YOLO Architecture

Input 448×448×3 (RGB Image)

Backbone (Feature Extractor) 24 Convolutional Layers

Extract visual features: edges → textures → shapes → parts → objects

Detection Head 2 Fully Connected Layers

4096 → 7×7×30

Output Tensor 7 × 7 × 30

Decode

decode

2 Boxes per cell [x, y, w, h, conf] × 2 = 10 values

20 Class probs P(dog), P(cat), ... = 20 values

Total: 30 per cell 7×7×30 = 1470