Ministry of Education of the Republic of Belarus
Educational Institution
“Belarusian State University of Informatics and Radioelectronics”

UDC 004.93`1004.93`1

Tang Yi

REAL-TIME OBJECT DETECTION ALGORITHM

ABSTRACT
For a Master of Science Degree

Advanced higher education
Speciality 1-40 80 02 System analysis, information control and processing (by industries)

Academic Supervisor
Alevtina B. Gourinovich
Ph.D., Associate Professor

Minsk 2024

INTRODUCTION

Video streaming is a continuous process of transmitting visual media that allows viewers to watch and analyze data in real-time while it is being transferred without interruption. This technology allows for the immediate reception of visual data from distant locations, such as outputs from surveillance cameras or live video inputs from events. Video streaming relies largely on compression techniques to facilitate rapid and efficient data transfer over a network, while aiming to uphold optimal image quality. Video streaming entails transmitting a sequence of static images, known as frames, which are played continuously. These frames are consistently sent at a predetermined frame rate to create a dynamic video. Typically, the standard frame rate ranges from 24 to 60 frames per second, although specific applications may utilize higher or lower frame rates.

A key application of video stream analysis is real-time object detection, and the development of real-time object detection techniques in computer vision has been a key driver of progress in several industries. Real-time object detection is a technique that swiftly and accurately identifies and locates multiple objects in a video stream, serving as a cornerstone in numerous applications. These include security surveillance, self-driving cars, face recognition, image retrieval, and augmented reality. Object detection algorithms aim to tackle challenges including multi-scale variations, occlusion, distortion, and category diversity. These algorithms must accurately recognize objects in different backgrounds, illumination conditions, and visual distractions. From a technology evolution perspective, object detection has transitioned from traditional image processing methods to deep learning-based approaches. Deep learning methods, particularly CNNs, have become the dominant technology in the field due to their powerful feature learning capabilities. They can automatically learn complex visual patterns from training data without the need for manual feature design. Deep learning methods for object detection can be broadly classified into two main groups: two-stage detection algorithms and one-stage detection algorithms. Two-stage algorithms, exemplified by R-CNN and its variants, initially generate region proposals and subsequently conduct classification and bounding-box regression. In contrast, one-stage algorithms like YOLO and SSD directly predict category probabilities and bounding box coordinates utilizing a single network model. These algorithms need to meet real-time performance requirements while ensuring high accuracy, as fast and timely detection results are crucial in application scenarios such as video surveillance or autonomous driving.

With the rapid advancement of IoT and mobile computing technologies, edge devices such as smartphones, drones, and various embedded systems have become ubiquitous in our daily lives and professional activities. However, these devices are inherently constrained by limited computational power, storage capacity, and energy availability. Consequently, deploying advanced, computation-intensive applications like object detection is a significant challenge. Therefore, research efforts have increasingly focused on minimizing model size and complexity without sacrificing performance. YOLOv4 represents a critical breakthrough in object detection, favored by many researchers and engineers for its outstanding detection speed and accuracy. Nevertheless, the high computational demand of the YOLOv4 model, necessary to maintain accuracy, restricts its deployment in resource-constrained settings.

This research endeavors to create object detection models that are both compact and efficient. Tailoring to the operational constraints of edge devices, the design of the model incorporates strategies such as low-precision computation and model compression to diminish arithmetic operations and storage demands. These strategies enable efficient object detection at endpoints and edge devices while optimizing resource utilization through effective modules and novel model architectures. The research enhances the YOLOv4 model through a series of deliberate refinements aimed at significantly reducing its parameter count and computational load with minimal accuracy trade-offs. Initially, we substitute the conventional backbone network with the lightweight, Transformer-based MobileViT, which effectively decreases model parameters while preserving performance. Additionally, the integration of parallel computation accelerates inference speeds, particularly for real-time object detection tasks. The adoption of self-attention mechanisms augments the model's global and local feature discernment. The implementation of depth-separable convolutions streamlines the detection head's architecture and bolsters training stability. In our data augmentation approach, we combine mosaic and mixup techniques to enhance sample diversity and minimize errors. An attention module is introduced to concentrate on pivotal regions, bolstering detection precision. We replace the standard IOU with CIOU to refine bounding box regression. Lastly, we recalibrate the loss function to equitably address different error types by utilizing BCE, CIOU, and focal loss. Through these collective enhancements, this article introduces the optimized algorithm MViT-YOLOv4, aiming for more efficacious and precise object detection on edge devices.

