HELMHOLTZ AI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

Helmholtz AI for Matter Research at HZDR

Helene Hoffmann, Mahnoor Tanveer, Steve Schmerler, Sebastian Starke, Peter Steinbach

HZDR / Onsite Consulting, Helmholtz-Zentrum Berlin, June 28, 2022

Helmholtz AI - a quick recap

What is Helmholtz AI?

- initiative by President of the Helmholtz Association, Prof. Otmar D. Wiestler

What is Helmholtz AI?

- initiative by President of the Helmholtz Association, Prof. Otmar D. Wiestler
- running over 7 years, 2019 - 2026

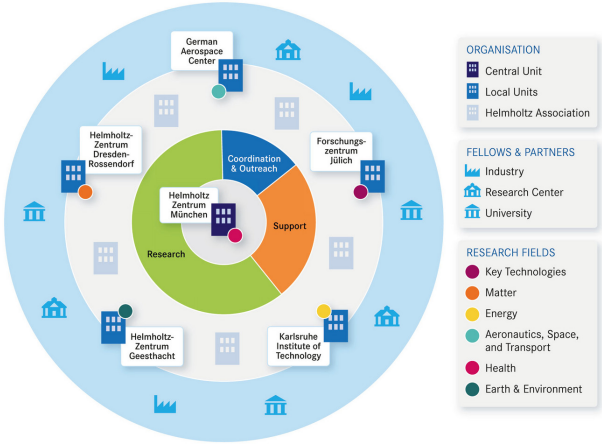
What is Helmholtz AI?

- initiative by President of the Helmholtz Association, Prof. Otmar D. Wiestler
- running over 7 years, 2019 - 2026
- 12 M€ per year
  (people + projects)

What is Helmholtz AI?

- initiative by President of the Helmholtz Association, Prof. Otmar D. Wiestler
- running over 7 years, 2019 - 2026
- 12 M€ per year (people + projects)
- central installation in Munich (universities and Helmholtz center)

Helmholtz AI: hub and spoke model

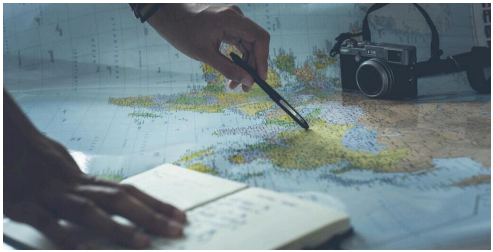


from www.helmholtz.ai

- five centers host a **local unit**
- each local unit:
  - **young investigator group**
  - **consultant team**
- central unit in Munich (universities and HMGU)
- planned staff:
  - 37 FTEs research
  - 35 FTEs consulting
  - 6 FTEs coordination, outreach, management

Two Funding Lines

Helmholtz AI Projects



unsplash.com:Glenn Carstens-Peters

- current call likely open until end of 2022
- max. 3 years, max. 200k € (must be matched)

Helmholtz AI Vouchers

unsplash.com:Dominik Scythe

- voucher submissions open anytime
- get in touch first: consultant-helmholtz.ai@hzdr.de

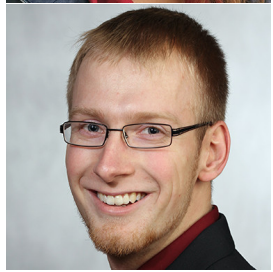
Helmholtz AI Local Unit For Matter At HZDR

Figure: Nico Hoffmann, YIG Lead

Figure: Peter Steinbach, Consultant Lead

Our Team at a Glance

Helmholtz AI Consultant Team at HZDR



- **reproducible automated** (ML) pipelines
- **inverse problems** & generative modelling
- (image) **denoising**
- **anomaly detection**
- regression & pattern recognition (object localisation, **image segmentation**)
- aspects of trustworthy ML (**uncertainties, robustness** and interpretability)

Jugend hackt

Lab Dresden:

- bi-weekly event (2h)
- mostly teenagers aged 12
- first contact with programming
- learning python using *pygame zero*
- developing ideas for projects

Goal:

- work on projects to make the world a better place
- get to use **ML** and know the risks

Opening 2021:

Anomaly Detection at Belle2 with KIT

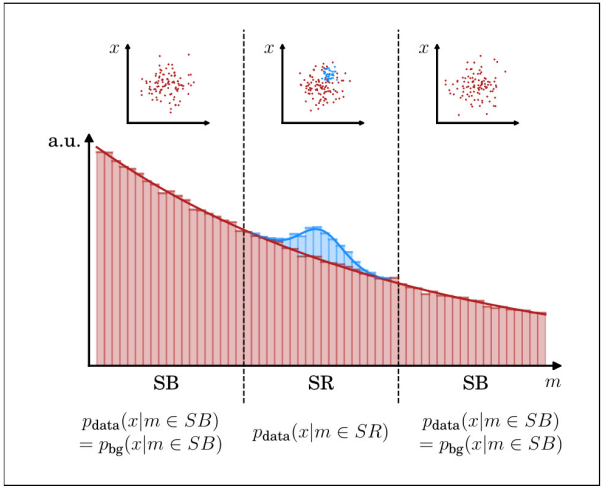


FIG. 1. Schematic view of the bump hunt. The signal (blue) is localized in the signal region (SR). The background (red) is estimated from a sideband region (SB).

- co-supervision of student (model suggested by KIT consultants)
- regular meetings
- code review
- discussion of results
- consulting in a narrow sense

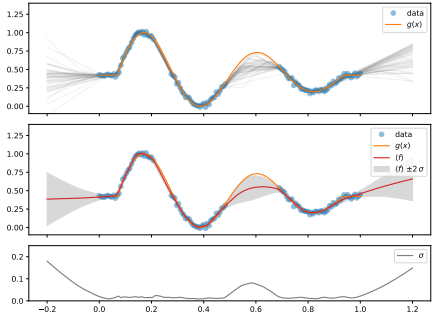
Uncertainty quantification and related methods

Voucher with A. Cangi, L. Fiedler, S. Kulkarni @CASUS

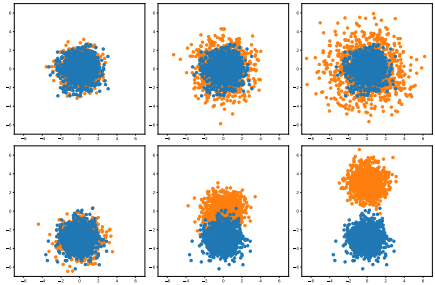
CASUS CENTER FOR ADVANCED SYSTEMS UNDERSTANDING

https://github.com/mala-project

Uncertainty quantification for NN surrogate models



Distribution shift detection for training and inference data sets (**visit our poster!**)

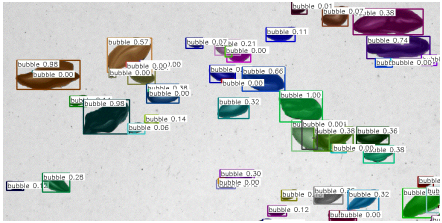


https://github.com/psteinb/c2st

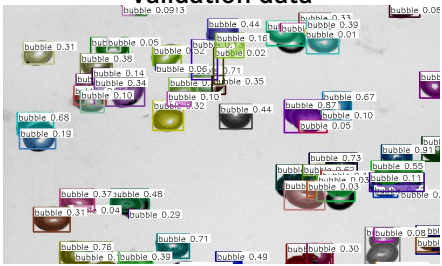
Video instance segmentation for bubbly flows

- accurate segmentation of bubble instances is a crucial first processing step in fluid dynamics applications
- extension of work on instance segmentation from static images
- Results look promising using the STEm-Seg algorithm [1]
- Goal: better handling of overlapping instances by exploiting temporal information (needs further evaluation)

Training data



Validation data



[1] https://github.com/sabarim/STEm-Seg

The Consulting Network at a Glance

Fostering *non-traditional* image processing

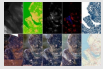
Challenge organization

DynamicEarthNet
EarthVision – CVPR
2021

Enhanced Sentinel 2 Agriculture
ESA Φ-lab
2021

AI4FoodSecurity
Planet hosted by ESA
2021/22

Helmholtz AI CountMeIn
Helmholtz Data Challenges
2022



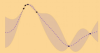
Benchmark data sets

So2Sat Global Urban LCZ
Urban local climate zone classification
Remote Sensing of Environment

