DOMAIN ADAPTATION FOR OBJECT DETECTION IN SPORTS

DAVIDE MICHELON

UNIVERSITI KUALA LUMPUR - MIIT

2023

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
NN	Neural Network
ROI	Region Of Interest
RPN	Regional Proposal Network
YOLO	You Look Only Once
UDA	Unsupervised Domain Adaptation

CHAPTER ONE

INTRODUCTION

1.1 INTRODUCTION

Object detection is a computer vision task that involves identifying and localizing objects of interest within an image. It can be used in various applications such as surveillance systems, robotics, and autonomous. The development of AlexNet – first general-purpose feature extractor CNN model - marked a major paradigm shift in computer vision for object detection (Krizhevsky et al., 2012). Recently, the performance of object detection has been significantly improved thanks to the use of advanced NN architectures such as YOLO (Joseph Redmon et al., 2016) and Faster R-CNN (Shaoqing Ren et al., 2016). However, the accuracy of object detection models heavily relies on the availability of labeled data. Collecting labeled data can be a challenging and time-consuming task, which limits the practicality of existing object detection models. Domain adaptation is a promising approach to overcoming the limitations of labeled data for object detection. It aims to transfer knowledge from a labeled source domain to an unlabeled target domain.

In this research proposal, I focus on unsupervised domain adaptation in object detection for multi-sport recognition. The goal is to develop a system that can detect players and balls from different sports, specifically football, and basketball, using a single model. To achieve this goal, I propose an unsupervised domain adaptation technique to transfer knowledge from a labeled source domain (football) to an unlabeled target domain (basketball) and vice versa.

The research objectives are twofold. First, I aim to investigate the effectiveness of unsupervised domain adaptation techniques in the context of object detection for multi-sport recognition. Second, I plan to evaluate the performance by comparing the system with the state-of-the-art results using benchmark dataset.

In the context of this research, the research questions might include to compare the performance of the proposed system to existing object detection systems using standard benchmark datasets and evaluation metrics. The second research question aims to understand what are the actual and optimal approaches that can accurately and efficiently detect players from multiple sports using a single model.

The scope of this research is to develop and evaluate an object detection system that can accurately and efficiently detect players from two sports, specifically football and basketball, using a single model. The study will use standard benchmark datasets and evaluation metrics to compare the performance of the proposed system. Furthermore, due to the significant time and effort required to gather new data, I have decided to rely solely on pre-existing datasets.

In this research, I propose an architecture for domain adaptation in object detection that combines two approaches to learn robust and domain-invariant features. Firstly, we employ a domain adaptation component at the image level, which is followed by training a domain classifier and incorporating the adversarial domain strategy. This helps in learning domain-invariant features that can be used for object detection in different domains. Secondly, we use Faster R-CNN to detect objects, starting from the CNN features shared with the previous loss. Since I have labels only for the source domain, I apply a different loss function depending on the image domain. If the image is from the source domain, I apply a loss with the ground truth labels. On the other hand, for target images, we generate pseudo labels using a different model trained on an Intermediate Domain created with CycleGAN (Jun-Yan Zhu et al., 2017), as shown in Figure 2. This approach helps in achieving domain adaptation in object detection with a minimal need for labeled data. An overview of the proposed system is shown in Figure 1.

The proposed architecture for domain adaptation in object detection has significant implications for real-world applications. One of the main challenges in object detection is the performance degradation when the trained model is tested on a different domain than the one it was trained on. Domain adaptation is critical to address this issue, especially when labeled data is scarce or expensive to obtain in the target domain. The architecture presented in this research provides a solution that can effectively adapt to different domains with minimal labeled data. Moreover, the proposed approach is flexible and can be easily extended to other tasks in computer vision.

