Addressing Generalization Failure in Deep Detection Models for Fishing Trawler Video Analytics

Birk Torpmann-Hagen[1,2], Pål Halvorsen[2,3], Michael Riegler[1,2] and Dag Johansen[1]

[1] *UiT The Arctic University of Norway, Norway*
[2] *SimulaMet, Norway*
[3] *University of Oslo, Norway*

Abstract

A problem with supervised machine learning algorithms is how to precisely predict outcome values for previously unseen data. In this paper, we evaluate conventional detection models trained on the Njord commercial fishing video surveillance dataset in this generalization context. Our results show that these models fail to generalize to a test-set consisting of samples with challenging lighting- and weather-conditions. To address this, a novel distributional-shift detector is introduced, exhibiting good performance and outperforming competing methods by a considerable margin.

[1]

1. Introduction

Out of Distribution (OOD) Generalizability is an often overlooked problem in deep learning [1, 2, 3]. Conventional methodologies assert that cross-validation or hold-out set evaluation is a suitable method of approximating a model's performance in deployment conditions [4], but multiple case-studies have shown that this is far from true. Deep Neural Networks (DNNs) tend to exhibit significant performance drops when deployed on data that is OOD from the training data, even in a manner imperceptible to a human observer [1, 2, 5, 6, 3]. This is known as *generalization failure*. It has for instance been shown that medical imaging systems fail to generalize to samples taken from different centers, demographics, or imaging equipment [1, 2, 7]. Similar behaviour has been demonstrated in Natural Language Processing models [1] and in context of large, general-purpose datasets such as ImageNet [6].

This is further complicated by the lack of transparency intrinsic to DNNs. Their predictive process is obscure, their confidence scores are often misleading, and generalization failure typically manifests in ways that are difficult to detect by inspection of the input data alone [8]. As a result, conventionally implemented DNNs lack the robustness and trustworthiness typically required in particularly sensitive or critical domains, therein as a component of an automatic anonymization pipeline or video analytics aboard fishing-trawlers as considered in this competition [9].

This quest for insight paper seeks to explore the generalizability of the deep learning pipeline in this domain. In particular, it considers the following questions:

1. To what extent does generalization failure play a role for the Njord dataset [10]?
2. Assuming generalization failure occurs, can the distributional shifts that induce it be successfully detected as it arises without requiring labels for evaluation?

[1] Data and code are available via https://github.com/BirkTorpmannHagen/NearestNeighbourDistributionalShiftDetection.

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

We observe that generalization failure does indeed occur in this domain, with all tested models exhibiting significant performance degradation when evaluated on OOD data. To address this, we introduce Nearest-Neighbour Distributional Shift Detection (NNDSD), a novel distributional-shift detector, capable of estimating when a distributional-shift and thus generalization failure is likely to occur during employment.

2. Related Work

Shen et al. survey the space of generalizable methods, including domain-adaptation and novel learning schemes [3], and conclude that there are still a number of unsolved challanges involved therein. Ye et al. demonstrate that a majority of said methods perform worse than conventional deep learning outside of specific conditions [6]. Rabanser, Gunnemann, and Lipton perform an empirical study of methods of detecting dataset shift [11], in particular methods involving performing statistical tests on various encodings of the data such as VAEs, PCA, and classifier-based encoding. Huang, Geng and Li investigate the use of gradients towards distributional shift detection for classification, with state-of-the-art results [12]. Liang, Li, and Srikant implement a shift-detector exploiting the discrepancy in softmax scores between In-Distribution (InD) and OOD data after gradient-based perturbation [13]. None of these works consider distributional-shift detection in the context of a practical dataset, however, nor on the object-detection task.

