CAP_6411_ Computer Vision Systems **Object Detection -Assignment2**

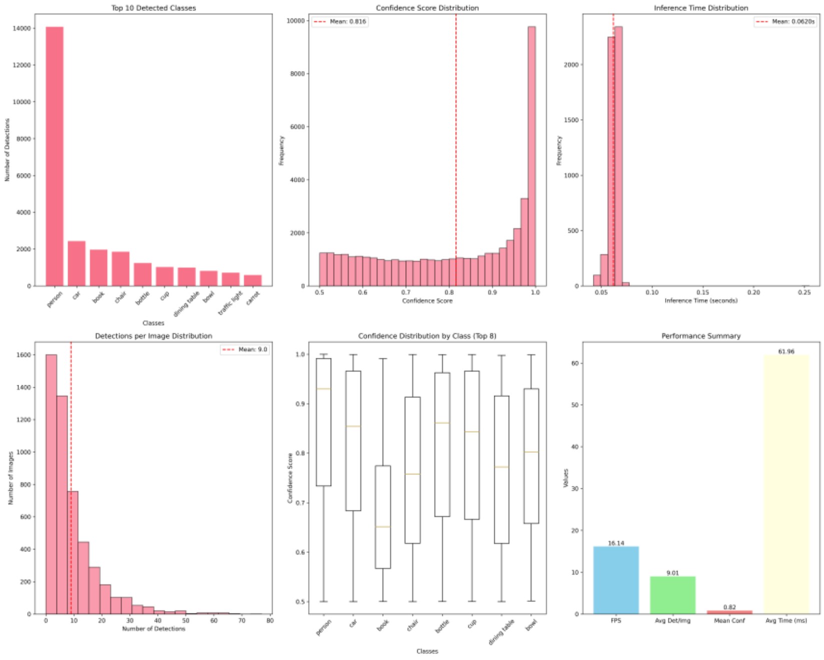
Karthika Ramasamy

==

My code is running object detection across the full COCO val2017 dataset (5000 images). And for each image performed I run inference through Faster R-CNN, DETR, Grounding-Dino and Dino model and collect detections based on confidence threshold then save bounding boxes, scores, labels for each detection. I'm measuring inference time. Then after all images summarizing the results which are total detections, detections per image, mean confidence, avg latency, FPS saved in JSON results file.

