https://doi.org/10.2352/EI.2023.35.8.IQSP-307

A framework for the metrification of input image quality in deep networks

Alexandra Psarrou, Sophie Triantaphillidou, Imran Feisal, Oliver van Zwanenberg, University of Westminster, London, UK

Abstract

Deep Neural Networks (DNNs) are critical for real-time embedded imaging applications including autonomous vehicles. DNNs are often trained and validated with images that originate from a limited number of cameras, each of which has its own hardware and image signal processing (ISP) characteristics. However, in most real-time embedded systems, the input images come from a variety of cameras with different optical components, sensors and ISP pipelines, and often include perturbations due to a variety of scene conditions. Data augmentation methods are commonly exploited to enhance the robustness of such systems. Alternatively, methods are employed to detect input images that are unfamiliar to the trained networks, including out of distribution detection. Despite these efforts DNNs remain widely systems with operational boundaries that cannot be easily defined. One reason is that, while training and benchmark image datasets include samples with a variety of perturbations, there is a lack of research in the areas of metrification of input image quality suitable to DNNs and a universal method to relate camera system performance to DNN robustness using appropriate quality metrics. This paper addresses this lack of metrification specific to DNNs systems and introduces a framework for systematic modification of camera system performance parameters that relate input image quality attributes to DNN performance.

Introduction

In recent years we have seen significant advances in image processing and computer vision applications based on Deep Neural Networks (DNNs). This is a critical technology for several real-time embedded imaging applications including autonomous vehicles, smart cities, and industrial computer vision.

However, even though deep neural networks are trained and validated based on a wide range of images, they have also been shown to be quite brittle when they come across artificial or natural adversarial examples. As early as 2014, it was shown that deliberate or natural adversarial changes occurring to the input images deteriorated the performance or classification decision of deep networks. For example, in [6], Goodfellow *et al.* demonstrated that by adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, they can change GoogleNet's classification of an image.

Other examples that cause deep neural network performance deterioration during downstream include the use of images that are taken in environmental or viewpoint conditions or have artifacts due to the imaging process that may not have been part of the training set [1]. Figure 1 shows an example of how the performance of AlexNet deteriorates when exposed to images that are corrupted with Gaussian noise, with distortion severity increasing from left to right, with the left-most image having no distortion (original).

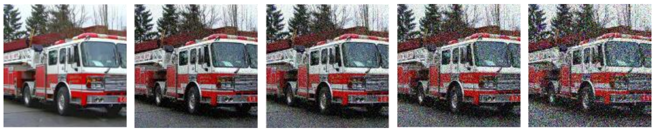


Figure 1*: In the presence of noise the recognition of the fire engine deteriorates as Gaussian Noise increases. In the original uncorrupted image (leftmost), top 1% the fire engine is recognized with 66% accuracy. As Gaussian noise is added to the image increasing progressively from standard deviation 0.08 to 0.12, 0.18 and 0.26 (rightmost image), top 1% accuracy for the fire engine is decreased respectively to 60%, 24%, 9% and 0%.*

In all such cases, the operational boundaries of DNNs cannot be explained or otherwise quantified. While widely studied in recent years, it is not yet clear the level of unseen distortions deep neural networks can tolerate, or the exact reasons for any network performance degradation. If machine learning models are to be trusted, particularly when deployed in critical applications, their level of robustness needs to be improved. This can only be possible if the underlying mechanisms are better understood.

In this paper we explore a methodology that would enable us to better understand such underlying mechanisms by introducing a framework for the metrification of camera image performance that relates input image quality attributes to DNN performance. The aim of our proposed framework is to provide information on the operational boundaries of deep neural networks with respect to the image artefacts produced by camera pipelines for varying image content and pipeline characteristics. Our aims are to: (a) inform changes to the design of the imaging pipelines to optimize embedded imaging systems; (b) examine breath of variability in the input data used in the training of networks; (c) inform changes to architectural structures that will account for the operational output of the image sensors used in embedded imaging systems. The rest of the paper is organized as follows. The section on Related Work presents a literature review on approaches taken to date to address the brittle characteristic of DNNs followed by a discussion on the issues related to DNNs robustness. In the section Domain Generalization, we discuss how current methods address the data shift issues inherent to the poor robustness of DNNs, whereas in the section Data Valuation using Reinforcement Learning we present a framework on how to use task specific reinforcement signal to model task specific fitness quality of images. In the section Observations using Imaging Performance Metrics we present our observations on DNNs robustness with respect to imaging performance metrics before we summarize in Conclusions.

