2020-06-12 7:27:55 AM

Main idea is to convert object detection into a set prediction problem using bipartite matching loss to enforce unique matches between the ground truth and predictions

A fixed number N of objects is predicted where N is apparently much greater than the maximum number of objects that can be expected in an image so the problem of the sparsity of the loss is not entirely solved

A special class “no object” is used for predictions corresponding to the background and a down–weighting factor of 10 is used to balance out the large abundance of such objects similar to the other OHEM style tricks used in existing detectors

A backbone CNN – resnet 50 or 101 – is first used to extract features of size d x H x W which are then spatially collapsed to form d x HW output that is used as input to the transformer component which expects a sequence and this is apparently seen as N vectors of size d from which it seems that N = HW though this apparently not actually confirmed anywhere in the paper but would mean that N >> Number of actual objects

A 3 layer feedforward / perceptron network takes the output of the transformer decoder and converts it into a class probability and bounding box coordinates of the N predicted objects

A crucial component of the representation seems to be the positional encodings that are added on to the output of the backbone before going into the transformer but not well explained

> these are apparently needed to solve the problem of absolute box prediction as compared to the anchor/grid relative predictions of other detectors which are much easier