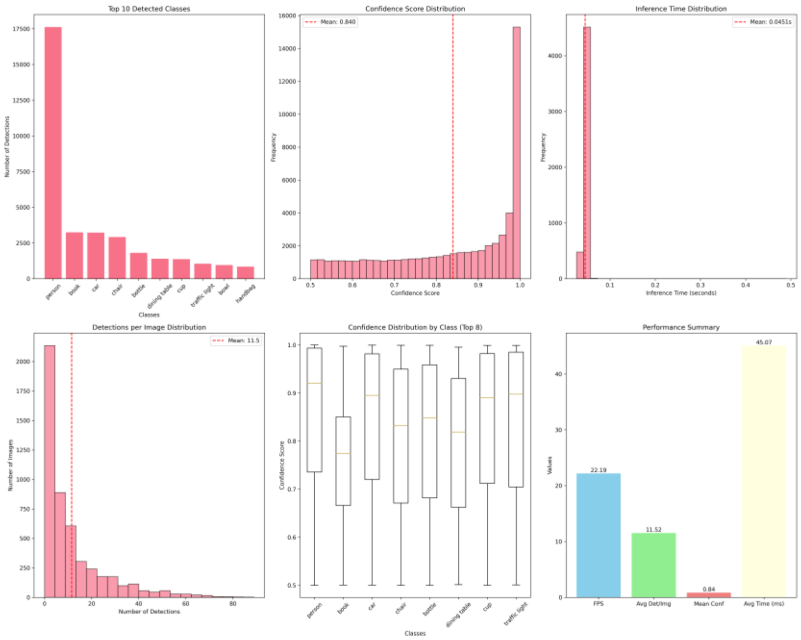
INSIGHTS

Faster R-CNN



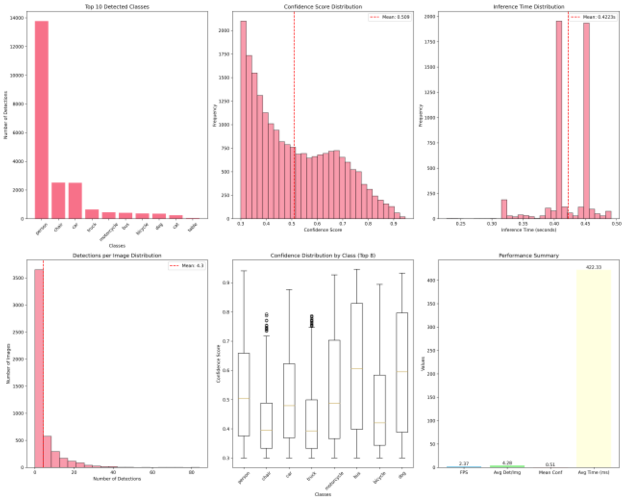
Faster R-CNN with ResNet-50 FPN produces solid, anchor-based detections. In evaluation, it achieved 9 detections per image with mean confidence 0.816 and inference time 62 ms (16 FPS). This makes it slower than DETR but faster than Grounding DINO. The histogram shows strong performance across frequent categories (*person, car, book*). The image sample demonstrates clear bounding boxes around laptops, monitors, keyboards, and persons, often with high confidence (0.75–1.0).

DETR:



DETR (facebook/detr-resnet-50) uses transformer-based global reasoning without anchors. It achieves ~11.5 detections per image, highest mean confidence **(0.84),** and fastest inference (45 ms, 22 FPS). Plots show its confidence scores skew strongly toward 0.9–1.0, indicating high decisiveness. The qualitative image displays crisp detections across persons, screens, and small accessories, with near-1.0 confidence. DETR balances speed, accuracy, and robustness in cluttered indoor scenes. It occasionally misses the tiniest items (like utensils) but overall demonstrates the most efficient performance in better.

GROUNDING-DINO:



Grounding DINO integrates detection with natural language prompts, enabling open-vocabulary detection. Its results show strong coverage of *persons* and *vehicles*, but average confidence is lower (mean 0.51). The model detects fewer objects per image (4.3) and inference is slower (422 ms per image, 2.37 FPS). The visualization shows correct bounding boxes but with moderate confidence scores (0.75–0.85), reflecting variability. This model shines when the task involves flexible categories or prompt-based detection, though at the cost of speed and precision.

Model	**Avg Time / Image**	**FPS**	**Avg Detections / Img**	**Mean Confidence**
Grounding DINO	422 ms	2.37	4.3	0.51
Faster R-CNN	62 ms	16.1	9.0	0.816
DETR	45 ms	22.2	11.5	0.84
Dino				

COMPARISION AMONTHE MODELS:

- For Faster R-CNN most of the detections are people, followed by cars, books, and chairs. It tends to stick to the main, frequent categories.
- In DETR, people are still the biggest group, but it also picks up smaller things like dogs, bottles, and traffic lights more often, so the spread is wider.
- Grounding-DINO is very narrow, and almost all detections are just people, chairs, and cars, with very few other classes.
- So, overall DETR shows the best variety across classes, Faster R-CNN is okay, and Grounding-DINO is the weakest when only looking at the COCO classes.

Confidence Score Distribution

- Faster R-CNN model average confidence is about 0.82, with most predictions scoring high (0.8–1.0). This shows it is steady and reliable.
- DETR average is slightly higher at 0.84, with scores packed even closer to the top end. It’s the most confident overall.
- Grounding-DINO average is lower at around 0.51. Scores are spread widely from 0.3 to 0.9, meaning less stable predictions.

