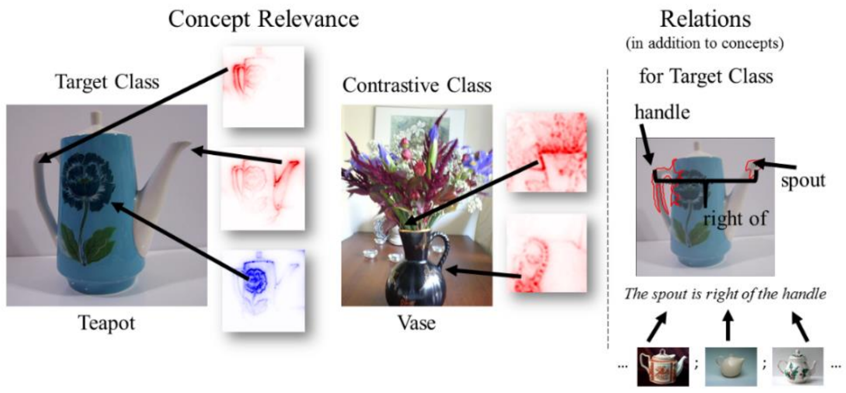
Generating Concept-based and Relational Explanations for Image Classification



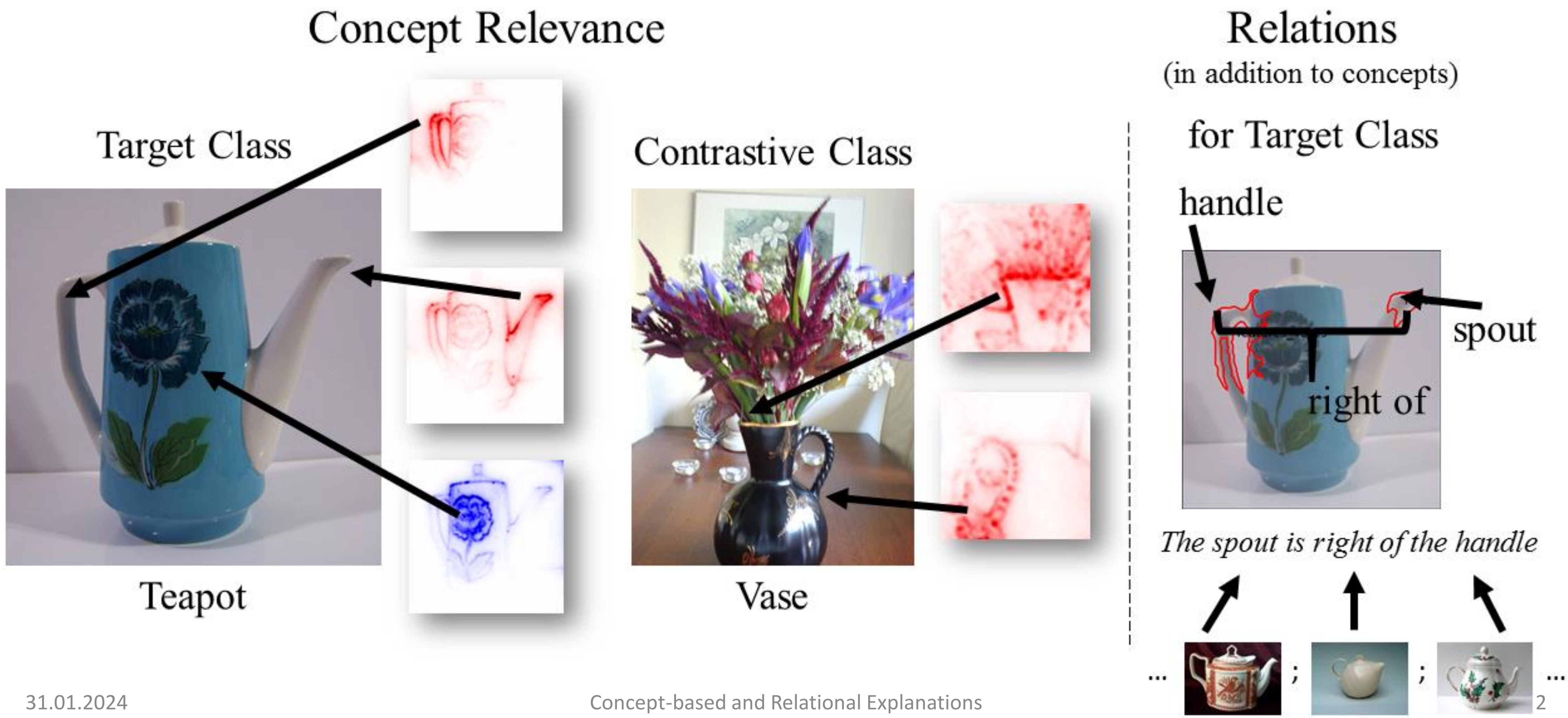
Based on a talk at the Dagstuhl Seminar 23442 in October 2023

Bettina Finzel

Cognitive Systems, University of Bamberg

(joint work with Patrick Hilme, Johannes Rabold and Ute Schmid)

Motivation



Concept Relevance
Target Class
Contrastive Class
Teapot
Vase
Relations
(in addition to concepts)
for Target Class
handle
spout
right of
The spout is right of the handle
... ; ; ...

