Improving Out-of-Distribution Detection with Markov Logic Networks

Konstantin Kirchheim[1] Frank Ortmeier[1]
Otto-von-Guericke University Magdeburg, Germany

Background

Out-of-Distribution Detection with Logical Reasoning [1]

Hypothesis: Current detectors rely too much on statistical patterns in neural representations and neglect high-level semantics

Idea

- Train DNNs to detect some human-understandable concepts in input
- Formulate constraints φ_i on plausible concept combinations for In-Distribution (ID) data, e.g.: *Stop-signs are red octagons*
- Inputs that violate a constraint are marked as Out-of-Distribution (OOD)

Limitations

- Strict logic too rigid for real-world applications where statistical associations dominate
- Instead, we seek a model in which frequently violated constraints contribute only marginally to the anomaly score

Markov Logic Networks (MLN) [3]

- Probabilistic generalization of First-order Logic (FOL)
- Can be seen as templates for large Markov Networks
- Each FOL formula φ_i is associated with a weight w_i
- For some input z, a MLN $\mathcal{M}$ predicts (simplified):

$$P_{\mathcal{M}}(z) = \frac{1}{Z} \exp\left(\sum_i w_i \varphi_i(z) \right) \tag{1}$$

Detection Approach

Standalone Markov Logic Network

- Train DNNs to approximate interpretation of FOL predicates $\{\mathcal{P}_n\}_{n=1}^N$
- Create constraint set $\{\varphi_i\}_{i=1}^N$ with these predicates
- Train MLN weights w_i by maximizing likelihood on ID training set
- Inference time outlier score:

$$D_{\mathcal{M}}(x) = -\sum_i w_i \varphi_i(\mathbf{x}) \tag{2}$$

- We do not need to compute partition function Z because $D_{\mathcal{M}}(\mathbf{x}) \propto P_{\mathcal{M}}(\mathbf{x}) \rightarrow$ Fast

Explainability

We know exactly by what amount a violated rule changed the outlier score

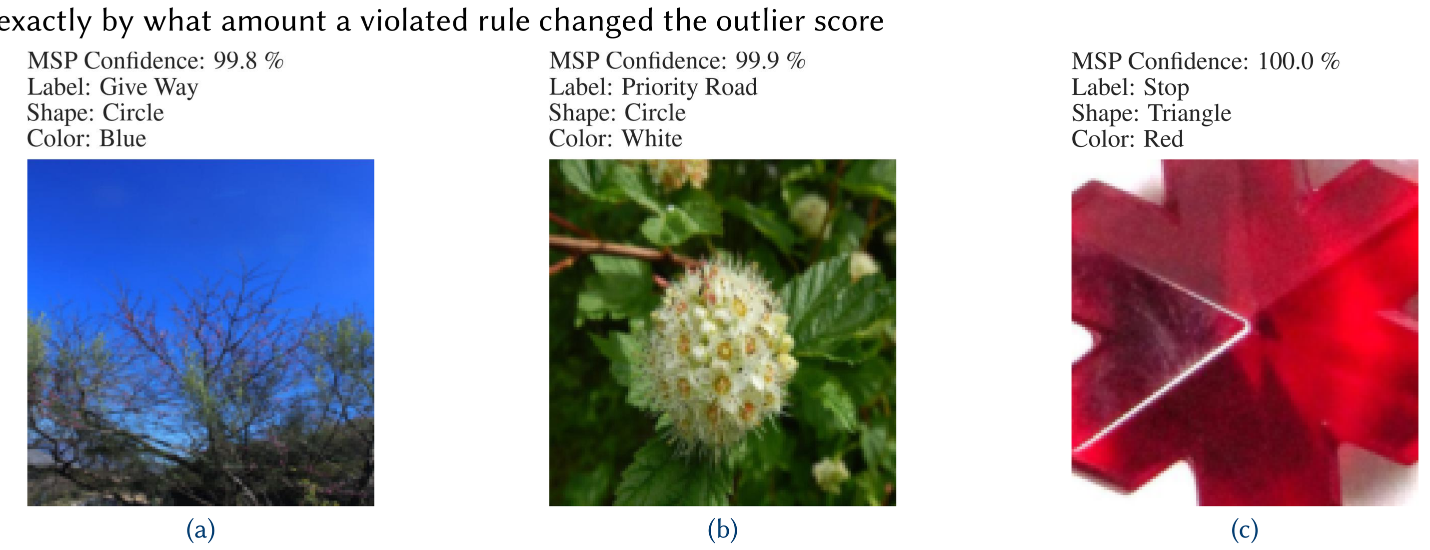


Figure: OOD samples with MSP confidence as predicted by a DNN trained on the GTSRB dataset

