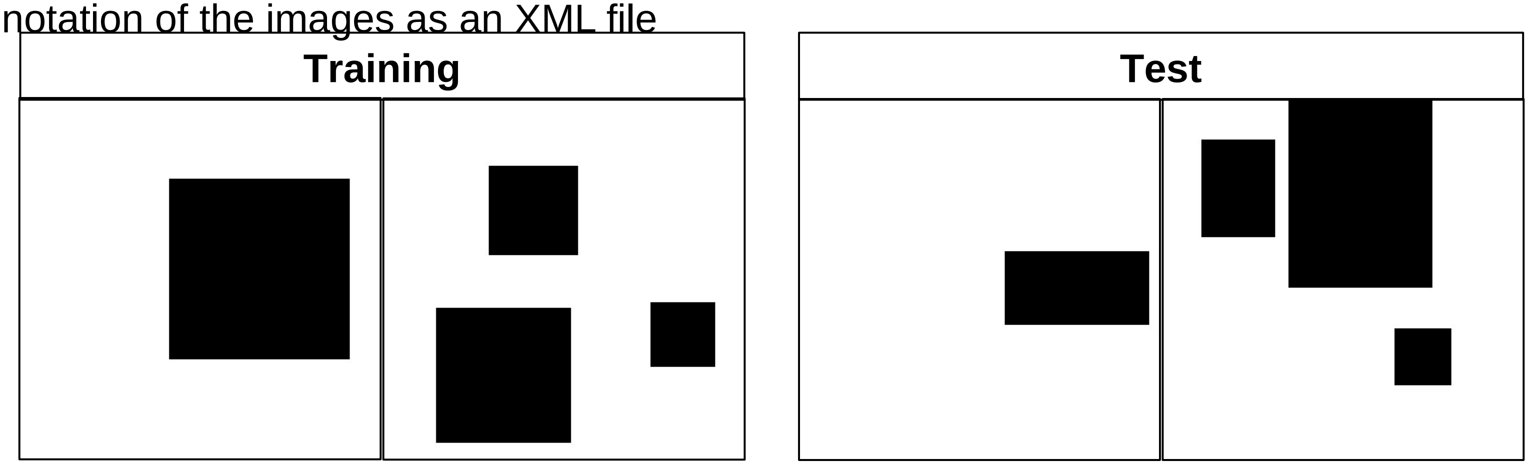
Topic 3: Understand AI: Finding traces in KI with geometric objects with Open Source

Introduction/Motivation: AI in the form of deep learning is being used more and more frequently. However, such approaches are mostly ‘black box’ approaches as it is often unclear how the algorithms work internally. In order to gain a better understanding, a simplified data set of geometric objects is to be used for training and testing both generative and predictive AI - see preliminary work in [SD24].

Information on the AMSL dataset:

- Resolution: 512x512 pixels
- Format: PNG
- Training: Images with squares of different sizes (8, 16, 32, 64, 128 pixels)
- Test: Images with rectangles of different sizes (varying based on {8, 16, 32, 64, 128} pixels)
- Annotation of the images as an XML file



Topic 3: Understand AI: Finding traces in KI with geometric objects with Open Source – Generative AI

Tasks:

- Literature review and, if necessary, further research on task understanding and selection of open source of different generative AI architectures
- Per person: Training and testing of a simplified open source autoencoder for the generation of quadrilaterals based on the AMSL dataset. Selection of possible autoencoder architectures based on open source (e.g. based on the pythae library, see [CVA22]):
  - Disentangling Variational Autoencoder [DKL+19]
  - Conditional Variational Autoencoder [KW14]
- Variation of the training configuration by means of:
  - Variation of the training duration (number of epochs)
  - Variation of the training data (size of the quadrilaterals, 1 vs. several quadrilaterals per image, amount of data)
  - Consideration of different sizes of the latent vector
- Evaluation of training and test images in comparison with the images reconstructed by the autoencoder
  - manual/visual comparison on a smaller test set of at least 10 images (per person)
  - Automated comparison, e.g. by means of a difference image, checking the shape and size of the quadrilaterals and the internal angle of the quadrilateral, performed on at least 1000 images
- XAI / explainability on the basis of the latent vector:
  - Visualise changes in the latent vector
  - Visualise the latent space with tSNE

