Fraunhofer IIS

Cognitive Systems Group
University of Bamberg · www.uni-bamberg.de/cogsys

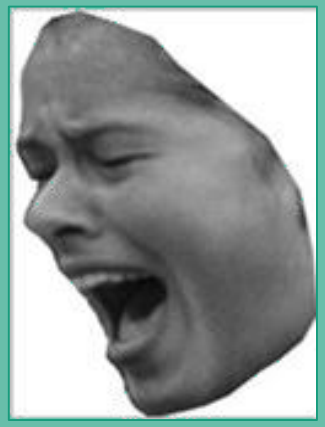
Towards Explaining Deep Learning Networks to Distinguish Facial Expressions of Pain and Emotions

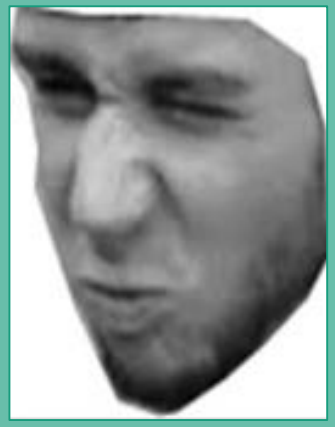
Katharina Weitz, Teena Hassan, Ute Schmid & Jens Garbas

Cognitive Systems Group
University of Bamberg · www.uni-bamberg.de/cogsys

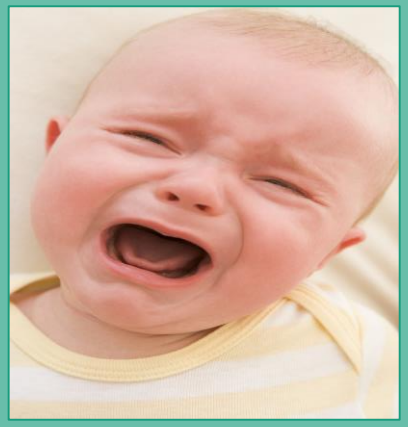
Fraunhofer IIS

Motivation



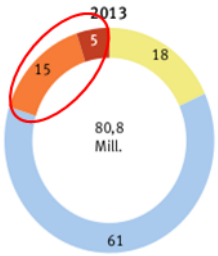


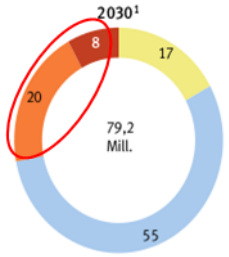
Aviezer et al. (2012)

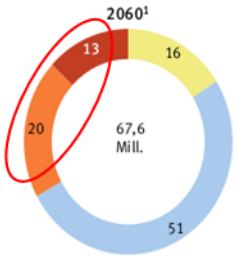


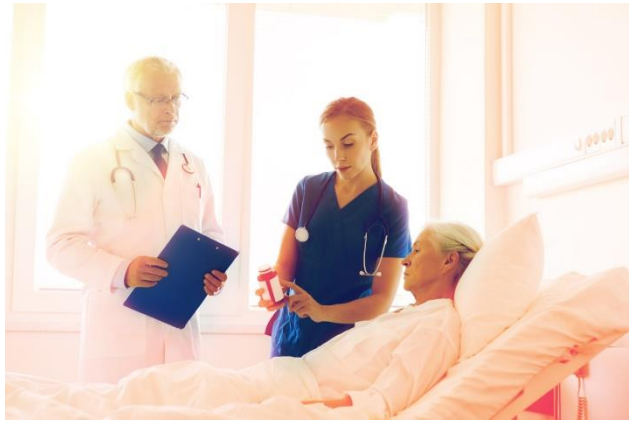
Brahnam et al. (2007)

