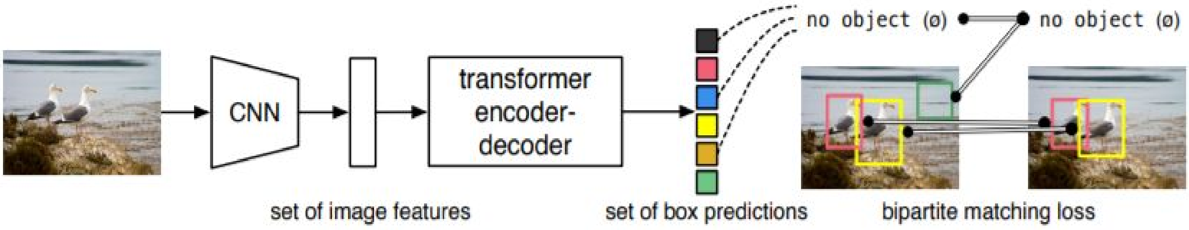
- These have to be hand-designed
- The final prediction depends heavily on these post-processing steps
- It makes the object detection pipeline complex

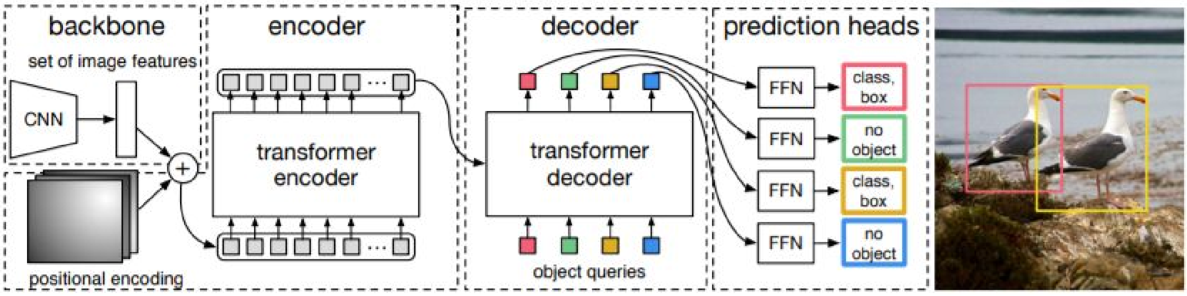
This paper converts the problem to a direct-set based prediction.
The main features of the paper are a

1. Set based global loss (forces unique prediction via bipartite matching)
2. Transformer encoder-decoder architecture

DETR simplifies the pipeline by dropping these post-processing steps and predicts all objects at once.

Architecture





(3xHxW) -> (Cxhxw) -> (dxhxw) -> (dx(h*w)) -> (dx(h*w)) -> (Nxd) -> (N tuples)

Set of box Predictions

- The model always outputs a fixed number of predictions for any given image (default 100) where each prediction is a set of **(class , BB)**.
- Assume N is the fixed number of prediction outputs of the model, i.e the maximum number of objects in an image.
- We even have a no-object class (background class)
- Since the number of sets in ground truth will be <N, we pad the sets with no object class and arbitrary BB. So we have an equal number of predictions and ground truth sets.

Loss Function (Hungarian)

- The target for loss function is to be high when there is a high difference between the two set inputs. It uses cross-entropy loss for class and a L1 + GIOU loss for BB.
- If the ground truth set has no-object class paired to prediction which predicts a class then only the cross entropy loss contributes, not the BB loss.
- To handle class imbalance, we scale the no-object class loss by a factor of 10.
- The training loss is the best bipartite match (one-to-one assignment) of these N sets from ground truth to prediction, such that the total **Hungarian Loss is minimized.**
- This optimum bipartite match can be found by the **Hungarian algorithm.**

Ex -

(c,b)	(phi,b)
(c,b)	(phi,b)
(c,b)	(phi,b)
(c,b)	(c,b)
(c,b)	(c,b)

Since the best match is found, the loss function is invariant by a permutation of predictions(i.e the order of predictions does not matter).

It encourages the model to give different predictions, and not predict the same object twice or else it will be penalized with the loss function.

Thus this loss function overcomes the problem of near-duplicates without any post-processing step used in modern detectors.