When and Why Test Generators for Deep Learning Produce Invalid Inputs: an Empirical Study

Summary Report

Advances in Testing Data-Intensive Software Applications
(IN0014, IN2107, IN45018)
Summer term 2023

Matteo Pinna
Technical University of Munich
Munich, Germany
{ge48fay}@tum.de

PAPER SUMMARY

Introduction

In the last decade, Deep Learning (DL) has achieved remarkable advancements leading to widespread adoption of DL-based systems in safety-critical and security-sensitive domains [13], such as self-driving vehicles [5] or robotics [24]. However, increase adoption also means new challenges, particularly in testing these large-scale, complex systems.

Traditional DL testing relies heavily on large, manually labeled datasets, which can be burdensome for humans and often fail to expose potential system misbehaviors. To address this, recent research efforts at the intersection of Software Engineering (SE) and DL have proposed and developed several Test Input Generation (TIG) techniques, which can generate large, labeled artificial datasets to be used for testing purposes. Nonetheless, TIGs introduce novel concerns as they may produce invalid inputs or disrupt the preservation of original labels, thereby compromising the reliability of the testing process. Since manual validation of these artificial inputs would require significant effort from human assessors, various Automated Validity Assessment techniques have been proposed, often leveraging Variational Auto-Encoders (VAEs) as out-of-distribution detectors [23]. These techniques aim to assess the validity of generated inputs without relying solely on manual validation.

Riccio and Tonella [17] conducted an empirical study to explore the effectiveness of TIGs in producing valid inputs, according to both automated validation approaches and human assessors.

In this report, we will provide an overview of the their study, as well as the assumptions, challenges and findings discussed in their paper. Finally, possible improvements and future directions for novel research efforts will be discussed.

Test Input Generation

Several Test Input Generators (TIGs) have been proposed in recent years. These techniques primarily aim to generate artificial, misbehavior-inducing inputs, i.e. inputs for which the label predicted by the DL system deviates from the expected label, to expose potential issues within the system under test. TIGs may also have additional objectives, such as optimizing specific coverage metrics that are relevant to the DL domain, e.g. neuron coverage, where coverage metrics serve the purpose of assessing the extent to which the DL system explores different regions of the input space.

TIGs can categorized based on their access to information about the system: *white-box* access, where they can obtain internal information such as neuron activation values, and/or *black-box* access, where they can only gather information about the inputs and outputs from the output layer.

These approaches can be additionally be classified based on the strategies employed to generate artificial inputs. Examples include pixel-level perturbation of existing images, manipulation of input representations from generative deep learning models, or modification of domain-specific model representations.

Although the benefits of leveraging these techniques are clear, new challenges are inevitably introduced as well. Specifically, how to define the expected label value for newly generated inputs, commonly known as the *oracle problem*. To address this, small perturbations that preserve the original label are applied to inputs for which a ground-truth label is already known. Furthermore, ensuring the validity of the outputs poses another potential issue that can impact the testing process. This issue arises due to the gap in size between the input space and the portion of the manifold which is semantically relevant for the task being considered, referred to as the *semantic manifold* problem. In fact, as TIGs explore this complex and extensive input space, they may generate inputs that no longer belong to the original input domain, consequently becoming invalid w.r.t. the classification task at hand.

In their study, the authors assessed TIGs employing three different techniques, including Raw Input Manipulation (RIM), Generative Deep Learning Model (GDLM), and Model-based Input Manipulation (MIM).

Raw Input Manipulation (RIM). RIM methodologies involve generating inputs by manipulating and perturbing existing images at the pixel level. Typically, a seed input with a known label, for which the system exhibits correct behavior, is employed. The seed image is subsequently modified by introducing imperceptible perturbations aimed at exposing potential misbehaviors. The way the initial seed inputs are chosen, as well as their quality, heavily impacts this family of methods, as regions of the input space that will be explored are only those accessible from said seeds.

Two examples of RIM-based methods that will be considered are *DeepXplore (DX)* [15] and *DLFuzz (DF)* [6]. While both have white-box access, the former employs occlusion, light, and blackout effects to elicit instances of misbehavior within the seed image, and the

latter is based on introducing noise to the seed image with the intention of enhancing the probability of encountering misbehavior, which is estimated from the output softmax layer. In both cases, maximization of the neuron coverage metric guides how perturbations and noise are applied, respectively.