Related Work

The number of safety-critical applications affected, and the volume of the current research are testimony to the need to provide a better understanding of the operational boundaries of DNN systems used in real-time imaging operations and thus enhance the interpretability and explainability of such systems.

DNNs are often trained and validated with images that originate from only a limited number of camera systems, each of which has its own hardware (optics, sensors) and image signal processing (ISP) limitations and artifacts. However, in most real-time embedded systems, the input images come from a variety of cameras systems with different ISP pipelines, and also include perturbations due to a variety of external (scene) conditions.

Data augmentation methods are often exploited to expand the subset of trained images to include images with distortions introduced by changing environmental conditions, such as modelled haze or rain, or by modelling changes in the imaging system, for example modelling varying camera parameters, by introducing Gaussian camera blur, Poisson noise or quantisation errors. Such data augmentation may result to improved performance of the network for artifact-specific images [6] but also degrades the performance for artifact-free images [7]. To tackle the problem from a different perspective, several methods have been researched to detect images from the target domain that are unfamiliar to the trained networks, such as out of distribution (OOD) detection, anomaly detection and open set recognition methods [13]. In addition, methods that use image quality assessment [12] or data valuation [11] aim to assign a valuation to the appropriateness of specific image samples. However, despite such efforts, DNNs remain widely systems whose operational boundaries cannot be explained or otherwise quantified [15].

One of the reasons that DNNs for imaging applications are failing to provide the required robustness in downstream, is that while augmented training and benchmark image datasets for DNNs include samples with a variety of natural and adversarial occurred perturbations, it is not possible to train or test for a model that includes all potential data shifts that may occur due to imaging artifacts. In addition, recent work suggests that not all samples used in training are equally useful to learn from, for example robustness errors may occur due to the inclusion of low-quality samples in datasets [4]. Out of distribution methods aim to address the former reason but they are mainly concerned with semantic shifts in the test datasets. Changes in appearances such as, for example changes in image contrast, are either excluded from the evaluation stage or treated as a sign of OOD, which contradicts with the primary goal in machine learning, i.e., to generalize beyond the training distribution [11]. The latter reason for the lack of DNN robustness is primarily tackled by approaches that use data valuation [11] and image quality assessment methods [12]. Their aim is to assign image quality values to train data and/or downstream individual data depending on the application. However, the majority of well-performing, established image quality metrics and relevant standards used in these methods are based on the knowledge that the recipient of images are human observers, which is not necessarily suitable for DNNs systems. Our objective is to study this problem by providing a framework to relate quality to DNN performance using meaningful camera system performance metrics.

Robustness Issues in DNNs

As mentioned earlier, the appearance of a captured scene can vary enormously, due to physical changes (in environment, viewpoint etc.) or camera variations and artifacts caused by the imaging processing pipelines. In our work we concentrate on attributes changes caused by capture and processing pipelines, and we use a subset from the ImageNet-C benchmark for some preliminary analysis [8].

ImageNet–C is a subset of ImageNet classes/images, to which a large number of corruptions have been added. In each image 25 corruption types have been applied, and for each corruption there are 5 levels of distortion with increasing severity. For our preliminary work we focus on only four corruptions: Gaussian Noise, Gaussian Blur, Motion Blur, Brightness Variations). These represent image degradations caused by sensor noise, lens blur, camera shaking and exposure errors respectively.

Figure 2, shows sample images of the ImageNet subset we used, representing images that automotive systems may encounter. These are sample uncorrupted images from each of the 28 classes we are considering and are part of the images AlexNet was trained on for object classification [10].

