TU WIEN Informatics

Attentional Neural Network based Dynamic Object Detection for Autonomous Multi-Agent Systems

Master Thesis Proposal

by

Daniel Scheuchenstuhl, BSc
Registration Number 01630368

Institute of Computer Engineering
Cyber-Physical Systems
Vienna University of Technology

September 21, 2022

Advisor: Univ.Prof. Dipl.-Ing. Dr.rer.nat. Radu Grosu
Assistance: Univ.Ass. Dott.mag. Luigi Berducci

1 Motivation & Problem Statement

Highly automated driving (HAD) agents arise from the third level of driving automation which require no human intervention for a car to be able to drive [8]. For this purpose, these agents must conform with complex system specification requirements such that an implementation of a reliable and safe autonomous driving system may be guaranteed. These system specification requirements include a robust communication network within the car's heterogeneous system as well as a solid perception of the environment through the car's sensory network. Especially a reliable and broadly covered perception of the environment is of major interest allowing to estimate a close-to-reality state of the environment and the best current action to be taken with respect to control. Most importantly, it enables the car to detect and recognize other participants and objects in the environment and react accordingly. Processing the image perceived by the car's camera yields promising results in order to reliably detect and recognize other participants and objects in the environment as well as further allows to estimate their pose using a stereo vision system. Classical state-of-the-art approaches in computer vision such as Scale-invariant feature transform (SIFT) or Hough Transform are cumbersome or quite computationally expensive though, practically difficult to implement for HAD agents beyond simulation [11] [6].

In order to overcome this performance issue while still achieving equal or even better results in terms of object detection and recognition accuracy, an alternative approach had to be tackled. This new approach lead to the emergence of Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) for object detection [2] [13]. Despite their rapid increase in performance once trained, DNNs require vast amounts of data due to their complex data models in order to prevent them from underfitting. Most CNNs are pre-trained on large auxiliary data sets designed for the object classification task favoring translation invariance which is not ideal for object detection. Generally, CNNs tend to have problems classifying images with different positions as well as they do not have coordinate frames which are a basic component of human vision.

Inspired by the humans attention mechanism used by the brain to reconfigure its focus on global and specific local information, Vaswani et. al. [17] proposed a new neural network architecture called Transformer which implements a self-attention mechanism and is used for machine translation tasks where it outperforms any model from the literature. The Transformer architecture has also been successfully applied to computer vision applications following [1] [3] where astonishing results have been achieved. Yet, the training and optimization of those architectures remains a challenge. The self-attention mechanism has further been successfully used for applications in computer vision following [7] [18]. However, due to the lack of ground truth annotations for the glimpse locations and shapes following [7] and the basic division and coupling of local and global information following [18], the question arises if the implemented concepts of attention for the object detection task can further be improved and in particular be efficiently integrated in a HAD agent context.

2 Aim of the thesis

In the context of this master thesis, a more advanced concept of attention in comparison to [17] based on human attention will be learned upon which an associated Attentional Neural Network (ANN) based dynamic object detection system for a multi-agent racing context will be developed. The ANN based dynamic object detection system will be further integrated in the Robot Operating System (ROS) in order to establish a collaboration with HAD agents. Specifically, human attention in a racing/driving context will be learned for which an ANN based dynamic object detection system for autonomous multi-agent racing is designed and trained. The ANN will be implemented in Python and the incorporated attention filter will be trained with a data set primarily focusing on the humans attention when driving given various scenarios. For the purpose of imitating human attention, the dataset is created by monitoring various drivers when manually driving F1Tenth cars [1] in different scenarios using eye tracking glasses. In order to monitor the attention of the subjects, a setup composed of a camera live-stream from a selected F1Tenth car and a steering wheel as well as a gas pedal for steering and controlling the velocity profile of the given F1Tenth car is built. All of this work necessary for the creation of the attention dataset used for training, validation and testing of the attention filter is done

[1] https://f1tenth.org/index.html

in collaboration with my colleagues Felix Resch and Stefan Ulmer.