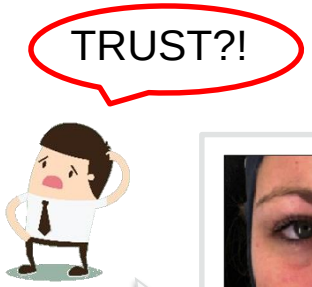
Motivation

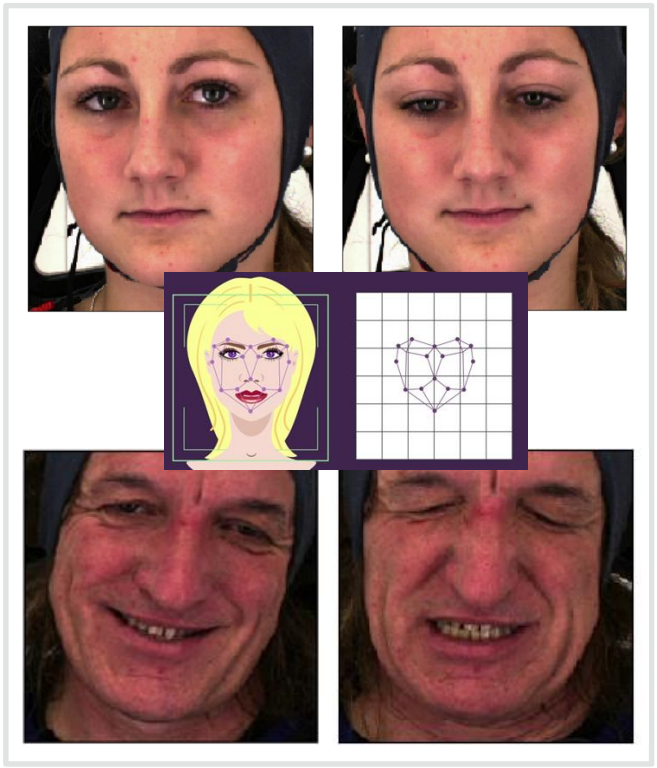












www.destatis.de

CogSys
Cognitive Systems Group
University of Bamberg · www.uni-bamberg.de/cogsys

Research Questions

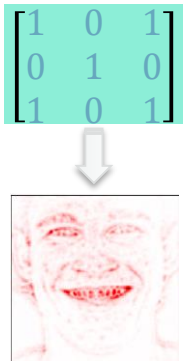
Predictive performance

Disgust

Pain

Happiness

Decision interpretation

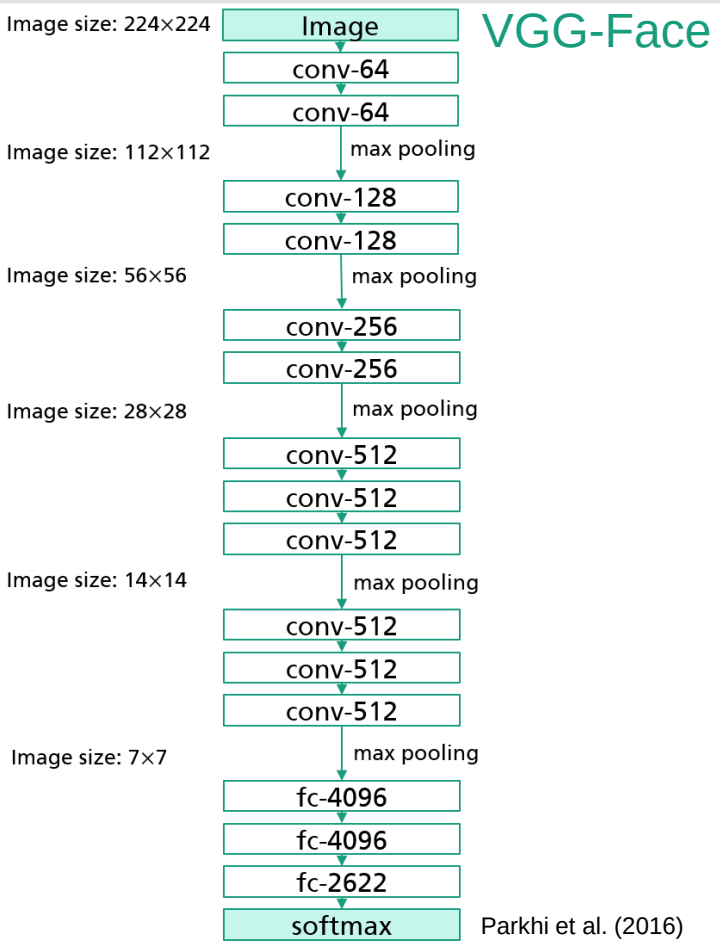


Feature explanation

Cognitive Systems Group
University of Bamberg · www.uni-bamberg.de/cogsys

Research Questions

Deep Learning



Explainable AI

Model-specific approaches

- Deconvnet
- Backpropagation
- Guided Backpropagation
- Grad-CAM
- Guided Grad-CAM
- **Layerwise Relevance Propagation (LRP)**

Model-agnostic approach

- Local Interpretable Model-agnostic Explanations (LIME)

