Domain Shift as a Confounding Variable in Unsupervised Pathology Detection

Felix Meissen[*1,2] FELIX.MEISSEN@TUM.DE
[1] *Technical University of Munich, Germany*
[2] *Klinikum Rechts der Isar, Munich, Germany*
Ioannis Lagogiannis[*1,2] I.LAGOGIANNIS@TUM.DE
Georgios Kaissis[1,2,3] G.KAISSIS@TUM.DE
[3] *Imperial College London, UK*
Daniel Rueckert[1,2,3] DANIEL.RUECKERT@TUM.DE

Editors: Under Review for MIDL 2022

Abstract

Unsupervised Pathology Detection (UPD) has recently received considerable attention in medical image diagnosis. However, the lack of publicly available benchmark datasets for UPD makes researchers fall back on datasets that were originally created for other tasks. These datasets may exhibit domain shift that acts as a confounding variable, fooling observers into believing that the models excel at detecting pathologies, while a significant part of the model's performance is detecting the domain shift. In this short paper, we show on the example of the Hyper-Kvasir dataset, how confounding variables can dramatically skew the actual performance of pathology detection methods.

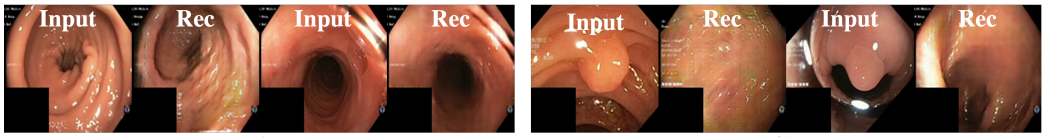
Keywords: Unsupervised Pathology Detection, Domain Shift, Confounders

1. Introduction

Recently, UPD was successfully applied to a multitude of different tasks in medical image diagnosis, such as lesion detection in brain MRI [1], finding pathologies in chest X-rays, or detecting polyps in colonoscopy videos [3; 6; 5; 7]. The reasons to use UPD are compelling, especially in the medical domain: UPD methods do not require difficult and costly to obtain labels and train with "normal" data only, which is theoretically vastly available. Publicly available data, however, is sparse, and the lack of benchmark datasets pushes researchers to instead use datasets that were originally not designed for UPD. These datasets might contain confounding variables that can skew the actual performance of evaluated models. One confounding variable that UPD methods are especially susceptible to is domain shift because samples from other domains are per definition anomalous as well.

Hyper-Kvasir (HK) [2] was used in several studies to evaluate new UPD methods [3; 6; 5; 7]. It is a large dataset for gastrointestinal (GI) endoscopy containing around one million images and video frames. In HK, normal samples are extracted from videos filmed while the endoscope navigates through the GI tract. The appearance of samples containing polyps, however, differs dramatically from the normal ones, with the endoscope camera capturing unique angles, focusing on the polyp, and the image exhibiting different lighting conditions. In this work, we show that this domain shift between images with and without polyps is the main reason for the good image-level performance of pathology detection models on this dataset.

[*] Contributed equally



a) b)

Figure 1: Normal (a) and anomal (b) images and their reconstructions by f-AnoGAN.

2. Experiments

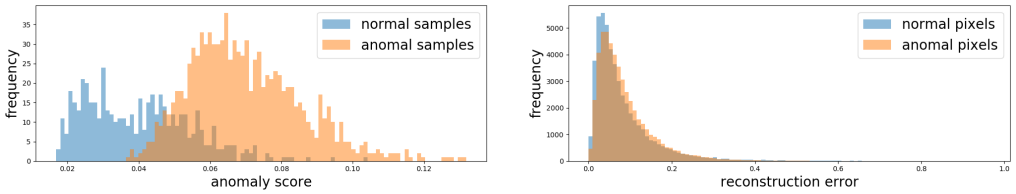
Data and Pre-processing We followed [3; 6; 5; 7] in their selection of training- and test-set samples from HK. The images are resized to 128×128 pixels and scaled into the range [0,1]. As described in [2], an image of the endoscope guidance system appears in the bottom left corner of many normal samples. We mask this part of every sample with black pixels as similar post-processing is already done in the anomalous test-set images.

Implementation Details We use the dense Autoencoder (AE) [1] and f-AnoGAN [4] architectures to perform our experiments. The implementation of f-AnoGAN is only slightly adapted to handle the increased resolution. We train both models for 10,000 iterations with the original optimization parameters. The anomaly map is computed as the pixel-wise absolute error $\mathbf{r} = |\mathbf{x} - \hat{\mathbf{x}}|$ between the input image $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$. As anomaly score we use the average over all pixels in the anomaly map.

Results Across 4 runs, AE and f-AnoGAN reach a competitive area under the receiver operating characteristics curve (AUROC) of 0.918 ± 0.002 and 0.912 ± 0.006 respectively. For the first experiment, we ignore polyp pixel reconstruction errors in calculating the anomaly scores. Performance drops only marginally to 0.899 ± 0.005 and 0.888 ± 0.009 respectively, showing that the residuals of these pixels barely affect performance. Next, we compare the anomaly scores of normal vs anomal images to the reconstruction errors of normal vs anomal pixels in polyp images. Figure 2 reveals that, while normal samples have overall lower anomaly scores, the polyp regions are only slightly worse reconstructed than the normal pixels. Lastly, we can see in Figure 1 that for anomal samples, f-AnoGAN does not reconstruct the input images without polyps, but generates images that are not related to the input in terms of their content.

3. Conclusion

The above experiments provide compelling evidence that the good image-level performance of unsupervised pathology detection methods on the Hyper-Kvasir dataset mainly relates to the domain shift between images with and without polyps and not to the actual presence of polyps. Especially, it is unclear if a better performing model is a stronger polyp detector, or simply better at detecting the domain shift. In clinical practice, a good-performing model on the HK dataset is still likely to produce many false positive and false negative predictions. We therefore urge the community to pay close attention to confounding variables when evaluating new methods.



(a) (b)

Figure 2: Histograms of (a) anomaly scores of normal and abnormal samples and (b) of reconstruction errors of equally sampled normal and anomal pixels of abnormal samples. Results from AE. We observed identical behaviour with f-AnoGAN.

References

[1] C. Baur, S. Denner, B. Wiestler, N. Navab, and S. Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. *Medical Image Analysis*, 69:101952, 2021.

[2] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, D. Johansen, C. Griwodz, H. K. Stensland, E. Garcia-Ceja, P. T. Schmidt, H. L. Hammer, M. A. Riegler, P. Halvorsen, and T. de Lange. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):283, 2020.

[3] Y. Chen, Y. Tian, G. Pang, and G. Carneiro. Deep one-class classification via interpolated gaussian descriptor. *arXiv preprint arXiv:2101.10043*, 2021.

[4] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Margarethe Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, 2019.

[5] Y. Tian, F. Liu, G. Pang, Y. Chen, Y. Liu, J. W. Verjans, R. Singh, and G. Carneiro. Self-supervised multi-class pre-training for unsupervised anomaly detection and segmentation in medical images. *arXiv preprint arXiv:2109.01303*, 2021.

[6] Y. Tian, G. Pang, F. Liu, Y. Chen, S. H. Shin, J. W. Verjans, R. Singh, and G. Carneiro. Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 128–140. Springer, 2021.

[7] Y. Tian, G. Pang, Y. Liu, C. Wang, Y. Chen, F. Liu, R. Singh, J. W. Verjans, and G. Carneiro. Unsupervised anomaly detection in medical images with a memory-augmented multi-level cross-attentional masked autoencoder. *arXiv preprint arXiv:2203.11725*, 2022.