This article is organized as follows: It begins with a detailed analysis of the background in object detection technology and a review of current detection methods. It then illustrates the successful application of this technology in various domains, including autonomous driving, video surveillance, and industrial automation, and discusses the challenges and bottlenecks in implementing real-time detection on resource-limited devices. The core processes of deep learning are subsequently reviewed, along with state-of-the-art techniques in feature fusion and extraction. In the experimental section, the article assesses the performance of several lightweight feature extraction networks, focusing on parameters and classification accuracy. Additionally, it elaborates on the data augmentation techniques utilized during the evaluation process. Moreover, controlled experiments substantiate the efficacy of these techniques in enhancing model performance. To visualize the model's predictive capabilities, post-prediction image samples are presented, showcasing the model's impressive performance. The discussion section highlights the benefits of the proposed algorithm, emphasizing its ability to strike a balance between accuracy, speed, and resource utilization. It also acknowledges potential limitations and provides constructive insights for future research directions.

The improved algorithm presented herein achieves the fastest inference speeds with minimal resource use, despite a slight reduction in detection accuracy. This advancement expands the application range of edge computing devices and provides robust support for real-time and mobility-centric applications, such as drones, autonomous vehicles, and remote monitoring smart cameras. In sum, this study offers a thorough and insightful investigation with significant theoretical and practical implications for enhancing real-time object detection in resource-constrained environments.

THESIS GENERAL DESCRIPTION

The study was conducted at the Educational Institution “Belarusian State University of Informatics and Radioelectronics”.

Research aim

The thesis aim is to develop a lightweight object detection algorithm to apply in resource-constrained environments.

Subject: Real-time Object Detection Algorithm

Object: To develop and evaluate a real-time object detection algorithm that provides accurate and fast detection of objects in images or videos.

Aim: The aim of this master thesis is to propose and implement an efficient real-time object detection algorithm for various applications such as autonomous driving, surveillance systems, and robotics.

Dissertation Key Provisions:

1. Review of existing object detection algorithms and techniques.
2. Design and development of the real-time object detection algorithm.
3. Integration of mosaic and mixup techniques for data augmentation.
4. Training network with enhanced datasets on graphic card.
5. Evaluation of the algorithm's performance in terms of accuracy, speed and parameter size.
6. Comparison with state-of-the-art object detection methods.

Author’s contribution

Developing MViT-YOLOv4, integrates MobileViT as a lightweight backbone, parallel computation for accelerated inference, self-attention mechanisms for improved feature discernment, and depth-separable convolutions for streamlined architecture. Mosaic and mixup techniques are employed for data augmentation, while attention modules and CIOU bounding box regression refine detection precision. The loss function is recalibrated to address different error types effectively. Datasets are fed into network which was trained on a graphic card for one week, and then tested on the testing datasets, and then analyzed using some metrics.

Testing and implementation of results

The main provisions and results of the dissertation work were reported and discussed at:

- Information Technologies and Systems 2022 (ITS 2022);
- Информационные технологии и управление : материалы 59-ой научной конференции аспирантов, магистрантов и студентов. (Минск, 2023);
- Новые горизонты - 2022 : сборник материалов IX Белорусско-Китайского молодежного инновационного форума (Минск, 2022);
- Беларусь-Китай: контуры инновационно-технологического сотрудничества: сборник материалов научно-практической конференции (Минск, 2023)

Author's publications

According to the presented results of the thesis research 11 articles were published:

1. Tang ,Yi Small object detection method / Yi. Tang, A. Gourinovitch // Информационные технологии и системы 2022 (ИТС 2022) = Information Technologies and Systems 2022 (ITS 2022): материалы Международной научной конференции, Минск, 23 ноября 2022 / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: Л. Ю. Шилин [и др.]. – Минск : БГУИР, 2022. – С. 169–170.
2. Tang, Yi. Real-time object detection based on CNN / Yi Tang, Di Zhao // Новые горизонты - 2022 : сборник материалов IX Белорусско-Китайского молодежного инновационного форума, 10-11 ноября 2022 года / Белорусский национальный технический университет. – Минск : БНТУ, 2022. – Т. 2. – С. 264-265.
3. Zhao Di, Tang Yi. Human physical activity recognition system // Новые горизонты - 2022 : сборник материалов IX Белорусско-Китайского молодежного инновационного форума, 10-11 ноября 2022 года / Белорусский национальный технический университет. – Минск : БНТУ, 2022. – Т. 1. – С. 200-201.
4. Tang Yi Zhao Di. The adaptive boosting algorithm in biomedical image segmentation // Информационные технологии и управление: материалы 59-ой научной конференции аспирантов, магистрантов и студентов, Минск, 17–21 апреля 2023 года / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: Л. Ю. Шилин [и др.]. – Минск, 2023. – С. 56.
5. Tang Yi. Weakly supervised object detection method / Tang Yi, Zhao Di // Информационные технологии и управление : материалы 59-ой научной конференции аспирантов, магистрантов и студентов, Минск, 17–21 апреля 2023 года / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: Л. Ю. Шилин [и др.]. – Минск, 2023. – С. 64.

6. Tang Yi. The Prospects of Constructing Cooperative Relationship between China and Belarus under the “Belt and Road” Initiative / Tang Yi // Актуальные векторы белорусско-китайского торгово-экономического сотрудничества : сборник статей III международной научно-практической конференции, Минск, 16 декабря 2022 г. / МО РБ, Белорусский государственный экономический университет, Республиканский институт китаеведения имени Конфуция БГУ ; [редакционная коллегия: Ю. А. Шаврук (главный редактор) и др.]. – Минск : Колорград, 2023. – С. 153-159.

7. Tang, Yi. Di Zhao, A. Gourinovitch. Mask-embedding and feature-fused network for medical image segmentation // Беларусь-Китай: контуры инновационно-технологического сотрудничества: сборник материалов научно-практической конференции (Минск, 19-20 октября 2023 г.) // Республиканское инновационное унитарное предприятие «Научно-технологический парк БНТУ «Политехник»; сост. М. А. Войтешонок. – Минск: БНТУ, 2023. – С. 65-66.

8. Zhao, Di. Tang, Yi, A. Gourinovitch, A. Liankova. Exploring the role of loss functions in biomedical image segmentation // Беларусь-Китай: контуры инновационно-технологического сотрудничества : сборник материалов научно-практической конференции (Минск, 19-20 октября 2023 г.) // Республиканское инновационное унитарное предприятие «Научно-технологический парк БНТУ «Политехник» ; сост. М. А. Войтешонок. – Минск : БНТУ, 2023. – С. 66-67.

9. Tang Yi, Zhao Di Small object detection algorithms for large-scale scenes// Сборник материалов XIII Международной научно-практической конференции профессорско-преподавательского состава, аспирантов, магистрантов и студентов «Актуальные проблемы правовых, экономических и гуманитарных наук», 20 апреля 2023 года, БИП, г. Минск. с. 258-259

10. Tang Yi, Zhao Di Enhancing medical image segmentation through advanced data augmentation techniques // Сборник материалов XIV Международной научно-практической конференции профессорско-преподавательского состава, аспирантов, магистрантов и студентов «Актуальные проблемы правовых, экономических и гуманитарных наук», 18 апреля 2024 года, БИП, г. Минск (The article is in the publishing queue)

11. Di Zhao, Yi Tang, Gourinovitch A.B. Effective algorithm for biomedical image segmentation // Доклады БГУИР, Reports of BSUIR, Minsk (The article is in the publishing queue).