Generative Deep Learning Models (GDLM). GDLMs generate new inputs by reconstructing the underlying distribution of the input data. They operate within a significantly reduce input space (known as *latent space*), and differently from RIM methods are able to produce inputs that deviate considerably from those in the original dataset. As a result, the quality of the generated inputs is closely tied to the quality of the training set and the generative model employed. Prominent generative models utilized within this context include Variational Auto-encoders (VAEs) and Generative Adversarial Networks (GANs).
Sinvad (SV) [9] and the *Feature Perturbation techniques (FPT)* [4] [8] are the two GDLM-based approaches used in the benchmark. SV has both white-box and black-box access, it is based on perturbations applied to the latent vector by adding a random value sampled from the standard normal distribution to a single element of the vector, and the exploration of the latent vector space can be guided by different metrics, for instance by surprise coverage. On the other hand, FPT is a method with black-box access that injects perturbations into the output of the first layer of the model, where the perturbations are guided by a metric which quantifies the distance from a misclassification and is computed from the softmax output layer.

Model-based Input Manipulation (MIM). MIM approaches generate test inputs by relying on a model representation of the input domain. The original inputs are transformed into an instance of an input model that abstracts its main features, then said model is manipulated and/or perturbed and transformed back to the original input format. Similar to GDLMs, MIM techniques operate within a reduced input space, and with appropriate model constraints, they can ensure the realism of the generated inputs. However, they heavily depend on the existence of a high-quality model representation for a given input domain.
DeepJanus (DJ) [16], with black-box access, takes a model instance, which is a representation of an input in the form of a bitmap image. It then converts this representation into Scalable Vector Graphics (SVG), a text-based format that describes two-dimensional vector graphics. Leveraging the SVG representation, DJ directly modifies the model instance by adjusting the parameters of the SVG and then transforms it back to the original input format.

Automated Validity Assessment

Given the potential labor-intensive nature of manually verifying large artificial test sets, development and research of automated validation assessment techniques have gained considerable attention. These approaches aim to address the challenges associated with assessing the validity of generated test sets in an automated manner, while also preserving ground-truth labels, consequently reducing the reliance on human testers.
Relevant approaches in this context are mostly VAE-based, since they can serve as effective detectors for out-of-distribution inputs. When a VAE is trained on a specific training set, they're expected to generate more precise reconstructions for inputs that closely resemble training data. On the opposite, when presented with inputs that deviate significantly from training data, which do not belong to the original input domain, VAEs produce less precise reconstructions. In their benchmark, Riccio and Tonella (2022) considered two different automated validators: *Distribution-Aware Input Validation (DAIV)* [3] and *SelfOracle* [21].

Distribution-Aware Input Validation (DAIV). DAIV is a VAE-based input validator designed for image classification tasks. Its VAE, which is distribution-aware, is trained on the same dataset used for training the evaluated system, enabling both models to learn from the same underlying distribution. Moreover, an anomaly set is required, containing out-of-distribution inputs.
DAIV's auto-encoder is trained to maximize the reconstruction fidelity metric, which measures the ability of the auto-encoder to faithfully reproduce input data. This measure is, in this context, quantified as the negation of two loss metrics: $loss_m$, based on reconstruction error, and $loss_a$, which quantifies the variance of the reconstructed distribution.
Once the training phase is completed, the VAE is expected to exhibit higher fidelity values for inputs belonging to the data distribution (nominal data), and lower values for out-of-distribution data (anomalous data). To establish a fidelity threshold for distinguishing between valid and invalid inputs, fidelity estimations are computed for both nominal and out-of-distribution data. Subsequently, multiple threshold values are tested, and the one that yields the highest F-measure, a popular measure of a test's accuracy based on precision and recall, is selected. Inputs with a reconstruction fidelity below this threshold are considered invalid.
One drawback of this approach is that the anomalous datasets has a crucial influence on the threshold selection outcome and therefore the validity assessment, meaning that it is crucial to choose it appropriately. For the same reason, DAIV cannot operate in scenarios where an anomalous dataset is unavailable. Another issue arises from the potential overlapping distribution between nominal and anomalous datasets, which would lead to a high rate of false alarms. To mitigate this, it is recommended to choose an anomalous dataset that satisfies two conditions: (1) shares the same input size as the nominal dataset and (2) encompasses different categories than those present in the nominal dataset. By doing so, the nominal and anomalous datasets are more likely to exhibit non-overlapping distributions.

