Towards Requirements Specification for Machine-learned Perception Based on Human Performance

Boyue Caroline Hu
Dept. of Computer Science
University of Toronto
Toronto, ON, Canada
boyue@cs.toronto.edu

Rick Salay
Dept. Electrical and Computer Engineering
University of Waterloo
Waterloo, ON, Canada
rsalay@gsd.uwaterloo.ca

Krzysztof Czarnecki
Dept. Electrical and Computer Engineering
University of Waterloo
Waterloo, ON, Canada
kczarnec@gsd.uwaterloo.ca

Mona Rahimi
Dept. of Computer Science
Northern Illinois University
DeKalb, IL, USA
rahimi@cs.niu.edu

Gehan Selim
Dept. of Computer Science
University of Toronto
Toronto, ON, Canada
gehan.selim@utoronto.ca

Marsha Chechik
Dept. of Computer Science
University of Toronto
Toronto, ON, Canada
chechik@cs.toronto.edu

***Abstract*—The application of machine learning (ML) based perception algorithms in safety-critical systems such as autonomous vehicles have raised major safety concerns due to the apparent risks to human lives. Yet assuring the safety of such systems is a challenging task, in a large part because ML components (MLCs) rarely have clearly specified requirements. Instead, they learn their intended tasks from the training data. One of the most well-studied properties that ensure the safety of MLCs is the robustness against small changes in images. But the range of changes considered *small* has not been systematically defined. In this paper, we propose an approach for specifying and testing requirements for *robustness* based on human perception. With this approach, the MLCs are required to be robust to changes that fall within the range defined based on human perception performance studies. We demonstrate the approach on a state-of-the-art object detector.**

I. Introduction

Systems that use machine learning (ML) to replicate or improve upon human competencies have become prevalent in different areas of science and engineering. For example, ML is used in automated driving systems for several purposes, including perceptual tasks such as pedestrian detection. The fast-paced development of such systems, e.g., Tesla Autopilot, has brought forth safety concerns because erroneous behaviors can lead to fatal accidents [23]. Guaranteeing safety for ML components (MLCs) of such systems is challenging because MLCs are designed to learn from training data rather than follow a list of carefully defined requirements [14]. In addition, it is often difficult to rigorously specify the tasks that MLCs are expected to perform. For example, socially-constructed concepts, like pedestrians, are hard to specify because there is no consensus on their definition [14].

Ashmore et al. identified a list of desired properties of MLCs that should be considered as requirements: performance, robustness, reusability and interpretability [1] (see top of Fig. 1). Robustness concerns the ability to handle stressful environmental conditions and unseen or unexpected data [22], both common in real world scenarios. Therefore, requirements that ensure robustness are crucial to assure that decisions made by ML can be trusted in safety-critical contexts. In this position paper, we identify a class of robustness requirements that are amenable to precise specification and propose a method for doing so.

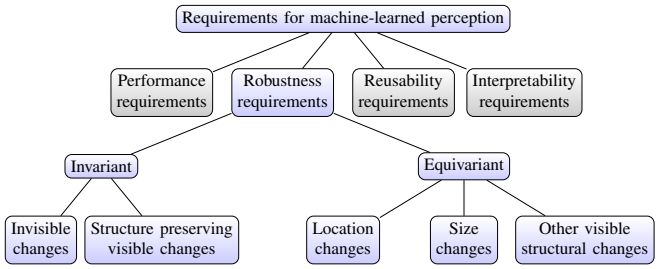

Fig. 1. Taxonomy of requirements for machine-learned perception.

It has been shown that ML systems are vulnerable to adversarial attacks, where minimal changes to the input image can cause misclassification [13]. Such behaviour is clearly undesirable for safety-critical systems. Thus, several studies investigated testing (e.g., [20]) and verifying (e.g., [8]) MLCs to guarantee their robustness within a small neighbourhood of the original image [8]. Yet, several questions require further investigation, e.g., how should this neighbourhood be defined? How should the resulting requirements specification be communicated to the user? How should we check whether the MLC satisfies this robustness requirement?