- So, DETR is the most confident, Faster R-CNN is also strong, while Grounding-DINO is inconsistent

Inference Time Distribution

- Faster R-CNN is about 0.062 seconds per image (≈16 FPS). Runs fast enough for real-time.
- DETR is faster at 0.045 seconds per image (≈22 FPS). Best for speed.
- Grounding-DINO much slower, around 0.42 seconds per image (≈2 FPS). This is because of its heavy text-prompt and transformer setup.
- Then the DETR is the fastest, Faster R-CNN is solid, and Grounding-DINO is too slow for real-time use without tuning.

Detections per Image

- Faster R-CNN Finds about 9 objects per image.
- DETR Finds the most, about 11–12 per image.
- Grounding-DINO is the lowest, about 4 per image.
- DETR gives the best recall, Faster R-CNN is in the middle, and Grounding-DINO detects the least.

Confidence Distribution by Class

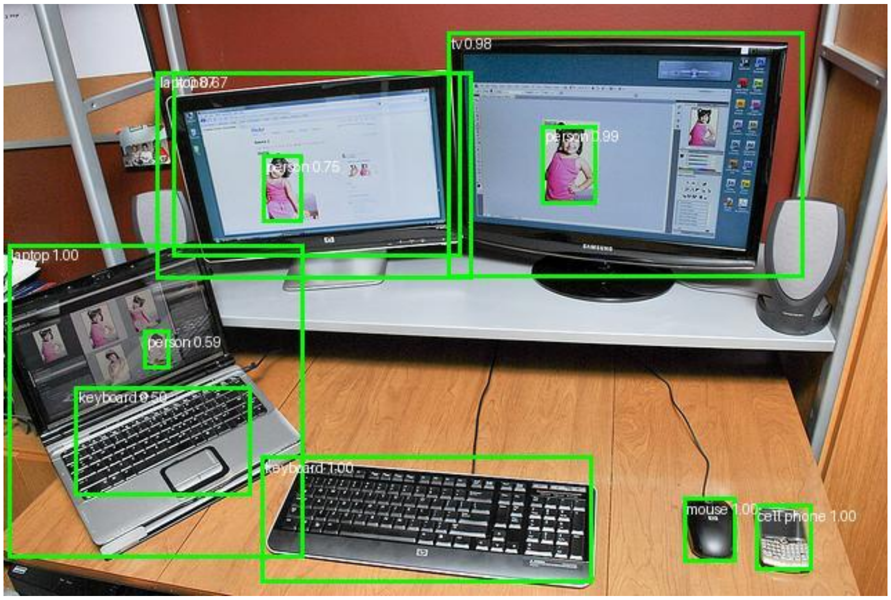
- Faster R-CNN gives high and consistent confidence across the main categories.
- DETR is also high, but better balance across different categories, even smaller ones like traffic lights and cups.
- The Grounding-DINO model is lower and more scattered, with confidence values jumping around a lot depending on the class.
- So, the model DETR has the best balance, Faster R-CNN is steady, Grounding-DINO is noisy.

Performance Summary

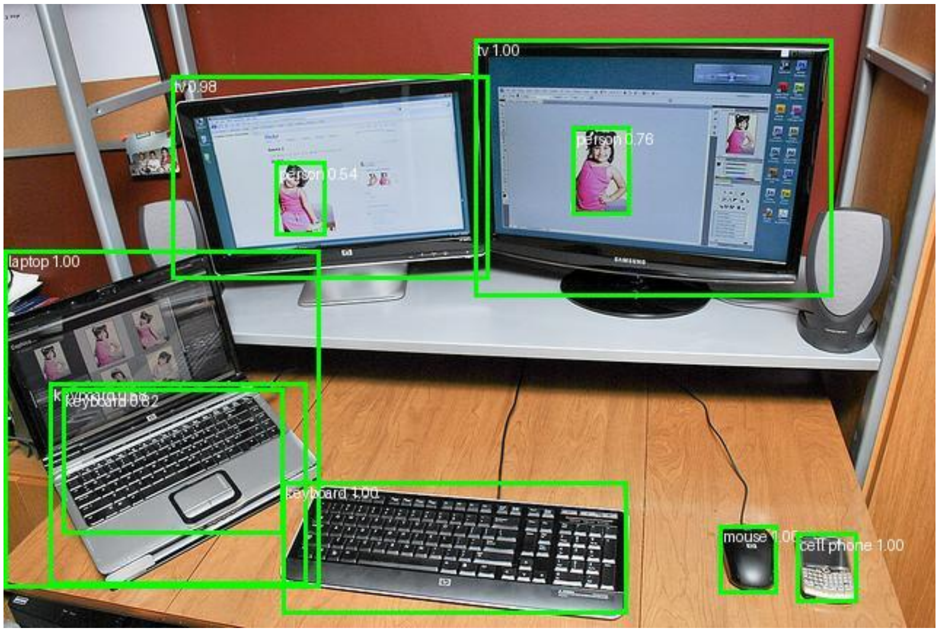
- Faster R-CNN gives 16 FPS, 9 detections per image, 0.82 confidence, 62 ms per image and DETR 22 FPS, 11.5 detections per image, 0.84 confidence, 45 ms per image.
- Grounding-DINO, 2 FPS, 4 detections per image, 0.51 confidence, 422 ms per image.