To test the effectiveness of domain adaptation, I will conduct experiments between Cityscapes (Marius Cordtset at., 2016), a popular urban driving scene dataset, and its synthetic counterpart, Foggy Cityscapes (Christos Sakaridis et at., 2018). This is a common benchmark in domain adaptation literature and will allow me to evaluate the performance of the model in adapting to different weather conditions. If the results show promising performance gains, I will further extend the evaluation to a sport dataset to assess the generalization capability of the proposed approach across different domains. This approach will help to validate the effectiveness of the domain adaptation technique in improving the model's performance in diverse real-world scenarios.

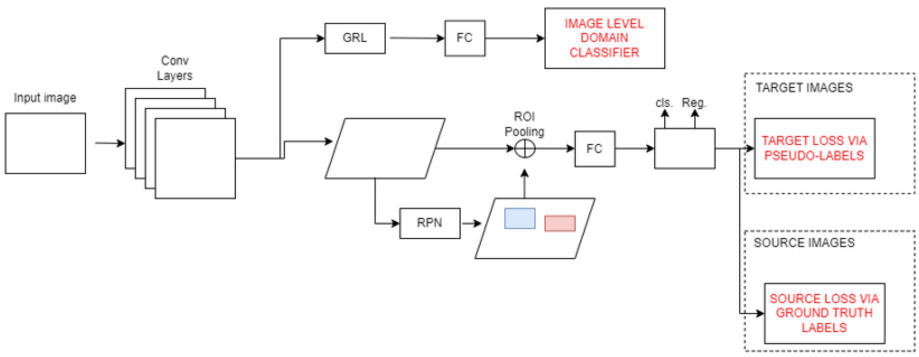

Figure 1 – An overview of the model.

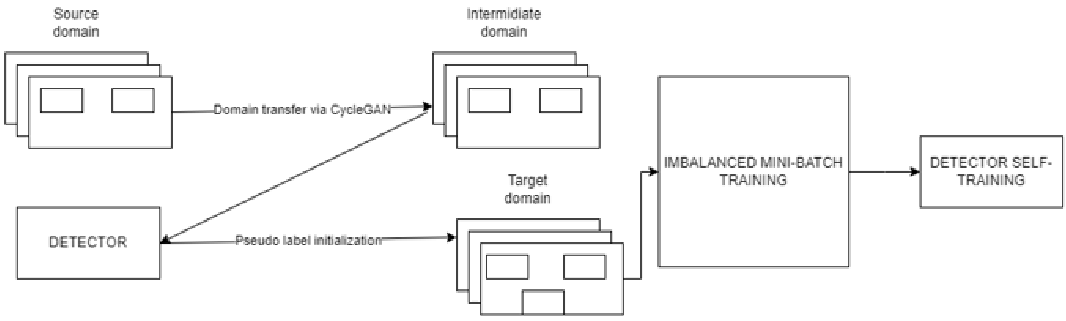

Figure 2 - Cross-Domain Semi-Supervised Learning (CDSSL) Framework Overview.

CHAPTER TWO

LITERATURE REVIEW

2.1 OBJECT DETECTION

Most of the current state-of-the-art approaches in object detection are based on convolutional neural networks (CNNs), which have shown remarkable performance in various computer vision tasks. One of these models is Faster R-CNN (Shaoqing Ren et al., 2016), which is a two-stage end-to-end detection framework that first proposes object regions using a region proposal network (RPN) and then classifies these proposals using a detection network. To reduce the computational need, single-shot approaches (W. Liu et at., 2016; Joseph Redmon et al., 2016) were designed. YOLO is one of these and it is a one-stage detector that directly predicts the bounding boxes and class probabilities of objects in a single shot through the network. These models require a large amount of labeled data to achieve high accuracy and the cost and effort of labeling data for each target domain can be prohibitive. This has led to the development of UDA methods, which aim to transfer knowledge from a labeled source domain to an unlabeled target domain.

