Research Proposal: Does RG-chromaticity and depth differ from RGB in distinguishing anomalous objects in stereo image datasets for self-driving vehicles?

Theoretical Issue

Self-driving vehicles are being tested live, on public streets, but when we attempt to validate the underlying image models' performance, there remains some unexplained underperformance, which prevents reliably benchmarking the performance of such systems before deployment. Therefore, further dependability research has been advocated. [Pradeep et al., 2023] This project is part of asking: What is causing these apparently capable models to be assessed as operating incorrectly?

The challenge of validating model performance is pervasive in the field of machine learning, as the previous software validation methodology of comparing expected output to actual output is of limited applicability. We might instead assess if the output is reliably reasonable, rather than reliably identical to expectation. This leaves the question, what is reasonable? We cannot assume that human comparison translates well to machine comparison. A human might expect a synthetic car image they have made to be identified as a car by a vision model, but the model not do so due to genuine details the model has learnt to associate with cars, thus producing the negative result "this is not a car" where the expected result was positive "this is a car". The model has failed to match actual output to expected output, yet not necessarily failed to produce a reasonable/correct output.

This research proposal is based on the hypothesis that the theoretical underperformance of image models is in part a flaw of the testing, by either mislabelling of the data, as the datasets are too large for every image to be manually validated, or the data being technically correctly labelled, but being such an unusual example of that object that it would be less reasonable to identify it as what it is than presume it must be something else; anomaly detection could bring these to human attention.

Practical Implications

This problem has severe implications, if we cannot validate model performance before deployment, then we cannot know whether the model will do harm to its users, bystanders, or the environment. This problem has already manifested with self-driving cars. This report being concerned with object classification in self-driving cars, let us recall the first pedestrian fatality by a self-driving car, in 2018, where the vehicle collided with a cyclist at night, after the self-driving system became confused about the nature of the object ahead. [Smiley, 2022] Despite advancements in self-driving, such dangerous underperformance in low light conditions remains prevalent. [Abdel-Aty & Ding, 2024] The impact of self-driving crashes is not only a matter of direct casualties or property damage, but also a loss of the public trust necessary to advance safer-than-human systems across sectors. Validation is becoming legally necessary, as UK law now requires that "[self-driving safety] is at least as high as careful and competent human drivers". [Harper et al., 2024]

Research Gap

In our understanding of what various vision models consider anomalous examples of objects in image datasets, the impact of what data channels are used has not been thoroughly explored. Specifically, previous work has neglected stereo disparity, an analogy of depth by the difference between two cameras' views, focussing solely on red, green, and blue light intensity channels (RGB). Therefore, this proposal is for an initial investigation into whether changing the three-channel encoding of images from RGB to RG-chromaticity and stereo disparity changes which images the vision models considers most anomalous.

Potential Results

By comparing which images each model finds most anomalous when encoded as RGB, and then when encoded as RG-chromaticity and stereo disparity, across a number of datasets and models to aid the generalisability of findings, it will be determined if the difference is significant. If so, further work will be encouraged on whether this difference can be exploited for the benefit of better handling anomalous images when assessing model performance, with assessment of output reasonableness more so than the existing methods of comparison to human expectation.

The work may find that a number of images are considered anomalous examples of an object across models, which could help highlight to a human mislabelling or highly anomalous examples that might not be expected to be identified in the manner they are labelled, in these datasets that are too large for manual validation of every image, improving the quality of data available for training models.

The work could alternatively find that there is little to no association in how anomalous each image is considered to be across models, which it is proposed would suggest these were more reflective of the individual model's process than a property of the image. This could advise against future qualitative research into what makes an image be considered anomalous, instead proposing the focus be on explainable AI, and investigation of the apparently importantly varied processes.

