NTNU
Kunnskap for en bedre verden

Department of Mechanical and Industrial Engineering

Master's Thesis - robotics & automation

Efficient, accurate, and privacy-preservant object detection in edge devices

Student:
Hallvard Enger Bjørgen

Supervisor at NTNU:
Amund Skavhaug

Trondheim/Esbjerg Spring 2024

Table of Contents

List of Figures

List of Tables

Abstract

What
Why
How
Principal contributions
Principal conclusion

1 Introduction

1.1 Background and Motivation

Attention data has been much exploited online for personalized advertising and displaying of information and content thought to be relevant for a specific persons online profile. In real world, however, museums and cultural spaces remain in the dark regarding what really draws the attention of visitors and enhance the visitation experience. The museum staff may obtain this information through questionnaires, but analyzing behavior rather than self-reported attention is inherently more accurate.

Intelligence regarding visitor attention in cultural facilities, such as museums or aquariums, could prove beneficial for multiple reasons, e.g. the list of following . One may switch out artwork or exhibitions that are not interesting to visitors, or tweak/enhance their appearance to be more attractive. This information may contain some answers to the following questions:

Questions

1. How do the exhibitions rate in terms of popularity?
2. At what exhibition do the visitors spend the longest?
3. What is the average/maximum number of people in front of a given exhibition? Is this maximum number of people often reached?
4. Is there a change in what exhibitions are popular depending on the time of the day?
5. How big of a factor is the localization of exhibitions to how much time is spent there?
6. What areas are most prone to lines forming?
7. Are all parts of the museum being visited?

These answers may be converted into concrete actions to augment the visitation quality. Some of these actions may include the following:

Actions

1. Put more lights in an exhibition that is relatively more popular during daylight than after dark.
2. Replace unpopular artwork or place the most popular/expensive artwork in areas where more people may notice it.
3. Automatically open a window if more than 10 people are present in a room on average for 30 minutes.
4. Automatically close off sections of a facility when it's close to closing time and no people are in that area.
5. Integrate with other systems. Notify someone if someone falls in the water, or if, during closing, someone is about to be locked inside, a door is open, or a light is still turned on.
6. Notify if the total count of people in the museum is higher than tickets sold + registered employees/workers for that time period (people has snuck in).

1.2 Problem Description

Recent advancements in object detection allows applications to achieve incredible things. Detecting in dark environments or where the objects are partly occluded remains challenging, however, and is one of the reasons to why well-lit facilities may have advanced crowd intelligence and safety features, while darker areas remain harder to understand.

Image sensors can see infrared light impossible for the human eye to process. Infrared sensors may be used to infer how many people are in an area and where there's movement, but individually segmenting individuals is challenging as the borders of an individual is fuzzy in the infrared spectrum, and thus the task of measuring the amount of time a person is within a given area, i.e. more detailed intelligence regarding the attention of e.g. museum visitors is hard to achieve without utilizing image sensors for more refined light data.

1.3 Project Scope

1.3.1 Objectives

The objectives are to:

Primary Objective 1
objective 1

1.3.2 Research Questions

Consequently, the questions which all are addressed in the following literature study and discussion are the following:

Research Question 1 [Challenges]
What are the current challenges and some ways of resolving these challenges in object detection?

1.4 Structure

2 Literature

Here I will add relevant finds from the prestudy. Maybe some interesting research, and to support my choice of topic. Hei.

3 Methodology

Made it work.d

4 Discussion

5 Reflections

6 Future work

The future work section is dedicated to outlining and proposing some potential pathways for further exploration. The idea of this section is to provide some ideas for several objectives of future work, which can be done either by themselves, or as a single project with multiple of the following proposed directions of research.