Position and contributions: Specifying full requirements of the expected behaviour for MLCs may not be feasible. Instead, we propose using human performance as a baseline

to express and formally specify a subset of the minimal expected input-output behaviour of MLCs. We investigate robustness requirements for MLCs designed for perception tasks, since robustness is crucial for the decisions made by the MLCs [5]. Human performance is used to bound the amount of changes that the MLCs are required to be robust to. We present a systematic method of generating such requirements and a method for testing whether the requirements have been satisfied. Our requirements can be used for verification and safety guarantees for MLCs in safety-critical systems.

The rest of the paper is organized as follows: Sec. II defines the proposed form of a robustness requirement for perception MLCs. We describe, and illustrate with an example, our proposed approach to specify and test the requirements based on human performance in Sec. III. Sec. IV summarizes the work related to requirement engineering, testing and verification for MLCs. We conclude in Sec. V with the summary of the paper and suggestions for future work.

II. Robustness Requirements

In this section, we discuss the form of *robustness* requirements for perception MLCs (see Fig. 1).

Using human performance as a baseline, we will assume that perception MLCs are required to be robust to any input modification that would not change human perception. For example, a pedestrian detector should still classify an input as a pedestrian even if a small amount of noise is added to the image or it is rotated by a limited amount. This is a common assumption made by work that explores robustness in the context of adversarial examples (e.g., [8]). For our purposes, we only consider modifications that can be formally defined as *transformations*. Investigating modifications in images that cannot be formally expressed as transformations, e.g., changing the clothes of a pedestrian, is left as future work. Some examples of transformations are:

- *Affine transformations* [10] such as scaling and rotating.
- Transformations modifying different aspects of the perceptual context [22]:
  - *Light sources*, e.g., changing the color or brightness of the light.
  - *Medium*, e.g., adding weather conditions like rain or fog.
  - *Objects*, e.g., changing position of the object.
  - *Observer (camera)*, e.g., different viewpoint and exposure of the camera, different amount of visual noise.

As shown in Fig. 1, we further refine requirements for robustness as *invariant* and *equivariant* requirements. An MLC can have multiple outputs and different outputs may be required to be invariant or equivariant with respect to given a transformation of the input. For example, an object detector produces a class label and a bounding box position and extent for each object it detects in the input image. With respect to a translation transformation that moves objects, we require that bounding box position is equivariant and moves a corresponding amount, while the class and bounding box extent is invariant. Another invariant here is that the set of detections is preserved — every object previously detected will still be detected (i.e, won't create false negatives) while every object now detected was also detected previously (i.e., won't create false positives). As another example, the transformation that adds noise to an image should leave all the outputs invariant.

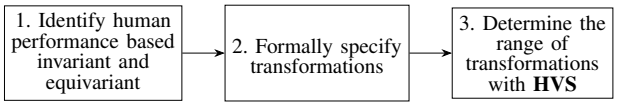

Fig. 2. Process for generating perception MLC robustness requirements.

We can further refine an invariant based on whether the modifications are (i) not visible to a human (e.g., adding slight noise), or (ii) visible but do not affect a human's ability to recognize the object as structural information in the images are preserved [25]. This distinction is important because if a transformation can be shown to be non-visible to humans then it must also be an invariant. On the other hand, the invariance of each type of visible transformation must be argued for separately.

Finally, to formally specify a robustness requirement based on a transformation, the *range of modifications* for which an MLC should be invariant or equivariant must be determined. For example, there is a bound on the amount of noise that can be added to an object image before a human will no longer be able to detect it. Thus, based on the parameters of the specific transformation, there is a range of parameter values determined by human performance limitations that the MLC must be robust to.

To summarize, a robustness requirement for a perception MLC consists of the following components: a formally defined transformation over the inputs of the MLC, a range of values for transformation parameters, and, an indication of the invariance or equivariance of each output of the MLC for this range of transformations.

III. Deriving and Checking Requirements

In Sec. II we defined the form of a robustness requirement for a perception MLC. In this section, we introduce a systematic method for deriving such requirements based on research about human performance in perception. We also discuss how to verify that an MLC satisfies such a requirement. We illustrate our method on the state-of-the-art object detector YOLO v4 [4].

