- Each grid predicts B bounding boxes with x,y,bx,by, and confidence score = P(obj)*IoU(truth,pred) and probabilities of different classes.
- Final output after conv layers and FCs is S*S*(5*B+C).
- Input resolution of image is doubled for fine details.
- L2 loss is used, as many grid cells don't have objects => high loss leads to low confidence scores fast => gradient flow cut down => Better to give less weightage to confidence score loss of boxes with no object.
- Big and small boxes are treated the same, but we want small box boundaries to be more accurate => we predict sqrt of bx and by.
- For each object, box with highest IoU with ground truth is predicted and only this cell is considered for loss.
- DPM(deformable parts model): It uses separate networks for each task like extracting features,predicting bounding boxes etc.
- Yolo faces more difficulty in localisation task maybe due to non-generalised shapes.
- Fast RCNN misses out many objects and classifies them as background maybe due to problem in RoI's proposals.
- Limitations:
    - As it has limited boxes for each cell, it may miss out small nearby objects.
    - It cannot generalise the size of bounding boxes.
    - It compromises quality as it uses multiple downsampling layers.
    - It treats errors in smaller and larger objects' boundaries.