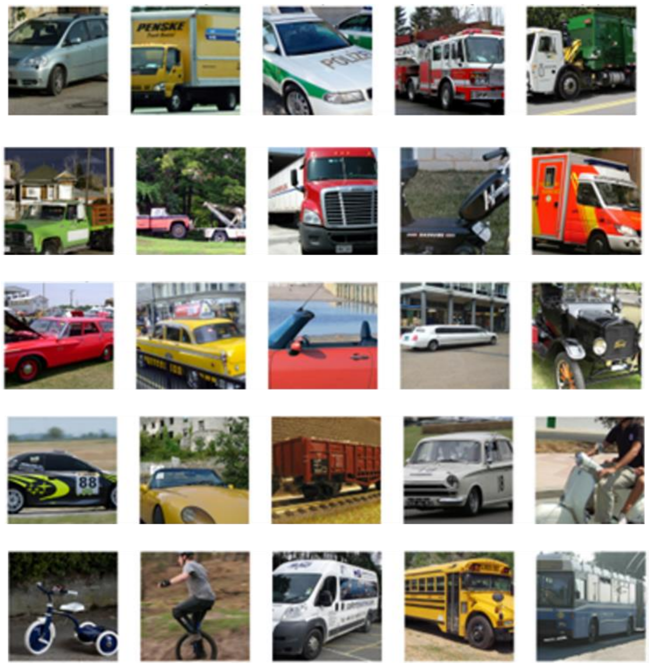


Figure 2: Sample images from 25 out of the 28 ImageNet classes used during our analysis.

Figure 3 shows the normalized median network performance change when the Gaussian Noise and Gaussian (Defocus) Blur corrupted images of these classes are input into AlexNet during testing. In other words, how much the network performance for the target dataset diminished compared to the uncorrupted / trained dataset. Figure 4 shows the normalized media performance change when the images are corrupted with Motion Blur and changes in Brightness.

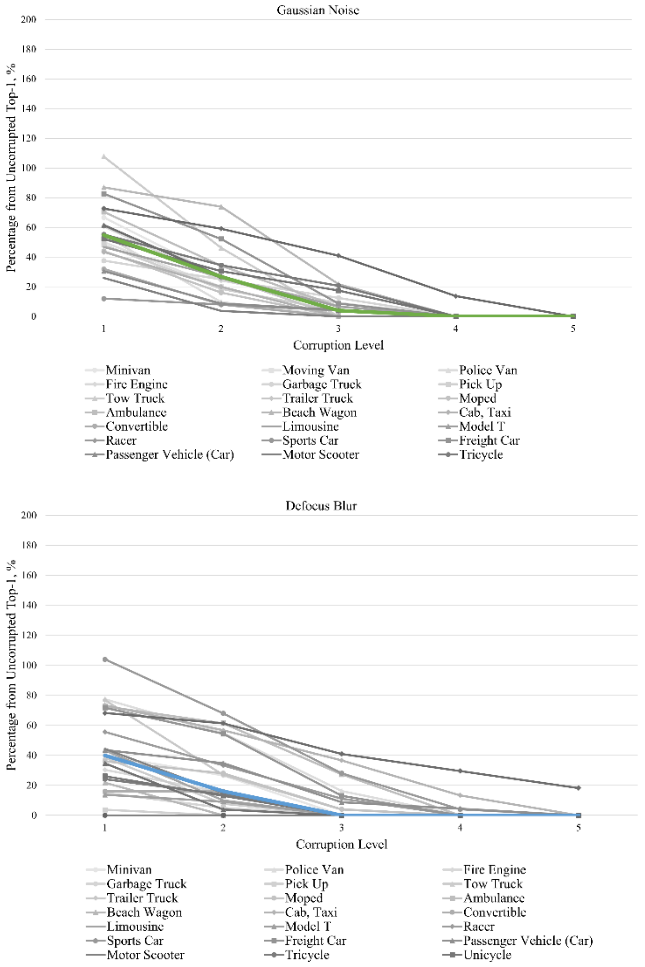


Figure 3: Normalized median network performance change for images corrupted with Gaussian Noise (top) and Defocus Blur (bottom) when input into AlexNet during downstream.

