b
UNIVERSITÄT
BERN

ARTORG CENTER
BIOMEDICAL ENGINEERING RESEARCH

Pretrained Features are Effective for Unsupervised Out-of-Distribution Detection in Medical Images

Lars Doorenbos[1], **Raphael Sznitman**[1], **Pablo Márquez-Neila**[1]

[1]{lars.doorenbos,raphael.sznitman,pablo.marquez}@artorg.unibe.ch

[1] ARTORG Center for Biomedical Engineering Research, Artificial Intelligence in Medical Imaging, Universität Bern, Switzerland

1 Introduction

Recent work by Rippel et al. [1] shows the effectiveness of pretrained features for anomaly detection on natural images. Despite impressive results, they expect worse performance on medical images. We show that their method works very well for out-of-distribution detection on medical scans from two domains: chest X-rays and OCT images. We consider both semantic and non-semantic shift, in the form of unseen pathologies and image corruptions, respectively. It achieves state-of-the-art results on all three cases evaluated, using the smallest EfficientNet, despite relying on features obtained from natural images. Additionally, we investigate the effect of model size, and find that in contrast to the results on natural images, bigger networks do not necessarily increase performance.

2 Method

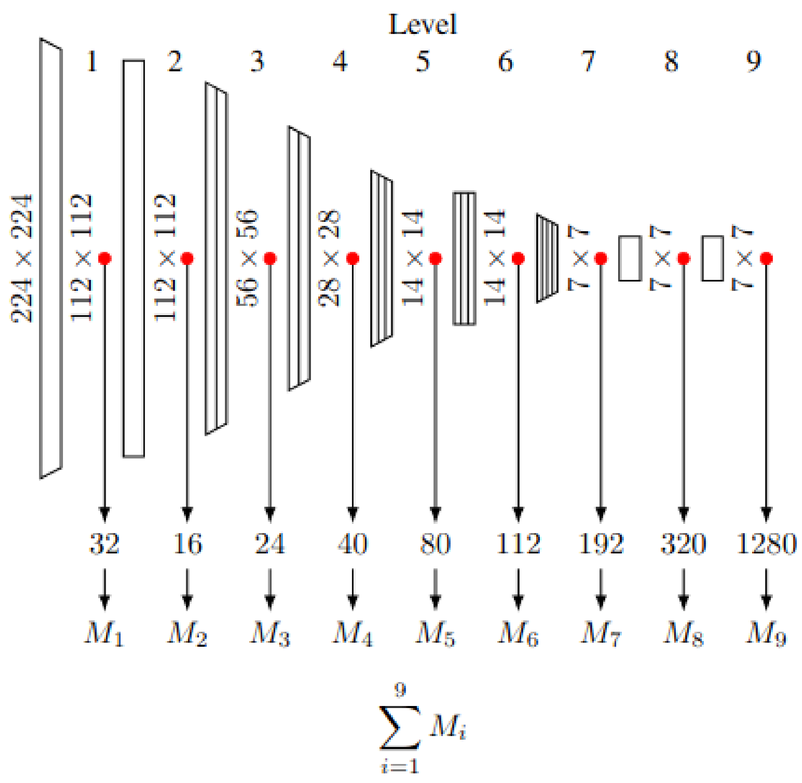

Figure 1: Samples are scored by the unweighted sum of Mahalanobis distances. Source: Rippel et al. [1].

The Mahalanobis anomaly detector (MahaAD) characterizes the training data at multiple layers in a pretrained network with a single multivariate Gaussian distribution [1]. It then performs detection by scoring samples based on the unweighted sum of the Mahalanobis distances to these layer-wise distributions. As the intermediate feature maps are too large to use directly, their dimensionality is reduced by channelwise average pooling. See Figure 1 for a visual representation. MahaAD is an unsupervised method, as no labels or additional validation data are needed.

3 Experiments

We perform a total of three evaluations on out-of-distribution problems. We first evaluate the method on two datasets, created by Márquez-Neila and Sznitman [2], where the task is to distinguish normal scans from corrupted ones. The third scenario involves separating healthy chest X-rays from pathological ones, using the dataset of Tang et al. [3], a subset of the NIH chest X-ray dataset.

We see from Table 1 that MahaAD outperforms all other methods, most notably on the chest X-ray corruptions.

	DDV	SVDD	Glow	OCGAN	MemDAE	MahaAD (b0)
OCT corr.	96.31	77.38	44.83	-	-	**97.54**
Chest corr.	82.38	66.63	54.62	-	-	**99.61**
Chest path.	70.32	-	-	83.35	87.78	**88.77**

Table 1: Performance of MahaAD on three medical cases in AUROC. Top two rows are corruptions (non-semantic shift), the bottom row concerns pathologies (semantic shift). Results of DDV, SVDD and Glow on the corruptions taken from Márquez-Neila and Sznitman [2], OCGAN and MemDAE results from Bozorgtabar et al. [4].