6.1 Highly specific training dataset

This idea is inspired by 2 elements: (1) Where's the Bear's (Elias et al., 2017) approach to artificially place many different bears collected online, in images captured by their cameras, to make more training data on bears for their specific application. (2) The fact that real world situations differ from the situations typically found in benchmark datasets, as mentioned in section sec:discussion ml techs. In section , we discuss research where a dataset with a different angle than what is typical in the benchmark datasets was created, which improved accuracy of the model.

For this idea, the main task will be to explore how much data from a specific setting is needed for a tiny/light ML models to surpass the accuracy of a large ML model. E.g. could a light YOLOv3 with an hours-worth of obtained specific data for a given setting be more accurate, faster, and smaller than a ML model trained on a larger general purpose dataset such as the COCO[1] dataset? Could these light ML models *surpass* the accuracy of SOTA machine learning models for low light settings? How does accuracy increase with the size of a specialized training dataset?

These insights will not only prove relevant for the specific ML models tested, but may also provide more general knowledge about the use of specialized datasets for ML applications.

It can be challenging to research the impact of a highly specialized dataset for ML model training on the accuracy and efficiency of models due to the difficulty of obtaining labeled data. Identifying correct model inferences, and thus recognizing improvements in the models, requires knowledge of the ground truth. This will likely be obtained by manually annotating data. E.g. to know the improvement of the accuracy of a human detecting model, one must know how many humans are actually in the image to be able to verify the model's count. This is a complex task as the images may contain natural persons who have not consented to being in the images. Therefore, the research will need consenting persons to build the dataset.

If the model is trained solely on specific data for the given use case, the ability to generalize the model to other use cases is lost, as it will only be highly efficient for the given elements in the images it has seen and may be confused by new shapes or objects it has not seen.

Proposed approach Mount a visual edge device (hereby "the device") and build a dataset of images from a specific setting. These images need to be compliant with GDPR, and need to be sharp as they are to be used for ML model training. Therefore, only consenting persons must be in the images.

Then, the edge device is deployed with a large, general purpose model (hereby "the large model") and inference is ran in a real-world setting in the same area. The individual privacy of the persons in these images may be preserved through blurring the images post-analysis.

While the large model is running, the obtained images from the specific setting may be annotated. This must be done in some way where a bounding box is set around each person in each image, i.e. the corner coordinates of each bounding box in the image is saved along with a reference to the image. The bounding boxes are what will be used to train the models, and the precision of the manual labor will thus influence the accuracy of the resulting model.

This specialized, newly obtained dataset is then incorporated into a light ML model, and deployed to the device.

[1] COCO: Common objects in context, introduced in section *sec:performance benchmark*.

The images and inferences from the large model on the device are now to be analyzed, finding the ground truth by looking at the blurred images and counting every person, and comparing these counts to the inferences of the model. Additionally, the bounding boxes of the blurred images should be manually evaluated if they seems sufficiently accurate or not (which will be a subjective measurement, which is okay as the measurement of model success will be mostly determined by achieving the correct count of people).

Finally, the data from the light, specialized ML model are collected and analyzed, and compared to the result of the large model.

Then, the experiment may be extended in multiple directions depending on the results. One would be to obtain more specific data, to see how more data will further improve the accuracies of the lighter version model. Another is to see how using less specific data would impact the accuracies. Another would be to try with a completely different model, i.e. switching from a Faster R-CNN to YOLO, or to drop the feature pyramid networks in a Faster R-CNN to evaluate the impact on performance and accuracy.

The primary objective of such an approach could be to deliver an assessment of the impact of usage of use case-specific data in machine learning model training.

6.2 Robustness of detectors

Another direction could be to assess the accuracies and throughputs of three different ML algorithms in various conditions, e.g. with regards to light, scale, occlusion and pose. The models should all be trained with the same data, and deployed to the same type of hardware and the same environment.

Some of the questions to be addressed would be: Will there be a difference in which detector is the most accurate, dependent on conditions such as light, scale, occlusion and pose? Will Faster R-CNN be more accurate but slower than YOLO? Are SOTA detectors overfitted to data commonly found in benchmark datasets, and less accurate on real-world data?