Based on the ImageNet-C subclasses used in this initial analysis we make the following observations regarding the effect of corruptions on DNNs robustness:

a) Based on the corruption level information alone, provided by the benchmark database, it is not possible to properly evaluate the effect that the corruptions/attributes have had on the images, and subsequently relate this appropriately to the performance of the network. This is because the corruption is presented as an arbitrary additive change (level) but does not suitably quantify the effect such a change has on the image quality.

b) There is a sharp drop in performance usually after corruption exceeds level 3. However, based on the current information in ImageNet-C we cannot explain why this drop exists due to the lack of standardized metrification for image quality.

c) Considering the network performance for each of the corruptions/attributes, we see that there is a lot of variation in the effect that they have. Brightness is shown to have the least effect whereas Gaussian Blur the biggest effect, but we do not know whether each level of individual distortion degrades the original image information equally. This is an important point which is often not accounted for in the relevant literature, and consequently wrong conclusions maybe drawn.

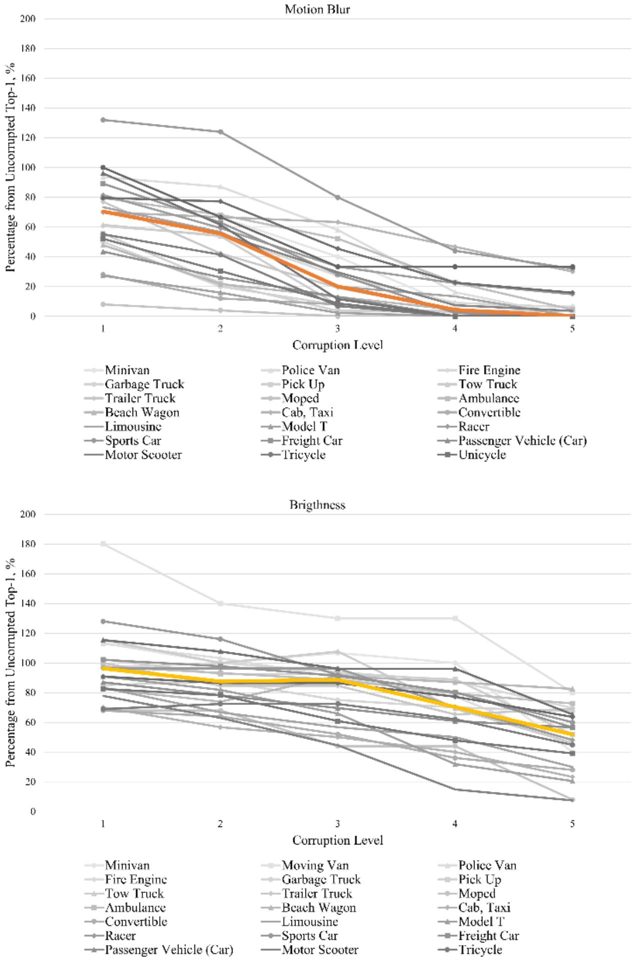


Figure 4: Normalized median network performance change for images corrupted with Motion Blur (top) and Brightness (bottom) when input into AlexNet during downstream.

Further, most of the results provided in the literature are based on average robustness network performances and therefore it is not clear how specific image characteristics and contents may affect DNNs' robustness. Figures 3 and 4 show the network performance for each class and highlight the median performance of the network for the chosen classes in this analysis. From Figures 3 and 4 we can see that the performance varies quite significantly from one class to the other. We call this scene dependent network performance. Most of this scene dependency is most probably due to variations in low level scene features (textures, colors, etc.), size of objects on the frame, as well as semantic variations. Another contributing factor, however, is variations in the original image quality of the difference classes/images since images in the ImageNet dataset have been originating from different camera systems with varying output quality. Any added artefacts may result to different levels of quality for each artifactual level.

Figure 5 summarizes the median performance of AlexNet for Gaussian Noise, Defocus Blur. Motion Blur and Brightness for the chosen ImageNet classes as a function of 'level' of corruption.