Once the attention filter has been designed and trained, the output of the attention filter is used to determine the current Region of Interest (RoI) for the object detection task. In a second step, the bounding box regression and classification of the objects in the RoI is performed. The ANN based dynamic object detection system will be validated and tested on an offline stream of the F1Tenth car's local camera in simulation first. Later on, the ANN based dynamic object detection system will be implemented in a ROS node to be able to collaborate with HAD agents. Finally, the resulting ANN based dynamic object detection system will be tested and used with F1Tenth cars in a multi-agent scenario.

Summarized, a proof of concept Attentional Neural Network (ANN) imitating human attention shall be designed and learned, capable of providing suitable Regions of Interest (RoI) for the object detection system such that objects that occur in a multi-agent racing context can be detected reliably and efficiently. In the end of this thesis, the following research questions shall be answered:

RQ1: How to design and learn a neural network imitating human attention for the object detection task in a multi-agent racing context?

RQ2: How does the performance and object detection accuracy compare to neural networks used for object detection without human attention?

RQ3: How does the performance and object detection accuracy compare to state-of-the-art approaches?

3 Methodology

This thesis applies a design and creation research strategy and builds upon academic literature as well as relevant specifications [14]. Specifically, the research method used for this thesis is implemented in the following steps:

1. A literature review is performed to establish the theoretical basis of the thesis.
2. Design of the ANN based Dynamic Object Detection System
   (a) Use eye-tracking devices in order to collect data of where humans look when driving given various driving scenarios.
   (b) Design and train an attention filter with the gathered data in order to imitate human attention when driving.
   (c) Incorporate the learned attention filter with a trained object detection neural network in an ANN based dynamic object detection system.
   (d) The designed ANN based dynamic object detection system will originally be implemented in Python.
3. The ANN based dynamic object detection system will be integrated in ROS to be able to be used in collaboration with HAD agents on real-world racecars.
4. Once the validation of the system was successful, simulation experiments as well as experiments on F1Tenth hardware cars will be performed.
5. Finally, the results of the experiments will be analyzed, interpreted and discussed.

4 Structure of the Thesis

We report a tentative structure of the thesis work:

1. Introduction
2. Scientific Background
   (a) Related Work
   (b) Technical Background
3. Scientific Methodology
4. Design of the ANN based Dynamic Object Detection System
   (a) Development of the Attention Filter
   (b) Development of the Object Detection Neural Network
   (c) Structure and Deployment of the ANN based Dynamic Object Detection System
5. Evaluation of the Dynamic Object Detection System based on an ANN
   (a) Single-Agent/Multi-Agent Simulation Evaluation
   (b) Multi-Agent Evaluation on F1Tenth Hardware Cars
6. Discussion of the Dynamic Object Detection System based on an ANN
7. Conclusion

5 State of the Art

Object detection is one of the fundamental problems of computer vision. With the rise of artificial intelligence and machine learning, a vast amount of techniques and methods based on DNNs and CNNs have been proposed such as [2] [13]. Most state-of-the-art object detection algorithms treat object detection as a regression problem, where classification and localization of the objects in the image are performed simultaneously. In general, state-of-the-art object detection algorithms can be categorized into one-stage methods such as You Only Look Once (YOLO) [10] and Single Shot Detector (SSD) [12] and two-stage methods such as Regions with CNN features (R-CNN) [5], Fast/Faster RCNN [4] [15] and R-FCN [9] depending on whether region proposals are computed. While two-stage methods provide a higher object detection accuracy, one-stage methods are superior in inference time. Moreover, Lin et. al. [16] proposed a simple and powerful framework for building feature pyramid networks to be used inside cnns for an efficient generation of region proposals.

