Ensemeble Methods for Visual Anomaly Detection in Manufacturing Settings

Toller Thesis Titel

Master thesis by Marc Saghir

Date of submission: February 8, 2024

1. Review: Super Supervisor
Darmstadt

Erklärung zur Abschlussarbeit gemäß § 22 Abs. 7 APB TU Darmstadt

Hiermit erkläre ich, Marc Saghir, dass ich die vorliegende Arbeit gemäß § 22 Abs. 7 APB der TU Darmstadt selbstständig, ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt habe. Ich habe mit Ausnahme der zitierten Literatur und anderer in der Arbeit genannter Quellen keine fremden Hilfsmittel benutzt. Die von mir bei der Anfertigung dieser wissenschaftlichen Arbeit wörtlich oder inhaltlich benutzte Literatur und alle anderen Quellen habe ich im Text deutlich gekennzeichnet und gesondert aufgeführt. Dies gilt auch für Quellen oder Hilfsmittel aus dem Internet.

Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Mir ist bekannt, dass im Falle eines Plagiats (§ 38 Abs. 2 APB) ein Täuschungsversuch vorliegt, der dazu führt, dass die Arbeit mit 5,0 bewertet und damit ein Prüfungsversuch verbraucht wird. Abschlussarbeiten dürfen nur einmal wiederholt werden.

Bei einer Thesis des Fachbereichs Architektur entspricht die eingereichte elektronische Fassung dem vorgestellten Modell und den vorgelegten Plänen.

Darmstadt, 8. Februar 2024

M. Saghir

Abstract

Abstract

Contents

Figures and Tables

List of Figures

List of Tables

Abbreviations, Symbols and Operators

List of Abbreviations

Notation	Description
DDPG	Deep Deterministic Policy Gradient
DQN	Deep Q Network
ML	Machine Learning
PPO	Proximal Policy Optimization
RL	Reinforcement Learning
SAC	Soft Actor Critic
TRPO	Trust Region Policy Optimization

List of Symbols

Notation	Description
A	continuous action space
S	continuous state space
$\mathcal{H}(\cdot)$	entropy
$\pi(a\|s_t)$	Policy

1. Introduction

This is a citation: [1]

This is a figure:

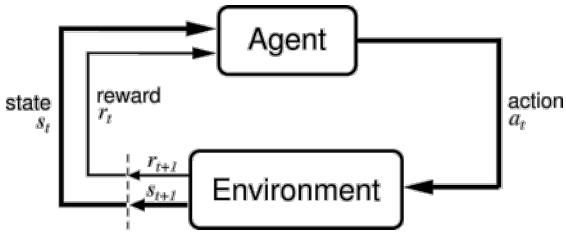

Figure 1.1.: I am a caption

- It is important to note somewhere in the paper that we are dealing with very high variance in our ensemble since we only have 5 models ish

1.1. Begin Intro

In recent years, image anomaly detection has become significantly more important among many scientific communities, especially in industrial applications. This is no surprise, considering the amount of mechanically manufactured parts in factories all over the world. Since in most parts of the world, manufactured items undergo more or less strict regulations and are expected to work in real case scenarios, there is a need for sufficient quality control, that is rising with the amount of produced components. A long time ago, it has come to a point where human based quality checks are not adequate anymore for the production volume, which has led to computer solutions for the problem. Generally speaking, anomaly detection has first been proposed in 1986 for intrusion detection systems. While the methods and modalities may change, the high level idea stays the same: Detecting data that deviates from a set standard to a degree that is becoming problematatic regarding the own requirements(letzer part maybe neu formulieren). Besides many approaches that were used over the years, deep learning approaches for image anomaly detection have become very popular lately. A likely reason for this are impressively high performance scores with state of the art models achieving an area under the receiver operateor curve of around 0.96 and sometimes even more. It is difficult to say what the first deep learning approaches to this topic were(dact checking), but a notable milestone is definetly Bergmann 2021(Referenz). Among blabla, they introduced the MVTecAD dataset which is used widely and serves as a dataset to benchmark on for nearly every IAD paper released afterwards. - Übergang benötigt

1.2. Contributions

This work builds upon the researched anomaly detection methods that were established as state of the art over the last years. paper1 and paper2 provide extensive overviews of all SOTA methods, including the ones mentioned here. The main contributions thus are: 1. Creating a pipeline for anomaly detection experiments and inference, utilizing existing IAD approaches. 2. Introducing three new categories for the mvtec(LOCO) dataset for anomaly detection experiments. 3. Researching anomaly detection performance on multi perspective datasets. 4. Experiment on the performance of different ensemble methods for IAD, including utilizing majority voting, stacking, CAWPE and (die anderen ensemble paper von discord zitieren).

The above contributions can be used as basis for industrial usage, aswell as a basis for future contributions on ensemble methods in the IAD space. Moreover it gives further insight on the efficiency of SOTA IAD methods on different kinds of data than the previous synthetic settings.

