KF5012- Poe Dameron- Project Ideation

The automation of transport vehicles like cars or delivery vans is an important part to the future of sectors like transport and business. In this context, drones are also important, as they offer a fast and manoeuvrable platform for tasks like surveillance or the delivery of small quantities of goods. Large companies like Amazon are set to start using drones for commercial uses soon [1], and other companies may follow shortly after. For a drone to be completely automated while it performs its given task, it must be able to know about and make decisions based on its surroundings, meaning that there is a need for object detection capable of being used in conjunction with drones for a variety of needs. This project aims to meet that need with the Glie-44 object detection model, trained to detect objects at ground level from the perspective of a drone.

The purpose of the Glie-44 is for conducting security tasks like monitoring an area or route and recording the information it sees. This makes it more efficient than most security cameras like CCTV, as it could cover a larger area, and the manoeuvrability and expendability of an unmanned drone can make it more favourable over human guards in certain areas. However, the drone itself most likely would not be able to act if it detected undesirable entities, so it would be recommended that it is used in conjunction with manned security.

Related work

Object detection models

i. Classic object detectors

Previously, object detectors would commonly use convolutional sliding window layers over the image to return input data. Early CNNs used this method and showed success in domains like face recognition in the case of Viola and Jones [2] and handwritten digit recognition with LeCun et al. [3]. Numerous other methods have also shown success like Deformable Part Models [4] and Histogram of Orientated Gradients [5], however the first experiments conducted using two stage detectors have mostly surpassed these types of detectors. Classic detectors like the previously mentioned show great success with tasks with narrow scopes, like recognising handwritten digits using the MNIST dataset [3] or faces from a particular point of view [2] but deviating from these tasks proved ineffective and challenging for the algorithms at the time.

ii. Two Stage object detectors

Models using this type of detector are most separated into feature extraction /object detection and classification stages. Early on, Selective Search [6] introduced a faster and more accurate method of detecting objects in images than previous models using exhaustive search methods, however used an SVM during the classification stage.

R-CNN [7] further developed on this method by exchanging the SVM for a neural network classifier. While R-CNN improved on earlier ConvNet models in terms of accuracy and precision, it still proved difficult and time consuming to train due to the model acting in two distinct feature extraction and classification segments and requiring two stages of training. Later developments in the R-CNN family like Fast R-CNN [8] and Faster R-CNN [9] would aim to solve these issues by combining the training stages into one and improving the loss function for region proposals.

Faster R-CNN also introduced Region Proposal Networks, which would suggest a list of regions of interest with accompanied objectness scores. RPNs have since became popular with other network

models and are used to reduce or distribute computations made by the model. Progressing even further, Mask R-CNN [10] extends upon the success of Faster R-CNN with the goal of improving instance segmentation (the task of classifying each pixel of an object). Previously, this task was attempted by Li et al. [11] with Fully Convolutional Instance Segmentation (FCIS) and showed promising results, yet their implementation led to errors when masks overlapped and issues with mask edges. FCIS and Mask R-CNN generate a "mask" over each object in parallel to the to bounding box detection and classification. The mask would highlight the object in frame, assigning a colour to the mask to distinguish it from objects detected in the same region. While this method can produce more accurate object classifications and is as fast as Faster R-CNN, training data for these models require both bounding boxes and mask data, which can be a significant requirement depending on the size of the training dataset.

iii. One stage object detectors

One stage detectors are a newer breed of models that differ from two stage models by consolidating both stages of the two-staged models into one network that can be trained all together. The OverFeat [12] model proposed an end-to-end training process for all stages that lead to improved accuracy for object detection and classification. YOLO [13] improved this method by using regression to create bounding box mappings and class probability maps, and vastly improving the speed at which bounding box and class predictions can be made, so much so that some implementations can be suitable in autonomous cars, however the accuracy was worse when compared to some two stage detectors and grew worse when achieving higher frames per second. SSD [14] soon surpassed YOLO, achieving even faster rates for detection, and achieving accuracy that was comparable to two stage detectors. It does this by placing bounding boxes of different scales over an object, then classifying the object in each box and refines each box. In recent years, novel innovations to the object detection stage have vastly improved the process, such as improved loss functions implemented in RetinaNet [15], designed to combat inefficient training phases caused by the imbalance of the small numbers of correct anchor boxes and vastly greater number of incorrect anchor boxes. New methods of creating anchor points for bounding boxes with the invention of corner pooling [16] and centre pooling [17] layers. These methods find objects by detecting sets of keypoints: CornerNet detects the top left and bottom right corner of an objects bounding box, while CentreNet detects a centre point in addition to the top left and bottom right corner of the object, making it like region of interest (RoI) pooling methods like those used in R-CNN.

iv. Others

Transformer networks have grown popular in recent years, like being used for action and posture classification with Action Transformer [18]. The architecture of Action Transformer is similar to earlier two stage object detectors, including a regional proposal network similar to that in Faster R-CNN.

Datasets

For computer vision, many datasets exist for training and testing object detection models in a variety of different settings. Datasets most commonly consist of images containing a variety of objects belonging to set classes like car, bicycle, person, or dog, from a variety of perspectives and viewpoints, and most often contain a training set with notations for every object in frame, including a class ID and bounding box co-ordinates. The PASCAL VOC challenges [19, 20] which ran between 2005 and 2012, provided as many as 11,000 images, 27,000 bounding box notations and 6,900 segmentations for objects. ImageNet [21], by far one of the largest datasets available, consists of

close to 1.2 million images in a similar setting. One of the most recent, Microsoft's COCO dataset [22] consists of 91 classes in 2.5 million object instances, present in 328,000 images, and claims that the objects in view "would easily by recognisable by a 4-year-old". While the datasets can be excessive and training a model on even a subset of a dataset, many frameworks like pyTorch and TensorFlow offer pre-trained two and one stage models that can significantly reduce the time spent training and can then be fine-tuned to a particular dataset. Because PASCAL, COCO and ImageNet are generally similar and do not differ in context drastically, a pre-trained model from any dataset would be suitable to use with Glie-44.

