Thesis for the Degree of Masters in Science

Gated YOLOv6 Dynamic Channel Gating for Efficient Object Detection

School of Computer Science and Engineering

The Graduate School

Hector Andres Acosta Pozo

June, 2024

The Graduate School
Kyungpook National University

Gated YOLOv6 Dynamic Channel Gating for Efficient Object Detection

Hector Andres Acosta Pozo

School of Computer Science and Engineering

The Graduate School

Supervised by professor Soon Ki Jung

Approved as a qualified thesis of Hector Andres Acosta Pozo

for the degree of Masters of Science

by the Evaluation Committee

June, 2024

Chairman: Prof. AAA ____________________

Prof. BBB ____________________

Prof. CCC ____________________

Prof. DDD ____________________

Prof. EEE ____________________

The Graduate School

Kyungpook National University

Contents

Acknowledgement

I would like to express my deepest appreciation to Professor Soon Ki Jung for his invaluable guidance, patience, and expertise throughout my research. His insights and encouragement were crucial to the success of this work.

I am deeply grateful to my family, whose unwavering support and belief in me have been my constant source of strength and motivation. To my mother, Soneyda Pozo, and my father, Andres Acosta, thank you for your endless love, encouragement, and sacrifices, which have not gone unnoticed. Your belief in my abilities has been a significant driving force in my pursuit of academic excellence.

I also wish to extend a special thanks to my sister, Diana Acosta, for her support, understanding, and companionship. Your presence and belief in my work have been a great source of comfort and encouragement.

To all of you, I am eternally grateful for your support and love. Thank you for being my guiding lights.

List of Tables

List of Figures

Abstract

In the evolving landscape of object detection, the advent of the YOLO (You Only Look Once) methodology has marked a significant leap forward for real-time applications, streamlining the process with its single-pass detection capabilities. Despite its advancements, the dynamic and often unpredictable nature of real-world environments, especially when operating on edge devices like security cameras, presents a pressing challenge for optimization. Addressing this, our research introduces the ”Gated Scene-Specific YOLO,” a novel adaptation of the YOLO framework, incorporating a dynamic gating mechanism aimed at enhancing computational efficiency without compromising the model's detection prowess.

Traditional YOLO architectures, while robust, are predisposed to processing vast amounts of data, much of which may be extraneous or irrelevant to the task at hand. This not only leads to unnecessary computational overhead but also hinders the deployment of such models in environments where resources are limited. Our Gated Scene-Specific YOLO methodology seeks to alleviate this issue by integrating a mechanism that dynamically adjusts the activation of neural pathways based on the relevance to the observed scene. Through a meticulous process of gate generation and analysis during the training phase, our approach identifies and deactivates neural pathways that are consistently inactive across specific environmental conditions. This strategic deactivation allows the model to shed redundant computational weight, thus becoming more streamlined and efficient for the task it is deployed to perform.

The core of our research lies in demonstrating the practicality of dynamically

tuning deep learning models to their operational context, significantly reducing the computational load while maintaining, and in some cases enhancing, detection accuracy. Our empirical results showcase that the Gated Scene-Specific YOLO not only elevates processing speeds but also upholds a high standard of accuracy, making it a compelling solution for real-time object detection across a diverse array of settings. This contribution is particularly relevant for the deployment of object detection models in resource-constrained devices, where optimizing the balance between efficiency and performance is paramount.

In summary, the Gated Scene-Specific YOLO represents a meaningful stride towards more adaptable and resource-efficient object detection solutions. By tailoring model processing pathways to the specific demands of the environment, this research paves the way for the development of highly optimized, context-aware deep learning models, thereby enhancing the applicability and effectiveness of real-time object detection systems in dynamic and varied settings.