SelfOracle. SelfOracle is a VAE-based, distribution-aware anomaly detection technique that stands apart from DAIV as it does not require an anomaly-containing dataset during training. This characteristic widens the potential scope of application for this approach. Its VAE is trained to minimize reconstruction error, hence higher reconstruction error values indicate invalid data.
In order to distinguish between nominal and anomalous inputs, the approach employs probability distribution fitting, i.e. it tries to identify the most suitable probability distribution that best describes the data. In this context, training data is utilized to fit a Gamma distribution using Maximum Likelihood estimation (MLE), then by minimizing the reconstruction error, the model can learn the parameters of the distribution that most accurately fit the data.

This information can be utilized to configure the threshold for the VAE's reconstruction error, to distinguish between nominal and anomalous data. The authors suggested setting the threshold based on the desired rate of false alarms, which determines the trade-off between true positives (i.e., valid data correctly classified as valid) and false negatives (i.e., valid data incorrectly classified as invalid).

Empirical study

The main goal of the experimental procedure was to compare various TIGs in terms of their reliability in producing valid inputs, according to both automated and human validators, and preserving original labels, by resorting to human judgment. Additionally, the authors aimed to assess the accuracy of automated validators in distinguishing between valid and invalid inputs for each of the considered TIGs, by comparing their judgments with human assessors.
Their study involves the previously introduced TIGs: DeepXplore (DX), DLFuzz (DLF), Sinvad (SV), Feature Perturbation Techniques (FPT), and DeepJanus (DJ). Two automated validity assessment techniques, Distribution-Aware Input Validation (DAIV) and SelfOracle, were considered. The investigation involved three classification tasks of increasing complexity: Modified National Institute of Standard and Technology (MNIST) [11], Street View House Numbers (SVHN) [14], and ImageNet-1K [2]. To evaluate the performance, the authors choose, for each dataset, a well-performing pre-trained DL classifier: LeNet [11] for MNIST, All-CNN-A [20], (with weights by Dola et al. [3]), for SVHN, and VGG16 [19] (Keras library implementation) for ImageNet-1K.
Finally, human assessors were hired through Amazon Mechanical Turk platform [1], a crowd-sourcing service providing access to independent individuals. Involving human testers in their study required extra effort in devising methods to increase confidence in the quality of the testers work. Specifically, candidates were required to possess a high reputation rating and to complete an attention check test.

TIGs setup and selection. The TIGs were selected based on their diverse test generation approaches and taking into account their availability as open-source implementations.
One significant challenge faced by the authors during this step was the limited coverage of the three chosen datasets by each technique. All TIGs originally covered only 2 out of 3 datasets, or, in case of DJ, only one. To overcome this limitation, the authors had to adapt the existing TIGs to any non-covered dataset.
DX and DLF originally covered only MNIST and ImageNet-1K, so they were extended to cover SVHN by using a DX variant developed by Dola et al. [3].
SV originally covered only MNIST and SVHN, so it was adapted to ImageNet-1K by integrating the pre-trained BigGAN by Brock et al.
FTP originally covered only MNIST and ImageNet-1K, it was applied to SVHN by retraining the GAN architecture used for MNIST.
DJ covered only one dataset, MNIST. The authors were able to adapt it to SVHN by applying the same vectorial model representation used for MNIST. However, finding an effective model-based representation for ImageNet-1K, or complex images in general, was not possible.

Automated Input Validators setup. In the case of DAIV, FashionMNIST [22] and CIFAR-10 [10] were used as the respective anomalous datasets for MNIST and SVHN, as suggested in DAIV's paper. Since DAIV didn't originally cover ImageNet-1K, there was no suggested anamalous dataset. To solve this, they followed the DAIV's authors suggested approach for selecting a suitable one, i.e. CelebA [12] .
For SelfOracle, a similar VAE architecture to that used in DAIV was adopted. As desired false alarm rate, the authors opted for a low value of 0.01%, since the majority of inputs in the original test sets represented nominal data.