6.3 Feasibility of implementing tiny ML models on cost-effective hardware

A third proposal would be to create optimized firmware for cost-effective hardware to perform the specific, repeating job of taking an image, performing model inference to count the number of people on the image, and sending the count to a remote server. How tiny in size may the models be for the inferences to be sufficiently accurate?

The exploration of hardware, which is in constant state of improvement and change, may be outdated in a couple of years. However, finding the status quo with regards to a in 2024 may still be relevant for future endeavours.

6.4 Heat map implementation

The last proposal is to create a visualization of the data. This can most likely be done as a part of a bigger project. To visualize the counts for various areas in the image may be best visualized through a heat map of where the detected persons have been. Such visualization, taking advantage of the stationary nature of the visual edge device, may be beneficial as much data can be integrated into one image in a completely anonymous and privacy preservant way.

7 Conclusions

8 Tanker

Skaffe en divers gruppe samtykkende mennesker med flere former for bekledning; luer, caps, jakke, tskjorte. Eller bare meg, med masse ulike klær.

Image segmentation, hente ut utsnittet av menneskene;

Plassere disse menneskene på tilfeldige steder rundt i et område;

Trene datasettet på disse utsnittene;

Gjør vi det da bedre for det gitte området?

Kan dette streamlines, prosedurebeskrivelse:

1. Sett opp enheten fysisk der den skal sitte.

2. Et bilde der det er ingen mennesker tas, og bruker verifiserer bildeområdet. Dette bildet beholdes og brukes senere i prosessen.

2. Bruker trykker på knapp: "start datainnsamling".

3. 10min: bruker/brukere beveger seg rundt, ulike posisjoner, ulike bekledning.

4. Datainnsamling fullført, enheten forsøker å hente ut menneskene i disse bildene, bruker korrigerer. Her vil det være nødvendig med et verktøy for bounding box der brukeren kan korrigere for hvert bilde som er blitt tatt.

5. Når utsnittene er korrekt satt, lagres de lokalt, og brukeren ferdig.

6. Enheten dupliserer det tomme bildet 100 000 ganger, og setter inn tilfeldige utsnitt på tilfeldige plasser i bildene.

7. ML modell (vision transformer?) trenes på generelt dataset, så fyller på med disse bildene.

Highlights:

1. The produced dataset will be highly specialized to a specific area, thus possibly greatly increasing the accuracy for difficult scenarios.

2. The dataset will not be optimized to different situations, but the segmented people may possibly be applicable to insertion in similar environments (with respect to light conditions).

3. Must be same light conditions in the whole operational period of the ML model.

4. Experiment to see if the produced data alone can suffice for similar accuracies on smaller models without the need of massive training, thus shifting the workload of machine learning to specialized data acquizition rather than computational heavy model training.

Questions: Do the cameras need to be RGB? Why need colors? More information for the space if we see greygscale? Better accuracy with colors?

Use Raspberry Pi Black camera?

Hypothesises regarding museums: Hvad er det der museumet er interessert i at finde ud? - Hvor lenge en person ser på hver fisk

Baseline creation: 1. Have the device function as normal, try to detect people. 2. Improve it so it has detection algorithm more optimized for short distance detection 3. If not well: specialized dataset. If well: heatmap and efficiency.

Specialized dataset: 1. Setup camera and make it send images to a google bucket. 2. Object detect persons on device, then send count and image to bucket. 3. Create code to obscure image before sending

Notes are taken on Work-Confluence as well for easier access and transparency with coworkers.

Bibliography

Elias, A. R., Golubovic, N., Krintz, C., & Wolski, R. (2017). Where's the bear? - automating wildlife image processing using iot and edge cloud systems. *2017 IEEE/ACM Second International Conference on Internet-of-Things Design and Implementation (IoTDI)*, 247–258. https://ieeexplore.ieee.org/document/7946882 (cit. on p. 7).