Combination with other Detectors

- Normalizing outlier scores is necessary
- For detector $D : \mathcal{X} \rightarrow \mathbb{R}$, fit some distribution to outlier scores for ID data
- Estimate survival function p_D over ID scores to transform outputs into calibrated $[0, 1]$ range
- Combined outlier score: $p_D(\mathbf{x}) \times -\sum_i w_i \varphi_i(\mathbf{x})$

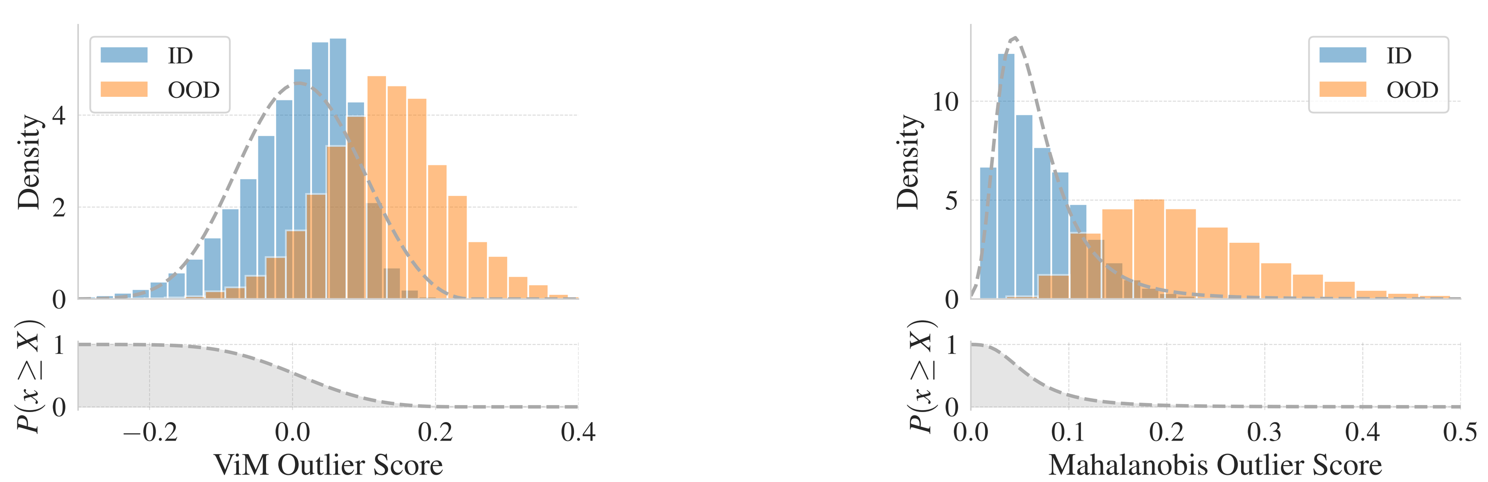


Figure: Approximating survival functions of outlier scores using GED

Constraint Search

Learning First-order Logic Constraints from Data

- For some datasets, no constraints available *a priori*
- Idea: take dataset with ID and OOD examples and optimize set of constraints by solving

$$\max_{\varphi \in \mathcal{P}(\mathcal{T})} \underbrace{\mathbb{E}_{(x_{\text{ID}}, x_{\text{OOD}})}\left[J(\varphi, x_{\text{ID}}, x_{\text{OOD}}) \right]}_{\text{Performance}} - \lambda \underbrace{C(\varphi)}_{\text{Complexity}} \tag{3}$$

  where $\mathcal{T}$ is the set of possible constraints and $\mathcal{P}$ is the powerset
- Exact computation is intractable

