- The paper streamlines the detection pipeline, effectively removing the need for many hand-designed components like a non-maximum suppression procedure or anchor generation that explicitly encode our prior knowledge about the task.
- Modern detectors use indirect ways, by defining surrogate regression and classification problems on a large set of proposals, anchors, or window centers.
- Self-attention layers model all pairwise interactions of elements of a sequence and help in removing duplicates etc. DETR parallely outputs all BBs. The bipartite matching loss assigns BB to each GT and is invariant to permutation of predicted objects which makes parallel decoding possible (non-autoregressive).
- Set prediction : One-vs-rest does not apply as there is an underlying structure between elements. To overcome near-duplicates, detectors use post-processing like non-max suppression. Set prediction needs to model interactions between all predicted elements to avoid redundancy. Hungarian algo is used to find bipartite matching.
- Most detection methods make predictions relative to some initial guesses on which performance depends. DETR directly predicts absolute box prediction w.r.t. the input image.
    - Set-based loss : Learnable NMS methods and relation networks explicitly model relations between different predictions with attention. Post-processing can be eliminated using direct set loss. But they use hand-crafted context features.
- Bipartite matching is done by minimizing cost for N predicted objects by changing their permutation. The matching cost takes into account both the class prediction and the similarity of predicted and ground truth boxes.
- Down-weight the log-probability term when $c_i = \varnothing$ by a factor 10 to account for class imbalance. In matching loss, -ve probs are used (not -ve log prob) as they did better. For BB Linear combi of IoU loss and L1 loss was used as it is scale-invariant.
- Arch. consists of CNN backbone, Transformer and then Feed forward network. CNN produces feature map whose no. of channels are reduced by 1x1 convolutions, and then its spatial dimensions are broken down before passing to Transformer encoder with fixed positional encoding as it is permutation invariant.
- Transformer decoder takes N learned positional encodings (object queries) as input which are added to input of each attention layer. It also uses encoder output to produce N output embeddings which are decoded separately into BBs by FFN.
- FFN outputs normalized center coordinates, height and width of the box w.r.t. the input image, and the linear layer predicts the class label using a softmax function. Auxiliary loss was found beneficial for decoder in predicting correct no. of objects for each class. Prediction FFNs and Hungarian loss are added at each decoder layer.
- DETR was trained on COCO and panoptic segmentation datasets. Eval metric was AP. Optimizer AdamW, Xavier initialisation. Backbone is Resnet (50/101) with frozen batchnorm layers. Higher feature resolutions were also used by dilations which doubled map size.It improved performance for small objects with 2x increase in overall cost.
- Scale augmentation was used.

Doubts : Linear projection layer for prediction;