In the context of images from flow-flying unmanned ariel vehicles, very few official options exist. The VisDrone [23] contest offers a large dataset of images and videos taken from drones on journeys in various cities in China. The dataset consists of 10,000 static images, 288 video clips which consist of a total of 260,000 frames, containing objects from 12 different classes of various objects and modes of transport, and people. The inD [24] dataset consists of video recordings taken from drones, but exclusively of motor vehicles, bicyclists, and pedestrians at German intersections. The context of this dataset is drastically different to the intended use of the project, so may be unsuitable for general object detection purposes.

Outline of the project plan

The VisDrone dataset should be used as it offers a large variety of high-quality images and videos close to the context that Glie-44 aims to cover. To improve the accuracy, widen the context that our model should be used in and drastically reduce training time, a pre-trained backbone should be included, being trained on any large dataset of common objects like MS COCO, PASCAL, or ImageNet, so that the model can be fine-tuned on VisDrone.

For our project, classic object detectors and models such as the Action Transformer should not be considered as both two stage and one stage detectors perform better in the context of our project. Most two stage detectors prior to models like Faster R-CNN would operate too slow to be considered for real-time object detection. However, models like Mask R-CNN or FCIS that include improved instance segmentation would require segmentation notations as input data, and as VisDrone does not include these, it would require our team to add the notations to each frame of video and image, which would total <270,000 notations to be created. Because of this, Mask R-CNN or any model that requires segmentation notations (or similarly keypoint notations) will not be considered. One stage detectors are considered a priority since they can achieve more frames processed per second than most two stage detectors. The exact model used should be determined in the Iterative Development stage, but models like Faster R-CNN, YOLO, SSD and RetinaNet should be considered.

References

[1] S. Keach, *AIR FARE Amazon is about to start making real DRONE deliveries – cutting delivery times to under 30 minutes,* The Sun, 2020.

[2] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision,* vol. 57, no. 2, pp. 137-154, 2004.

[3] Y. LeCun, L. Bottou and Y. H. P. Bengio, “Gradient-Based Learning Applied to Document Recognition,” *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278-2324, 1998.

[4] P. F. Felzenszwalb, R. B. Girshick and D. McAllester, “Cascade object detection with deformable part models,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[5] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005.

[6] J. R. Uijlings, K. E. Sande, T. Gevers and A. W. Smeulders, “Selective Search for Object Recognition,” *International Journal of Computer Vision,* vol. 104, no. 2, pp. 154-171, 2013.

[7] R. Girshick, J. Donahue, T. Darrell and J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” in *CVPR '14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[8] R. Girshick, “Fast R-CNN,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[9] S. Ren, K. He, R. Girshick and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 39, no. 6, pp. 1137-1149, 2017.

[10] K. He, G. Gkioxari, P. Dollar and R. Girshick, “Mask R-CNN,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 42, no. 2, pp. 386-397, 2020.

[11] Y. Li, H. Qi, J. Dai, X. Ji and Y. Wei, “Fully Convolutional Instance-Aware Semantic Segmentation,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus and Y. LeCun, “OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks,” in *ICLR 2014 : International Conference on Learning Representations (ICLR) 2014*, 2014.

[13] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu and A. C. Berg, “SSD: Single Shot MultiBox Detector,” in *14th European Conference on Computer Vision, ECCV 2016*, 2016.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollar, “Focal Loss for Dense Object Detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 42, no. 2, pp. 318-327, 2020.

[16] H. Law and J. Deng, “CornerNet: Detecting Objects as Paired Keypoints,” *International Journal of Computer Vision,* vol. 128, no. 3, pp. 642-656, 2020.

[17] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang and Q. Tian, “CenterNet: Keypoint Triplets for Object Detection,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[18] R. Girdhar, J. J. Carreira, C. Doersch and A. Zisserman, “Video Action Transformer Network,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[19] M. Everingham, L. van Gool, C. Williams, J. Winn and A. Zisserman, “The PASCAL VIdeo Object Classes Homepage,” 2010. [Online]. Available: http://host.robots.ox.ac.uk/pascal/VOC/. [Accessed 01 May 2021].

[20] M. Everingham, L. van Gool, C. Williams, J. Winn and A. Zisserman, “The Pascal Visual Object Classes (VOC) Challenge,” *International Journal of Computer Vision,* vol. 88, no. 2, pp. 303-338, 2010.

[21] A. Krizhevsky, I. Sutskever and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of The ACM,* vol. 60, no. 6, pp. 84-90, 2017.

[22] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, “Microsoft COCO: Common Objects in Context,” in *European Conference on Computer Vision*, 2014.

[23] P. Zhu, L. Wen, X. Bian, H. Ling and Q. Hu, “Vision Meets Drones: A Challenge,” *arXiv preprint arXiv:1804.07437,* 2018.

[24] J. Bock, R. Krajewki, T. Moers, S. Runde and L. Vater, “The inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections,” *2020 IEEE Intelligent Vehicles (IV),* pp. 1929-1934, 2019.