Topic 3: Understand AI: Finding traces in KI with geometric objects with Open Source – Predictive AI

Tasks:

- Literature review and, if necessary, further research on task understanding and selection of open source of various generative AI architectures
- Per person: Training and testing of a simplified open source deep learning approach for object recognition of quadrilaterals, trained on the basis of the AMSL dataset. Selection of possible architectures based on open source:
    - SSD300, ResNet50, etc.
- Variation of the training configuration by means of:
    - Variation of the training data (size of the quadrilaterals, 1 vs multiple quadrilaterals per image, amount of data)
    - Variation of the classification target (detection of quadrilaterals, classification of size)
    - Consideration of different sizes of the latent vector
- Evaluation of training and test images in comparison with the images reconstructed by the autoencoder
    - manual/visual comparison on a smaller test set of at least 10 images (per person)
    - Automated comparison (e.g. using overlapping areas between detected bounding box and annotation or confusion matrix for classification of variables) on at least 1000 images
- Strengthening explainability using existing open source tools:
    - Consideration of layer-based XAI methods (cf. e.g. LRP, Grad-Cam, Captum https://github.com/pytorch/captum)
    - Consideration of counterfactual explanations (e.g. DiCE https://github.com/interpretml/DiCE)

Topic 3: Understand AI: Finding traces in KI with geometric objects with Open Source – both teams

Tasks:

- Literature review and, if necessary, further research on task understanding and selection of open source of various generative AI architectures
- Per person: Training and testing of a simplified open source deep learning approach for object recognition of quadrilaterals, trained on the basis of the AMSL dataset. Selection of possible architectures based on open source:
  - SSD300, ResNet50, etc.
- Variation of the training configuration by means of:
  - Variation of the training data (size of the quadrilaterals, 1 vs multiple quadrilaterals per image, amount of data)
  - Variation of the classification target (detection of quadrilaterals, classification of size)
  - Consideration of different sizes of the latent vector
- Evaluation of training and test images in comparison with the images reconstructed by the autoencoder
  - manual/visual comparison on a smaller test set of at least 10 images (per person)
  - Automated comparison (e.g. using overlapping areas between detected bounding box and annotation or confusion matrix for classification of variables) on at least 1000 images
- Strengthening explainability using existing open source tools:
  - Consideration of layer-based XAI methods (cf. e.g. LRP, Grad-Cam, Captum https://github.com/pytorch/captum)
  - Consideration of counterfactual explanations (e.g. DiCE https://github.com/interpretml/DiCE)

Topic 3: Understand AI:
Spurensuche in KI mit geometrischen Objekten mit Open Source

Expected result:

- Executable open source instance of the AI architecture under consideration (as source code), including all successfully trained models or approaches for explainability (executable on the AMSL GPU computer, possibly as a Docker instance)
- Scientific elaboration into a presenation and a scientific report, including process-accompanying documentation of all steps, well-founded selection of a concept taking into account and naming alternatives

Basic knowledge:

Programming skills (e.g. Python); basics of image processing; basic understanding of machine learning; motivation to familiarise yourself with new topics

Supervisors: Stefan Seidlitz (AI), Dennis Siegel (XAI)

Team: 4-8 (2-4 per subtask)

References:

[SD24] Stefan Seidlitz, Jana Dittmann: ‘Forensic Analysis of GAN Training and Generation: Output Artifacts Assessment of Circles and Lines’. Proceedings of the SECURWARE 2024, The Eighteenth International Conference on Emerging Security Information, Systems and Technologies, IARIA, 2024

[DKL+19] Yann Dubois, Alexandros Kastanos, Dave Lines, Bart Melman: ‘Disentangling VAE’. Online https://github.com/YannDubs/disentangling-vae 2019.

[CVA22] Clément Chadebec, Louis J. Vincent, Stéphanie Allassonnière: ‘Pythae: Unifying Generative Autoencoders in Python - A Benchmarking Use Case’. In Advances in Neural Information Processing Systems, vol 35, 2022

[KW14] Diederik P. Kingma, Max Welling: Auto-Encoding Variational Bayes. ICLR 2014, or the online source https://github.com/unnir/cVAE