Cognitive Systems Group
University of Bamberg · www.uni-bamberg.de/cogsys

XAI: Idea

Interpretation

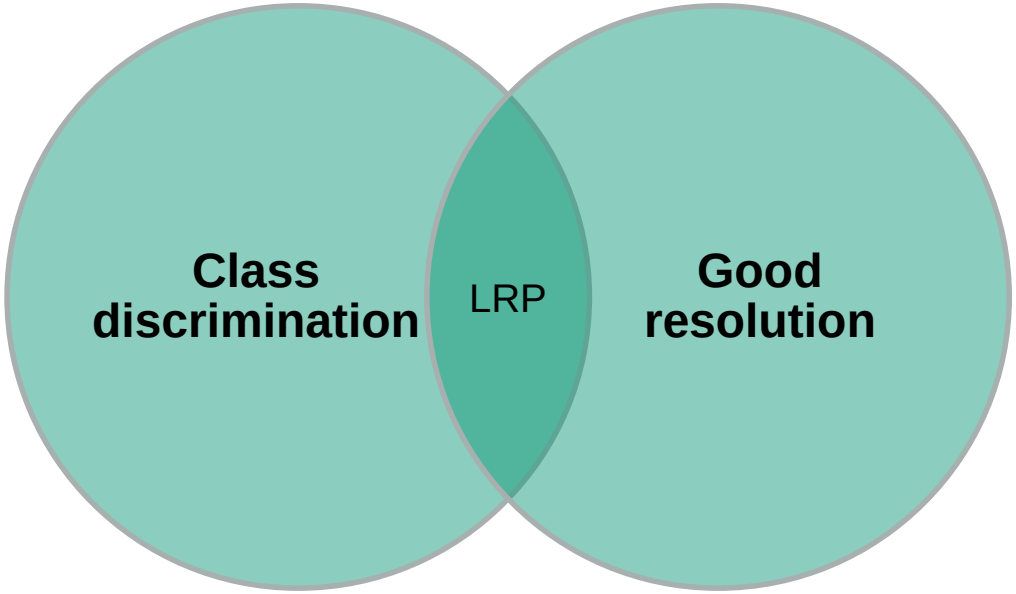
"The mapping of an abstract concept into a domain that the human can make sense of."

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Explanation

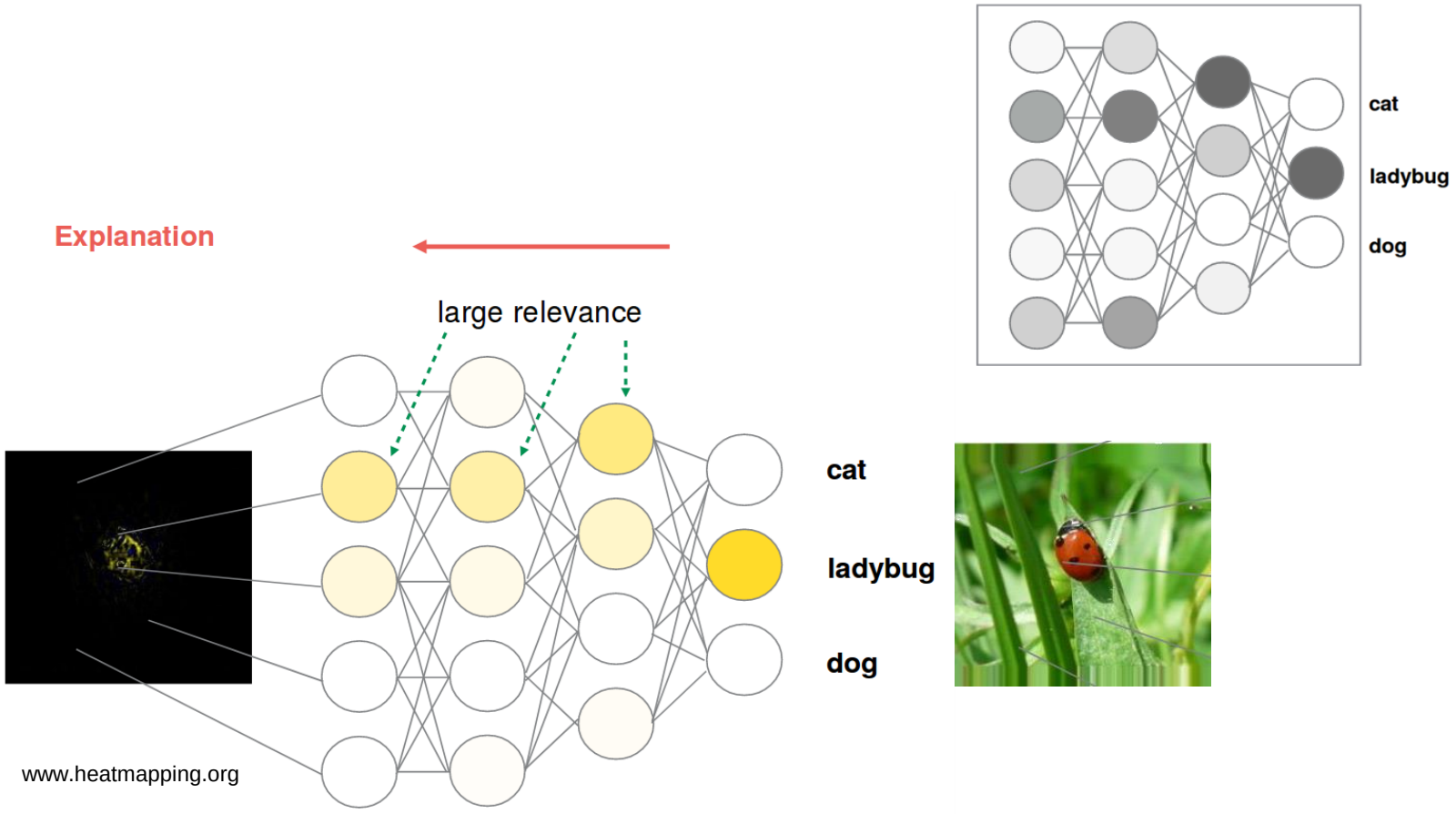
"The collection of features of the interpretable domain, that have contributed for a given example to produce a decision."