Thesis structure and size

The dissertation paper consists of introduction, general description of the paper, four chapters with conclusions after each chapter, conclusion, bibliography, one appendix.

The total amount of the thesis is _75__ pages, of which _52_ pages of text, 16 figures on 6 pages, 6 tables on 4 pages, a list of used bibliographic sources (46 titles on 4 pages), a list of the author's publications on the subject of the thesis (8 titles on 2 pages), 1 appendixes on 8 pages, graphic material on 3 pages.

Plagiarism

An examination of the dissertation «Real-time Object Detection Algorithm» by Tang Yi was carried out for the correctness of the use of borrowed materials using the network resource «Antiplagiat» (access address: https://antiplagiat.ru) in the online mode 08.04.2024. As a result of the verification, the correctness of the use of borrowed materials was established (the originality of the thesis is 86.27 %)

SUMMARY OF THESIS

The **introduction** addresses the problems of real-time object detection in resource-constrained environments, where conventional deep learning models face challenges due to computational limitations. It emphasizes the importance of efficient algorithms for object detection, particularly in the context of video streaming applications where real-time analysis is crucial. Furthermore, the introduction provides an overview of the evolution of object detection technology, from traditional image processing methods to deep learning-based approaches, highlighting the transition from manual feature design to automatic feature learning using CNNs. It also underscores the significance of real-time object detection in various domains such as surveillance, autonomous driving, and industrial automation, emphasizing the need for lightweight and efficient models for deployment on edge devices.

The **general description of work** establishes the connection between the research and the priority areas of scientific research, outlining the aim and tasks of the study. The research aims to enhance real-time object detection by developing a lightweight and efficient algorithm suitable for resource-constrained environments. The proposed algorithm, MViT-YOLOv4, leverages the MobileViT backbone, data augmentation techniques, attention mechanisms, and loss function modifications to achieve a balance between accuracy and efficiency. The study contributes to the field of computer vision by addressing the challenges of deploying deep learning models on edge devices and facilitating real-time object detection in practical scenarios.

In the first chapter, we embark on an exhaustive exploration of object detection within resource-constrained environments. Kicking off with an elucidation of artificial intelligence's evolutionary journey, we trace its trajectory from conceptualization to practical realization, particularly within the realm of computer vision. This historical backdrop sets the stage for a comprehensive review of contemporary object detection methodologies, wherein we categorize them into two distinct paradigms: two-stage and one-stage approaches. Delving deeper, we unearth the inherent trade-offs between these methodologies; while two-stage methods boast unparalleled accuracy and resilience, they often falter under computational constraints. Conversely, one-stage approaches prioritize speed but may compromise accuracy in their pursuit. As we navigate through the diverse application landscapes of real-time object detection across industries, a glaring need for lightweight algorithms tailored to edge computing environments emerges. Finally, we delineate a specific endeavor: the enhancement of the YOLOv4 algorithm. Through the strategic integration of MobileViT backbone, data augmentation techniques, attention mechanisms, and refined loss functions, we forge ahead to conceptualize the MViT-YOLOv4 model—a beacon of hope for achieving the elusive balance between accuracy and efficiency in real-time object detection within resource-constrained settings.

In the second chapter, it begins with an exploration of deep learning as a field, categorizing learning tasks into supervised, semi-supervised, and unsupervised categories. The chapter then delves into the theoretical foundations of deep learning, covering essential concepts such as the MLP, activation functions, optimizers, and backpropagation. Additionally, it discusses multiscale fusion techniques, attention mechanisms, and their roles in enhancing the capabilities of deep learning models. Finally, the chapter focuses on lightweight feature extraction networks, tracing their evolution from hand-designed descriptors to automatic feature learning using deep CNNs. It highlights the significance of these lightweight networks in resource-constrained environments and emphasizes their ability to balance model size, computational efficiency, and performance. Furthermore, the chapter underscores the importance of transfer learning and multi-task learning in enhancing the adaptability and versatility of feature extraction networks across different domains and tasks.