Experimental Procedure. Artificial datasets were generated for each dataset using the proposed TIGs. Subsequently, groups of two assessors were proposed a survey containing images from all TIGs, where no image was contained in another survey. They were tasked with labeling and classifying the proposed images for each dataset, and specifically they had to determine whether an image was valid or out of the domain of the respective classification task. If the two assessors didn't agree on the validity/invalidity of an image, it was excluded from the benchmark.
Finally, the same set of images was evaluated using automated validators. Statistical analysis was conducted utilizing Fisher's exact test to determine the significance of the results. Comparisons were made between pairs of TIGs to assess the validity of the generated inputs and their ability to preserve labels. A p-value below 0.05 was considered statistically significant. The same statistical test was employed to compare the results of the human study with the assessments of the automated validators. If the p-value was greater than or equal to 0.05 and the effect size was negligible, it indicated a lack of statistically significant differences.

Results

In this section of our summary, we will provide a brief and concise analysis of the experiment results and highlight the main contributions and findings of the study. We will first analyze the validity of inputs generated by different TIGs, as assessed by both automated validators and human assessors. Next, we will compare the TIGs in terms of their label preservation reliability, as determined by human assessors. Lastly, we will discuss the level of agreement between automated and human validity assessments.

TIGs comparison in terms of input validity. The comparison showed varying degrees of input validity across the different TIGs (Table II from [17]).
For MNIST, the GDLMs (Sinvad and FPT) and DeepJanus produced a higher number of valid inputs, as determined by automated validators, compared to RIM-based approaches. These results were in good agreement with human validators, who considered the majority of artificially generated images to be valid, with the exception of those generated by DeepXplore. However, there was considerable disagreement between automated and human judgement as far as RIMs were concerned
For SVHN, the majority of tests produced by GDLMs and DLFuzz

[1] https://www.mturk.com/

were considered valid by automated validators. On average, human assessors determined a lower percentage of valid inputs (15%) compared to MNIST. This difference can be attributed to the higher complexity of the SVHN dataset and its lower resolution. Notably, DLFuzz emerged as the best-performing TIG with a substantial margin.
In case of ImageNet-1K, all inputs generated by each TIG were deemed valid by automated validators. However, human assessors exhibited varying levels of agreement. DLFuzz and FPT achieved identical agreement rates, while DX had a slightly lower agreement rate (90%) valid inputs). The agreement for the SVHN dataset was even lower, with only 60% of inputs being deemed valid by human assessors.

TIGs comparison in terms of label preservation. For MNIST, DLFuzz achieved the highest label preservation ratio, by a significant margin, reaching 99% of valid inputs preserving their ground-truth label. On the other hand, GDLMs showed the poorest performance in terms of label preservation, $< 58\%$
In case of SVHN, DeepXplore (DX) performed the best, with a label preservation ratio of 79%. However, both Sinvad and DJ had particularly low label preservation ratios, with only 9% of valid inputs preserving their original labels. It is worth noting how SV, despite having the second best result in terms of input validity, transformed most of the original ground-truth labels into a different one.
For ImageNet-1K, all TIGs showed good results, with label preservation ratios above 83%. Among them, Feature Perturbation Techniques (FPT) achieved the highest label preservation rate, successfully preserving the labels of all inputs.

Automated and Human validity comparison. Automated validators assessment showed a good match with human judgment, with 78% accuracy (Table III from [17]). It is worth noting how SelfOracle achieve dhigher accuracy than DAIV across all datasets (7% higher on average).

Contributions and Findings

The contributions and findings of the authors are significant, as their study was the first to propose a comprehensive analysis of TIGs and automated validators while taking into consideration label preservation.
The study showed how all TIGs are able to generate valid inputs, although the preservation of ground-truth labels is not guaranteed and is not necessarily correlated to input validity performance, as seen in the case of Sinvad for the SVHD dataset.
Different degrees of input validity among the TIGs were identified, each facing distinct challenges. For example, RIM encountered issues related to corrupted pixels, while GDLMs lacked continuity in the latent space. The Model Inversion Method (MIM), on the other hand, struggled with obtaining high-quality representations of the input model.
The study highlights a good level of agreement between automated validation and human judgment, underscoring the reliability of automated validation approaches in evaluating image classification datasets of varying complexity. Despite that, there is room for improvement and the authors highlighted how distribution-aware validators are mostly challenged when dealing with feature-rich datasets such as ImageNet-1K. In fact, these validators heavily rely on low-level image features from the original training set, making it more difficult for them to accurately distinguish between valid and invalid inputs, especially in complex tasks.