DynamicEarthNet
Semantic change segmentation
CVPR 2022

DENETHOR
Daily crop monitoring
NeurIPS 2021

Analysis tools



Optimization
methods

Performance
prediction

AutoML research

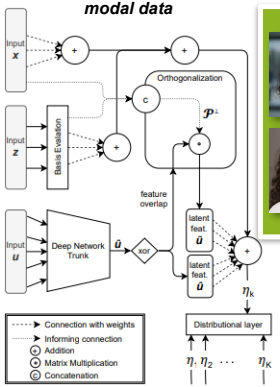
Deep
learning
research

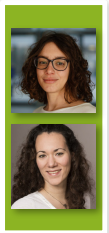
[5] Camero, A., Wang, H., Alba, E. and Bäck, T., 2021. Bayesian neural architecture search using a training-free performance metric. Applied Soft Computing, 106, p.107356.
[6] Kondmann, L., et al., 2021. DENETHOR: The DynamicEarthNET dataset for Harmonized, inter-Operable, analysis-Ready, daily crop monitoring from space. In Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track
[7] Traore, K.R., Camero, A. and Zhu, X.X., 2021. Lessons from the Clustering Analysis of a Search Space: A Centroid-based Approach to Initializing NAS. Workshop on 'Data Science Meets Optimisation' at IJCAI 2021
[8] Traoré, K.R., Camero, A. and Zhu, X.X., 2021. Compact Neural Architecture Search for Local Climate Zones Classification. In 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning.
[9] Traoré, K.R., Camero, A. and Zhu, X.X., 2021. A Data-driven Approach to Neural Architecture Search Initialization. arXiv preprint arXiv:2111.03524.
[10] Zhu, X.X., Qiu, C., Hu, J., Shi, Y., Wang, Y., Schmitt, M., and Taubenböck, H., 2021. The Urban Morphology on Our Planet - Global Perspectives from Space. Remote Sensing of Environment, 269, p.112794.
[11] Toker, A., et al., 2022. DynamicEarthNet: Daily Multi-Spectral Satellite Dataset for Semantic Change Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[12] Traoré, K.R., Camero, A. and Zhu, X.X., 2022. HPO: We won't get fooled again. In First Conference on Automated Machine Learning (Late-Breaking Workshop).

A glimpse into our vouchers

Supporting the community with open-source tools

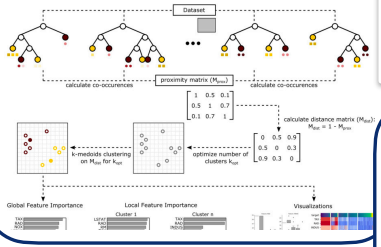
PySDDR: for high-dimensional multi-modal data



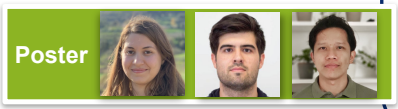


D Ruegamer, *etal*. "deepregression: a Flexible Neural Network Framework for Semi-Structured Deep Distributional Regression." *arXiv:2104.02705*.

Forest-guided clustering: Interpretability of random forests



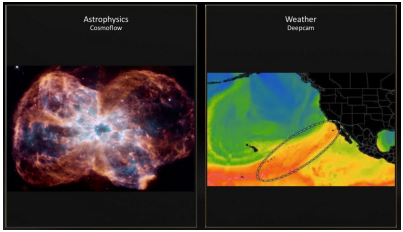
ML pipeline template: for clean and reproducible research code



MLPerf Training HPC

Highly Parallel Training of Neural Networks

- Active contributions to the MLCommons working group
- Training on up to 3072 GPUs on JUWELS Booster
- Based on containerized solution from NVIDIA
- IO Optimization (HDF5)
- Hacking around PyTorch problems for 2048+ GPUs



Karlsruher Institut für Technologie

JÜLICH Forschungszentrum

1.0-1110	HelmholtzAI	horeka_gpu_n512_pytorch1.10	Intel Xeon Platinum 8368	256	NVIDIA A100-PCIE-40GB	512	PyTorch 1.10		4.36
1.0-1111	HelmholtzAI	juwelsbooster_gpu_n1024_mxnet1.9	AMD EPYC 7402	512	NVIDIA A100-SXM4-40GB	1024	MXNet 1.9	16.73	
1.0-1112	HelmholtzAI	juwelsbooster_gpu_n1024_pytorch1.10	AMD EPYC 7402	512	NVIDIA A100-SXM4-40GB	1024	PyTorch 1.10		2.73
1.0-1113	HelmholtzAI	juwelsbooster_gpu_n2048_pytorch1.10	AMD EPYC 7402	1024	NVIDIA A100-SXM4-40GB	2048	PyTorch 1.10		2.56

Some of Our Research

Approximated Uncertainties $\hat{\sigma}$, arxiv:2204.05173

Approximate Accuracy as a Bernoulli probability

$$\mu_{\text{ACC}} \pm \hat{\sigma}_{\text{ACC}} = \mu_{\text{ACC}} \pm z \sqrt{\frac{1}{n_{holdout}} \text{ACC}_{holdout} \left(1 - \text{ACC}_{holdout}\right)}$$

In the limit of large numbers, this converges to a normal distribution. Use z to construct confidence interval assuming normality.

