Visual Recognition

Object Detection, Recognition and Segmentation

Aleix Pujol, Diana Tat, Georg Herodes, Gunjan Paul
Master in Computer Vision

***Abstract*— This week came with the challange of working with the Detectron2 framework and understanding how to deal with annotation formats. We have applied Faster R-CNN and Mask R-CNN models to track and segment the cars and pedestrians, using the KITTI-MOTS dataset and COCO weights. We have also tried the newest version of YOLO v9 and see its performance.**

I. Introduction

The first task of this module was to become familiar with the PyTorch, converting our code from Keras and see the differences. We have have made the transition of our best models from the previous module to PyTorch and PyTorch Lightning to compare the results and see the differences.

PyTorch is a very adaptive and dynamic method of building models. Because of that, we can modify and debug the code easy, as well as transform it to PyTorch Lightning. On the other hand, Keras, has a more high-level abstraction and computational graphs that facilitate the efficient deployment and optimisation of the code.

This week we have been using Detectron2 framework which is made by Facebook Artificial Intelligence Research's. It is based on PyTorch and it features panoptic segmentation, bounding box and instance segmentation mask object recognition. It provides a wide range of models, such as Panoptic FPN, Dense Pose, TensorMask and the ones that we focused on, Mask R-CNN and Faster R-CNN.

Faster R-CNN provides accurate object localisation by using Region of Interest (RoI) pooling. It suggest regions for examination and then it uses Convolutional Neural Networks (CNNs) to extract the information, simplifying the detection process. Faster R-CNN produces generally accurate object localisation, but it can't work with masks, which are useful in situations when we need more precise information of the object boundaries. For this reason, Mask R-CNN was introduces as an extension of Faster R-CNN. It overcomes its drawback by having an extra branch that is responsible for segmenting the prediction mask. It uses pixel-wise segmentation and it more precisley in object identification.

Also, as an optional task we have tried the new version of YOLO, YOLO v9. YOLO is an algorithm that performs real-time object detection by dividing the image into a grid and predicting bounding boxes and class probabilities. We have evaluated and compared its performance with the Faster R-CNN and Mask R-CNN models.

II. KITTI-MOTS Dataset

The dataset comprises 12 training sequences, contributing to a wealth of 8,073 pedestrian masks and 18,831 car masks, along with 9 validation sequences, incorporating 3,347 pedestrian masks and 8,068 car masks. An additional 29 sequences are allocated for testing purposes.

Fig. 1. KITTI MOTS Dataset(left) and Annotations(right).

Car instances are sequentially labeled as 1000, 1001, 1002, and so forth, while pedestrian instances follow a similar pattern starting from 2000.

Additional annotations are provided in a text format, in $instances\_txt$, and they are sent as segmentation masks in PNG format.It is then separated into subsets for training and testing, which provides a realistic and varied range of circumstances for thorough model training and assessment. We can observe how a frame from the dataset looks like in figures 1.

III. Related Works

A. Neural Networks in Computer Vision

In the field of computer vision, as [1] says, deep learning has become indispensable for how machines perceive and process visual data. Convolutional neural networks (CNNs), in particular, are deep learning models that have demonstrated unmatched performance in tasks like object detection, instance segmentation, and image categorization. They are able to recognise complex patterns and features, going beyond conventional computer vision techniques, thanks to their autonomous learning of hierarchical representations from unprocessed visual data.

B. Object Detection

1) Objectives and Uses: Deep learning has brought to an important shift in object detection, a fundamental aspect of computer vision. In important studies like Zhao et al.'s [14] from 2018, the writers thoroughly examine this paradigm change by diving into the complexities of object detection enabled by deep learning techniques. A more recent study, published in 2020 by Xiao et al. [12], offers a modern

viewpoint by showcasing the developments in deep learning-based object identification and its numerous uses in multimedia tools and applications. Moreover, Kaur and Singh's paper [4] from 2023 provides a thorough examination, highlighting how object detection is changing in the context of digital signal processing.

2) Two-Shot Detectors: Two-Shot Detectors are an important advance in the field of object detection since they enable more precise and nuanced identification of visual objects. 2015 saw the introduction of Faster R-CNN[2]. By integrating region proposal networks, this paper study transformed object identification and greatly enhanced real-time performance. The Faster R-CNN framework built on this basis to create the framework for later advancements in two-shot detection techniques.

One key illustration of this trajectory's evolution is the Faster Mask R-CNN refinement by Ren et al. [9] that same year. This iteration presents Region Proposal Networks (RPNs), which are meant to make region proposals more quickly and efficiently while also accelerating the object detection process. These Two-Shot Detectors are very useful in improving the accuracy of object detection and laying the groundwork for future improvements like the incorporation of instance segmentation as shown by Mask R-CNN [3]. This is especially true of Faster R-CNN and its variants.