Proposed Greedy Algorithm

- Add a constraint if it improves performance by at least $\delta_{\min}$

Input: Training set $\mathcal{D}_{\text{train}}$, validation set $\mathcal{D}_{\text{val}}$, baseline performance J_0, rule set $\mathcal{T}$
Output: Selected constraints φ
Initialize $\varphi \leftarrow \emptyset$
Initialize $J \leftarrow J_0$
for all $\varphi_i \in \mathcal{T}$ **do**
  $\varphi' \leftarrow \varphi \cup \{\varphi_i\}$
  Train MLN detector with φ' on $\mathcal{D}_{\text{train}}$
  $J' \leftarrow$ Evaluate detector on $\mathcal{D}_{\text{val}}$
  if $J' > J + \delta_{\min}$ **then**
    $J \leftarrow J'$
    $\varphi \leftarrow \varphi'$
  end if
end for
return φ

Experiments

Traffic Sign Recognition (GTSRB) [4]

- We have 43 constraints over the predicates: class, shape and color
- Statistically significant performance gains, e.g. MLN+Ensemble reduces FPR95 by 37% (relative)
- Across detectors, MLN consistently enhances performance

Face Attribute Prediction (CelebA) [2]

Constraint search on CelebA yields the following result:

$$\forall \mathbf{x} \quad \textsc{young}(\mathbf{x}) \tag{4}$$
$$\forall \mathbf{x} \quad \textsc{heavy\_makeup}(\mathbf{x}) \Rightarrow \textsc{gray\_hair}(\mathbf{x}) \tag{5}$$
$$\forall \mathbf{x} \quad \textsc{wearing\_lipstick}(\mathbf{x}) \Rightarrow \textsc{gray\_hair}(\mathbf{x}) \tag{6}$$
$$\forall \mathbf{x} \quad \textsc{wearing\_lipstick}(\mathbf{x}) \Rightarrow \textsc{no\_beard}(\mathbf{x}) \tag{7}$$
$$\forall \mathbf{x} \quad \neg\textsc{male}(\mathbf{x}) \Rightarrow \textsc{no\_beard}(\mathbf{x}) \tag{8}$$

- Since constraints are human-understandable, we can manually curate them
- E.g. for MLN+Ensemble, FPR95 is reduced by 20% (relative)
- Overall, combination with MLN improves performance of all tested detectors

Table: AUROC for different detectors on **GTSRB** using a pattern-based baseline, combination with MLN, and a supervised MLN-based detector. All values in percent, averaged over ten seeds. Δ indicates the gain relative to the preceding column.

Detector	Baseline	+MLN	+Supervision
MSP	98.96	99.60 Δ 0.64	99.90 Δ 0.30
Ensemble	99.80	99.88 Δ 0.08	99.96 Δ 0.08
EBO	99.05	99.50 Δ 0.45	99.77 Δ 0.27
DICE	99.04	99.50 Δ 0.46	99.77 Δ 0.27
SHE	84.13	95.04 Δ 10.91	99.83 Δ 4.79
ReAct	96.85	99.09 Δ 2.24	99.92 Δ 0.82
Mahalanobis	99.23	99.72 Δ 0.49	99.96 Δ 0.23
ViM	99.47	99.80 Δ 0.33	99.96 Δ 0.16

Table: AUROC for different detectors on **CelebA** using a pattern-based baseline, combination with MLN, and a supervised MLN-based detector. All values in percent, averaged over ten seeds. Δ indicates the gain relative to the preceding column.

Detector	Baseline	+MLN	+Supervision
MSP	48.68	60.72 Δ 12.04	71.10 Δ 10.38
Ensemble	83.43	90.42 Δ 6.99	97.42 Δ 7.00
EBO	45.24	73.89 Δ 28.65	89.89 Δ 16.00
DICE	46.83	74.98 Δ 28.16	90.31 Δ 15.32
SHE	39.78	71.54 Δ 31.76	89.75 Δ 18.21
ReAct	44.84	72.06 Δ 27.22	89.55 Δ 17.49
Mahalanobis	95.12	96.01 Δ 0.89	97.86 Δ 1.85
ViM	84.94	91.75 Δ 6.82	97.12 Δ 5.37

Ablation Studies

Omitting Rules

- As expected, omitting constraints decreases performance
- Some constraints contribute more to performance than others

