Enhancing YOLOV5 for Drone and UAV Detection

Ben-Gurion University - IEM Department

Dr. Ari Pakman

Introduction to Deep Learning

2024 Semester B Final Project - Group 10 15/7/2024

By Alon Bar Koter, Noa Gerson, Shahar Shpitaler, Andrei Plotkin

Motivation

- Overall market share of 30.2B USD in 2024 (5.42M units)
- Civil and commercial applications: delivery, security, agriculture
- Photography and videography, mapping
- Data transmission
- Military operations
- Challenges criminal activities

Project Goals

1 **Train a custom YOLOv5**
Establish baseline for drone and UAV detection performance.

2 **Enhance Architecture**
Attempt to increase inference speed and accuracy by modifying model architecture (C3tr/Ghost/ Focus).

3 **Hyperparameter tunning**
Translate, Shear, Mixup, Scale.

4 **Future work –Pruning**
Reduce computational costs while maintaining accuracy.

Data Preparation

1

Download dataset

Obtain 4.5k images multipole sources

2

Preprocess images

Clean and format data for for training

3

Annotate images

Ensure proper labeling for accurate detection

4

Validate data

Verify dataset quality and and consistency

Dataset

	Drones	UAV'S
Train	3155	450
Test	788	112

0.3212962962962963 0.2208333333333333 0.6652777777777777 0.2953125 1

Yolo (You Only Look Once)

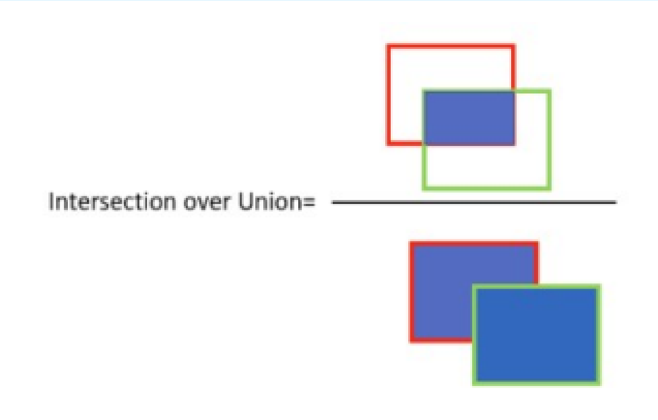
What is Yolo?

A real-time object detection system.
How it works:

- Residual blocks- divide the image into N*N grid. Each is responsible for localizing and predicting the class it covers.
- Bounding box regression
- Intersection Over Unions or IOU for short
- Non-Maximum Suppression. keep only the boxes with the highest probability score of detection.

Our target measurements

- MAP50, 50-95 (Mean Average Precision)
- Speed- GFLOPS – Giga Floating Point Operations Per Seconds
- Model Size/ Complexity (# parameters)



YOLOv5 enhancement's

C3TR (Transformer Enhanced C3)

A variation of the C3 layer that integrates transformer-based mechanisms for capturing long-range dependencies and enhancing feature extraction.

GhostConv

An efficient version of the convolutional layer that reduces the number of parameters and computational cost by generating more feature maps with fewer computations.

C3Ghost

Combines the C3 layer with Ghost Convolutions to reduce computational complexity while maintaining efficient feature extraction.

Focus

A layer that slices the input feature map into smaller patches and concatenates them along the channel dimension. This helps in capturing fine-grained details.

Optimize Hyperparameters

Shear

Introduces geometric deformations by tilting the images along the x or y-axis.

Mimics real-world situations where objects may appear tilted due to camera angles.

Scaling

Resizing the input to different scales or dimensions.

Enable to handle both small and large objects effectively.

Mixup

Combine pairs of images and their corresponding object labels to create new training examples.

Enhances the model’s ability to handle variations in object appearances.

Translation

Involves shifting or moving the objects within the image.

Improves accuracy in detecting objects even when they are not centered or located at expected positions.


Model comparison and conclusion

Model	Map50	Map50-95	GFLOPS	Model size
Baseline	**0.96**	**0.65**	**4.2**	**1,761,871**
Baseline – with HP	0.82	0.34	4.1	
Ghost c3 and conv	**0.86**	**0.48**	**2.3**	**939,275**
Ghost c3 and conv Baseline – with HP	0.54	0.2	2.3	
C3TR instead of c3	0.88	0.49	4.1	1,762,063
C3TR instead of c3 – with HP	0.6	0.21	4.1	
Focus	0.9	0.49	Na	1,761,871
Focus – with HP	0.55	0.21	Na	
Ghost c3 and conv with focus	0.86	0.48	2.3	943,683