

November 19, 2020

Related work

1 Object detection overview

Object detection has attracted a lot of attention in recent years because it has a wide range of applications and technological breakthroughs. The task of object detection has been explored extensively in academia as well as real-world applications like monitoring security, transportation surveillance, autonomous driving, drone scene analysis, robot perception etc. There are many factors and efforts that have lead to the enhancement of object detection techniques. Some of the prominent ones are the development of deep convolutional neural networks and enhancement of GPUs computing power. In the current scenario, deep learning model has been extensively used in the field of generic as well as domain-specific object detection. The current state of the art object detection models use deep convolutional neural networks(CNNs) as their backbone as well as detection network for extracting features from input images or videos and solve the task of classification as well as localization respectively. [14]

The problem of object detection could be divided into two categories:

- One stage approach
- Two-stage approach

2 One stage approach

The one stage approach could be divided into two categories :

- 3D object detection

- 2D object detection

2.1 3D object detection

The single-stage 3D object detection approaches [41, 46, 43, 17, 36, 19] parse the given 3D data to a compact representation like voxel grid or bird eye view image and use CNN to predict the bounding box. This enables single-stage approaches to become simple as well as efficient. A significant drawback of this approach is that these approaches downscale the feature maps progressively. Hence, the spatial resolution of the feature maps get lost and thus, the structural information of point cloud could not be considered explicitly. Therefore single-stage approaches are less accurate when it comes to processing the sparse point clouds.

In single-stage approaches [18, 42], they slice the dense 3D data to 2D slices and are fed to 2D object detection models to get the prediction.

2.2 2D object detection

Unified one stage 2D object detection approach refers to architectures which directly predict class probabilities as well as bounding box offsets from images with single feed-forward Convolutional Neural Network(CNN) in a monolithic setting which does not involve generation of proposal region or post classification that encapsulates all computation using a single network.

YOLO [30] divides the input image into $M \times M$ grid cells and utilizes CNNs to get the bounding box regression, confidence scores as well as class probabilities of each grid cell. YOLO0000 [31] and YOLOv3 [18] further improve the performance. Even though YOLO is fast, it misses small objects because of the coarse segmentation of input images. These drawbacks were addressed by SSD [21] by utilizing feature pyramids for single stage object detection. In SSD for every feature map locations anchor boxes of various aspect ratios and scales are generated. In RetinaNet [20] they proposed focal loss in order to handle the imbalance between target and background object bounding boxes.

3 Two-stage approach

The two-stage approach could be divided into two categories :

- 3D object detection
- 2D object detection

3.1 3D object detection

The advantage of two-stage 3D object detection approach [28, 33, 4, 44, 27, 34] over one-stage 3D object detection approach is that two-stage approaches can leverage spatial information in the second stage, that focuses on the region of interest (ROI's) which are predicted by the first stage and hence predicts bounding box more accurately. This shows that accurate localization can be achieved when fine-grained spatial information is leveraged. Operating on each point and re-extracting features on each ROI increases the computational cost substantially. Hence it becomes hard for two-stage approaches to reach real-time speed. [11]

In two-stage approaches [47, 16], the dense volume data is fed to the 3D version of R-CNN models to get the prediction. In [38] slice of dense volume data, i.e. slice of 3D CT scans are fed to Mask R-CNN to get the prediction.

3.2 2D object detection

Two-stage 2D object detection approaches are region-based frameworks. In the case of two-stage approach region proposals which are category independent are generated from an image. CNN features are then extracted from these regions. After that category specific classifiers are utilized to determine the label of the categories for the proposals.

The two-stage 2D object detection algorithms are best represented by the R-CNN family [10, 32, 12]. Faster R-CNN introduced the Region Proposal Network (RPN). A substantial number of background candidates are filtered out by RPN, and a different network is used to predict bounding box co-ordinates and class labels for each proposal. In R-FCN [6] position-sensitive feature maps are extracted. These feature maps are fed to RPN to get class scores. Mask R-CNN [13] extends Faster R-CNN to instance segmentation, they first find the bounding box coordinates and crop and segment the bounding box region to get the refined mask.

4 Datasets

The datasets for 3D object detection can be broadly classified into two categories :

- Sparse 3D point cloud
- Dense volume data

4.1 Sparse 3D point cloud

4.1.1 LiDAR data

LiDAR (Light Detection and Ranging) is a remote sensing technique in which the distance to the target is gauged by illuminating the target utilizing a laser light and using a sensor in order to measure the reflection of light [29]. The difference in wavelengths and laser return times are then used to obtain a 3D representation of the target. LiDAR data has several applications in surveying, geomatics, geography, archaeology, seismology, forestry. The KITTI Vision Benchmark Suite [9] is one the datasets for 3d object detection in the field of autonomous driving. It contains LIDAR data taken using a sensor mounted in the front of the car.