3. Nearest Neighbour Distributional Shift Detection

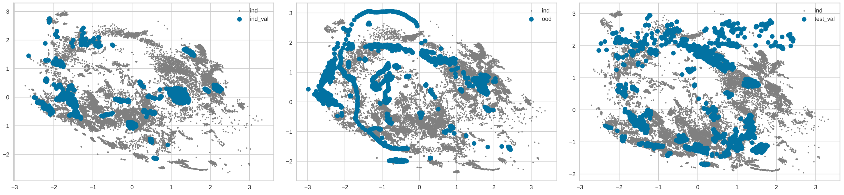
Nearest-Neighbour Distributional Shift Detection (NNDSD) is based on the work by Rabanser, Gunnemann, and Lipton [11], where distributional shifts are identified by performing statistical tests on various forms of encodings of the data. In our application of NNDSD, we use one of the best reported performing configurations, consisting of a trained Variational Autoencoder (VAE) with Kolmogorov-Smirnoff testing adjusted with Bonferroni correction.

We observe that this approach is not able to distinguish between InD and OOD images. This can be understood by visually inspecting the encoding space, through, for instance, PCA as shown in Figure 1. Though it is clear that the validation set, for instance, is largely contained within the bounds of the training distribution, it is evidently sampled from a limited region on the distribution. Due to the high similarity between subsequent frames of a video feed used to obtain data samples, sampling bias is unavoidable. This form of sampling bias does not typically induce generalization failure.

To address this, two steps are added to the procedure. Firstly (i), the encodings are transformed to a two-dimensional space through PCA. Beyond facilitating visualization and decreasing computing costs, this also ensures there is sufficient variability along each dimension to perform viable statistical tests. Secondly (ii), the tests are performed with a stricter null-hypothesis. Whereas regular testing asserts that both populations are identically distributed, NNDSD instead asserts that the data to test is distributed identically to any region within the training distribution. In more practical terms, this involves testing the transformed deployment encodings against their nearest neighbours in the bank of transformed training encodings. In practice, this should potentially eliminate the effect of sampling bias on the tests.

4. Experiments and Results

Two experiments were performed: a simple evaluation of the generalizability of the YoloV5 pipeline across model sizes, and evaluation of NNDSD on three different folds of the Njord



(a) InD (b) OOD (c) test

Figure 1: Latent encoding of the training data against the validation data, OOD-data, and test-set.

dataset. We demonstrate that the models fail to generalize when deployed on data with lighting- and weather-conditions not present in the training set, but that this failure can be successfully detected with NNDSD given sufficient sample sizes.

4.1. Experimental Setup

To determine the generalizability of a given model, it is necessary to evaluate it on OOD samples. Whereas other domains may have several independently curated datasets to facilitate such evaluation, the Njord dataset is the only publicly known one of its kind. As a result, it is necessary to partition the dataset such that the test-set can be considered OOD to the training and validation set. To this end, the dataset was inspected for samples that had particularly foggy, dark, or otherwise challenging lighting and weather conditions. These samples were then reserved for the test-set - referred to as the OOD-set - with the remaining data being split into training and validation sets. The original test set as provided by the task organizers was also included in the evaluation. Training samples were extracted from the videos at a rate of one image per 25 frames.

For the generalizability experiment, ten models were trained for each size - extra small, small, medium, and large - on the InD partition of the dataset, using the default hyperparameters as provided by the Ultralytics YoloV5 library [14]. These hyperparameters can be found on the GitHub.

To evaluate the efficacy of the distributional-shift detector, we opted to follow the methodology described in [11]. The VAE was trained with no augmentations on the training set, and the resulting shift detector was evaluated at a range of different sample sizes on the OOD-set, the test-set, and the InD validation set. The detection accuracy was estimated by randomly selecting samples of the given size from the test-set 1000 times, with the deployment data being considered OOD if testing yielded $p<0.05$ after Bonferroni correction, and InD otherwise.