With the introduction of Mask R-CNN in [3], instance segmentation has made substantial progress. This model adds a second branch to the Faster R-CNN architecture, which predicts segmentation masks in addition to bounding boxes and class labels. Mask R-CNN works by first producing region suggestions, which it then refines to provide precise object masks. A critical layer of information is added by integrating pixel-level segmentation, which allows the model to offer detailed spatial awareness of object instances inside an image.

3) Single-Shot Detectors: In the field of object detection, single-shot detectors (SSDs) have become an important category because they provide real-time processing capabilities without sacrificing accuracy.

The YOLO series of single-shot detectors has made a substantial contribution to the area. By treating object identification as a regression problem and explicitly predicting bounding box coordinates and class probabilities in a single network pass, YOLO constitutes a paradigm change. The subsequent iterations of YOLO—YOLOv1 [8], YOLOv2 (YOLO9000) [6], YOLOv3 [7], YOLOv7 [10], and most recently, YOLOv9 [11]—showcase the continuous attempts to enhance single-shot object detection precision, effectiveness, and adaptability.

4) Segmentation: In computer vision segmentation is important because it allows to understand the input data better. A in depth analysis of object detection using deep learning for segmentation is explained in [15], enhancing the precision of the segmentation results. In the paper [13] the focus is on exploring different models and approaches, highlighting the effectiveness of trying different methods and their applicability in different contexts. Furthermore, [5] provides an examination of object detection using deep learning in the context of digital signal processing, illuminating the significant developments and difficulties in the area.

REFERENCES

[1] Junyi Chai et al. "Deep learning in computer vision: A critical review of emerging techniques and application scenarios". In: *Machine Learning with Applications* 6 (2021), p. 100134. ISSN: 2666-8270. DOI: https://doi.org/10.1016/j.mlwa.2021.100134. URL: https://www.sciencedirect.com/science/article/pii/S2666827021000670.

[2] Ross B. Girshick. "Fast R-CNN". In: *CoRR* abs/1504.08083 (2015). arXiv: 1504.08083. URL: http://arxiv.org/abs/1504.08083.

[3] Kaiming He et al. "Mask R-CNN". In: *CoRR* abs/1703.06870 (2017). arXiv: 1703.06870. URL: http://arxiv.org/abs/1703.06870.

[4] Ravpreet Kaur and Sarbjeet Singh. "A comprehensive review of object detection with deep learning". In: *Digital Signal Processing* 132 (2023), p. 103812. ISSN: 1051-2004. DOI: https://doi.org/10.1016/j.dsp.2022.103812. URL: https://www.sciencedirect.com/science/article/pii/S1051200422004298.

[5] Ravpreet Kaur and Sarbjeet Singh. "A comprehensive review of object detection with deep learning". In: *Digital Signal Processing* 132 (2023), p. 103812. DOI: https://doi.org/10.1016/j.dsp.2022.103812. URL: https://www.sciencedirect.com/science/article/pii/S1051200422004298.

[6] Joseph Redmon and Ali Farhadi. "YOLO9000: Better, Faster, Stronger". In: *CoRR* abs/1612.08242 (2016). arXiv: 1612.08242. URL: http://arxiv.org/abs/1612.08242.

[7] Joseph Redmon and Ali Farhadi. "YOLOv3: An Incremental Improvement". In: *CoRR* abs/1804.02767 (2018). arXiv: 1804.02767. URL: http://arxiv.org/abs/1804.02767.

[8] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: *CoRR* abs/1506.02640 (2015). arXiv: 1506.02640. URL: http://arxiv.org/abs/1506.02640.

[9] Shaoqing Ren et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *CoRR* abs/1506.01497 (2015). arXiv: 1506.01497. URL: http://arxiv.org/abs/1506.01497.

[10] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. *YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors*. 2022. arXiv: 2207.02696 [cs.CV].

[11] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. *YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information*. 2024. arXiv: 2402.13616 [cs.CV].

[12] Youzi Xiao et al. "A review of object detection based on deep learning". In: *Multimedia Tools and Applications* 79.33–34 (June 2020), pp. 23729–23791. ISSN: 1573-7721. DOI: 10.1007/s11042-020-08976-6. URL: http://dx.doi.org/10.1007/s11042-020-08976-6.

[13] Xiao Youzi et al. "A review of object detection based on deep learning". In: *CoRR* 79.33-34 (2020), pp. 23729–23791. URL: http://dx.doi.org/10.1007/s11042-020-08976-6.

[14] Zhong-Qiu Zhao et al. "Object Detection with Deep Learning: A Review". In: *CoRR* abs/1807.05511 (2018). arXiv: 1807.05511. URL: http://arxiv.org/abs/1807.05511.

[15] Zhong-Qiu Zhao et al. "Object Detection with Deep Learning: A Review". In: *CoRR* abs/1807.05511 (2018). URL: http://arxiv.org/abs/1807.05511.