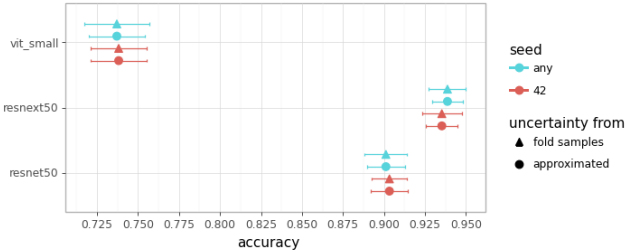
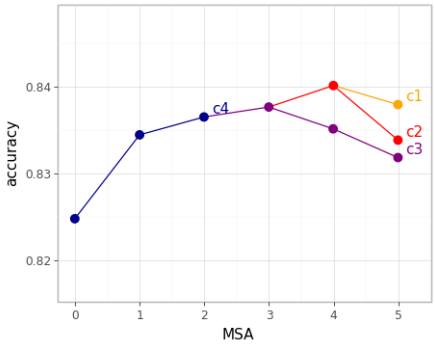


Figure 5: Comparison of fold sample based uncertainty with approximated uncertainty using eq. (1) [2]. Each estimate was obtained for one seed (42) or any seed available (total 6 seeds). The uncertainty plotted for seed 42 was obtained using the approximation in eq. (1). The uncertainty plotted for all seeds was obtained using the sample standard deviation.

How Do Vision Transformers Work? [1]



How Do Vision Transformers Work? [1]

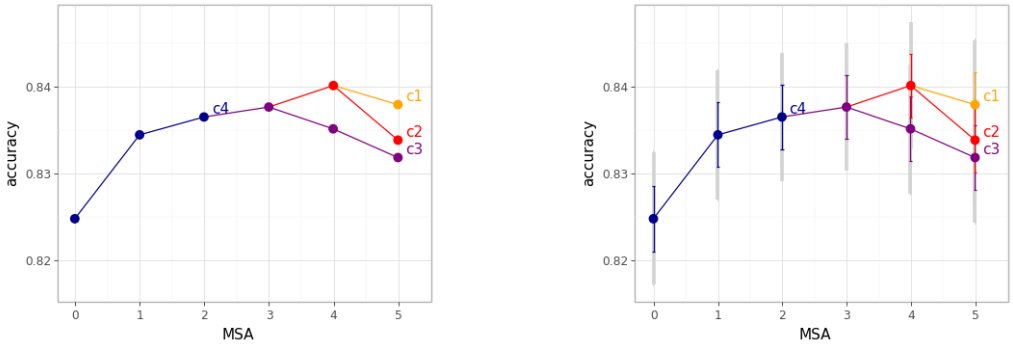


Figure 2: Reproduction of figure 12a from [1] (left). Augmentation of the same figure with estimated accuracy calculated using eq. (1) using a one-sigma 68.2% (colored) and two-sigma 95% (grey) confidence interval (right). Data to reproduce these figures was obtained by using [3] on the figures from the preprint PDF.

Summary

Conclusions

- diverse HZDR team ready for ML in matter research!

Conclusions

- diverse HZDR team ready for ML in matter research!
- our track record demonstrates our impact: happy to continue

Conclusions

- diverse HZDR team ready for ML in matter research!
- our track record demonstrates our impact: happy to continue
- putting ML consulting to the field is challenging:
  communication, communication, communication

Conclusions

- diverse HZDR team ready for ML in matter research!
- our track record demonstrates our impact: happy to continue
- putting ML consulting to the field is challenging:
  communication, communication, communication

Questions, Comments, Vouchers or Concerns are highly welcome!

References

References I

[1] Namuk Park and Songkuk Kim. How do vision transformers work?, 2022.

[2] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. 2018.

[3] Ankit Rohatgi. Webplotdigitizer: Version 4.5, 2021.