2.2 DOMAIN ADAPTATION

Unsupervised Domain Adaptation (UDA) techniques give the advantage of saving time by avoiding the manual labeling of data in the target domain. The most used technique to fill the gap between the two domains is adversarial learning, it is composed of a domain discriminator trained to distinguish between the two domains while the main model is trained to fool the discriminator. Some papers implemented the paradigm of Domain-Adversarial Neural Networks (Y. Ganin et al., 2016; Y. Chen et at., 2018; Han-Kai Hsu et at., 2020), which proposed a method for jointly learning domain-invariant features and a domain classifier using adversarial training. Another approach (Fuxun Yu et at., 2021) is to create an intermediate domain that is closer to the target distribution. This can be achieved by generating synthetic samples that reduce the gap between the source and target domains in the feature space. This can be done by implementing CycleGAN (Jun-Yan Zhu et at., 2017) architecture.

CHAPTER THREE

METHODOLOGY

3 OVERVIEW

The system is trained using two different loss functions, each serving a specific purpose. The first loss function is designed to identify the domain of the image (whether it is from the source or target domain), while the second loss function incorporates labeled data to further refine the network's predictions. However, since labeled data is not available for the target domain, I implement a cross-domain semi-supervised learning technique that utilizes pseudo-labels to train the network on the target data. By combining these methods, we aim to achieve a more accurate and robust model that can effectively generalize to new, unseen data. An overview of my framework is shown in Figure 1.

3.1 IMAGE LEVEL DOMAIN CLASSIFIER

After the convolutional layers, the feature map output is fed into an Image Level Domain Classifier. The aim of this classifier is to distinguish whether an image is from the source or target domain. The implementation of the classifier is shown in the upper part of Image 1.

To make the features domain-invariant, we use the Gradient Reversal Layer (GRL) (Yaroslav Ganin et at. 2015) approach, which involves inverting the sign of the gradient during backpropagation. At training time, the aim is to maximize the loss of the domain classifier by finding domain-invariant features, while at the same time, I also aim to minimize the loss of the domain classifier by adjusting the parameters of the classifier.

3.2 CROSS-DOMAIN SEMI-SUPERVISED LEARNING (CDSSL)

The initial image is transformed to a different domain using CycleGAN (Jun-Yan Zhu et al., 2017) architecture, which is a type of generative adversarial network that can learn to translate between two different image domains without having paired images. The self-training detector then uses these transformed images to generate pseudo labels for the target domain. These pseudo labels are used to compute a loss function. By generating pseudo labels in this way, the system can potentially improve its performance on the target domain, even if labeled data is limited or unavailable.

REFERENCES

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. (2012). *ImageNet classification with deep convolutional neural networks.* In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12). Curran Associates Inc., Red Hook, NY, USA, 1097–1105.

Joseph Redmon, Santosh Divvala, Ross Girshick. Ali Farhadi. (2016). *You Only Look Once: Unified, Real-Time Object Detection.*

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun (2016). *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*

Lee, Dong-Hyun. (2013). *Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks*. ICML 2013 Workshop : Challenges in Representation Learning (WREPL).

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros (2017). *Unpaired image-to-image translation using cycle-Consistent adversarial networks*. In Proceedings of the IEEE international conference on computer vision, pages 2223–2232.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg (2016). *SSD: Single shot multibox detector*. In ECCV.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky (2016). *Domain adversarial training of neural networks*. JMLR.

Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. V. Gool (2018). *Domain adaptive faster r-cnn for object detection in the wild*. In CVPR.

Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih, Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan, Yang (2020). *Progressive domain adaptation for object detection*. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pages 749–757.

Fuxun Yu, Di Wang, Yinpeng Chen, Nikolaos Karianakis, Tong Shen, Pei Yu, Dimitrios Lymberopoulos, Sidi Lu, Weisong Shi and Xiang Chen (2021). *Unsupervised Domain Adaptation for Object Detection via Cross-Domain Semi-Supervised Learning*. In CVPR.

Yaroslav Ganin, Victor Lempitsky(2015). *Unsupervised domain adaptation by backpropagation.* In International conference on machine learning, pages 1180–1189. PMLR.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool (2018). *Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision*, 126(9):973–992

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele (2016). *The cityscapes dataset for semantic urban scene understanding.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3213–3223.