High resolution object detection algorithm based on parallel

Han Gao[1,2], Ling Xu[1,2], Yixuan Zhao[1,2], Tanbao Yan[3,4]

1. Aviation Key Laboratory of Science and Technology on Airborene And Missile-borne Computer, Xi'an 710000, China.

2. Xi'an Aeronautics Computing Technology Research Institute, Xi'an 710000, China.

3. 18th research department, Xi'an 710000, China.

4. Aeronautics Computing Technology Research Institute, Xi'an 710000, China.

Abstract: With the continuous development and improvement of the aviation information warfare system, target detection technology has also become a key part of the airborne system to perceive the environment. Traditional target detection technology is now difficult to meet the requirements of high precision and high real-time performance in airborne scenarios. With the continuous development of deep learning technology, neural network has become the latest method to deal with object detection task, which greatly improves its accuracy and processing efficiency. However, due to the different targets detected in airborne scenes, the scale of data needed to be processed by using neural networks to process target detection tasks also expands dramatically, and the computing resources provided by single chip are already difficult to meet the needs of target detection algorithm execution in airborne environment. This paper proposes a set of high-precision target detection algorithms based on parallelism, which greatly improves the precision and processing efficiency of the target detection algorithm in airborne scenarios.

Keywords: Object Detection; Data Parallel; Space Parallel; High Resolution Image

1. Introduction

With the deepening of the concept information warfare system and the continuous development of airborne environment perception system, the importance of aviation information warfare is also increasing. target detection technology, as a key part of the airborne system to perceive the environment, provides critical technical guidance for decision-making and behavior control of combat missions. It has a wide range of application value and development needs in the fields of aerial reconnaissance, early warning detection and border patrol.

In airborne environment, the target scenario is more complex, which brings great challenges to the application of NN[1]. target detection in airborne scenarios is mainly aimed at air-to-ground shooting, such

as cruise survey, key target search and confirmation, etc. Due to the overhead perspective, the target recognized by the camera has complete rotation invariance (rotation invariance refers to the feature of the target has nothing to do with its orientation in space). Besides, the target is small and dense, and has a certain distance from the aircraft body. In order to make the targets captured clearer and easier to identify, the resolution of the airborne camera will be very high. The images collected are all ultra-high resolution images, which are different from the typical large and prominent subject targets in the representative ImageNet[2] data set. And there are significant directional features (e.g. trees are definitely vertical in ImageNet data and people are mostly upright walking targets). Therefore, the continuous increase of image resolution and the difference of target recognition make the additional information brought by NN network further expand, and the computational complexity of target recognition task increases sharply.

Besides, from the perspective of airborne environment perception system, target recognition is in the service of the whole decision-making, control system, so the real-time demand of target recognition is high. But the development and performance improvement of airborne computing equipment (computing equipment for processing NN tasks) has a large gap compared with the speed of target recognition algorithm iteration[3] and sensor technology development. Most of the existing high-precision target recognition algorithms are difficult to meet the high real-time requirements of the airborne environment. The gap between the computing requirements of target recognition algorithms and the computing power provided by a single chip in airborne computing resources is growing[4]. The computing resources provided by a single chip cannot meet the needs of ultra-high-precision image recognition algorithms. Therefore, the introduction of a multi-chip-based[5,6] image parallel acceleration processing algorithm in the target recognition process to improve the target recognition efficiency and real-time response speed has great research significance and development prospects for airborne computer systems, which is also an indispensable part of the airborne intelligent processing nowadays.

According to the current demand and development gap, this paper proposes a set of high-speed intelligent target detection technology for high-resolution images. Firstly, it divides the super-high-resolution image into multiple low-resolution images. Then, a parallel acceleration technology is designed for the algorithm model dealing with the low-resolution images, which can simultaneously perform accelerated recognition processing on multiple groups of images after segmentation. Finally, by stitching the low-resolution images and restoring them to the recognized original images, high-speed intelligent recognition of ultra-high-resolution images can be realized.

2. Preprocessing and post-processing of image segmentation

In the airborne environment, most of the shooting scenes are air-to-ground or air-to-air scenes. Due to the long detection distance, in order to make the captured target clearer, the camera is equipped with high resolution, and the collected images are all ultra-high-resolution pictures. If these pictures are directly recognized, the recognition efficiency and accuracy are difficult to meet the required standards. In order to identify more accurately and efficiently, it is necessary to split the image.