4.1.2 RGB-D data

RGB-D data is a combination of RGB data along with its corresponding depth data. One of the RGB-D object detection datasets is SUN RGB-D [37].

4.2 Dense volume data

Dense volume data is also referred to as dense 3D data in [7]. Computed Tomography (CT), Magnetic Resonance Imaging (MRI) are some of the examples of dense volume data.

4.2.1 Computed Tomography (CT)

The term “Computed Tomography” refers to computerized X-ray imaging process in which narrow beam of X-rays are aimed at the patient and rotated quickly around the body which produces signals that are utilized by

machine's computer to get cross-sectional images or slices of the body. These slices are referred to as tomographic images and contains more detailed information in comparison to conventional x-rays. [24] CT scans are not limited to medical domain; they are also used in industries. In industries, they are used for detection of flaws like cracks and voids as well as particle analysis of materials. They are used in metrology for the measurement of internal and external geometry of complex parts. Some of the CT scan object detection datasets available are RSNA pneumonia detection challenge [26], COVID 19 CT scans [15], DeepLesion dataset [2], covid-19-chest-xray-lung-bounding-boxes-dataset [5]

4.2.2 Magnetic Resonance Imaging (MRI)

Magnetic Resonance Imaging(MRI) is a non-invasive imaging technique that creates three-dimensional anatomical images. It is used for diagnosis and detection of diseases as well as treatment monitoring. It relies on sophisticated technology which excites as well as detects the change in the direction of the rotational axis of protons that are found in water which makes up living tissues. MRI scanners are well suited to get the image of the non-bony parts as well as soft tissues of the body. They are different from CT scans in the way that they don't use damaging ionizing radiation of the x-rays. The spinal cord, brain, nerves, ligaments, muscles, tendons are visible more clearly with MRI in comparison to regular x-rays and CT, this is the reason why MRI is used for imaging the knee as well as shoulder injuries. [25]

In [45, 3], they use MRI data as an input for the task of object detection. There are no open dataset available for object detection in MRI scans.

5 Resolution of image

Resolution means the number of pixels present in an image. Resolution is identified by the width and height of the image. Resolution also refers to the total pixels an image contains, for example, if an image is 2048 pixels wide and 1536 pixels high. Total pixels = $2048 \times 1536 = 3145728$ pixels or 3.1 Megapixels. So the resolution image is 3.1 Megapixels.[23]. One of the factors that impact the performance of object detection models is the resolution of images [22].

5.1 Advantage of using low resolution image

5.1.1 Low memory requirement

The memory required to store the images reduces when we use the image of lower resolution; this is well illustrated in figure 1.

Inch size (changed)	Resolution (changed)	Pixel dimensions (you set)	File size
2 x 2 in	100 ppi	200 x 200 px	117.2 KB
3 x 3 in	100 ppi	300 x 300 px	263.7 KB
6 x 6 in	100 ppi	600 x 600 px	1.03 MB

Figure 1: memory requirements [1]

5.1.2 Frame Per Second(Fps) increases

As the resolution of the image decreases, the fps of the object detection model increases [39].

5.2 Disadvantage of using low-resolution image

Downsampling image leads to too much information loss and affects the performance of the model [35].

6 Discussion

3D data could be sent to 3D object detection models, or they could be sliced and fed to 2D object detection techniques for prediction. Slicing the image leads to loss of information and hence 3D object detection models perform better than 2D object detection models. The 3D object detection technique, whether it is one stage approach or two-stage approach has these limitations:

- A major drawback of these 3D algorithms is that they are computationally expensive [42].
- 3D data are quite difficult to collect, annotate and store and hence there are fewer datasets for deep learning models. [40, 8]

As discussed in section 5, reducing the resolution of the image has advantages like reduction in memory requirement, increasing the fps of the model. But the papers discussed in the sections one stage approach 2 and two stage approach 3 don't provide the details about how these models would perform when the resolution is varied.

References

- [1] Adobe. <https://helpx.adobe.com/photoshop/kb/advanced-cropping-resizing-resampling-photoshop.html>. Accessed on 22 October 2020.
- [2] NIH Clinical Center. <https://nihcc.app.box.com/v/DeepLesion/folder/50715173939>. Accessed on 22 October 2020.
- [3] Hao Chen, Qi Dou, Lequan Yu, and Pheng-Ann Heng. Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation. In *arXiv:1608.05895v1 [cs.CV]*, 2016.
- [4] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Fast point r-cnn. 2019.
- [5] covid-19-chest-xray-lung-bounding-boxes dataset. <https://github.com/GeneralBlockchain/covid-19-chest-xray-lung-bounding-boxes-dataset>. Accessed on 22 October 2020.
- [6] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.
- [7] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, and I. Posner. Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks. 2017.
- [8] Di Feng, Xiao Wei, Lars Rosenbaum, Atsuto Maki, and Klaus Dietmayer. Deep active learning for efficient training of a lidar 3d object detector. In *arXiv*, 2019.
- [9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [10] Ross Girshick. Fast r-cnn. In *In Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