Having accounted for such variations in the outputs as may be attributed to which model is selected, the significance attributable to RGB vs RG-chromaticity and depth can be assessed. An example to illustrate why we might expect depth information - where disparity encodes depth, as its inverse - to be beneficial to identifying anomalous examples of objects, is if an object is depicted with appropriate perspective on a flat surface, a perfect classification algorithm assessing only colour (RGB), should identify it as the object depicted, whereas with depth information, a perfect model might identify that real examples of the object depicted are characteristically not flat, and thus be able to distinguish this case in a way impossible with RGB.

Risk Mitigation

The principal risk to this project is suspected to be over scoping, as I cannot be certain how challenging each stage will be. However, the scope may not be reducible, as there is a single series of necessary actions to gather the necessary data, these laid out in the table here. What is flexible in scope is the analysis stage, with the number of Potential Results mentioned above, in how much of the answering of these is left to the potential further work suggested there.

As the probability of the scope being too large cannot be much mitigated, I have decided to reserve four additional weeks to mitigate the severity of its consequences if this does occur, thus reducing the total risk.

Week (the Monday of)	Activity
1 (24th of June)	Familiarise self with Anomalib library and datasets
2 (1st of July)	Write RG-chromaticity and stereo disparity encoder
3 (8th of July)	Implement Anomalib to train a given model on a given dataset
4 (15th of July)	Train models on datasets, and save image rankings
5 (22nd of July)	Data analysis in accordance with the Potential Results discussed
6 (29th of July)	Report findings
7 (5th of August)	Reserved as a redundancy, project may take longer than expected
8 (12th of August)	Reserved as a redundancy, project may take longer than expected
9 (19th of August)	Reserved as a redundancy, project may take longer than expected
10 (26th of August)	Reserved as a redundancy, project may take longer than expected

Ethical issues are not anticipated, ethical approval has not been sought.

Technical difficulties may occur, given my lack of familiarity with the specific software resources used, which is mitigated by working in a PhD office with colleges using the same tools for other projects, and weekly team meetings where challenges will be discussed.

Required Resources

This project will use publicly available image datasets and vision models.

Programming will be conducted with python, and the computer vision library Anomalib will be used for its range of opensource vision models.

There are a number of datasets available, and I expect to find more while some models are busy training during the data collection of week 5, the ones I currently expect to use are DurLAR, KITTI, and CityScapes. I will get it working on one dataset first, at which point expanding to others is only the minimal task of downloading the images and changing which files the python script accesses.

A computer with the Linux operating system will be required, which is satisfied by working in the PhD office that accompanies the heavy-duty computer vision lab associated with vehicle automation research at Durham University.

References

Abdel-Aty, Mohamed; Ding, Shengxuan (2024) "A matched case-control analysis of autonomous vs human-driven vehicle accidents". Nature Communications 15, Article 4931. DOI: 10.1038/s41467-024-48526-4

Harper, Mark (The Rt Hon); Department for Transport; Centre for Connected and Autonomous Vehicles (2024) "Self-driving vehicles set to be on roads by 2026 as Automated Vehicles Act becomes law". Published under the 2022 to 2024 Sunak Conservative government of the United Kingdom.

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.

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

Appendix: Explanation of RG-chromaticity and depth, as an alternative to RGB

RGB is the form of image encoding the public are most likely to be familiar with, each pixel of an image has three values, the intensity of red, green, and blue light, which is easily translated to display on a matrix of micro-LEDs, where many red, green, and blue light-emitting diodes (LEDs) are positioned in a grid, and each supplied with power appropriate to the intensity of the light of that colour for that pixel, creating the impression of a colour to any human viewer that perceives colour with red, green, and blue photoreceptors (cone cells).

Chromaticity instead considers colour independently of intensity, as the proportion of the total intensity belonging to each colour. For example: red intensity / (red intensity + green intensity + blue intensity). As the red and green chromaticity is sufficient to imply the blue chromaticity, equal to 1 – (red + green), the third channel is typically used for total light intensity, thus encoding all the same information as with RGB. This project will instead use the third channel for stereo disparity, which implies depth from the difference between two cameras, and neglects total light intensity.