In the process of selecting segmented sub-images, the algorithm sets an overlap window to avoid the reduction of the accuracy caused by the situation that the target is at the segmented boundary,Although the

setting of the overlapping window has significantly improved the recognition accuracy, a big overlapping area will increase the workload of post-processing and affect the execution efficiency of the algorithm. Therefore, it is necessary to make a trade-off between the accuracy and the execution efficiency, and choose a suitable size of overlapping windows. After some experiments and various investigations, we choose to set the overlap window to 15%.

In the pre-processing stage, the rotation invariance of the target in the overhead image mentioned above is solved by rotating and increasing the data. The training images are augmented by rotating around the unit circle to ensure that the class is independent of the orientation of the objectAs for algorithm design, an optimized object detection framework for overhead images is implemented, extending the existing neural network framework, optimizing for small and dense objects, supporting the analysis of spatial images through the flexibility of Python and a large user community, and simplifying the operation procedures of pre-processing and post-processing.

3. Parallel policy analysis

When processing ultra-high-precision images, we choose to segment the original image, which will result in a huge collection of images after processing. Therefore, we need to process massive input data in a parallel mode. The existing mainstream parallel strategies are divided into three categories: model parallelism, data parallelism and space parallelism.

3.1 Model Parallelism

Model parallelism firstly decomposes a large network model into multiple parts, then provides data samples and sends the divided sub-models to different processing cores, so as to train these divided sub-models in parallel. Finally the sub-models are integrated and processed to obtain a processed image.

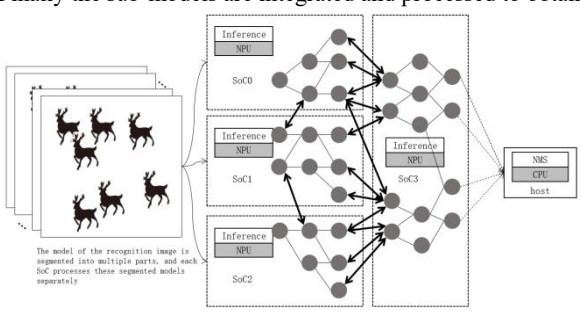

Fig.1. Schematic diagram of model parallelism

3.2 Data parallelism

Data parallelism refers to multiple processing cores taking turns to process data in an instruction pipeline-like pattern. We send different pictures (or video streams frame by frame) into different computing cores, so that each SoC independently completes the end-to-end target detection process. The segmentation of the ultra-high-resolution image is performed by the CPU on the SoC, and then the image is sent to the NPU for processing. Each SoC independently executes the end-to-end algorithm, and the results are finally sent to the main CPU to run the NMS algorithm to post-process the image.

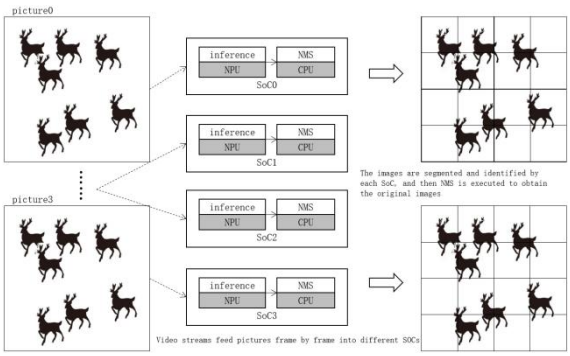

Fig.2. Schematic diagram of data parallelism

3.3 Spatial Parallelism

Spatial parallelism refers to dividing the original data so that each processing core processes a part of the original data, and finally restores the original data. That is, the original image is segmented in the core CPU, and the segmented sub-images are sent to different SoC to complete the identification of each group of sub-images. Then the processed sub-images are sent back to the core CPU for post-processing and finally processed original image is obtained.