Motivation

- Concept- and Relation-based Explanations

  → Classify and evaluate models on classes that share concepts, however, in different spatial configuration (contrastive explainability)

- Taking advantage of the benefits of
  - relevance-based explanations to extract concepts from Convolutional Neural Networks
  - comprehensible rule induction with Inductive Logic Programming

Concept Extraction and Relational Learning

- Concept Relevance Propagation (CRP) by Achtibat et al. (2023)
  - Relevance Maximization: Searching for most relevant samples

$$R_i^{(l)} = \sum_j \frac{a_i w_{ij}}{\Sigma_i a_i w_{ij}} R_j^{(l+1)} \Rightarrow R_i^{(l)}(x|\theta \cup \theta_l) = \frac{a_i w_{ij}}{\sum_i a_i w_{ij}} * \delta_{jc_l} * R_j^{l+1}(x|\theta)$$

$$\Rightarrow \tau_{max}^{rel}(x) = \max_i R_i(x|\theta)$$

- Inductive Logic Programming (ILP), e.g., Cropper et al. (2022)

$$\forall e \in E^+ : B \cup H \models e \text{ and } \forall e \in E^- : B \cup H \not\models e.$$

Experiment and Results

Table 2: The pre-trained models used for contrastive classification (CE is cross entropy; BCE is its binary equivalent).

	#Train	#Test	Train F1	Test F1	#Class	Batch Size	Max. Epochs	Optimizer	Loss Function	Learning Rate
VGG16-Picasso	18,002	1998	0.9933	0.9924	1	32	20	Adam	BCE	0.0001
VGG16-Adience	9942	2252	0.9913	0.8702	2	32	20	Adam	CE	0.0001
VGG16-Teapot-Vase	231	100	1.0000	0.9200	2	32	20	Adam	CE	0.0001
ResNet50-PathMNIST	25,765	2462	0.9974	0.9709	2	128	10	Adam	CE	0.001

Table 3: Explainer faithfulness and train/test data metrics for the networks after masking of rule + non-background-knowledge (BK) concepts compared to only masking non-background-knowledge concepts. Bold values indicate an expected drop in performance.

		Rule + Non-BK-Masking		**Non-BK-Masking**	
Experiment	Explainer Faithfulness	Amount of Masked Concepts	F1 score	Amount of Masked Concepts	F1 score
onTrain-VGG16-Picasso	0.9860	208	**0.2985**	164	0.9921
onTrain-VGG16-Adience-FM	1.0000	18	**0.9437**	8	0.9913
onTrain-VGG16-Adience-MF	1.0000	17	**0.9896**	8	0.9913
onTrain-VGG16-Teapot-Vase	0.9970	24	1.0000	14	1.0000
onTrain-VGG16-Vase-Teapot	0.9970	27	1.0000	14	1.0000
onTrain-ResNet50-PathMNIST	1.0000	1384	**0.9872**	1378	0.9873
onTest-VGG16-Picasso	0.9980	193	**0.6689**	162	0.9924
onTest-VGG16-Adience-FM	0.9975	15	**0.8079**	5	0.8708
onTest-VGG16-Adience-MF	0.9975	11	**0.8673**	5	0.8708
onTest-ResNet50-PathMNIST	0.9980	1374	0.9367	1367	0.9367

Experiment and Results

Figure 5: Teapot (top left), false positive female (top middle), outlier female (top right) and rule cluster for smiling persons (bottom row).