Based on the major advances in Natural Language Processing (NLP), document summarization and machine translation tasks by implementing the concept of self-attention following [17], Attention has also been considered and successfully applied in computer vision tasks such as object detection [7] [18]. Specifically, Hara et. al. [7] proposed an augmenting deep neural network with an attention mechanism for visual object detection. When comparing to the Fast R-CNN [4], a consistent performance improvement of the incorporated attention mechanism to the DNNs can be determined. Furthermore, Zhu et. al. [18] proposed a fully convolutional network, named as Attention CoupleNet, to incorporate the attention-related information and global and local information of objects to improve the detection performance. The Attention CoupleNet achieves state-of-the-art performance on the PASCAL VOC and COCO datasets for object detection. Moreover, the Transformer architecture has also been successfully applied to computer vision applications such as image recognition [3] and object detection [1]. Both approaches achieve competitive results in their respective domain of application.

However, all of the discussed methods that incorporate an attention mechanism implement the basic concept of self-attention while a more advanced attention concept imitating human attention may proof beneficial in terms of training time, optimization and performance. Additionally, for most object detection approaches, the PASCAL VOC and COCO benchmarks are the defacto standard in object detection performance comparison while they cannot be used as measure for quantifying the object detection performance in an autonomous driving context.

6 Relevance to the Curricula of Computer Engineering

The proposed thesis is related to as well as part of research areas such as computer vision, artificial intelligence, robotics and autonomous driving. Therefore, a study program primarily focusing on the design and development of rigorous cyber-physical systems and deepening the knowledge in those specific research areas fits best to design, implement and evaluate the proposed idea in a master thesis accordingly. Based on the curriculum for the master's program in computer engineering, the key areas Automation, Digital Circuits and Systems and Cyber-Physical Systems have been chosen. Most of the attended courses are either related to topics in machine vision, artificial intelligence, robotics or autonomous driving including:

- 376.054 Machine Vision and Cognitive Robotics
- 182.763 Stochastic Foundations of Cyber-Physical Systems
- 183.660 Mobile Robotics
- 191.119 Autonomous Racing Cars
- 182.762 Project Computer Engineering, where I participated in the IROS 2021 in Prague
- 182.753 Internet of Things

Thus, the study program perfectly conforms and relates to the topics of the proposed thesis.

References

[1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020.

[2] D. E. Christian Szegedy, Alexander Toshev. Deep neural networks for object detection. 2013.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[4] R. Girshick. Fast r-cnn, 2015.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2013.

[6] A. Goldenshluger and A. Zeevi. The Hough transform estimator. *The Annals of Statistics*, 32(5):1908 – 1932, 2004.

[7] K. Hara, M.-Y. Liu, O. Tuzel, and A. massoud Farahmand. Attentional network for visual object detection, 2017.

[8] S. International. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. 2021.

[9] K. H. J. S. Jifeng Dai, Yi Li. R-fcn: Object detection via region-based fully convolutional networks. 2016.

[10] R. G. A. F. Joseph Redmon, Santosh Divvala. You only look once: Unified, real-time object detection. 2016.

[11] E. Karami, M. S. Shehata, and A. J. Smith. Image identification using SIFT algorithm: Performance analysis against different image deformations. *CoRR*, abs/1710.02728, 2017.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot MultiBox detector. In *Computer Vision – ECCV 2016*, pages 21–37. Springer International Publishing, 2016.

[13] Q. Lu, C. Liu, Z. Jiang, A. Men, and B. Yang. G-cnn: Object detection via grid convolutional neural network. *IEEE Access*, 5:24023–24031, 2017.

[14] B. J. Oates. *Researching Information Systems and Computing*. SAGE Publications, Ltd., 2012 edition, 2005.

[15] R. G. J. S. Shaoqing Ren, Kaiming He. Faster r-cnn: Towards real-time object detection with region proposal networks. 2015.

[16] R. G. K. H. B. H. Tsung-Yi Lin, Piotr Dollar and S. Belongie. Feature pyramid networks for object detection. 2016.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[18] Y. Zhu, C. Zhao, H. Guo, J. Wang, X. Zhao, and H. Lu. Attention couplenet: Fully convolutional attention coupling network for object detection. *IEEE Transactions on Image Processing*, 28(1):113–126, 2019.