

Language Models as Reasoners for Out-of-Distribution Detection

Konstantin Kirchheim Frank Ortmeier

Chair of Software and Systems Engineering Faculty of Computer Science Otto-von-Guericke-University Magdeburg, Germany

16.07.2024



Synopsis

- Out-of-Distribution (OOD) Detection
- OOD Detection with Logical Reasoning
- Language Models as Reasoners
- Results **GTSRB** Fruits
- Discussion **Advantages** Limitations
- Conclusion



Out-of-Distribution (OOD) Detection



...

Language Models as Reasoners

Results

Motivation



Figure: Training Data



...

Language Models as Reasoners

Results

Motivation



Figure: Training Data



Figure: Prediction: Priority Road, Confidence 100.0%



OOD Detection

- DNNs make high-confidence errors on OOD data
- In safety critical applications: OOD detection crucial

Formal

- Classifier $f: \mathcal{X} \to \mathcal{Y}$ maps inputs to per-class probabilities
- Detector $D_f : \mathcal{X} \to \mathbb{R}$ maps inputs to outlier scores
- Detection: apply threshold

Baseline

- Use (negative) maximum softmax probability (MSP) of f [1]
- For multiple classifiers f_1, \ldots, f_n : take average



OOD Detection with Logical Reasoning



Results

sion Referer

OOD Detection with Logical Reasoning



K. Kirchheim, T. Gonschorek, and F. Ortmeier, "Out-of-distribution detection with logical reasoning," in Proceedings of the IEEE/CVF Winter Conference on

Applications of Computer Vision, 2024, pp. 2122-2131



lesults

ion Referen

OOD Detection with Logical Reasoning

Perception

 Several DNNs *f*₁, ..., *f_n* detect human-understandable attributes



K. Kirchheim, T. Gonschorek, and F. Ortmeier, "Out-of-distribution detection with logical reasoning," in Proceedings of the IEEE/CVF Winter Conference on

Applications of Computer Vision, 2024, pp. 2122–2131



Results

lusion Refer

OOD Detection with Logical Reasoning

Perception

 Several DNNs f₁,...,f_n detect human-understandable attributes

Reasoning

- First-Order Logic Knowledge base
- E.g.: $\forall x \text{ stop}(x) \rightarrow \text{red}(x) \land \text{octagon}(x)$
- If any constraint violated \rightarrow OOD
- Otherwise: average MSP over all DNNs



Applications of Computer Vision, 2024, pp. 2122-2131





Traffic Sign Classification (GTSRB) [3]

Logical OOD

0

Language Models as Reasoners

Knowledge Base

• One constraint for each class: e.g. $\forall x \quad stop(x) \rightarrow red(x) \land octagon(x)$

Results

• Cuts SOTA error rate by $\approx 75\%$



Figure: GTSRB



Figure: OOD Data





ion Conclusion

n References

Large Language Models (LLM) as Reasoners

- LLM recent breakthrough in Natural Language Processing [4], [5]
- Creating and maintaining formal knowledge base is cumbersome
- LLMs possess extensive knowledge about the world
- Can be given instructions in natural language
- Can we use LLMs instead of a formal knowledge base?



Logical OOD Lan

Language Models as Reasoners

sults Di

Conclusion R

References





Results



usion Reference

Traffic Sign Classification (GTSRB) [3]

- Attributes: Sign Type, Color, Shape
- Test against OOD data from 5 unrelated datasets
- Average results over 10 training runs

Variants

- ▶ Naive LLM ("Is this OOD? Yes/No")
- Naive LLM + in-context domain knowledge
- LLM (with MSP as outlier scores)
- LLM + in-context domain knowledge



Results 000

Traffic Sign Classification (GTSRB) [3]

DETECTOR	AUROC ↑	AUPR-IN ↑	AUPR-OUT ↑	FPR95 \downarrow
MSP [1]	$99.04 \pm 0.07 $	$98.35 \pm 0.14 $	$99.29 \pm 0.05 $	2.54 ± 0.15
EBO [6]	$99.03 \pm 0.11 $	$98.76 \pm 0.14 $	$99.08 \pm 0.12 $	$2.26 \pm 0.27 $
MaxLogit [7]	$99.01 \pm 0.11 $	$98.73 \pm 0.14 $	$99.07\pm$	$2.29 \pm 0.27 $
Entropy [8]	$99.15 \pm 0.07 $	$98.64 \pm 0.13 $	$99.33 \pm 0.06 $	$2.46 \pm 0.15 $
ReAct [9]	99.04 ± 0.10	$98.77 \pm 0.13 $	$99.08 \pm 0.12 $	2.21 ± 0.24
Mahalanobis [10]	$99.70 \pm 0.02 $	$99.40 \pm 0.06 $	$99.83 \pm 0.01 $	1.11 ± 0.05
ViM [11]	$96.96 \pm 0.08 $	$95.95 \pm 0.10 $	$99.75 \pm 0.02 $	$6.08 \pm 0.16 $
Ensemble [12]	$99.77 \pm 0.03 $	$99.58 \pm 0.05 $	$99.86 \pm 0.01 $	$0.99 \pm 0.07 $
LogicOOD [2]	99.85 ± 0.01	99.74 \pm 0.02	99.92 ± 0.01	0.60 \pm 0.04
Naive Mistral 7B	50.42 ±0.82	40.72 ±2.60	78.54 ±0.24	100.00 ±0.00
Nave Mistral 7B + KB	$62.45 \pm 1.39 $	$77.85 \pm 0.82 $	82.66 ±0.45	100.00 ± 0.00
Mistral 7B	$95.82 \pm 0.07 $	$84.73 \pm 0.47 $	98.17 ± 0.03	5.15 ± 0.11
Mistral 7B + KB	99.85 \pm 0.01	$\textbf{99.74} \pm \ 0.02$	$\textbf{99.92} \pm \ 0.01$	0.60 \pm 0.04



Results ○○● nclusion Referen

Other Language Models?