In the third chapter, a detailed exploration of the experimental results and their implications takes center stage. The chapter commences with an exposition of the architectural intricacies of the MViT-YOLOv4 algorithm, elucidating how it seamlessly amalgamates the efficiency of YOLOv4 with the transformative capabilities of the MobileViT backbone. Attention is devoted to the meticulous selection and preprocessing of datasets, ensuring their reflective of real-world complexities and conducive to effective model training. Experimental setups are thoroughly detailed, encompassing parameters, configurations, and hardware considerations, to ensure the reproducibility and reliability of results. Evaluation metrics such as precision, recall, and mAP serve as the yardstick for assessing the algorithm's performance, providing quantifiable insights into its accuracy and efficiency. Through rigorous experimentation and analysis, the chapter unearths the strengths and limitations of the MViT-YOLOv4 algorithm, thereby paving the way for future advancements and applications in the domain of real-time object detection in resource-constrained environments.

CONCLUSION

This thesis is mainly based on the YOLOv4 algorithm for research and improvement. Overall, the improvement is mainly divided into the following aspects:

1. Replacing the backbone network with MobileViT. MobileViT, as a Transformer-based lightweight visual model, is characterized by high efficiency and light weight, which can reduce the number of parameters while maintaining high model performance. Second, MobileViT's Transformer structure allows parallel computation, which accelerates the inference process of the model and is particularly suitable for real-time object detection tasks. In addition, MobileViT utilizes a self-attention mechanism to learn the relationship between global and local features in an image, which improves the accsuracy and robustness of object detection. Meanwhile, the detection head uses depth-separable convolution with fewer parameters, which is easy to train and adjust, making the training process more stable and reliable.

2. In terms of data processing, two data enhancement techniques, fusion Mosaic and Mixup are used. Specifically, the probability of fusion data enhancement is set to 0.5, while the Mixup data enhancement technique is further applied to the fusion-enhanced data, and the probability is also set to 0.5. In this way, more diversified training samples can be generated, which can help the model better adapt to different scenes and data changes. At the same time, it can also reduce the error introduced by fuzzy labeling and enhance the interference of complex background on detection.

3. Adding the attention module makes the network pay more attention to regions of interest and less attention to useless regions when processing images, which makes the model have stronger detection performance. Meanwhile, in the detection head part, the depth separable convolution is used to replace the traditional convolution to reduce the number of parameters to accelerate the model computation.

4. Optimize the model loss function and use CIOU instead of IOU , CIOU takes into account the distance, overlap, scale, and penalty between the object and the anchor point, which makes the object bounding box regression more stable, and in the process of training, no problems such as scatter, which has better results compared to IOU and GIOU. In addition, the composition of the loss function is modified. For the classification loss, the model uses BCE loss, which can effectively handle multi-label classification tasks. The regression loss is computed using CIOU, which considers the similarity between the prediction bounding box and the truth bounding box, as well as their sizes, aspect ratios, and overlaps. object confidence loss is determined using focal loss, giving higher weight to

difficult samples and helping the model to focus on challenging regions. By appropriately combining the classification, object and regression losses, the proportion of weights for the different losses is calculated.

In summary, this paper focuses on the research of lightweight object detection algorithms, and successfully proposes and implements the MViT-YOLOv4 algorithm in response to the limitations of the computing power of terminal devices and the urgent need for the response speed of detection algorithms. Through rigorous experimental validation on the Pascal VOC2007+2012 dataset, this article fully demonstrates the superior prediction capability of the model. The quantitative experimental results show that while maintaining the lightweight advantage, MViT-YOLOv4 achieves significant improvement in accuracy compared to existing lightweight models. And compared with traditional heavyweight models, a significant reduction of model parameters is realized with only a minor loss of detection accuracy. The results of the qualitative experiments visualize the predictive ability of the model and its ability to focus attention on key regions by showing the processed images and heat maps. In addition, through comparative experiments, this article also verifies the effectiveness of the data enhancement strategy after combining mosaic and mixup techniques in improving the model performance.

The proposed MViT-YOLOv4 model contributes significantly to the advancement of the lightweight object detection field with its excellent performance and practicality.