4.2. Baseline Generalizability

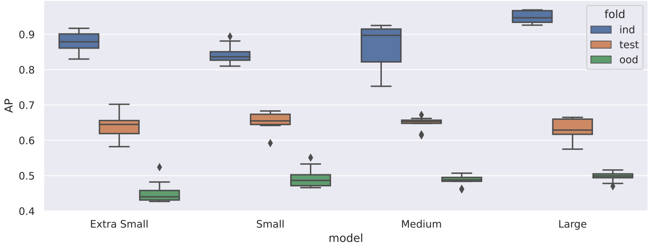


Figure 2: AP across model sizes and folds (blue:InD, orange: test, green: OOD)

Figure 2 illustrates that all of the tested models are subject to generalization failure. The OOD dataset induces the most significant performance drops, with the test set being situated between the InD and OOD sets performance-wise.

4.3. Distributional Shift Detection

.

Table 1
Shift detection accuracy per fold for NNDSD

Samples	100	500	1000	1500	2000	2500
IND-val	0.999	0.893	0.836	0.890	0.879	0.919
OOD	0.052	0.999	1.000	1.000	1.000	1.000
Test	0.000	0.002	0.062	0.439	0.867	0.999
Total acc.	0.350	0.631	0.776	0.819	0.915	0.972

Table 1 presents the detection accuracy of NNDSD across the three folds. With a batch-size of 500, the detector manages to correctly identify that the OOD partition indeed is OOD and conversely that the validation set is InD in 99.9% and 89.3% of cases respectively. The test set was also often identified as OOD at larger sample sizes, which is consistent with the findings as shown in Figure 2, where the test set induces a smaller but nevertheless significant degree of generalization failure as evident by the performance drop over the InD-set.

5. Concluding Remarks

The results from Section 4.2 confirm that generalization failure is indeed a factor in the fishing trawler surveillance video domain. Each of the tested models exhibited considerable performance degradation when evaluated on the OOD partition. This reaffirms the findings elsewhere in the literature that generalization failure is ubiquitous in deep learning, and highlights the need for research towards increasing the generalizability of DNNs. NNDSD demonstrates significant performance potential and may have considerable utility in practical deployment.

This utility is hampered somewhat by the fairly large sample-size requirements. In practical scenarios, distributional shifts may not last long enough for a sufficient amount of samples to be collected. Further work is required towards reducing the sample size requirement to more reasonable levels.

Another limitation of this work is the method by which the shift-detectors were evaluated. Throughout this paper, the OOD-detection problem was treated as a classification task. In reality, it is more likely to be a regression-type problem - i.e. that there are degrees of severity for distributional-shift, as evidenced by the difference in performance on the OOD set and the test set. For sufficient proof of performance, NNDSD needs to be evaluated on multiple test-sets with varying degrees of distributional shift. A more sophisticated system could also leverage the p-values directly to estimate the severity of the shift by analyzing the correlation between the p-values and the performance drops of the system prior to deployment.

Overall, our work confirms the conjecture that special consideration for generalizability needs to be made when designing deep learning systems, in particular in domains characterized by high degrees of complexity, dynamicity and potential for distributional shifts, such as onboard fishing trawlers. Though there do not currently exist any methods capable of endowing DNN with suitable levels of generalizability for practical deployment in performance-critical domains, detecting generalization failure through NNDSD or similar methods may be a sufficient workaround and endow deep learning systems with an increased degree of trustworthiness and transparency and thus significantly reduce the many risks associated with generalization failure.

References

[1] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F. A. Wichmann, Shortcut learning in deep neural networks, Nature Machine Intelligence 2 (2020) 665–673. URL: http://dx.doi.org/10.1038/s42256-020-00257-z. doi:10.1038/s42256-020-00257-z.

