Laidlaw Scholar: Tobias Threlfall-Holmes
Supervisor: Toby Breckon
June-September, 2024

Research Report: Does an RG-chromaticity and stereo-Disparity (RGD) image encoding differ from RGB in distinguishing anomalous objects in image datasets for self-driving vehicles?

Introduction

Vehicles are dangerous, massive objects that operate at speed near exposed members of the public. Self-driving technologies seek to reduce the probability of collisions, by removing both the inattention and unreliable analyses of drivers. However, neither are assuredly addressed, as partial automation of vehicles worsens human inattention, despite advancements to mitigate this effect [Monsaingeon et al., 2021], and the computational analysis it is replaced with is challenging to validate as reliable, for which further dependability research has been advocated [Pradeep et al., 2023].

In 2018, the first pedestrian fatality by a self-driving vehicle was reported; The collision involved both operator inattention, and undependable object classification. [Smiley, 2022]

This project seeks to investigate the hypothesis that some part of the consistent underperformance across vision models for object classification in pre-deployment testing is due to either mislabelling of the data, or so highly anomalous examples of technically correctly labelled objects that they might reasonably be labelled differently. This hypothesis arises from image datasets being too large to manually validate every image. It is proposed that, if various unsupervised anomaly detection algorithms were made to rank images by how anomalous they are of an example of what they are labelled as, the highest ranked could be reviewed by researchers to assess how to improve the quality of the dataset, such as by relabelling, or the method by which the model is assessed on predicting image labels.

The impetus for this research, the expectations of it, and the plan of work, are further detailed in the associated research proposal. [Threlfall-Holmes, 2024]

Methodology

This section concerns the actual course of work undertaken, which differed somewhat from the plan of work described in the proposal document for this project.

I decided to complete the planned stages with a single model first, to confirm each as functional during development, and then switch to using a variety of models for the data collection. I initially tried PatchCore, but had issue getting it to complete these tests due to limited memory (RAM). Therefore, I switched to DFKDE, a model that was very quick to train locally on my computer, and thus easy to prototype with. This testing used Anomalib's synthetic anomaly test split mode, with which the dataset is automatically split into training, validation, and test data, and Perlin noise is added to some of the test images to create ones that should be identified as anomalous, outside the distribution of normal images observed in the training set. The classification task of deciding whether or not an image is anomalous involved first training, modelling the distribution of normal images, then detection, giving each image an anomaly score to represent how different to that model of normal images it

is, and finally applying a threshold of a high enough score for it to be classed as anomalous, which is determined by trialling the model on validation data to pick the threshold that gives the highest AUROC, which represents how good the classification was. An AUROC of 100% identifies a perfect classifier, and at 50% the classification decision in meaningless.

A python program was written to transform a given dataset of RGB, or RGB and greyscale, stereo image pairs into RGD images. The first (RGB) image is split into its three colour channels, red, green, and blue, and RG-chromaticity calculated as: red / (red + green + blue) and green / (red + green + blue), providing a two-channel version of the colour image, neglecting light intensity. Stereo disparity was calculated by a program produced by Toby Breckon as an example piece for computer science students at Durham University, which converts both images to greyscale, then uses the OpenCV python library's modified H. Hirschmuller algorithm, with configurable parameters, citing [Hamilton et al., 2013]. The two channels of RG-chromaticity are then merged with the stereo-disparity channel for RGD. For virtual datasets, a depth channel may be directly available for merging with RG-chromaticity, in which case the stereo disparity calculation is neglected as redundant.

The attempts to implement Anomalib that followed were unsuccessful.

The final test in preparation of ranking how anomalous images were, and comparing those ranked most anomalous when the dataset was encoded as RGD or RGB, according to the anomaly score given to each image by several models, was to validate that the anomaly scores produced were meaningful. I used the VKITTI2 dataset [Cabon et al., 2020], a virtual dataset based on the KITTI dataset, extracting all images of cars – across all camera, but not weather, variations - as rectangular crops ten pixels wider on each side than each group of pixels separated by less than ten pixels labelled as car in the semantic segmentation version of each frame. These were converted from RGB to RGD. I used both the left and right camera images, as the depth channel was exactly provided for both of these, uniquely possible in a virtual dataset, meaning stereo disparity was not calculated, avoiding any noise that calculation may have otherwise introduced.

