On Challenging Aspects of Reproducibility in Deep Anomaly Detection

K. Kirchheim, M. Filax, F. Ortmeier

Dept. Computer Science
Otto-von-Guericke University Magdeburg
Germany

August 23, 2022

Deep Anomaly Detection

Reproducibility

Challenging Aspects

- Nondeterminism in Network Optimization
- Sensitivity to Hyperparameters
- Complexity of Experiments
- Dataset Selection
- Resource Limitations
- Dependencies

Complexity-Evidence Trade-off

Anomaly Detection with Deep Neural Networks

- Assumption: normal data points $\mathbf{x} \in \mathcal{X}$ drawn from $p_{in}(\mathbf{x})$
- Anomalies: $\mathcal{A} = \{\mathbf{x} : p_{in}(\mathbf{x}) < \alpha\}$
- Learn $f_\theta : \mathcal{X} \rightarrow \mathbb{R}$

$$\text{outlier}(\mathbf{x}) = \begin{cases} 1 & \text{if} \quad f_\theta(\mathbf{x}) > \tau \\ 0 & \text{else} \end{cases} \tag{1}$$

Testing:

- Test ability of f_θ to distinguish between points drawn from p_{in} and several p_{out}^i

Types of Reproducibility [1]

Method Reproducibility:

- Reproducibility of the numerical results when the same code gets executed

Results Reproducibility:

- Reproducibility of statistically similar results when a method is reimplemented.

Inferential Reproducibility:

- Reproducibility of findings or conclusions in different experimental setups.

Challenging Aspects

Reproducible Research
in Pattern Recognition

Nondeterminism in Network Optimization

Performance of DNNs depend on random seed [1, 2, 3, 4]

- Initialization, Data Ordering, Data Splitting
- Randomness in Algorithms (Dropout)
- Randomness in low-level libraries (CUDA)
- ...

Mitigation:

- ~~Fix random seed~~
- Conduct repeated experiments with different random seeds, varying all sources of nondeterminism
- use e.g. statistical tests

Reproducible Research in Pattern Recognition

Nondeterminism in Network Optimization

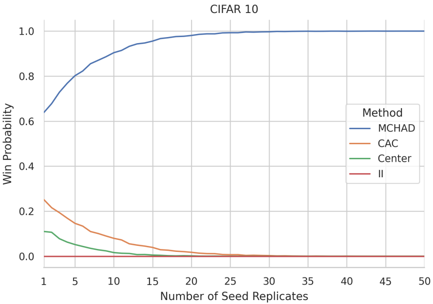
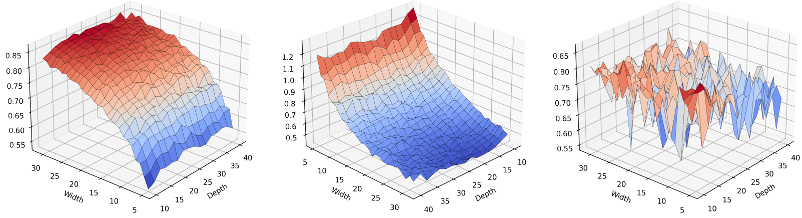


Figure: Estimated probability of method having the highest average AUROC over a specific number of seed replicates of experiments on the CIFAR10 dataset.

Sensitivity to Hyperparameters



(a) Accuracy (b) Loss (c) AUROC

Figure: Classification accuracy and AUROC of MCHAD for varying network widths and depths for a ResNet architecture [5], trained on the CIFAR-10 dataset.

Sensitivity to Hyperparameters

Mitigation:

- Perform sweeps to investigate influence of hyper parameters
- Allocate equal computational budget to each tested method [1]

Complexity of Experiments

- (Code) complexity increases likelihood of errors

Target leakage

- E.g., by overlap between datasets
- In pre-trained weights
- Inconsistent pre-processing

Mitigation:

- *“Outsource”* complexity to third parties
- Scrutinize training-scripts of pre-trained models

Reproducible Research in Pattern Recognition

Dataset Selection

- Performance between datasets might differ

Mitigation:

- Test on large variety of different distributions p_{in}/p_{out}

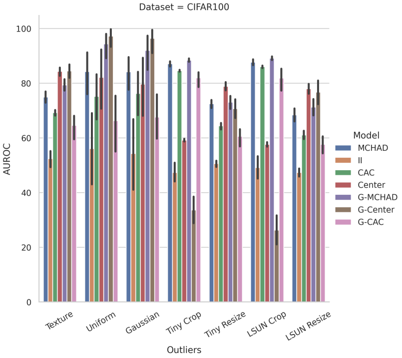


Figure: Anomaly Detection performance of models on different OOD datasets over 21 training runs with error bars.

Resource Limitations

- Optimization of DNNs is a resource-intensive process
- Resource requirements limit the number of individuals that can reproduce a method

Mitigation:

- use pre-trained models
- train for fewer iterations
- do fewer experiments

Dependencies

Software/Code, Data, pre-Trained models

- Sometimes difficult to set up
- Might be taken down at some point, e.g., [6]

Mitigation:

- Virtualization
- Reduce number of dependencies
- Copy dependencies to own code repository

Reproducible Research in Pattern Recognition

Complexity-Evidence Trade-off

Complexity-Evidence Trade-off

- Increase Inferential reproducibility $\rightarrow$ increase complexity
- Increase Method reproducibility $\rightarrow$ reduced complexity
- Results reproducibility: it depends

Aspect	**Inferential**	**Results**	**Method**
Nondeterminism	More Experiments	More Experiments	
HP-Sensitivity	More Experiments		
Complexity		Decrease Complexity	Decrease Complexity
Dataset Selection	More Experiments		
Resource Limitations	Decrease	Decrease	Decrease
Dependencies			Reduce Dependencies

Conclusion

- Complexity of experiments decreases the reproducibility
- Strength of the empirical evidence increases the reproducibility
- Trade-off
- Inferential Reproducibility more important

MCHAD (GitLab)

Reproducible Research in Pattern Recognition

References I

Xavier Bouthillier, César Laurent, and Pascal Vincent.
Unreproducible research is reproducible.
In International Conference on Machine Learning, pages 725–734, 2019.

Konstantin Kirchheim, Tim Gonschorek, and Frank Ortmeier.

Addressing randomness in evaluation protocols for out-of-distribution detection.
2nd Workshop on Artificial Intelligence for Anomalies and Novelties at IJCAI, 2021.

Cecilia Summers and Michael J. Dinneen.
Nondeterminism and instability in neural network optimization.
In International Conference on Machine Learning, pages 9913–9922. PMLR, 2021.

References II

- Prabhat Nagarajan, Garrett Warnell, and Peter Stone.
  The impact of nondeterminism on reproducibility in deep reinforcement learning.
  In 2nd Reproducibility in Machine Learning Workshop at ICML, 2018.

- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.
  Deep residual learning for image recognition.
  In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

- Abeba Birhane and Vinay Uday Prabhu.
  Large image datasets: A pyrrhic win for computer vision?
  In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1536–1546. IEEE, 2021.

Reproducible Research in Pattern Recognition