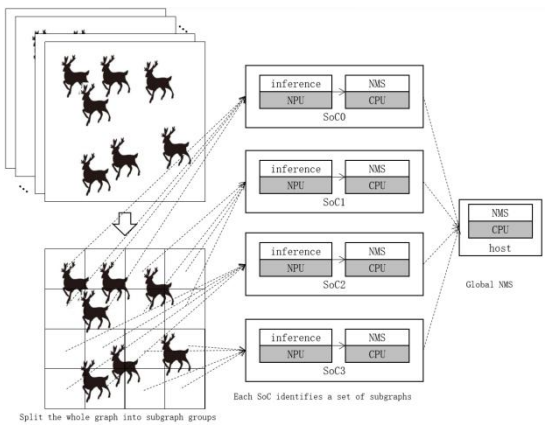

Fig.3. Schematic diagram of spatial parallelism

3.4 Analysis and selection of parallel strategy

In summary, the algorithm of model parallel needs to make adjustments to the neural network model. Different networks need different adjustment, the implementation is difficult, the portability is poor, and there is no good application value in airborne environment.

Data between the characteristics of parallel strategy is more chips use similar water processing mode, so using the strategy for the close degree of the input data has certain requirements, when the input data is not populated or the amount of data to be processed is very small, there will be a piece of system load imbalance problem, the throughput of the proposed algorithm will be significantly affected. At the same time, under this strategy, each chip independently executes an end-to-end algorithm process, so the NPU needs to have an ARM CPU core that can be used to execute the scalar computing unit in the algorithm. Figure 4 shows the execution procedure of the data parallel strategy.

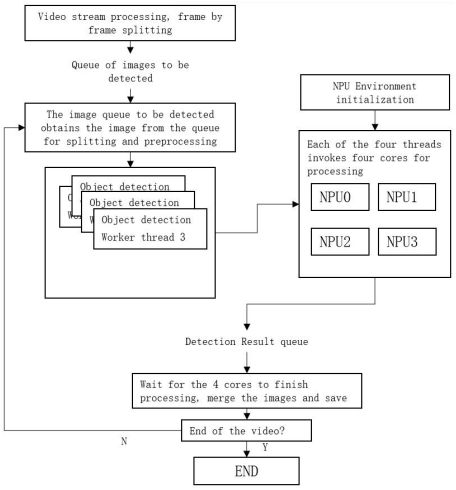

Fig.4.Data parallel object detection execution flow chart(video streaming)

The spatial parallel strategy is characterized by low single-frame delay in processing image pairs and clear division between scalar and tensor computing units. Therefore, the algorithm strategy has no restrictions on hardware resources and input data. However, the implementation efficiency of this strategy is limited by the processing speed of the pre-processing and image segmentation module, and the performance bottleneck is in the scalar processing unit (namely the core CPU module). Figure 5 shows the execution process of the spatial parallel strategy.

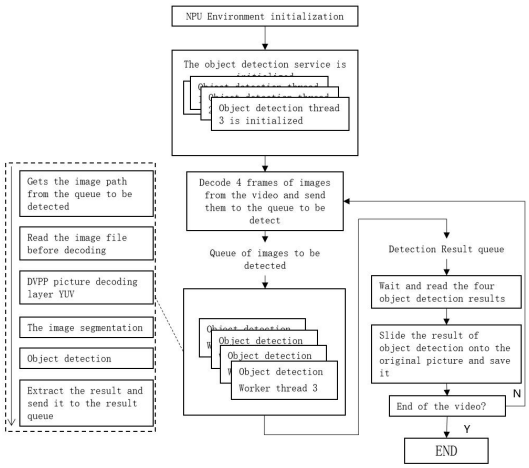

Fig.5. Spatial parallel object detection execution flow chart(video streaming)

4. Experiments and Analysis

This paper deploys the algorithm framework on the Huawei Atlas300I development board, which is equipped with 4 Shengteng 310 chips, each 310 chip contains a set of ARM CPU and NPU chip. In NP mode, the Atlas300 is connected to the workstation as a slave device through the PCIe interface, providing 4-core NPU resources. In SoC mode, the ARM CPU core on the Atlas300 can be used as an independent core CPU. Therefore, this paper chooses to deploy the data parallel algorithm strategy on the Ascend 310 chip for testing. The on-chip CPU can independently complete the pre-processing and post-processing

processes, which improves the utilization of on-chip resources and improves the parallel efficiency.