CRITICAL REVIEW

The study conducted by Riccio and Tonella (2022) presents solid assumptions and reasoning, including appropriate dataset selection and the use of widely adopted DL models. The open-source nature of these datasets and models guarantees the reproducibility of the experiments and enables further validation by other researchers. Measures to ensure the quality of human judgments have been taken, implementing a rigorous selection process. Nonetheless, it is worth mentioning certain aspects of the paper that could be subject to critical discussion.

Considerations on Human and Platform Bias. The study's statistical relevance may be questioned due to the limited number of human assessors, which consisted of only 220 individuals from the same crowd-sourcing platform. To enhance the validity of the results and mitigate potential human biases, which may be transferred into training data and the model learning process [1], it would have been beneficial to increase the number of human testers involved in the evaluation process.
Moreover, it is crucial to acknowledge that the environment and conditions under which human workers complete their tasks can introduce uncontrolled variables that may impact the results. To address this concern, incorporating a percentage of in-person testers could have provided a more controlled and standardized environment for the assessments, particularly in complex datasets like ImageNet-1K where human judgment may also be prone to errors. Exploring alternative crowd-sourcing platforms, such as TaskRabbit [2] or Appen [3], could have been advantageous in mitigating platform-related bias and uncovering potential unexpected implications related to a specific service. Incorporating a diverse range of platforms would have been of great interest to see if there are statistically relevant difference in the judgment of different testing pools.

Background and Motivation of Human Testers. Notably, the study didn't discuss essential aspects that can impact the quality of the work carried out by human testers. Demographic distribution, educational background, and the hourly-pay-rate/task-pay-rate were not explicitly mentioned, yet these factors are known to possibly influence the quality of annotations[7] [18].
In terms of demographic distribution, it is important to ensure a wide range of cultural and linguistic backgrounds among the testers. This diversity helps mitigate biases and ensures that the annotations are not limited by a narrow perspective.
Furthermore, considering the educational background of the testers, particularly whether they hold a degree or not, is crucial, especially when dealing with non-trivial, university-level tasks like the ones addressed in this study. This is more relevant when working with complex datasets such as ImageNet-1K.

[2] https://www.taskrabbit.com/
[3] https://appen.com/

Lastly, the study did not clarify whether the pay rate provided to the testers was appropriate or above average. The pay rate plays a vital role in motivating and attracting skilled individuals, ensuring their commitment and dedication to providing high-quality annotations.

Human disagreement. In cases where two independent human assessors disagreed on a survey, the authors' decision to exclude those instances may have overlooked an opportunity for valuable insights. It could have been beneficial to investigate the reasons behind such disagreement by involving additional assessors. This approach would have provided a deeper understanding of the factors that led to the assessors having different opinions in terms of input validity and potentially reduced the amount of discarded inputs during the experiment.

Overall, the study provides valuable insights into the limitations and strengths of different TIGs, sheds light on the importance of exploring the latent space, and highlights the challenges faced by distribution-aware validators in complex datasets. These contributions make it a valuable foundation for enhancing existing TIGs, automated validators, and the development of novel approaches.
However, it is important to acknowledge that considering the additional factors we discussed earlier would have further enriched the study. Taking into account the expertise, potential biases, and diversity of the human testers involved in the labeling tasks could have provided a more comprehensive understanding of the results. Addressing aspects such as demographic distribution, educational background, and pay-hourly rates would have enhanced the relevance and applicability of the findings. By incorporating these considerations, future research and development efforts in this field could build upon a stronger foundation, leading to improved TIGs and automated validators that are more robust and reliable.