[2] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, F. Hormozdiari, N. Houlsby, S. Hou, G. Jerfel, A. Karthikesalingam, M. Lucic, Y. Ma, C. McLean, D. Mincu, A. Mitani, A. Montanari, Z. Nado, V. Natarajan, C. Nielson, T. F. Osborne, R. Raman, K. Ramasamy, R. Sayres, J. Schrouff, M. Seneviratne, S. Sequeira, H. Suresh, V. Veitch, M. Vladymyrov, X. Wang, K. Webster, S. Yadlowsky, T. Yun, X. Zhai, D. Sculley, Underspecification presents challenges for credibility in modern machine learning, 2020. URL: https://arxiv.org/abs/2011.03395. doi:10.48550/ARXIV.2011.03395.

[3] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, P. Cui, Towards out-of-distribution generalization: A survey, 2021. URL: https://arxiv.org/abs/2108.13624. doi:10.48550/ARXIV.2108.13624.

[4] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016. http://www.deeplearningbook.org.

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

[6] N. Ye, K. Li, H. Bai, R. Yu, L. Hong, F. Zhou, Z. Li, J. Zhu, Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization, 2021. URL: http://arxiv-export-lb.library.cornell.edu/abs/2106.03721. arXiv:2106.03721.

[7] S. Ali, N. Ghatwary, D. Jha, E. Isik-Polat, G. Polat, C. Yang, W. Li, A. Galdran, M.-A. G. Ballester, V. Thambawita, S. Hicks, S. Poudel, S.-W. Lee, Z. Jin, T. Gan, C. Yu, J. Yan, D. Yeo, H. Lee, N. K. Tomar, M. Haithmi, A. Ahmed, M. A. Riegler, C. Daul, P. Halvorsen, J. Rittscher, O. E. Salem, D. Lamarque, R. Cannizzaro, S. Realdon, T. de Lange, J. E. East, Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge, 2022. URL: https://arxiv.org/abs/2202.12031. doi:10.48550/ARXIV.2202.12031.

[8] A. Jacovi, A. Marasović, T. Miller, Y. Goldberg, Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai, 2020. URL: https://arxiv.org/abs/2010.07487. doi:10.48550/ARXIV.2010.07487.

[9] T.-A. S. Nordmo, A. B. Ovesen, H. D. Johansen, M. A. Riegler, D. Johansen, Njordvid: A fishing trawler video analytics task, in: MediaEval'22:, 2022.

[10] T.-A. S. Nordmo, A. B. Ovesen, B. A. Juliussen, S. A. Hicks, V. Thambawita, H. D. Johansen, P. Halvorsen, M. A. Riegler, D. Johansen, Njord: A fishing trawler dataset, in: Proceedings of the 13th ACM Multimedia Systems Conference, MMSys '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 197–202. URL: https://doi.org/10.1145/3524273.3532886. doi:10.1145/3524273.3532886.

[11] S. Rabanser, S. Günnemann, Z. C. Lipton, Failing loudly: An empirical study of methods for detecting dataset shift, 2018. URL: https://arxiv.org/abs/1810.11953. doi:10.48550/ARXIV.1810.11953.

[12] R. Huang, A. Geng, Y. Li, On the importance of gradients for detecting distributional shifts in the wild, 2021. URL: https://arxiv.org/abs/2110.00218. doi:10.48550/ARXIV.2110.00218.

[13] S. Liang, Y. Li, R. Srikant, Enhancing the reliability of out-of-distribution image detection in neural networks, 2017. URL: https://arxiv.org/abs/1706.02690. doi:10.48550/ARXIV.1706.02690.

[14] G. Jocher, A. Stoken, J. Borovec, NanoCode012, ChristopherSTAN, L. Changyu, Laughing, tkianai, A. Hogan, lorenzomammana, yxNONG, AlexWang1900, L. Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, F. Ingham, Frederik, Guilhen, Hatovix, J. Poznanski, J. Fang, L. Yu, changyu98, M. Wang, N. Gupta, O. Akhtar, PetrDvoracek, P. Rai, ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, 2020. URL: https://doi.org/10.5281/zenodo.4154370. doi:10.5281/zenodo.4154370.