Montavon et al. (2017)



Cognitive Systems Group
University of Bamberg · www.uni-bamberg.de/cogsys

XAI: Layerwise Relevance Propagation (LRP)

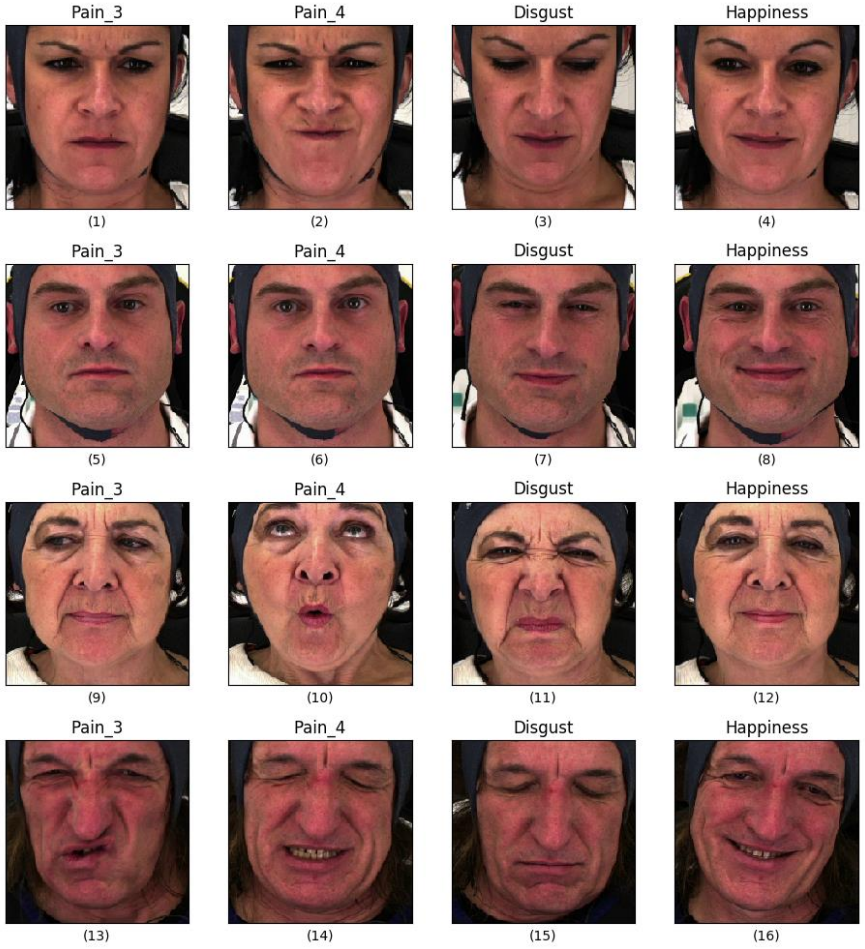


Variations: ε, α, β, preset, flat-rule

Bach et al. (2015), Bach et al. (2016), Kohlbrenner (2017), Lapuschkin et al. (2017)