This should be an easy task, as with these synthetic images there is no alternative source of noise or phenomena of similar appearance to these synthetic anomalies. A low score would demonstrate a model's inability to generate a meaningful assessment of how anomalous an image is, such that it could not provide a meaningful ranking of the images, and must be unsuitable for generating the proposed ranking of images. This was the last stage completed.

I tested DFKDE, which I had been using exclusively until now, and found its classification to be extremely poor. This motivated me to try another model that takes longer to train, Cflow, which I left training for six epochs (taking around forty hours on my local machine), which also performed poorly. Finally, Deep Feature Modelling (DFM) was chosen to be trialled. DFM was expressly designed to improve the accuracy of out-of-distribution confidence scores, which is exactly the task of putting a meaningful score to how anomalous an anomalous image is, rather than just how anomalous non-anomalous images are, and to better handle adversarial samples, images designed to reduce model performance, which I suspect came into effect as better handling of the images being padded. [Ahuja et al., 2019]

Results and Discussion

DFKDE and Cflow both performed very poorly at classifying the synthetic anomalies with the padded car images, with AUROCs of 54%-68% and 64% respectively. As these AUROCs show the models to be poorly guessing in generating their anomaly scores for images, including them in the originally proposed image ranking task would not have yielded meaningful rankings. The range in results from DFKDE of 54%-68% was yielded by comparing padding the car images with a constant background (68%) and with a background of uniform random noise across all three channels (54%), which was re-randomised every two hundred images. This showed that the cases it was succeeding at were when synthetic anomalies appeared on the constant padding, being trivial to detect, rather than the model capturing the variation of the constituent car image. Cflow and DFM were tested with the noisy padding.

DFM produced an AUROC of 94% for the padded car image task of classifying between normal images and those with synthetic anomalies, which indicates its ranking of how anomalous images are has the potential to be meaningful.

This shows that, while all the models produce an anomaly score, this does not mean it can be assumed to be a meaningful representation of how anomalous the image is, nor thereby useful for ranking the dataset. In this case, multiple models greatly underperformed.

Conclusion

The data collection originally proposed, that is a comparison of the images ranked most anomalous by various algorithms when encoded as RGD versus RGB, was not conducted. However, the preliminary testing with the VKITTI2 padded car images revealed that not all models were suitable for creating such a ranking, which was not anticipated.

It is recommended to consider in future works that not all models are appropriate for the task of ranking images by anomaly score, at least for the varying aspect ratios arising from object-level images, rather than whole scenes, and that it may be of benefit to precede such ranking by a similar validation of the models' anomaly score outputs as was conducted here.

For a larger scope repetition of this work, one could address all-weather operation, when the quality of image data is reduced, such as by low light, partial refraction of images by droplets on lenses, and phenomena occluding different objects in each stereo camera view. One might begin with the still relatively clean data of VKITTI2's weather variations, but creation of a new dataset of images from a real vehicle would likely be necessary to validate the attempted improvement to dependability in more realistic unfavourable conditions.

Laidlaw Scholar: Tobias Threlfall-Holmes
Supervisor: Toby Breckon
June-September, 2024

Reflection

The main issue encountered, preventing the completion of this work, was the allocated time.

The risk of the project taking substantially longer than the six-week minimum was identified before commencing the research, particularly as I approached the research with the goal of making it worthy of publication, however remote than potential result was discussed as being, and expected this may take some additional weeks of refinement after a conducting a minimum completion as a proof of concept. However, I do not think having approached it more conservatively would have resolved this issue. I would instead suggest that there was not sufficient time to both become familiar with the specific tools and materials of the project, and complete it, within the allotted time. While I was familiar with python programming, I struggled to implement the Anomalib library. The result of this issue was that the project was only approximately half completed after ten weeks, having failed to conduct the data collection stage, originally scheduled for the fourth week.

Once the ten weeks allotted had run out, I had to leave Durham, and attempted to resume work, so that it could be progressed to at least data collection. However, this required a second version of the initial week of familiarisation, figuring out how to get the software working for my new hardware situation. It was possible to fit other commitments around this, those that had been scheduled on the assumption the research would be concluded. However, it was not possible to clear more than one week additional to that originally planned as the maximum time the project might occupy. This was primarily due not to anticipated commitments, but rather urgent matters unrelated to the project, and the nature or duration of these delays could not have been anticipated.