1 Motivation and Objectives

1.1 Introduction

Object detection stands as a foundational pillar within the field of computer vision, influencing an extensive range of applications from advanced surveillance systems to the dynamic realm of autonomous driving technologies. The rapid evolution of deep learning methodologies has significantly propelled the field forward, introducing architectures capable of not only enhancing the accuracy of object detection but also its efficiency and adaptability in real-time processing environments. Among these innovations, the YOLO (You Only Look Once) architecture, introduced by Redmon et al., has emerged as a seminal contribution, revolutionizing the way real-time object detection is approached by enabling swift and efficient processing without the need for iterative detection stages. This architecture underscores the potential of modern object detection systems, demonstrating remarkable versatility across a variety of computing environments, from high-powered servers to more constrained devices such as smartphones and embedded systems.

Despite the advancements brought forth by YOLO and its subsequent iterations, challenges remain, particularly in the context of edge computing where computational resources are limited, and the demand for real-time processing is paramount. Recognizing the potential for further optimization, our study delves into the exploration of YOLOv6, a variant designed with a keen focus on hardware efficiency and optimized for real-time applications. The architecture of YOLOv6 serves as an ideal foundation for our investigation, given its emphasis on performance in hardware-constrained environments. Within such contexts, minor enhancements can lead to substantial gains in processing speed and overall system efficiency, thereby improving the applicability of YOLO-based models in edge devices.

To advance the capabilities of YOLOv6, our research introduces the ”Gated Scene-Specific YOLO,” a novel framework that marries the concept of dynamic gating with the principle of model pruning specifically tailored to the YOLO architecture. Model pruning, a technique aimed at reducing the computational burden of neural networks, achieves this by eliminating superfluous or insignificant parameters, thereby refining the network's structure with minimal detriment to its performance. Our approach innovates beyond traditional model pruning by implementing a dynamic gating mechanism that adjusts in real-time to the distinctive characteristics of the input scene, thereby optimizing the efficiency of the object detection process without sacrificing accuracy.

A cornerstone of our methodology is the adaptation of Improved SemHash, a technique initially proposed by Kaiser and Bengio and further refined by Chen et al. This approach facilitates the generation of binary gates during the model training phase, enabling selective activation or deactivation of network filters in response to variations in the input. Through dynamic gate generation and subsequent analysis tailored to specific scenes, our method identifies filters that consistently remain inactive. These filters are then statically pruned from the network for deployment, allowing the model to operate more efficiently by focusing computational resources on active, scene-relevant pathways. The Gater Network, integral during the training phase for gate determination, is thus rendered unnecessary during actual deployment, replaced by the pre-determined, statically applied gates that ensure both efficiency and specificity in detection.

Our contributions to the YOLO architecture, through the integration of a gating network and the innovative use of Improved SemHash, not only enable precise control over network activity but also herald significant improvements in computational efficiency and detection accuracy. The effectiveness of our approach is substantiated through rigorous experimental validation, focusing on key performance metrics such as floating-point operations per second (FLOPs), frames per second (FPS), and mean Average Precision (mAP@0.5:0.95). Our findings reveal a notable increase in FPS for the Gated Scene-Specific YOLO model in comparison to its YOLOv6 counterpart, without any compromise on detection robustness, as evidenced by stable mAP scores. This research thus presents a significant step forward in optimizing deep learning models for object detection, especially in scenarios where computational resources are at a premium.

1.2 Motivation

The relentless pursuit of advancements in computer vision, specifically within the domain of object detection, has been driven by the escalating demands of modern applications. These range from enhancing public safety through sophisticated surveillance mechanisms to propelling the future of mobility with autonomous vehicles. The inception of deep learning architectures like YOLO has significantly narrowed the gap between theoretical possibility and practical implementation, offering a glimpse into the potential of real-time object detection systems. Yet, as these technologies are increasingly deployed in real-world scenarios, particularly on the edge, the limitations of current models under resource-constrained conditions become apparent. The motivation behind our work is rooted in the desire to transcend these limitations, pushing the boundaries of what is possible with existing object detection frameworks.

Our focus on YOLOv6, known for its balance of speed and accuracy, stems from a recognition of the critical need for optimization in edge computing scenarios where resources are scarce yet the demand for high-performance computing is incessant. The drive to refine and enhance the efficiency of such models without compromising their detection capabilities underlines our research. We are particularly inspired by the potential impact of our work on a wide array of applications, from low-power IoT devices to mobile applications requiring real-time analysis and feedback, envisioning a future where advanced object detection is not only possible but also practical and pervasive, regardless of computational limitations.