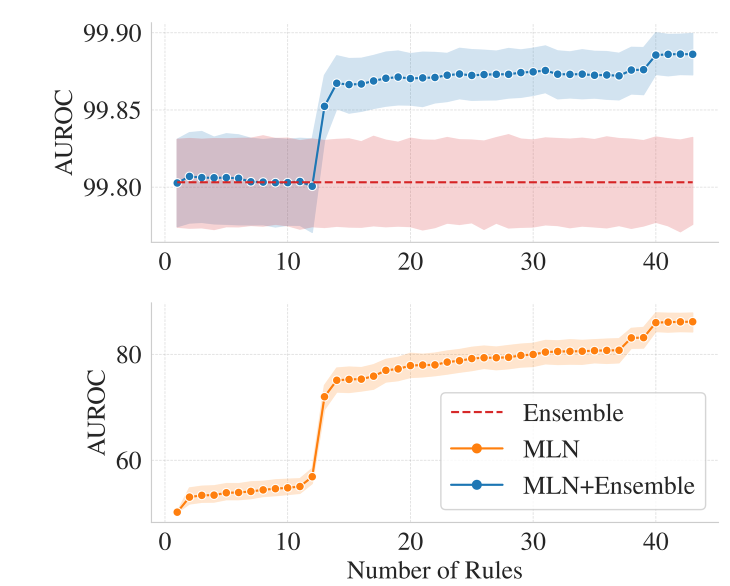


Figure: Ablation on constraints for GTSRB

Constraint Search Regularization

- Regularizing constraint optimization improves results
- No regularization leads to large number of rules
- Strong regularization leads to small number of rules, may degrade generalization

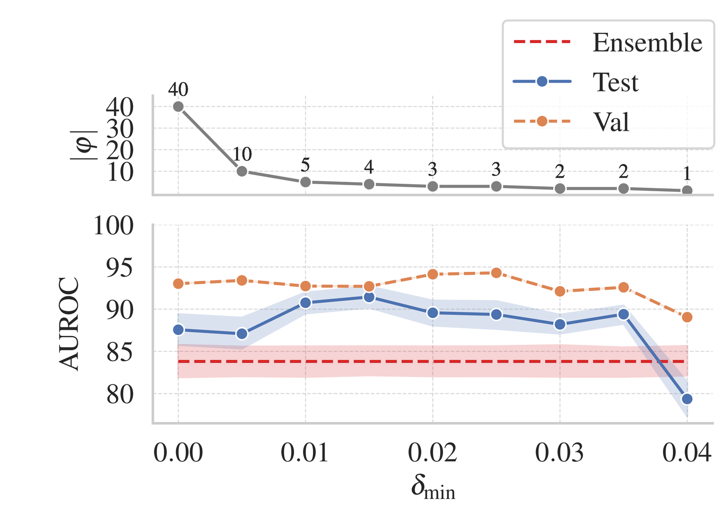
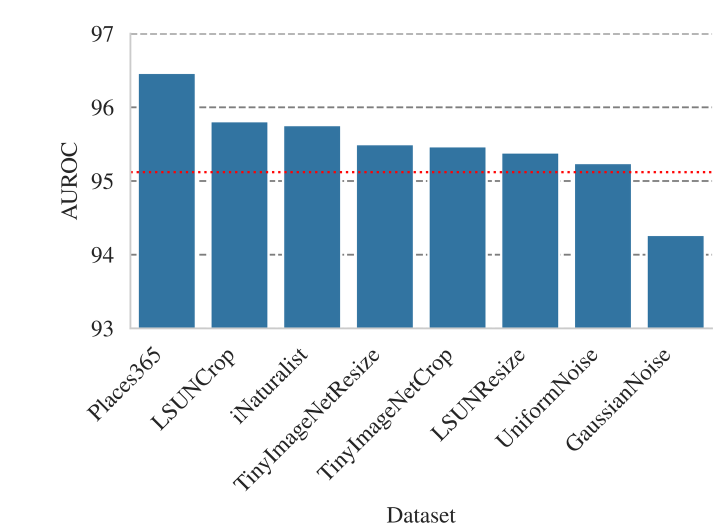


Figure: Number of constraints and performance for varying $\delta_{\min}$

Constraint Search Dataset

- Found constraints depend on OOD dataset used for optimization
- Sufficient variability seems beneficial
- Noise only provides a weak signal



(a) MLN-OOD repository

(b) PyTorch-OOD repository

Figure: GitHub Repositories

References

[1] Konstantin Kirchheim, Tim Gonschorek, and Frank Ortmeier. Out-of-distribution detection with logical reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, page 2122–2131, 2024.
[2] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, page 3730–3738, 2015.
[3] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62(1):107–136, 2006.
[4] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012.