- Finally, the DETR leads in speed, recall, and confidence. Faster R-CNN is a reliable middle option. Grounding-DINO performs poorly on COCO numbers but is special because it can handle new categories through text prompts.

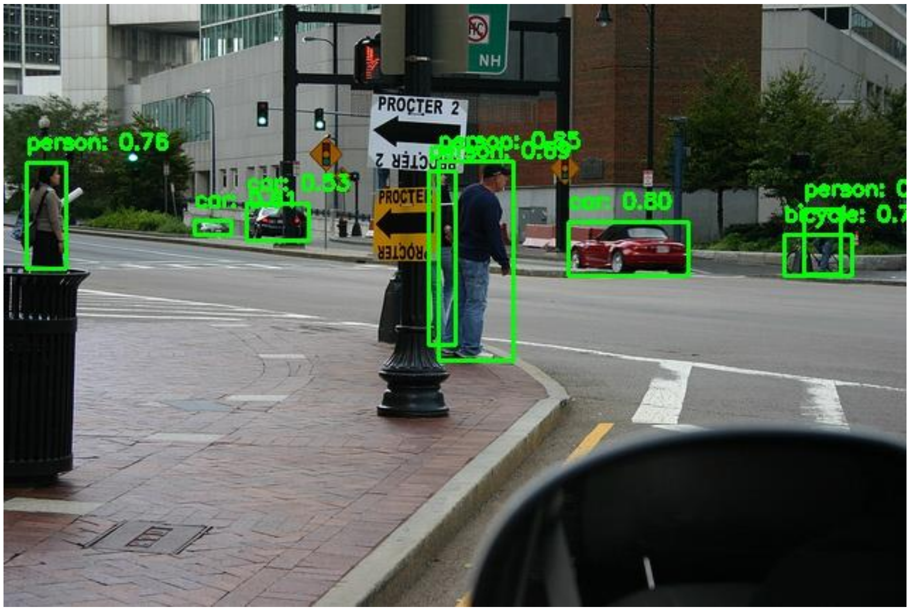
DISPLAYING THE IMAGES OF OBJECT DETECTION:



Faster R-CNN model: The Indoor setup with laptops, monitors, keyboard, and person, the bounding boxes are accurate with high confidence, showing its robustness for structured objects.



DETR Model: Same indoor scene, with sharper boxes and higher confidences (close to 1.0). The model is decisive and fast, capturing most items without overlap or duplicates.



Grounding DINO Model: Detects persons and vehicles at a crosswalk with moderate confidence (0.75–0.85). Shows real-world utility in traffic/pedestrian scenarios, aligning with its text-prompt design.

ERROR TACKLED WHILE RUNNING MODEL:

1. I tried to Run the inference for the Dino model but I'm not able to run the Dino model to get the actual inference comparison. I know the Dino would perform better than other models. I got Build issue: with dino loading model and I tried to fix it. I'm working on it.