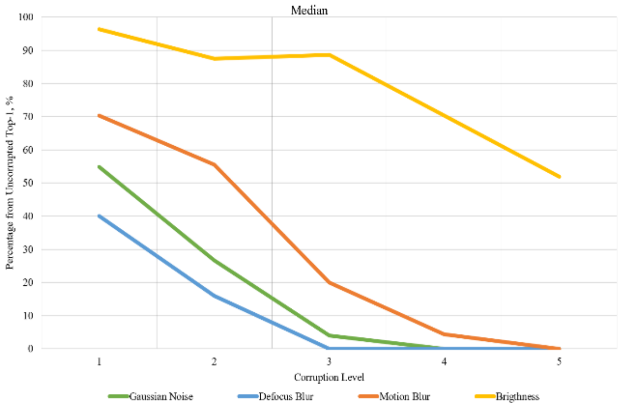


Figure 5: The median performance of AlexNet for Gaussian Noise, Defocus Blur, Motion Blur, Brightness for the chosen ImageNet classes.

Figure 6 focuses in two of the classes in ImageNet, "taxis" and "fire engine". We observe that even though both classes start with a very similar top-1% network performance of 0.60 and 0.66 respectively, the addition of the corruption causes significant network performance differences. Gaussian blur affected the fire engine class severely, while the taxis class less so. On the other hand, changes in brightness had almost no effect on the fire engine class, while affected the taxis class significantly.

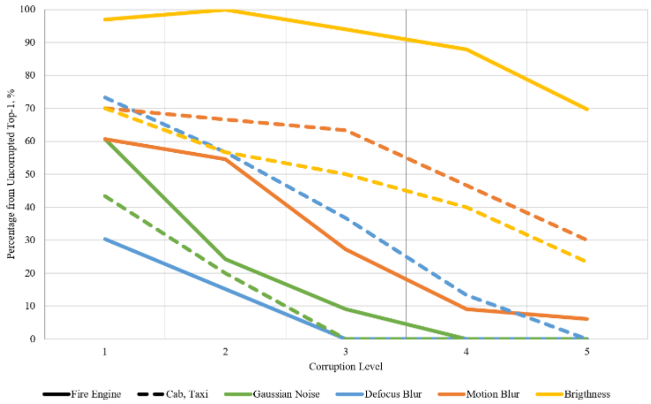


Figure 6: The median performance of AlexNet for two classes Fire Engine and Cab, Taxi under the four corruption types: Gaussian Noise, Defocus Blur, Motion Blur, Brightness.

It is important to note that, based on the ImageNet-C benchmark dataset, we cannot evaluate the networks' performances when more than one corruption is applied on the images. However, this is not a reflection of the image capture process faced by real-time embedded imaging applications. The effect that each additional artifact has to the distribution of a given dataset is multiplicative, thus requiring a much larger number of samples to model it [9].

Our observation from these experiments is that current benchmark datasets and methods do not sufficiently analyze the reasons behind the performance degradation of deep neural networks and do not account for variations observed in different classes or types of images.

Domain Generalization: Issues and Proposal

The underlying reason for the drop in performance of deep networks is that the source and target data are not independent and identically distributed [2]. When the deep networks models are deployed in real-life scenarios, out-of-distribution data are encountered, and that exposes specific biases in the databases.

Database biases give rise to a problem that is commonly called domain shift [13]. The main term that is used to describe the methods that are developed to address the performance deterioration that deep networks models face due to domain shift is broadly covered by domain generalization. Examples include data augmentation methods and domain alignment methods, just to name a few. Evaluation of domain generalization methods are mainly based on metrics that report the average performance of the models in the tested domain shift scenarios.

While this is a widely studied area, there are still doubts on the efficacy of the existing methods to generalize across datasets. The relative performance between methods varies across datasets and shifts [4]. In addition, there has been little work in defining the underlying mechanisms that cause these shifts and variations of network performance due to capturing system bias.

In order to address the problem of domain shift, in particular caused due to capturing system bias, we need an insight on the appropriateness of images for a given task. To achieve this, we are developing a framework that:

a) Introduces systematic imaging system parameter (attribute) variations that are representative of (model) real camera system variations.
b) Develops evaluation metrics for a set task.
c) Tests the abovementioned metrics for their suitability to describe image attributes/characteristics in an appropriate manner for neural networks, relevant tasks and architectures.

When discussing imaging metrics, image quality (IQ) assessment metrics are typically based on human perception [10] and therefore cannot necessarily work. IQ metric results are developed to correlate with the human visual system but not necessarily with performance variations in deep network performance. There are examples where image quality is learned through deep networks based on experts labelled assessment [3]. However, the reported results are also not always positive.