– in my work i contribute the following things: - pipeline to infer new images on different algorithms and compare them -> pipeline is industry focussed for benefits of the guys where i write my thesis

- research on multi perspective detection

- research of ensemble output learning to enhance individual network performance -> simple network over 5-6 outputs

- introduction of very new dataset categories in style of mvtec LOCO dataset

2. Background

This is an algorithm

2.1. Classes of Anomaly detection

When trying to understand the choices of IAD approaches for the pipeline and ensemble, one first has to learn about a few important distinctions of models on this topic. The deep learning approaches that have established themselves as state of the art in image anomaly detection are almost exclusively unsupervised approaches. This partiall stems from the fact that naturally anomalous images occurr far less than normal images, hence the word "normal". This is especially true in industrial settings, due to the high performance of production factories nowadays. Therefore if one were to consider using a supervised learning approach to detect anomalies, either a strong class imbalance or an unrepresentative class distribution would occur. While there are some solutions for this, they often are either not goo enough for imbalances this high(synonym klänge cool) or far to extensive. Some papers like (supervised papers zitieren) utilize supervised approaches with some success, but still yield a worse performance than the popular unsupervised approaches generally used. Consequently the biggest model distinction is between unsupervised and supervised ones. Here it has to be said that there are technically also other settings of IAD one could talk about at this level of observation, but since we are also directing our focus to to RGB images, they will not be talked about. Moreover one has to make some simplifications to allow such sharp categorizations of partially interwoven approaches.

The supervised learning category could also further be split up into sub-categories at a lower level. But seeing as the performances of unsupervised approaches dominantely outweigh the performance and cost of the former, this work will solely focus on the latter kind of approaches. In the unsupervised IAD setting we then normally distinguish between reconstruction and representation based models. One of the key differences between those two is(hier dringend auch paper zitieren die das untersuchen),

...

If we now consider the classification of algorithms above, aswell as figure x, we can see that there are quite a lot of unique models and approaches to the same end. To ensure that the built pipeline is able to help experiment on images from different points of view, so to say, aswell as ensure that our ensemble approaches cover as various different aspects as possible, it is crucial to select approaches from majorly different branches. Here it may be noted that the performance of the single models is not completely disregarded, as those models may prove themselves not very useful in the ensemble setting or even as a point of view for experimentation. Therefore certain approaches from the survey papers ...., which yielded performances that were not remotely comparable with the highest performing models, were not considered, even if they might cover a previously unrepresented class of IAD setting. The main choices were: - patchcore + paper - DRAEM + paper - CSFlow + paper

With this choice we still represent reconstruction and representation based settings somewhat comparably, aswell as providing different examples for a variety of subclasses, namely distribution maps, autoencoder, memory banks, teacher-student models, diffusion models and ...

- there are different kinds of approaches to IAD - look at tree picture

- First important distinction is between supervised and unsupervised -> we focus on unsupervised -> list problems with supervised approaches and thus advantages of unsupervised ones

- briefly touch on other IAD settings like few shot, along with references

- among unsupervised approaches, there are two more fundamental distinctions -> reconstruction based vs representation/feature embedding based -> explain difference with lots of references

- for reconstruction based touch on 2-3 base categories like GANs etc and link fundamental papers for GANs etc - for representation based important to explain memory bank, teacher student, and distribution map - explain normalizing flow somehow somewhere in there

- maybe say which algos we chose and what we covered with that

2.2. The Datasets

The datasets used in image anomaly detection are scarce, especially when it comes to anomaly detection in a manufacturing setting. There are a few that specialize on certain textures(references) and some that can be used for wide ranging categories. What currently stands out as a gold standard among IAD datasets is the MVTecAD(referenz) dataset. It was designed by Bergman et al.(referenz) as a highly representative and standardized set of anomalous images along with training images. It has 15 classes from (some examples) to (...). It provides image labels aswell as segmentation ground truths, making it versatile and applicable for multiple algorithms. The masks come as black and white grayscale images, while the iamge labels are given through its folder structure. Its paradigmatic structure tree can be seen in figure xy.(hier ein satz der die ordner struktur beschreibt) Example images of the dataset are to be seen in figure z. They typically are of a rectangular shape and their resolutions range from blabla to blabla. More specifications can be found in (mvtec reference) and the whole dataset is publicly available at (dataset link).
The MVTecAD(referenz) dataset is regarded as the go to dataset(wissenschaftlich formulieren) among IAD papers, and has since its introduction been used in nearly every