Cognitive Systems Group
University of Bamberg · www.uni-bamberg.de/cogsys

Material & Procedure



BioVid Heat Pain Dataset

Walter et al. (2013)

- Part A: Pain, 5.5 sec videos
- Part D: Basic emotions, 1 min videos

Before data cleaning

Part	Name	Subjects	Frames
Part A	Pain intensity 3	87	12,006
	Pain intensity 4	87	12,006
Part D	Disgust	76	114,076
	Happiness	75	112,575

After data cleaning

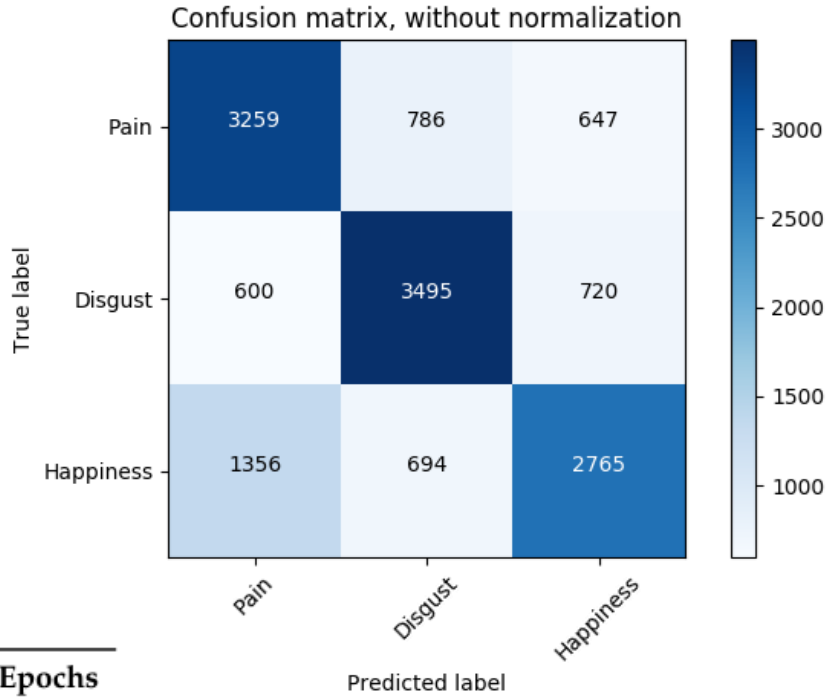
Part	Name	Subjects	Frames
Part A	Pain intensity 3	87	12,006
	Pain intensity 4	87	12,006
Part D	Disgust	75	24,075
	Happiness	75	24,075

CSYS
Cognitive Systems Group
University of Bamberg · www.uni-bamberg.de/cogsys

Results: CNN

Parameters

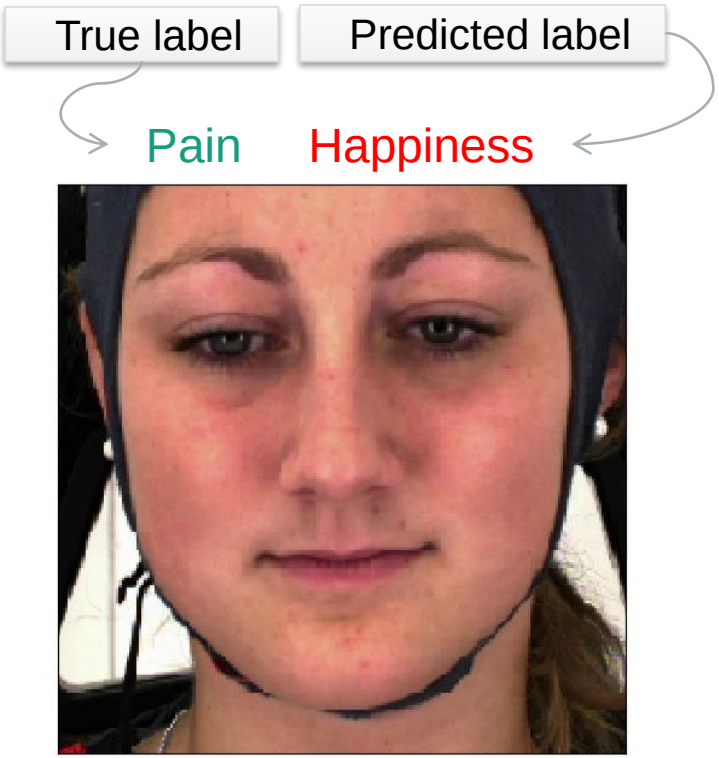
- Rescaling image pixel values between 0 -1
- Optimizer: ADAM, learning rate: 0.00001
- Loss function: Categorical entropy
- Regularization L2, constant: 0.0001
- Early stopping