- [11] Chenhang He, Hui Zeng, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Structure aware single-stage 3d object detection from point cloud. In *CVPR*₂₀₂₀, 2020.
- [12] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Facebook AI Research (FAIR)*, 2018.
- [13] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *arxiv.org*, 2018.
- [14] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, and Rong Qu. A survey of deep learning-based object detection. In *arXiv:1907.09408v2 [cs.CV]*, 2019.
- [15] Kaggle. <https://www.kaggle.com/andrewmvd/covid19-ct-scans>. Accessed on 22 October 2020.
- [16] Evi Kopelowitz¹ and Guy Englehard. Lung nodules detection and segmentation using 3d mask-rcnn. In *arXiv*, 2019.
- [17] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019.
- [18] Andréanne Lemay. Kidney recognition in ct using yolov3. In *In Advances in neural information processing systems, pages 91–99*, 2019.
- [19] Bo Li. 3d fully convolutional network for vehicle detection in point cloud. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.
- [20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *In Proceedings of the IEEE international conference on computer vision, pages 2980–2988*, 2017.
- [21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multi-box detector. In *European conference on computer vision, pages 21–37. Springer*, 2016.

- [22] Jerubbaal John Luke, Rajkumar Joseph, and Mahesh Balaji. Impact of image size on accuracy and generalization of convolutional neural networks. In *IJRAR*, 2019.
- [23] microscope.org. <https://microscope-microscope.org/microscope-info/image-resolution/>. Accessed on 22 October 2020.
- [24] NIH. <https://www.nibib.nih.gov/science-education/science-topics/computed-tomography-ct>. Accessed on 22 October 2020.
- [25] NIH. <https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri>. Accessed on 22 October 2020.
- [26] Radiological Society of North America. <https://www.kaggle.com/c/rsna-pneumonia-detection-challenge>. Accessed on 22 October 2020.
- [27] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. 2019.
- [28] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, , and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. 2018.
- [29] AZO quantum. <https://www.azoquantum.com/News.aspx?newsID=7018>. Accessed on 22 October 2020.
- [30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788*, 2016.
- [31] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *In CVPR*, 2017.
- [32] Shaoqing Ren, Ross Girshick Kaiming He, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *In Advances in neural information processing systems, pages 91–99*, 2015.

- [33] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointtrcnn: 3d object proposal generation and detection from point cloud. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–779, 2019.
- [34] Shaoshuai Shi, Zhe Wang, Xiaogang Wang, and Hongsheng Li. Part-a2net: 3d part-aware and aggregation neural network for object detection from point cloud. 2019.
- [35] Connor Shorten and Taghi M. Khoshgoufar. A survey on image data augmentation for deep learning. In *Springer*, 2019.
- [36] M. Simony, S. Milzy, K. Amende, and H.-M. Gross. Complex-yolo: an euler-region-proposal for real-time 3d object detection on point clouds. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.
- [37] S. Song, S. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015)*, 2015.
- [38] Tang, Y., Yan, K., Tang, Y., Liu, J., Xiao, J., Summers, and R.M. Uldor: A universal lesion detector for ct scans with pseudo masks and hard negative example mining. In *arXiv:1901.06359*, 2019.
- [39] towards data science. <https://towardsdatascience.com/no-gpu-for-your-production-server-a20616bb04bd>. Accessed on 22 October 2020.
- [40] Qian Wang, Neelanjana Bhowmik, and Toby P. Breckon. Multi-class 3d object detection within volumetric 3d computed tomography baggage security screening imagery. In *arXiv*, 2020.
- [41] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. In *MDPI*, 2018.
- [42] Anqi Yang, Aswin Sankaranarayanan, Srinivasa Narasimhan, David Held, and Jen-Hao Chang. 3d object detection from ct scans using a slice-and-fuse approach. In *Carnegie Mellon University*, 2019.

- [43] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Realtime 3d object detection from point clouds. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7652–7660, 2018.
- [44] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: sparse-to-dense 3d object detector for point cloud. In *ICCV*, 2019.
- [45] B. Zhao, J. Soraghan, G. Di-caterina, L. Petropoulakis, D. Grose, and T. Doshi. Automatic 3d segmentation of mri data for detection of head and neck cancerous lymph nodes. In *2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 298–303, 2018.
- [46] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [47] Wentao Zhu, Chaochun Liu, Wei Fan, and Xiaohui Xie1. Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. In *arXiv:1801.09555v1 [cs.CV]*, 2018.