PRE-REPORT: You Only Look Once: Unified, Real-Time Object Detection

SongSeungWu

1 ABSTRACT

YOLO treats object detection as a regression problem, predicting bounding boxes and class probabilities with a single neural network. Its unified architecture is optimized for real-time performance. YOLO processes images at 45 FPS, while Fast YOLO achieves 155 FPS. YOLO makes fewer false positives in the background and learns more generalized object representations compared to other methods.

2 INTRODUCTION

Like the human visual system, YOLO detects objects in a single glance. Unlike traditional methods that use complex pipelines, YOLO uses a simpler, faster process with a single neural network.

- Speed: YOLO processes images at 45 FPS on a GPU, and Fast YOLO reaches over 150 FPS, making it ideal for real-time streaming.
- Global Reasoning: YOLO reduces background errors by considering the whole image during training and testing, with half the background mistakes of Fast R-CNN.
- Generalization: YOLO performs well in new domains, such as artwork, and handles unexpected inputs better.

3 UNIFIED DETECTION

YOLO unifies object detection components into a single neural network.

- Grid-based Detection: The input image is divided into an $S \times S$ grid, where each grid cell is responsible for detecting objects within it.
- Bounding Box Prediction: Each grid cell predicts B bounding boxes and confidence scores.
- Class Probabilities: Each grid cell predicts conditional class probabilities, applied only if an object is present.
- Final Prediction: Class probabilities and confidence scores are multiplied to generate final class-specific confidence scores.

4 NETWORK DESIGN

YOLO is implemented as a CNN and evaluated on the PASCAL VOC dataset.

- Architecture: Inspired by GoogLeNet, YOLO has 24 convolutional layers followed by 2 fully connected layers.
- Fast YOLO: A smaller version with fewer convolutional layers, producing a $7 \times 7 \times 30$ tensor as output.

5 TRAINING

YOLO pretrains on the ImageNet 1000-class dataset.

- Output: The final layer predicts class probabilities and bounding box coordinates, with width and height normalized to fall between 0 and 1.
- Training Setup: Trained on the PASCAL VOC dataset for 135 epochs with a learning rate schedule.
- Regularization and Data Augmentation: Dropout and random scaling, translation, and color adjustments prevent overfitting.

6 LIMITATIONS OF YOLO

YOLO's spatial constraints limit its ability to predict nearby objects and struggle with small objects in groups.

- Generalization Issues: YOLO struggles to generalize to new aspect ratios and configurations due to coarse features from downsampling layers.
- Loss Function: Errors in small boxes have a larger impact on IOU. Incorrect localizations are the main source of errors.

7 COMPARISON TO OTHER DETECTION SYSTEMS

- DPM: Uses sliding windows and static features, but YOLO replaces this with a single neural network for faster and more accurate processing.
- R-CNN: Generates region proposals with Selective Search and processes in multiple stages. YOLO uses grid cells to predict 98 boxes, reducing duplicate detections.
- Fast/Faster R-CNN: Improves speed and accuracy but still lacks real-time performance. YOLO is inherently fast without a complex pipeline.
- Single-Class Detectors: Optimized for specific objects like faces, but YOLO handles multiple objects and classes simultaneously.
- Deep MultiBox: Predicts regions without Selective Search, but unlike YOLO, it's not a complete detection system.
- OverFeat: Uses sliding windows but lacks global context reasoning, requiring post-processing. YOLO integrates global reasoning.
- MultiGrasp: Predicts a single graspable region, whereas YOLO predicts bounding boxes and class probabilities for multiple objects.

8 EXPERIMENTS

- YOLO is the fastest detection system on PASCAL VOC 2007. Fast YOLO achieves 52.7% mAP, more than twice as accurate as previous real-time systems.
- YOLO struggles with localization errors but makes fewer background mistakes than Fast R-CNN. Combining YOLO with Fast R-CNN improves performance, raising mAP from 71.8% to 75.0%.
- On VOC 2012, YOLO scores 57.9% mAP, performing better than R-CNN in some categories but struggling with small objects.
- YOLO generalizes better to artwork than other detectors, effectively modeling object size, shape, and relationships.

9 CONCLUSION

YOLO is a unified model for object detection that is simple to build and trained on full images. Unlike classifier-based methods, YOLO uses a loss function directly tied to detection performance, with the entire model trained jointly. Fast YOLO is the fastest general-purpose object detector and sets the state-of-the-art in real-time detection. It generalizes well to new domains, making it ideal for applications requiring fast and robust detection.