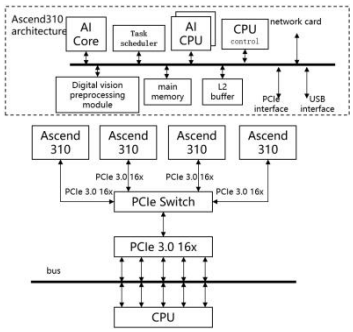

Fig.6. Atlas300I logical architecture

The following Table 1 shows the comparison of target detection efficiency under single-core and 4-core data parallelism. It can be seen that the use of spatial parallelism strategy significantly shortens the single-frame image processing delay. Due to the full use of the computing resources of 4 cores, the total time required to process 4 tasks in parallel is only 15%~20% higher than that of a single task. The execution performance of the 4NPU data parallel algorithm has been improved nearly linearly.

Object detection mode	Image resolution	Number of tasks	Reasoning time(ms)	total time (ms)
1NPU	3840x3392	1	6280	7170
4NPU Spatial parallel	3840x3392	1	1560	2370
4NPU Data parallel	3840x3392	4	6840	8730

Table.1.Performance evaluation of single NPU and 4NPU parallel strategies

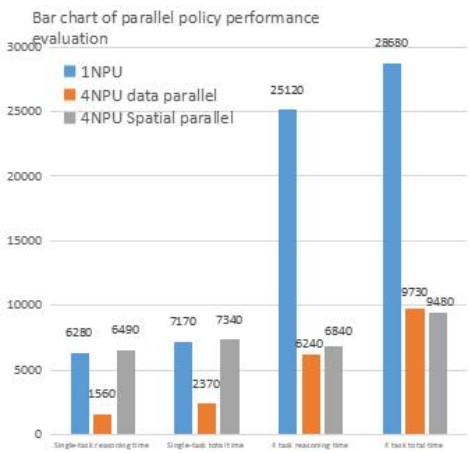

Fig.7. Performance evaluation bar charts of single NPU and 4NPU parallel strategies

Comparing the spatial parallel strategy and the data parallel strategy, it can be clearly seen that the advantage of the spatial parallel strategy is that the processing efficiency of a single frame image has been improved nearly linearly, but the tasks that require the core CPU such as image segmentation and pre-processing have become the performance bottleneck of the task. When processing video streams, the conflict of multi-frame images in CPU calls will also affect the processing efficiency of image algorithms. The advantage of the data parallel strategy is that when the amount of data to be processed is large enough, multi-frame images can be scheduled into 4 cores in a pipelined manner for image processing, which can make full use of multi-core resources for processing. The total task time has been improved nearly linearly. However, when the input data is not dense enough or the amount of data to be processed is small, the throughput of the algorithm will be greatly affected.

Conclusion

In order to effectively improve the practicability of target detection tasks of neural network algorithms in airborne environment and further improve the accuracy and execution efficiency of the algorithm, this paper proposes a high-precision target detection algorithm based on data parallelism and spatial parallelism. The algorithm is analyzed from the perspectives of image pre-processing, post-processing and parallel strategy selection, which also optimizes the algorithm currently deployed on a single core. In this paper, two parallel strategies are selected for experimental analysis. Future research can focus on deploying the algorithm in a more complex multi-core heterogeneous resource system.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geo.rey E Hinton. 2012. ImageNet Classication with Deep Convolutional Neural Networks[J]. In Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bo.ou, and K. Q.Weinberger(Eds.). Curran Associates, Inc., 1097–1105.

[2] Olga Russakovsky, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 115(3):211–252, 2015.

[3] Joseph Redmon and Ali Farhadi. 2016. YOLO9000: Be.er, Faster, Stronger[J]. CoRRabs/1612. 08242 (2016).

[4] Adam Van E.en, CosmiQ Works, In-Q-Tel. You Only Look Twice: Rapid Multi-Scale target Detection In Satellite Imagery[J]. 24 May 2018.

[5] V´ıt R°uˇziˇcka and Franz Franchetti,Department of Electrical and Computer Engineering, Carnegie Mellon University. Fast and accurate target detection in high resolution 4K and 8K video using GPUs[J]. 24 Oct 2018.

[6] Radway, R.M., Bartolo, A., Jolly, P.C. et al. Illusion of large on-chip memory by networked computing chips for neural network inference. Nat Electron 4, 71–80 (2021).