OPEN CHALLENGES AND POTENTIAL IMPROVEMENTS

The capabilities of TIGs in generating valid inputs for deep learning, as well as the ability of automated validators in distinguishing valid and invalid inputs, was comprehensively discussed. Nonetheless, there are still potential areas for improvement as well as potential avenues for further development.
As already mentioned in previous sections, there were different reasons for which each TIG had invalid outputs (e.g, pixel corruption or lack of continuity in the latent space). Further research efforts could focus into deeper investigating these issues and explore potential solutions to address them effectively, consequently improving reliability and robustness of TIGs approaches in general.
Another limitation observed in the study was the inconsistent preservation of ground-truth labels. For instance, only 38% of the generated inputs retained their original label for the SVHN dataset. Improving label preservation is essential to have reliable inputs that align with the expected label. Further research is required to investigate methods that can enhance the label preservation performance of TIGs, ensuring that artificial inputs consistently reflect the desired class or category.
It is also worth noting that TIGs could be analyzed by taking into account additional aspects such as input diversity. Being able to generate representative inputs across the entire input domain is essential for training robust models that are able to generalize well to unseen data.
Finally, as also mentioned in the paper, extending the evaluation of TIGs to a wider range of deep learning systems, including industrial ones, would enhance the applicability of the findings. Understanding the performance and limitations of TIGs on real-world DL models employed in practical settings can be crucial for adoption and further development.

REFERENCES

[1] Natã M Barbosa and Monchu Chen. 2019. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
[3] Swaroopa Dola, Matthew B Dwyer, and Mary Lou Soffa. 2021. Distribution-aware testing of neural networks using generative models. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 226–237.
[4] Isaac Dunn, Laura Hanu, Hadrien Pouget, Daniel Kroening, and Tom Melham. 2020. Evaluating robustness to context-sensitive feature perturbations of different granularities. *arXiv preprint arXiv:2001.11055* (2020).
[5] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. 2020. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics* 37, 3 (2020), 362–386.
[6] Jianmin Guo, Yu Jiang, Yue Zhao, Quan Chen, and Jiaguang Sun. 2018. Dlfuzz: Differential fuzzing testing of deep learning systems. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 739–743.
[7] Tyler Hamby and Wyn Taylor. 2016. Survey satisficing inflates reliability and validity measures: An experimental comparison of college and Amazon Mechanical Turk samples. *Educational and Psychological Measurement* 76, 6 (2016), 912–932.
[8] Nathan Inkawhich, Wei Wen, Hai Helen Li, and Yiran Chen. 2019. Feature space perturbations yield more transferable adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7066–7074.
[9] Sungmin Kang, Robert Feldt, and Shin Yoo. 2020. Sinvad: Search-based image space navigation for dnn image classifier test input generation. In *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops*. 521–528.
[10] Alex Krizhevsky, Geoffrey Hinton, and others. 2009. Learning multiple layers of features from tiny images. (2009).
[11] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
[12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*. 3730–3738.
[13] Lei Ma, Fuyuan Zhang, Jiyuan Sun, Minhui Xue, Bo Li, Felix Juefei-Xu, Chao Xie, Li Li, Yang Liu, Jianjun Zhao, and others. 2018. Deepmutation: Mutation testing of deep learning systems. In *2018 IEEE 29th international symposium on software reliability engineering (ISSRE)*. IEEE, 100–111.
[14] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning. (2011).
[15] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. 2017. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*. 1–18.
[16] Vincenzo Riccio and Paolo Tonella. 2020. Model-based exploration of the frontier of behaviours for deep learning system testing. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 876–888.
[17] Vincenzo Riccio and Paolo Tonella. 2022. When and Why Test Generators for Deep Learning Produce Invalid Inputs: an Empirical Study. *arXiv preprint arXiv:2212.11368* (2022).
[18] Zahra Shakeri Hossein Abad, Gregory P Butler, Wendy Thompson, and Joon Lee. 2022. Crowdsourcing for machine learning in public health surveillance: lessons learned from Amazon Mechanical Turk. *Journal of Medical Internet Research* 24, 1 (2022), e28749.
[19] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[20] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).

[21] Andrea Stocco, Michael Weiss, Marco Calzana, and Paolo Tonella. 2020. Misbehaviour prediction for autonomous driving systems. In *Proceedings of the ACM/IEEE 42nd international conference on software engineering*. 359–371.

[22] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).

[23] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2021. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334* (2021).

[24] Fangyi Zhang, Jürgen Leitner, Michael Milford, Ben Upcroft, and Peter Corke. 2015. Towards vision-based deep reinforcement learning for robotic motion control. *arXiv preprint arXiv:1511.03791* (2015).