An alternative is to derive Image Fitness (IF) for input to network tasks by using reinforcement learning. Relevant models measure fitness for the purpose of images for a given network task. Examples include work by Yoon [15], where the model is domain agnostic and Saeed [11] where the model is applied in biomedical imaging. However, such metrics do not specify *why* any image may or may not be fit for purpose, as they don't relate the decision to the image's inherent information and attributes that in turn depend on the performance of the camera systems that have produced them.

We argue that using such models is insufficient to provide a learned image assessment that best fits a task. Subsequently, we can then investigate how learned image fitness metrics are related to imaging performance metrics that describe generic imaging system characteristics for average test signals as well as scene specification metrics that describe individual scene contents.

Data Valuation using Reinforcement Learning

The architecture in Figure 7 shows the data valuation model developed by Yoon [15] and represents the general principle behind the use of reinforcement learning to quantify the value of data.

Yoon proposed a reinforcement learning framework that learns data values jointly with a target task predictor model, therefore it integrates data valuation with the training of the target task predictor.

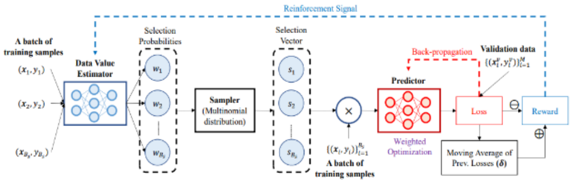


Figure 7: Data valuation using reinforcement learning [15]

The network has two components: the data value estimator (in blue) that is trained based on reinforcement learning with a sampling process, and the task predictor (in red), which can be any feed forward deep learning predictor model for a given task.

The data value estimator, modelled by deep reinforcement learning, learns how likely each datum is used in the training of the predictor model and is trained using a reinforcement signal of the reward obtained on a small validation set that reflects performance on a target task.

What is of most interest to us here, is that the output of the data value estimator provides a set of selection probabilities that rank the input samples according to their importance to maximize the performance of the network for the given task. This data valuation can correspond to the learned image fitness values.

In our framework shown in Figure 8, we propose to use a dataset of uncorrupted images and introduce systematic changes by employing physical camera system parameters (optics, sensor) and ISP models, and feed these systematically changed images to a data valuation model that is trained for a given task. We can use the model to rank the fitness for purpose of the corrupted images relevant to a given task.

Once we have received the fitness for purpose for the images, we can then relate them with the imaging performance metrics and scene metrics, therefore closing the loop and relating fitness for purpose to the imaging process.

A sample of imaging performance metrics that we aim to use for this purpose include metrics relevant to exposure and tone reproduction, edge frequency content, optical resolution, texture reproduction, optical aberrations, information content/capacity, signal-to-noise. For scene content our metrics include contrast, brightness, energy (busyness, complexity), coherence, colorfulness, and dominant color palettes.

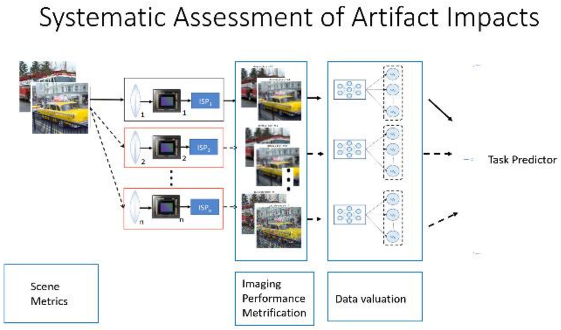

Systematic Assessment of Artifact Impacts
1
1
ISP$_1$
2
2
ISP$_2$
n
n
ISP$_n$
Task Predictor
Scene Metrics
Imaging Performance Metrification
Data valuation