1.3 Objectives

The primary objective of our research is to develop an optimized version of the YOLOv6 architecture, termed ”Gated Scene-Specific YOLO,” which incorporates a dynamic gating mechanism to enhance computational efficiency in object detection tasks, particularly in edge computing environments. To achieve this, we aim to:

Implement Dynamic Gating: Integrate a dynamic gating mechanism that adapts to the unique characteristics of input scenes, thereby selectively activating relevant neural pathways and improving model efficiency. Optimize Through

Model Pruning: Apply model pruning techniques in conjunction with dynamic gating to eliminate redundant parameters and streamline the model, focusing computational resources on critical tasks. Leverage Improved SemHash: Utilize the Improved SemHash technique for effective gate generation during training, allowing for precise control over network activity and further optimization of the model for specific scenarios. Demonstrate Practical Efficacy: Validate the effectiveness of the Gated Scene-Specific YOLO model through extensive testing, focusing on key performance metrics such as processing speed (FPS), computational efficiency (FLOPs), and detection accuracy (mAP@0.5:0.95). Enhance Real-World Applicability: Ensure that the optimized model maintains high detection accuracy while significantly reducing computational load, making it suitable for deployment in real-world, resource-constrained environments. Through these objectives, our research seeks to address the pressing challenges of deploying sophisticated object detection models in edge computing scenarios, offering a pathway to more efficient, accurate, and accessible real-time object detection technologies.

2 Related Work

The exploration of efficiency within neural network architectures, particularly for object detection in computationally constrained environments, constitutes a significant area of research. This section delves into various methodologies and developments that have shaped the current landscape of efficient neural network design, highlighting the relevance and novelty of our approach within this context.

2.1 Sparsity and Conditional Computation

A fundamental concept in the pursuit of neural network efficiency is the integration of sparsity and conditional computation mechanisms. Sparsity in neural networks refers to the idea that not all neurons or connections (weights) are necessary for every input. By identifying and activating only a subset of the neural pathways for a given input, significant reductions in computational overhead can be achieved without sacrificing the model's ability to represent complex functions. This principle is akin to how decision trees operate, where at each node a decision is made that leads to a subset of the next possible states, thus not exploring all branches of the tree for a given input.

The concept of conditional computation extends this idea further by introducing mechanisms that allow a neural network to adapt its computation pathways dynamically based on the input. This adaptability ensures that only the most relevant parts of the network are engaged during the forward pass, which not only saves computational resources but also helps in reducing overfitting by limiting the effective capacity of the model based on the complexity of the input.

Introduced by Bengio et al. **bengio2013**, the idea of selectively activating neural pathways has been a cornerstone in the development of more efficient neural network architectures. The authors suggest that such mechanisms could allow for deeper and more complex models by allocating computational resources more judiciously. By simulating sparsity and conditional computation, networks can potentially achieve a balance between depth and width, optimizing the computational cost without compromising the benefits of distributed representations.

Several approaches have been proposed to implement sparsity and conditional computation in neural networks. One method involves pruning, where connections between neurons are removed based on their importance to the model's performance. Another approach is the use of gating mechanisms, where gates control the flow of information in the network, effectively turning on or off certain pathways based on the input.

Furthermore, research in this area has explored the use of dropout as a form of introducing dynamic sparsity during training, encouraging the network to develop more robust and efficient representations. Additionally, recent advancements have introduced techniques such as dynamic routing between capsules in Capsule Networks, where the network learns complex spatial hierarchies in data by activating only the relevant parts of the network for a given input.

In summary, the integration of sparsity and conditional computation into neural networks offers a promising avenue for enhancing computational efficiency. By leveraging these concepts, researchers aim to develop models that can process information more effectively, adapting their structure to the demands of the input while preserving the representational power that characterizes deep learning. This ongoing exploration holds the potential to significantly impact the development of neural network architectures, especially in contexts where computational resources are limited.