paper as a dataaset to benchmark ones approaches on. This is also likely to remain the trend, since many important algorithms in the recent years have primarily been benmarked on it, forcing new approaches to also be benchmarked on this dataset to be comparable to the current SOTA approaches. Due to its importance MVTecAD is one of only two datasets relevant to this work, and serves as a comparison to investigate SOTA algoithm performances of the second dataset, which will be our main focus.
Later in (loco year) Bergman et al(referenz) has introduced another IAD dataset that is loosely related to their original MVTecAD dataset, namely the MVTecAD LOCO dataset(reference). This dataset works with the same ground ideas as their original MVTecAD set, but extends the conceptual contents of the dataset by logical anomalies(neu formulieren das klingt scheiße). It consists of five class:(class names). The difference to the other dataset is that the anomalous categories for each class are only seperated into good images, images with structural anomalies and images with logical anomalies. Structural anomalies being visible damages to the objects, similar to the MVTecAD dataset. Logical anomalies denote violations against arbitrary restrictions imposed by Bergmann et al.(referenz). To illustrate this by an example: The class of pushpins represents a birds view of a compartmentised box of pushpins(see figure a). A rule added was, that each compartment is only to contain one pushpin. This means that if one region were to miss their contents, or contain two pushpins, it would constitute a logical anomaly. If on the other hand a pushpin would have a crooked or broken tip, it would be a structural anomaly. The addition of logical constraints opened an interesting area of research, since the high performance of current SOTA algorithms were only measured on structural anomalies so far. Yet it would be insightful to see if those models could also detect logical anomalies, since those also ocurr in real life settings, such as manufacturing settings. (Noch ansprechen dass LOCO eine neue metric -> sPRO ermöglicht und die saturation configs ansprechen) Bergmann et al.(referenz) also released a new IAD model together with the new dataset. The model uses autoencoders(bissi besser beschreiben hier). Unfortunately the code has not been made public. Aside from approaches tailored specifically towards the detection of logical anomalies, it would be interesting to see how SOTA methods of structural anomaly detection perform on the LOCO dataset. The performance of previous methods on the LOCO dataset is already partially evaluated in some papers like(referenzen von benchmark papers),

but will comprehensively be investigated later in this work. Moreover the novel dataset categories introduced later are composed of structural aswell as logical anomalies and formatted in the MVTecAD LOCO dataset style. Aside from the conceptual differences in the two datasets, there are slight changes to the structure tree aswell. The anomaly classes are only changed by name, since it is irrellevant for the models whether the anomaly name is ""

anmerkungen für text oben: - beschreiben was mvtec neu bringt: zb dass es näher an real world ist - saturation thesholds ansprechen

2.3. metrics

- show metrics from survey papers - explain which metrics we used and where the other ones are used - explain also why we used the ones we used, and what disadvantages of other ones where

- touch on paul bergmann paper for sPRO score, say how it is better than pixel auroc and normal pro score, also explain saturation thresholds

- some math formula for calculating the important metrics

2.4. description of patchcore algo

2.5. description of simplenet

2.6. description of AST

2.7. description of DRAEM

2.8. description of another reconstruction based algo

3. Related Work

4. Method

This is an table:

4.1. Our own Dataset

- repeat motivation why we added additional data in mvtec style - say that we went with loco mvtec flair(maybe give reasons) - say that we came up with a set of structural and logical anomalies for each category - list categories(flat connector, angle and special construct)

- 3 sub sections for the three categories

- flat connector - link the exact one we used(or examples of some) - give structural anomalies - give logical anomalies - for both briefly touch on how we produced them - show image examples for each

- repeat same for other categories

- also when describing angle: - touch on how there is a special case with multi perspective detection

m	$\Re\{\underline{\mathfrak{X}}(m)\}$	$-\Im\{\underline{\mathfrak{X}}(m)\}$	$\mathfrak{X}(m)$	$\frac{\mathfrak{X}(m)}{23}$	A_m	$\varphi(m)$ / °	φ_m / °
1	16.128	8.872	16.128	1.402	1.373	-146.6	-137.6
2	3.442	-2.509	3.442	0.299	0.343	133.2	152.4
3	1.826	-0.363	1.826	0.159	0.119	168.5	-161.1
4	0.993	-0.429	0.993	0.086	0.08	25.6	90
5	1.29	0.099	1.29	0.112	0.097	-175.6	-114.7
6	0.483	-0.183	0.483	0.042	0.063	22.3	122.5
7	0.766	-0.475	0.766	0.067	0.039	141.6	-122
8	0.624	0.365	0.624	0.054	0.04	-35.7	90
9	0.641	-0.466	0.641	0.056	0.045	133.3	-106.3
10	0.45	0.421	0.45	0.039	0.034	-69.4	110.9
11	0.598	-0.597	0.598	0.052	0.025	92.3	-109.3

Table 4.1.: Table Caption

4.2. pipeline

- explain brief structure of the pipeline - ???

4.3. Ensemble network

- network architecture - specifics

4.4. Different ensemble approaches

- weighted, random forest etc - specifics

5. Experimental Setup

6. Experimental Results

- analysis on how methods worked on own dataset individually -> if poor performance error analysis and also address different subclasses

- analysis of how ensemble model worked and if it improved performance

7. Conclusion and Future work

Bibliography

[1] K. Roth, L. Pemula, J. Zepeda, B. Scholkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 14318–14328, Jun 2022.

A. Appendix

Appendix here