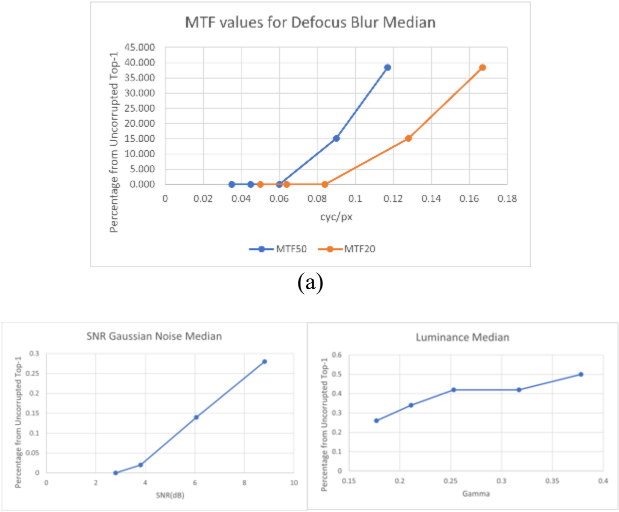
Figure 8: Framework for the systematic assessment of artifact impacts

Observations using Imaging Performance Metrics

Imaging performance metrics are routinely used by the imaging industries for imaging system design and optimization, since they can relate image attributes (such as resolution, noise, color and tone reproduction) to specific system properties (optical, sensor, color filter array, etc.). For example, performance metrics extracted from the camera Modulation Transfer Function (MTF), which describes the ability of the camera for reproducing contrast at different spatial frequencies, can be related to image sharpness (MTF50) or resolution (MTF10, MTF20). Figure 9 shows examples on how three corruption types applied on ImageNet-C images can be evaluated through imaging performance metrics derived from Imatest© test charts that have been subjected to the same levels of corruption as the ImageNet-C images, using the same corruption filters. The resulting metrics values are then related to the performance of the AlexNet network. Figure 9(a) shows how the MTF50 and MTF20 values extracted from the Imatest© charts corrupted by Gaussian Blur vary with the median performance of the network. MTF50 provides a measure of image sharpness, whereas MTF20 relates to vanishing resolution. In Figure 9(a) we observe that sharpness in images reduces the performance of the network deteriorates rapidly, whereas network deterioration is slower as image resolution is reduced.

Figure 9(b) shows the network performance with respect to signal-to-noise (SNR) ratio, where SNR (dB) = 20*$\log_{10}$ (signal/noise) measured using ISO 15739. We observe that network performance decreases linearly as SNR increases. Finally, Figure 9(c) shows how changes in global contrast, through the measure of gamma value, affects the performance of the network. Here we notice that relative changes in brightness and contrast have a relatively small effect on the robustness of the DNN, an observation also supported by previous studies [9]. But although the contrast range covers most camera contrast variations, it does not cover extreme image contrasts.

Figure 9 provides observations on the performance of the network based on example camera performance metrics that can potentially be used in our framework. They certainly require further evaluation and validation to draw meaningful conclusions on their suitability. Nonetheless, such descriptors are internal (relatable) to the quality of the image, as opposed to arbitrary values commonly used in evaluations found in literature, such as the standard deviation of the noise or blur filters. The latter describes changes external to the image and are therefore unrelatable.



(a)

(b) (c)

Figure 9: Median network performance variation with respect to scene metrics (a) MTF50 & MTF20, (b) SNR (c) Gamma value.

Conclusion

In this paper we report on the robustness of deep neural networks, based on sample images provided by ImageNet-C [9]. We observed that the current benchmark datasets cannot fully explain the underlying reasons for the brittle performance of deep networks. To address the problems discussed in the paper, we propose a framework based on metrics that relate camera system performance of cameras in embedded systems to the robustness of deep neural networks. The framework will employ: (a) systematic imaging variations to images as well as to test charts, from which imaging performance measurements can be derived. We do this using modelling imaging functions that describe physical camera system parameters (optics, sensor) and ISP models (rather than arbitrary models that are often found in the computer vision literature); (b) reinforcement learning to associate task depended valuation to images and derive image fitness metrics; (c) scene descriptors/metrics that can be used to differentiate between different original scene contents that may affect differently network performance (tackle network scene dependency); (d) association of image fitness metrics with imaging performance and scene description metrics. We aim to test the validity of the framework and proposed performance metrics with a number of different network architectures and tasks.

References

[1] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. Advances in neural information processing systems, 32:9453–9463, 2019.