Tested Models

- Llama3 7B
- Gemma 7B
- Mistral 7B (best performance)
- Mixtral 8x7B [13]

Findings

- Naive approach usually does not outperform random guessing
- In-context domain knowledge usually provides performance increase
- Larger models do not necessarily provide better performance
- Results highly dependent on details of prompt



Fruits Dataset

- Attributes: Fruit Type, Color
- Domain knowledge e.g.: "All cherries are red"
- 5 OOD Datasets, averaged over 10 training runs

Findings

- LogicOOD: state-of-the-art
- Naive approach outperforms random guessing
- However: no LLM improves performance over baseline



Discussion



ults Discussion

sion References

Advantage: Explainability

- Raw outlier score often not informative
- LLMs can provide human-understandable explanations



esults Discussion

nclusion Referen

Advantage: Explainability

- Raw outlier score often not informative
- LLMs can provide human-understandable explanations
- Example observation: white, circle, go left or straight



Figure: Normal go-left-or-straight sign



Discussion

Advantage: Explainability

- Raw outlier score often not informative
- LLMs can provide human-understandable explanations
- Example observation: white, circle, go left or straight



Figure: Normal go-left-or-straight sign

Based on the provided background knowledge, a 'go left or straight' sign is expected to be blue in color. However, the observation states that the sign is white in color, which is not in line with the expected behavior.



Advantage: Bias Mitigation

Language Models as Reasoners

Discussion

 Extensive background-knowledge from Internet can help to reveal bias in dataset

[...] the observation of a yellow cherry is not unusual or anomalous, as yellow is a typical color for cherries.



Results

Discussion

usion Reference

Advantage: Bias Mitigation

 Extensive background-knowledge from Internet can help to reveal bias in dataset

[...] the observation of a yellow cherry is not unusual or anomalous, as yellow is a typical color for cherries.



Figure: Rainier Cherries



Discussion

Advantage: Flexibility

Provided in-context knowledge usually improves results

- Knowledge base does not have to be complete
- Can contain contradictions
- Constraints can be easily added or removed



800

Limitation: Resource Requirements

Advances in algorithms, hardware, quantization: faster LLM inference

Language Models as Reasoners

Still $\approx 1.6s$ per input for 7B Model on Nvidia A100 (fp16)

Possible Mitigation Strategies

- Smaller, domain specific LLMs
- Precompute all possible observations
- Wait for algorithmic/hardware improvements



Discussion ŏŏ

Limitation: Hallucinations

Detected: red, square-shaped, no-trucks sign

```
{"score": 0.3}
```

Explanation: Although the color of the sign is consistent with the given knowledge base, the shape being 'square' is not a valid shape for a 'no trucks' sign according to the provided background knowledge. However, since all observations are assumed to be correct and not misreported, it is possible that there is a new type of traffic sign with a square shape and 'no trucks' meaning that is not included in the given knowledge base. Therefore, the probability of anomaly is relatively low, as the sign's color is still consistent with the knowledge base.



Conclusion



Conclusion

- Domain constraints can be formulated in natural language \rightarrow user-friendly
- Explanations in natural language
- Can be as effective as formal knowledge bases

Possible Future Research Directions

- In-context examples?
- Multimodality?
- Combination with RAG for very large or evolving databases?



Reasoners Resul

Discussion

Conclusion References

Acknowledgment

This research has received funding from the *Federal Ministry for Economic Affairs and Climate Action (BMWK)* and the *European Union* under grant agreements 19I21039A.



Federal Ministry for Economic Affairs and Climate Action



References I

- [1] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks." International Conference on Learning Representations, 2017.
- [2] K. Kirchheim, T. Gonschorek, and F. Ortmeier, "Out-of-distribution detection with logical reasoning," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 2122–2131.
- [3] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," Neural networks, vol. 32, pp. 323-332, 2012.
- T. Brown, B. Mann, N. Ryder, et al., "Language models are few-shot learners," [4] Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.
- [5] OpenAI, "Gpt-4 technical report," ArXiv, vol. abs/2303.08774, 2023.



References II

- [6] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," Advances in Neural Information Processing Systems, vol. 33, 2020.
- D. Hendrycks, S. Basart, M. Mazeika, M. Mostajabi, J. Steinhardt, and D. Song, [7] "Scaling out-of-distribution detection for real-world settings," International Conference on Machine Learning, 2022.
- [8] R. Chan, M. Rottmann, and H. Gottschalk, "Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation," in Proceedings of the ieee/cvf international conference on computer vision, 2021, pp. 5128–5137.
- [9] Y. Sun, C. Guo, and Y. Li, "React: Out-of-distribution detection with rectified activations," Advances in Neural Information Processing Systems, vol. 34, pp. 144–157, 2021.



References III

- K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting [10] out-of-distribution samples and adversarial attacks," Advances in Neural Information Processing Systems, vol. 31, 2018.
- H. Wang, Z. Li, L. Feng, and W. Zhang, "Vim: Out-of-distribution with virtual-logit [11] matching," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4921–4930.
- B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive [12] uncertainty estimation using deep ensembles," in Advances in neural information processing systems, 2017, pp. 6402–6413.
- A. Q. Jiang, A. Sablayrolles, A. Roux, et al., "Mixtral of experts," arXiv preprint [13] arXiv:2401.04088, 2024.