As the first step of our method, shown in Fig. 2, we identify types of modifications that humans are invariant and equivariant to when recognizing objects in images. To do that, we propose to use studies of the *Human Visual System* (HVS). The HVS includes the eyes, the connecting pathways to the visual cortex, and the visual cortex [11]. Different versions of the HVS model can be designed to include the relevant human vision characteristics for different vision applications to improve visual quality [12], such as image enhancement, segmentation, coding, and image quality assessment [3]. For example, research on the HVS has shown that spatial shifting and rotations should have little effect on visual fidelity [6]

and therefore be equivariant. And noise in image is shown to preserve local structure of an image well [25]; thus, human performance should be invariant to certain amount of noise in images.

In the next step, we identify transformations to formally define the modifications in images discussed in the first step. For example, a noise transformation can be modeled as adding Gaussian noise to images, which has a probability density function equal to that of the Gaussian distribution [2] and takes two parameters: mean and variance. Thus, since noise preserves the structure of the image, human performance should be invariant to additional Gaussian noise within a range of values for the parameters.

In the last step, we determine the range of parameters for the transformations that MLCs are required to be invariant or equivariant to. To do this, we propose to use a search (e.g., binary search) through the space of parameter values along with an oracle for determining whether the transformation with a given parameter value is invariant/equivariant for humans. For example, Gaussian noise with different variance values (we assume the mean is 0) are added to an image and the result is checked with the oracle.

One possible choice for an oracle is an Image Quality Assessment (IQA) metric based on HVS research. An IQA metric is the algorithmic evaluation of the objective quality of an image for a human viewer. With access to the original (distortion-free) image, IQA metrics are able to determine the effect of image distortions on visual quality [11].

A large variety of IQA metrics with different capabilities exist. Some check the perceivability of distortions by computing computing contrast thresholds for detection and some include overarching i.e., what the HVS attempts to achieve when the human is shown a distorted image for example, some relate loss of information to perceptual loss of quality [6]. Therefore, we aim to determine which IQA metrics are well suited to obtain the range of distortions that do not affect human's detection of an object. Further use these metrics to generate requirements for an MLC's robustness.

Continuing the example, an IQA like Signal-to-Noise Ratio (VSNR) [6] that can check visibility of distortions would be appropriate to determine the range of values for the variance of Gaussian noise that is not visible to human. The search process can use VSNR to check each noisy image generated with a particular variance value to determine whether the noise is visible. Fig. 3 shows an image obtained from Berkeley DeepDrive [21]. For this image, the range of values for the variance of the Gaussian noise resulting in invisible changes, obtained using a simple linear search and VSNR as the oracle is the interval $[0, 0.067]$. Therefore, MLCs are required to be robust to additional Gaussian noise with mean 0 and variance within this interval.

To demonstrate the testability of our example robustness requirement, we use it to evaluate the state-of-the-art object detector, YOLO v4 [4] with pre-trained weights. Fig. 3 and Fig. 4 show YOLO's output for the same image without and with the additional Gaussian noise, respectively. The added

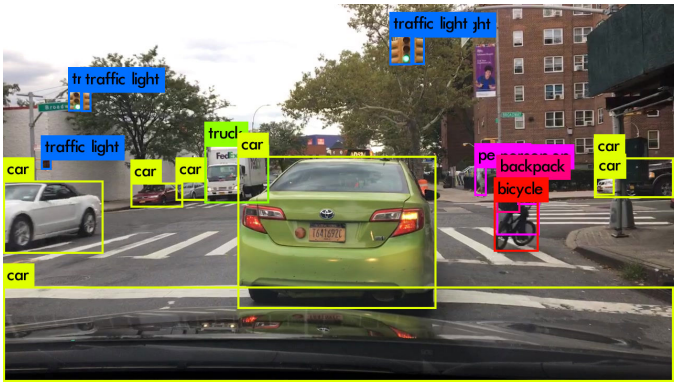

Fig. 3. YOLO detection of the original image.