[2] Dina Bashkirova, Dan Hendrycks, Donghyun Kim, Haojin Liao, Samafth Mishra, Chandramauli Rajagopalan, Kate Saenko, Kuniati Saito, B. Tayyab, P. Teterwak, B. Usman. VisDA-2021 Competition: Universal Domain Adaptation to Improve Performance on Out-of-Distribution Data. Proceedings of Machine Learning Research.volume 176, pp 66—79, 2022.

[3] L.S. Chow and R. Paramesran. “Review of medical image quality assessment”. In: Biomed. Signal Processing and Control 27 (2016), pp. 145–154.

[4] N Drenkow, N Sani, I Shpitser, M Unberath, “Robustness in Deep Learning for Computer Vision: Mind the gap?”, arXiv preprint arXiv:2112.00639, 2021

[5] E.W.S Fry, S. Triantaphillidou, R.B. Jenkin, R.E. Jacobson, and J.R. Jarvis. Noise Power Spectrum Scene-Dependency in Simulated Image Capture Systems, Electronic Imaging, 2020

[6] Ian Goodfellow, Shlens Jonathon, and Szegedy Christian. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

[7] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, Wieland Brendel, “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness”, arXiv preprint arXiv:1811.12231, 2018

[8] D Hendrycks, T Dietterich, Benchmarking neural network robustness to common corruptions and perturbations, arXiv preprint arXiv:1903.12261, 2019

[9] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021. Unsolved problems in ML safety. CoRR, abs/2109.13916.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. NIPS, 2012

[11] Shaheer U Saeed, Yunguan Fu, Vasilis Stavrinides, Zachary Baum, Qianye Yang, Mirabela Rusu, Richard E Fan, Geoffrey A Sonn, J Alison Noble, Dean C Barratt, Yipeng Hu, in International Workshop on Advances in Simplifying Medical Ultrasound: “Adaptable image

quality assessment using meta-reinforcement learning of task amenability", 2021

[12] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, Mohammad Sabokrou, "A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges", arXiv preprint arXiv:2110.14051, 2021

[13] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, Ludwig Schmidt, "Measuring robustness to natural distribution shifts in image classification", Advances in Neural Information Processing Systems, 2020

[14] M. Toneva., A. Sordoni, A, R. des Combes, A. Trischler, A, Y. Bengio, and G. Gordon. An empirical study of example forgetting during deep neural network learning. In ICLR, 2019

[15] J Yoon, S Arik, T Pfister, in , International Conference on Machine Learning: "Data valuation using reinforcement learning, International Conference on Machine Learning", 2020

Author Biography

Alexandra Psarrou is a Reader in Computational Vision at the University of Westminster. She received her BSc in Computer Science (1987) and PhD in Computer Vision (1996) from Queen Mary, London. Her research background is in machine learning with particular emphasis in neural networks for analysis of visual behavior. Most recently she has been applying computational techniques in modelling image quality and systems performance for mobile phone cameras and automotive applications.

Sophie Triantaphillidou is a Professor in Imaging Science at the University of Westminster, UK, and the Director of Computational Vision and Imaging Technology research group. She graduated with a BSc in Imaging Science (1995) and a PhD in the area of photographic digitization (2001). Her research is interdisciplinary, exploring interrelationships between imaging systems engineering, image contents and perception by human and machine vision systems. She is currently serving as IS&T Vice President for Conferences.

Imran Feisal is currently an undergraduate student pursuing a bachelor's degree in computer science at The University of Westminster. His passions lie within artificial intelligence, machine learning, and data analysis. Outside of academia, Imran is involved in programming competitions, and actively seeks opportunities to apply his skills to real-world challenges.

Oliver van Zwanenberg received his BSc at the University of Westminster, London, in 2017. Then moved on to pursue his PhD in that same year. His PhD, titled 'Camera Spatial Frequency Response Derived from Pictorial Natural Scenes', was awarded in 2022. In his thesis, a methodology was established that adapted the standard ISO12233 e-SFR to utilize captures of natural scenes to estimate the camera system performance. He is currently working as an Image Quality Engineer at Onsemi and is a keen photographer and videographer.