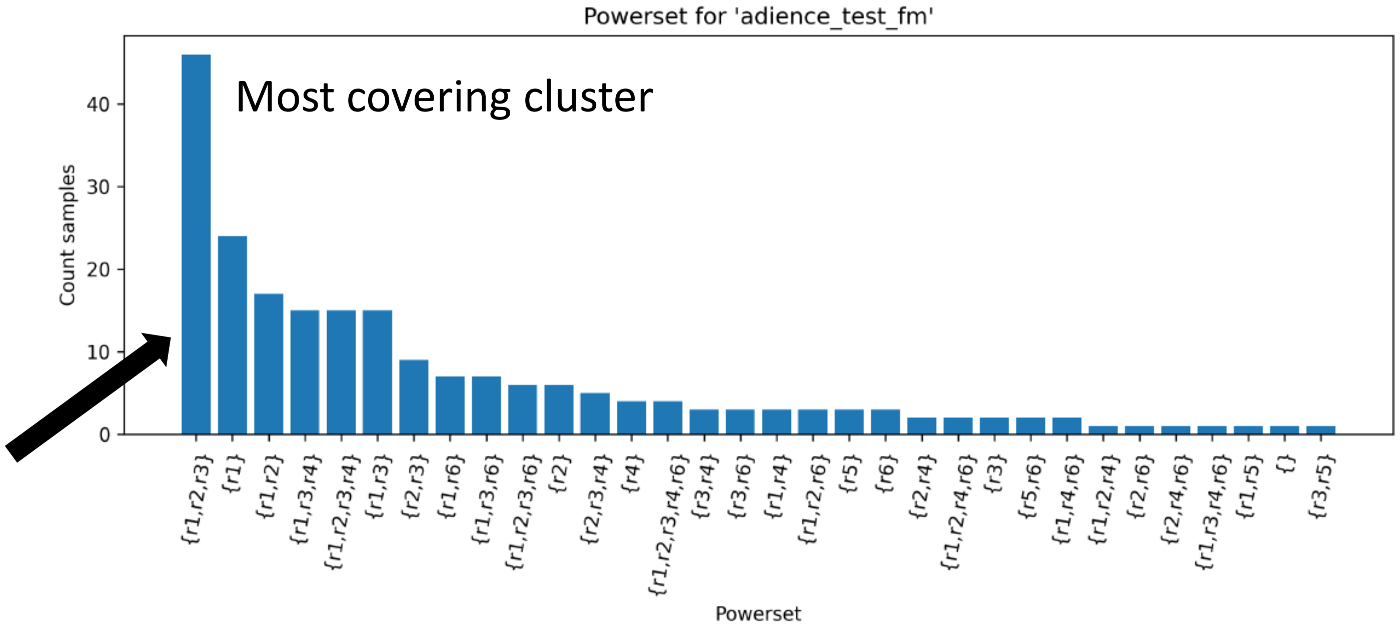


Figure 8: Rule clusters with most coverage for experiment Adience-Test-FM.

Summary

1. We introduced an explanation method that combines
   - relevance-based concept extraction with
   - interpretable relational learning
   - to validate concepts learned from images against general, domain-relevant spatial relations
2. In order to facilitate the exploration and adaptation of a model’s predictions, we enhanced our method by human-understandable contrastive explanations and by outlier detection

Summary

3. For a collection of data sets (abstract, real-world, scientific), we showed quantitatively that our ILP-based surrogate model is faithful to the CNN model’s predictive outcomes
   - when the CNN is enforced to use the most relevant concepts
   - when the CNN is permitted to use the most relevant concepts

References (Selection from paper)

- Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin, ‘From "where" to "what": Towards human-understandable explanations through concept relevance propagation’, CoRR, abs/2206.03208, (2022).
- Sebastian Bruckert, Bettina Finzel, and Ute Schmid, ‘The next generation of medical decision support: A roadmap toward transparent expert companions’, Frontiers in Artificial Intelligence, 3, 507973, (2020).
- Andrew Cropper, Sebastijan Dumancic, Richard Evans, and Stephen H. Muggleton, ‘Inductive logic programming at 30’, Machine Learning, 111(1), 147–172, (2022).
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, ‘Imagenet: A large-scale hierarchical image database’, in Proc. of the IEEE CVPR, pp. 248–255. IEEE, (2009).
- Eran Eidinger, Roee Enbar, and Tal Hassner, ‘Age and gender estimation of unfiltered faces’, IEEE Transactions on Information Forensics and Security, 9(12), 2170–2179, (2014).
- Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller, ‘Unmasking clever hans predictors and assessing what machines really learn’, Nature Communications, 10(1), 1–8, (2019).
- Johannes Rabold, Gesina Schwalbe, and Ute Schmid, ‘Expressive explanations of DNNs by combining concept analysis with ILP’, in German Conference on Artificial Intelligence (Künstliche Intelligenz), pp. 148–162. Springer, (2020).
- Jochen Renz, Qualitative spatial reasoning with topological information, Springer, 2002.
- Ashwin Srinivasan, The Aleph Manual, 2007.
- Stefano Teso and Kristian Kersting, ‘Explanatory interactive machine learning’, in Proc. of the AAAI/ACM AIES, pp. 239–245. ACM, (2019)
- Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni, ‘MedMNIST v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification’, CoRR, abs/2110.14795, (2021)