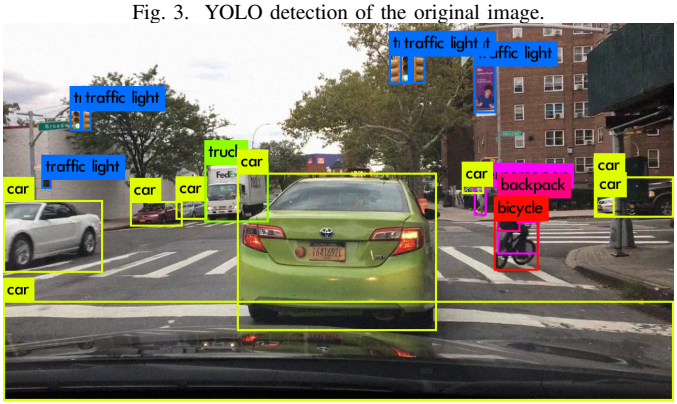

Fig. 4. YOLO detection of the image with additional Gaussian noise added (mean=0, variance=0.001). The sign between the traffic lights and the red building is misclassified as a traffic light. In addition, a car below this sign, which was not detected in the original image, is now detected.

noise is invisible to humans, but it caused YOLO to misclassify the sign between the traffic lights and the red building as a traffic light. In addition, a car below this sign, which was not detected in the original image, is now detected. lead to misclassification by YOLO (details are in Fig. 4). These differences clearly show that the specific YOLO model is not robust with respect to the addition of the Gaussian noise that is not visible to humans. These mistakes can lead to deadly consequences in safety-critial systems like automated driving and thus the use of this version of YOLO would be unsafe.

In general, there are a number of systematic approaches to determine whether a robustness requirement is satisfied by a given MLC. One of these is using *metamorphic testing* [24]—a technique to generate follow-up test cases based on existing ones [20]. The first step of this method is to identify metamorphic relations (MRs) that relate multiple pairs of inputs and outputs of the software being tested. For invariant robustness requirements, the transformation relates inputs and the IQA metric is used as an equivalence relation on the corresponding outputs. This MR is then used to generate new test cases by generating new images from the original one by applying the transformation, and then checking whether the results are considered identical to the original by the IQA metric.

While testing can reveal counter-examples and thus show that the requirement is not satisfied, it is not sufficient to prove requirement satisfaction—formal verification would be needed for this task.

IV. Related Work

In this section, we list previous work on requirements for MLCs, testing or verifying robustness of MLCs and adversarial examples generation using HVS.

For requirements specifications for MLCs, Vogelsang et al. [19] conducted interviews with data scientists to know their opinions about the types of requirements that are necessary for MLCs, but they did not proceed with detailed specifications of the requirements mentioned in the paper. Rahimi et al. [14] suggest a method for creating MLC requirements by explicitly specifying domain-related concepts, whereas we focus on robustness requirements using human perception studies. There are other attempts by different communities to specify requirements for MLCs in different types of software systems by creating component-level specifications [18], dataset specifications [9], model specifications [17] and development process specifications [16] of MLCs. We proposed a different approach to robustness requirements specification using human performance as a base line.

Testing and verification of the robustness of MLCs have been explored by the software engineering community. Xie et al. [20] gave partial specifications of robustness of ML models with respect to mutations in the training set or test set or both. In contrast, we aim to study mutations of test images and measure the similarity of the original and mutated images based on how humans perceive the difference. Huang et al. [8] explored safety verification for the robustness of MLCs. They focused on verifying that adding small perturbations to an image, e.g., changes to the values of a few pixels, should not affect the output of MLCs. Our approach differs in the definition of the "small range"—we define the range based on human perception as a range that can be added without changing a human interpretation of the objects.

Previous work has explored generation of adversarial examples using HVS. Ho et al. [7] demonstrated the idea of incorporating HVS models into adversarial AI to produce adversarial examples with small visual difference. But the work did not clearly define the notion of the perceptual distance, did not include a mathematical interpretation of the HVS, and focused on invisible changes only. Rozsa et al. [15] proposed calculating an *adversarial similarity score* to quantify the differences between the original image and adversarial images using SSIM [25], a type of IQA.

V. Conclusion

In this paper, we emphasized the importance of requirements specification for machine-learned perception in safety-critical systems when making safety guarantees. We proposed a method to specify and test the robustness requirements of MLCs based on human performance approximated using HVS models. A demonstration on YOLO v4 showed that it is not robust against changes, e.g., Gaussian noise that are not visible to humans.