It was originally supposed, when considering this risk, that it was not possible to reduce the scope of the project. However, I would now advise that the data collection be considered in itself a complete project, as a contribution towards future analysis, suitable to the scope of a six-week internship. Further, for even that scope to be robust to unexpected delays, which one cannot presume will not occur, nor that they will be insignificant, I would suggest familiarity with Anomalib itself may be a prerequisite, perhaps by completing a small test project before the internship, to save a week of focussed familiarisation, and so speed further work as to save perhaps even a couple of weeks more. If I were to reattempt this project, such a preliminary work, not seeking to contribute to the project, only to my understanding of the tools, would have been a priority to me earlier in the year.

One week was spent on attending the British Machine Vision Association and Society for Pattern Recognition's (BMVA) Computer Vision Summer School. However, I would not attribute this as a delay to the project, as it usefully furthered by understanding and curiosity, that may have positively contributed to efficiency. Though I cannot reasonably determine its the overall effect on the project, it was an essential part of the internship, which is ultimately more so for learning and personal development than research outputs.

Laidlaw Scholar: Tobias Threlfall-Holmes
Supervisor: Toby Breckon
June-September, 2024

Acknowledgements

Central to this project was Prof. Toby Breckon's computer vision group at Durham University, and I extend my thanks to the team for the environment created, and the wider insights into computer vision that our weekly meetings provided me. Toby proposed and supervised this project, and I thank him for his patience and advice throughout this internship.

An important aspect of my learning in computer vision with this internship was the BMVA Computer Vision Summer School, and I thank the organisers of this (Amir Atapour-Abarghouei, Jingjing Deng, Ulrik R. Beierholm, Stuart James, Toby Breckon, and Andrew Gilbert), as well as the many fellow volunteers whom I found it a pleasure to work with.

This internship was funded by the Laidlaw Foundation, through their granting me the Laidlaw Leadership and Research Scholarship, and I thank them for affording me this opportunity, in addition to the initial leadership training preceding the internship, where I had valued conversations with fellow scholars whom I am unlikely to have otherwise met.

References

Ahuja, Nilesh A.; Ndiour, Ibrahima J.; Kalyanpur, Trushant; Tickoo, Omesh (2019) "Probabilistic Modeling of Deep Features for Out-of-Distribution and Adversarial Detection". Presented at the 4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

Cabon, Yohann; Murray, Naila; Humenberger, Martin (2020) "Virtual KITTI 2". Naver Labs Europe. Downloaded from: https://europe.naverlabs.com/research/computer-vision/proxy-virtual-worlds-vkitti-2/, July 2024.

Hamilton, Oliver K.; Breckon, Toby P.; Bai, Xuejiao; Kamata, Sei I. (2013) "A foreground object based quantitative assessment of dense stereo approaches for use in automotive environments". IEEE. Presented at the 20th IEEE International Conference on Image Processing (ICIP 2013), Melbourne, Victoria, Australia, 15-18th September, 2013, pp 418-422.

Monsaingeon, Noé; Caroux, Loïc; Mouginé, Axelle; Langlois, Sabine; Lemercier, Céline (2021) "Impact of interface design on drivers' behavior in partially automated cars: An on-road study". Elsevier. Transport Research Part F: Traffic Psychology and Behaviour, Volume 81, Pages 508-521.

Pradeep, Aneesh; Bakoev, Mironshokh; Akhroljonova, Nazokat (2023) "A Reliability Analysis of Self-Driving Vehicles: Evaluating the Safety and Performance of Autonomous Driving Systems". IEEE. Presented at the 15th International Conference on Electronics, Computers, and Artificial Intelligence (ECAI), Bucharest, Romania, 29-30th June, 2023, pp 1-5.

Smiley, Lauren (2022) "'I'm the Operator': The Aftermath of a Self-Driving Tragedy". Wired.

Threlfall-Holmes, Tobias (2024) "Research Proposal: Does RG-chromaticity and depth differ from RGB in distinguishing anomalous objects in stereo image datasets for self-driving vehicles?". Laidlaw Scholars Network.