2.2 Dynamic Sparse Training

Dynamic Sparse Training (DST), as proposed by Liu et al. **liu2017**, represents a significant leap forward from static sparsity models towards a more flexible and efficient neural network architecture. DST addresses one of the primary concerns in neural network efficiency—how to use the minimal number of parameters without sacrificing the model's ability to learn complex patterns. Unlike traditional training methods that rely on a fixed architecture, DST allows the network to adjust its architecture dynamically during the training process. This is achieved by periodically redistributing the network's connections to focus more on those that are most beneficial for learning the current task.

The key innovation of DST lies in its ability to identify and leverage the most informative connections within a network based on the training data. By doing so, it optimizes both the model capacity and computational resources, directing them towards the aspects of the data that are most crucial for performance. This method not only improves the efficiency of the network but also has the potential to enhance its generalization ability by preventing overfitting to less relevant features. However, implementing DST effectively requires sophisticated mechanisms for deciding when and how to adjust the network's sparsity. This dynamic adjustment process introduces additional complexity and computational overhead, which can be challenging to manage, particularly in environments where computational resources are strictly limited.

2.3 Dynamic Filter Selection

Building upon the concept of dynamic adjustment in neural networks, Chen et al. **chen2018** introduced an approach specifically designed for Convolutional Neural Networks (CNNs), dubbed GaterNet. This method revolutionizes the way CNNs are structured by implementing a dynamic filter selection mechanism that adapts in real-time to the input. GaterNet operates by employing a gating network that runs parallel to the main network. The gating network analyzes the input and determines which filters in the main network are most relevant for processing that particular input. By activating only a subset of the filters, GaterNet significantly reduces the computational load required for each forward pass through the network.

The brilliance of dynamic filter selection lies in its capacity to maintain high levels of accuracy while dramatically enhancing computational efficiency. This is especially pertinent for applications requiring real-time processing, such as video analysis or mobile computing, where resource constraints are a major consideration. Moreover, the adaptability of GaterNet enables the CNN to focus its computational power on the most informative features of the input, potentially leading to better performance on complex tasks. Despite its advantages, the design and training of a gating network that can accurately predict the most effective filters for any given input pose substantial challenges, requiring careful consideration of the trade-offs between complexity, efficiency, and accuracy.

2.4 Neural Network Pruning

Parallel to dynamic selection techniques, neural network pruning strategies focus on reducing network complexity without severely impacting performance. Zhang et al. **zhang2018** presented a methodical approach to pruning using gradient descent, streamlining the network in a manner that minimizes performance degradation. This technique aligns with the overarching goal of developing efficient, robust neural network models suitable for a wide range of applications.

2.5 Efficient Object Detection Models

The quest for efficiency extends into the domain of object detection, with MobileNet YOLO by Howard et al. **howard2017** setting a precedent. By leveraging depth-wise separable convolutions, MobileNet YOLO offers a significant reduction in computational demands while sustaining performance levels, demonstrating the viability of real-time object detection on embedded systems. This integration of MobileNet with the YOLO framework exemplifies the practical application of efficiency-driven design principles in object detection models.

2.6 Adverse Conditions Object Detection

Emerging research by Kalwar et al. **kalwar2023**, titled 'GDIP: Object-Detection in Adverse Weather Conditions Using Gated Differentiable Image Processing,' explores object detection under challenging environmental conditions. Utilizing a gated image processing technique, this work provides a novel perspective on efficiency and adaptability, presenting a potential area for future comparative studies with our gated scene-specific approach.

2.7 Our Contribution

Building upon these foundational advancements, our work introduces the "Gated Scene-Specific YOLO," which incorporates dynamic gating and model pruning tailored specifically for the YOLO architecture. Unlike existing methodologies, our approach leverages Improved SemHash to establish static gating configurations post-training, significantly reducing computational requirements during inference. This innovation sets our work apart in the pursuit of efficient, real-time object detection, particularly in environments where computational resources are at a premium.

3 Methodology

4 Experiments and Results

5 Conclusion

Appendices

A

B

Abstract in Korean