Our next steps are to conduct more experiments with different types of transformations and IQAs; to implement the specification and testing procedure for the requirements; to verify the requirements and to validate our approach.

Other questions remain open for future investigation: How to specify requirements related to changes that cannot be expressed as transformations, e.g., changing clothes on a pedestrian? How to specify requirements related to changes that are not structure preserving but where humans are still able to detect the objects, e.g., shadows that cover parts of an object? Which IQA metrics better suit different types of requirements? Are there other studies that can be used to assess human performance besides HVS models?

References

[1] R. Ashmore et al. Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. *ArXiv*, abs/1905.04223, 2019.

[2] T. Barbu. Variational Image Denoising Approach with Diffusion Porous Media Flow. *Abstract and Applied Analysis*, 2013, 2013.

[3] A. Beghdadi et al. A survey of perceptual image processing methods. *Signal Processing: Image Communication*, 28(8):811–831, 2013.

[4] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. *ArXiv*, abs/2004.10934, 2020.

[5] M. Borg et al. Safely Entering the Deep: A Review of Verification and Validation for Machine Learning and a Challenge Elicitation in the Automotive Industry. *JASE'19*, 1:1–19, 2019.

[6] D. M. Chandler and S. S. Hemami. VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images. *IEEE Trans. on Image Processing*, 16(9):2284–2298, 2007.

[7] Y. Ho and S. Wookey. The Human Visual System and Adversarial AI. *ArXiv*, abs/2001.01172, 2020.

[8] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu. Safety Verification of Deep Neural Networks. In *CAV'17*, pages 3–29, 2017.

[9] M. Kohli, R. Summers, and J. Geis. Medical Image Data and Datasets in the Era of Machine Learning. *JDI*, 30(4):392–399, 2017.

[10] R. R. Mekala et al. Metamorphic Detection of Adversarial Examples in Deep Learning Models with Affine Transformations. In *MET'19*, MET '19, page 55–62. IEEE Press, 2019.

[11] A. Moorthy, Z. Wang, and A. Bovik. *Visual Perception and Quality Assessment*, pages 419–439. 2011.

[12] M. Nadenau et al. Human Vision Models for Perceptually Optimized Image Processing – A Review. *Proc. of the IEEE*, 2000.

[13] N. Papernot et al. The Limitations of Deep Learning in Adversarial Settings. In *Euro S&P'16*, pages 372–387, 2016.

[14] M. Rahimi et al. Toward Requirements Specification for Machine-Learned Components. In *AIRE'19*, pages 241–244, 2019.

[15] A. Rozsa, E. M. Rudd, and T. E. Boult. Adversarial Diversity and Hard Positive Generation. *CVPRW'16*, pages 410–417, 2016.

[16] R. Salay and K. Czarnecki. Using Machine Learning Safely in Automotive Software: An Assessment and Adaption of Software Process Requirements in ISO 26262. *ArXiv*, abs/1808.01614, 2018.

[17] S. A. Seshia et al. Formal Specification for Deep Neural Networks. In *ATVA'18*, pages 20–34. Springer, 2018.

[18] S. A. Seshia and D. Sadigh. Towards Verified Artificial Intelligence. *ArXiv*, abs/1606.08514, 2016.

[19] A. Vogelsang and M. Borg. Requirements Engineering for Machine Learning: Perspectives from Data Scientists. In *AIRE'19*, pages 245–251, 2019.

[20] X. Xie et al. Testing and Validating Machine Learning Classifiers by Metamorphic Testing. *Journal of Systems and Software*, 84(4):544–558, 2011.

[21] F. Yu et al. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *CVPR'20*, June 2020.

[22] O. Zendel et al. CV-HAZOP: Introducing Test Data Validation for Computer Vision. In *ICCV'15*, pages 2066–2074, 2015.

[23] M. Zhang et al. DeepRoad: GAN-based Metamorphic Autonomous Driving System Testing. *ArXiv*, abs/1802.02295, 2018.

[24] Z. Q. Zhou et al. Metamorphic Testing and its Applications. In *ISFST'04*, page 346–351, 2004.

[25] Zhou Wang et al. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. on Image Processing*, 13(4):600–612, 2004.