Table 2 shows that, especially for the pathologies, bigger networks do not significantly increase performance. On the other hand, on a natural image dataset such as MVTEC, performance only starts dropping for the largest network. Note that the method achieves near perfect scores on the OCT corruptions, hence the small differences there could simply be a consequence of statistical noise.

	b0	b1	b2	b3	b4	b5	b6	b7
MVTEC	90.6	93.3	93.6	94.0	94.8	95.2	**95.3**	94.2
OCT corr.	97.5	98.0	98.2	98.0	97.8	97.8	97.9	**98.4**
Chest corr.	99.6	99.7	**99.8**	99.3	96.9	93.2	90.4	86.8
Chest path.	**88.8**	85.2	85.6	85.6	85.5	86	83.4	85.1
ImageNet top-1 [5]	76.3	78.8	79.8	81.1	82.6	83.3	84.0	84.4

Table 2: Results in AUROC with EfficientNets of different sizes. MVTEC are natural images, the others are medical scans. MVTEC results taken from Rippel et al. [1]. The top-1 classification performance of the EfficientNets on ImageNet is included for reference.

Example images from each dataset are given in Figure 2.

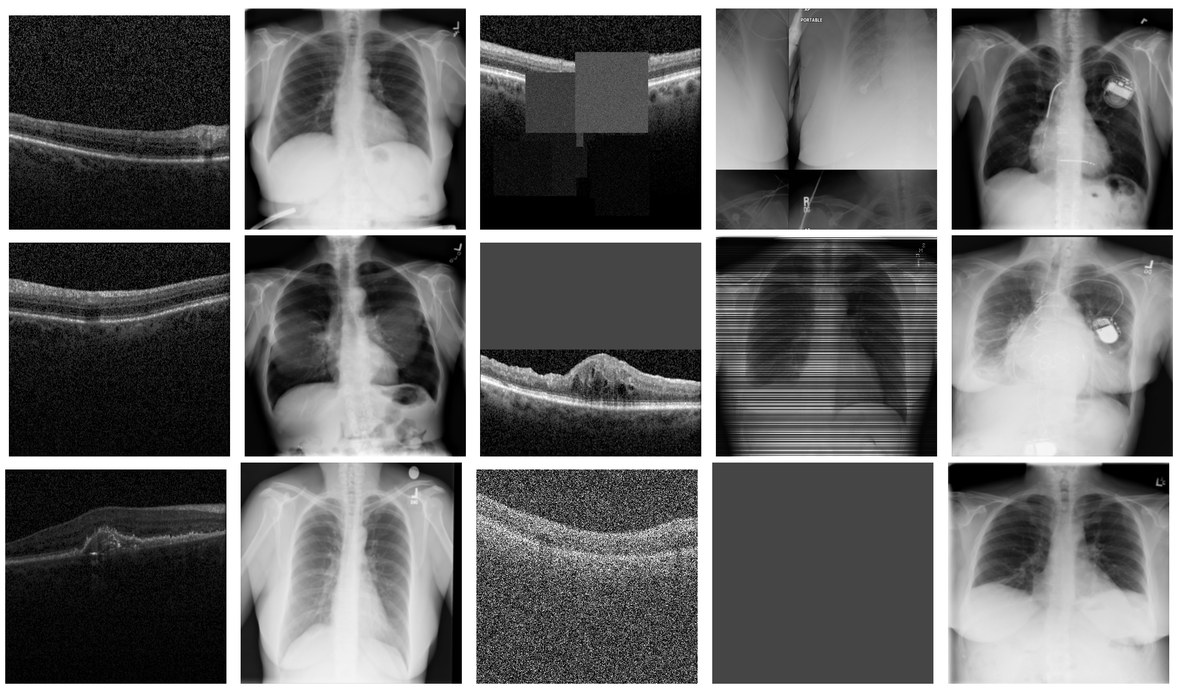

Figure 2: Examples images from the different datasets. From left to right: normal OCT scans, normal chest X-rays, corrupted OCT scans, corrupted chest X-rays, diseased chest X-rays.

4 Conclusion

We have shown that the simple MahaAD method improves upon the state-of-the-art in the three medical cases considered, despite being trained solely on natural images, demonstrating the general applicability of these features. Furthermore, it is very fast, as it requires no training, and the results are deterministic. As a result, we believe it to be an excellent choice for unsupervised out-of-distribution detection in the medical domain, when dealing with unimodal data.
Future work could investigate whether the performance remains impressive when the training data follows a multimodal distribution, as a single multivariate Gaussian might not be enough to characterize it, as well as if the features can be adapted to the medical domain in some way to further increase results.

References

[1] O. Rippel, P. Mertens, and D. Merhof, "Modeling the distribution of normal data in pre-trained deep features for anomaly detection," *arXiv preprint arXiv:2005.14140*, 2020.

[2] P. Márquez-Neila and R. Sznitman, "Image data validation for medical systems," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 329–337, Springer, 2019.

[3] Y.-X. Tang, Y.-B. Tang, M. Han, J. Xiao, and R. M. Summers, "Abnormal chest x-ray identification with generative adversarial one-class classifier," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1358–1361, IEEE, 2019.

[4] B. Bozorgtabar, D. Mahapatra, G. Vray, and J.-P. Thiran, "Salad: Self-supervised aggregation learning for anomaly detection on x-rays," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 468–478, Springer, 2020.

[5] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, pp. 6105–6114, PMLR, 2019.