Fold	Training		Validation		Testing		Epochs
	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy	
1	0.86	0.998	2.08	0.623	2.34	0.537	2
2	0.85	0.998	2.11	0.608	2.17	0.610	2
3	0.85	0.998	1.99	0.637	2.28	0.592	2
4	0.86	0.997	2.18	0.600	2.85	0.562	2
5	0.55	0.999	1.81	0.625	1.58	0.665	4
Average					*2.24*	*0.593*	

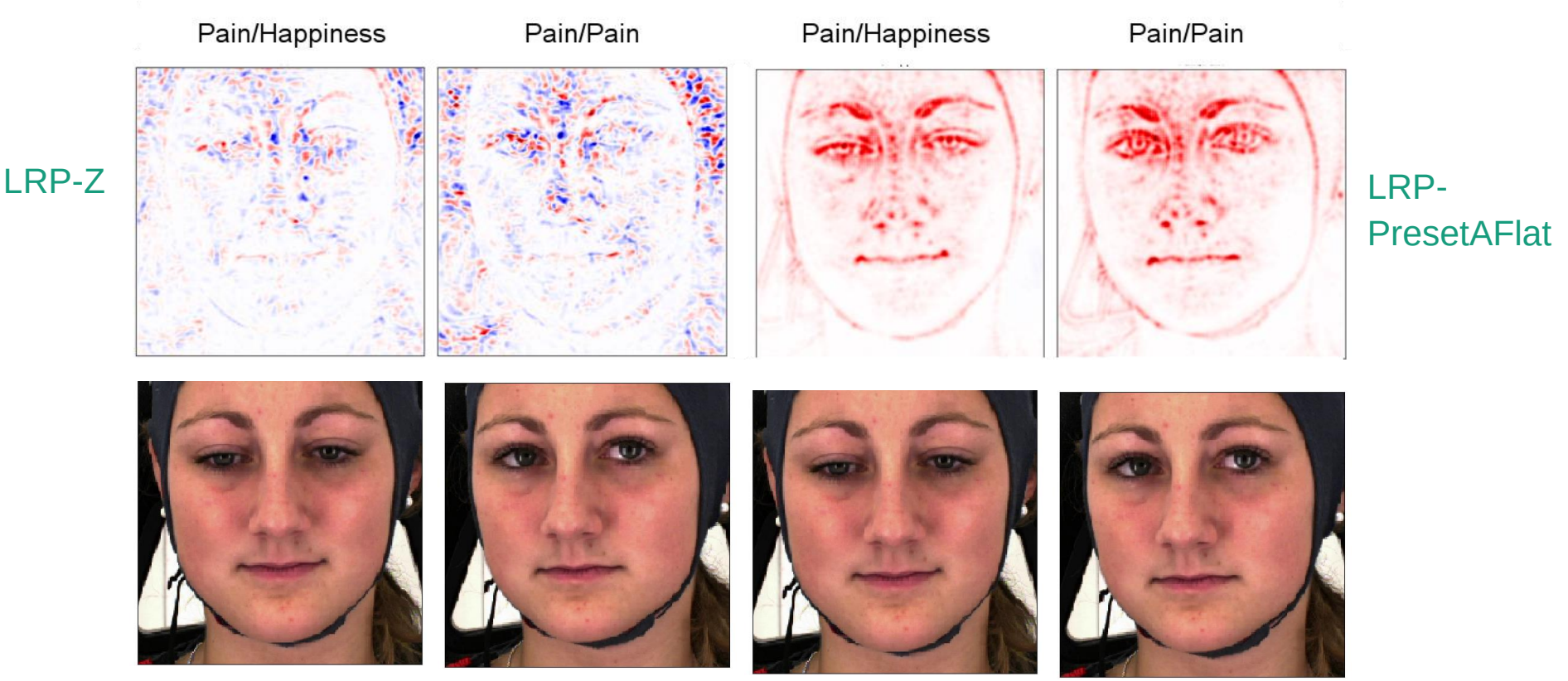
Cognitive Systems Group
University of Bamberg · www.uni-bamberg.de/cogsys

Results: CNN – Best Fold



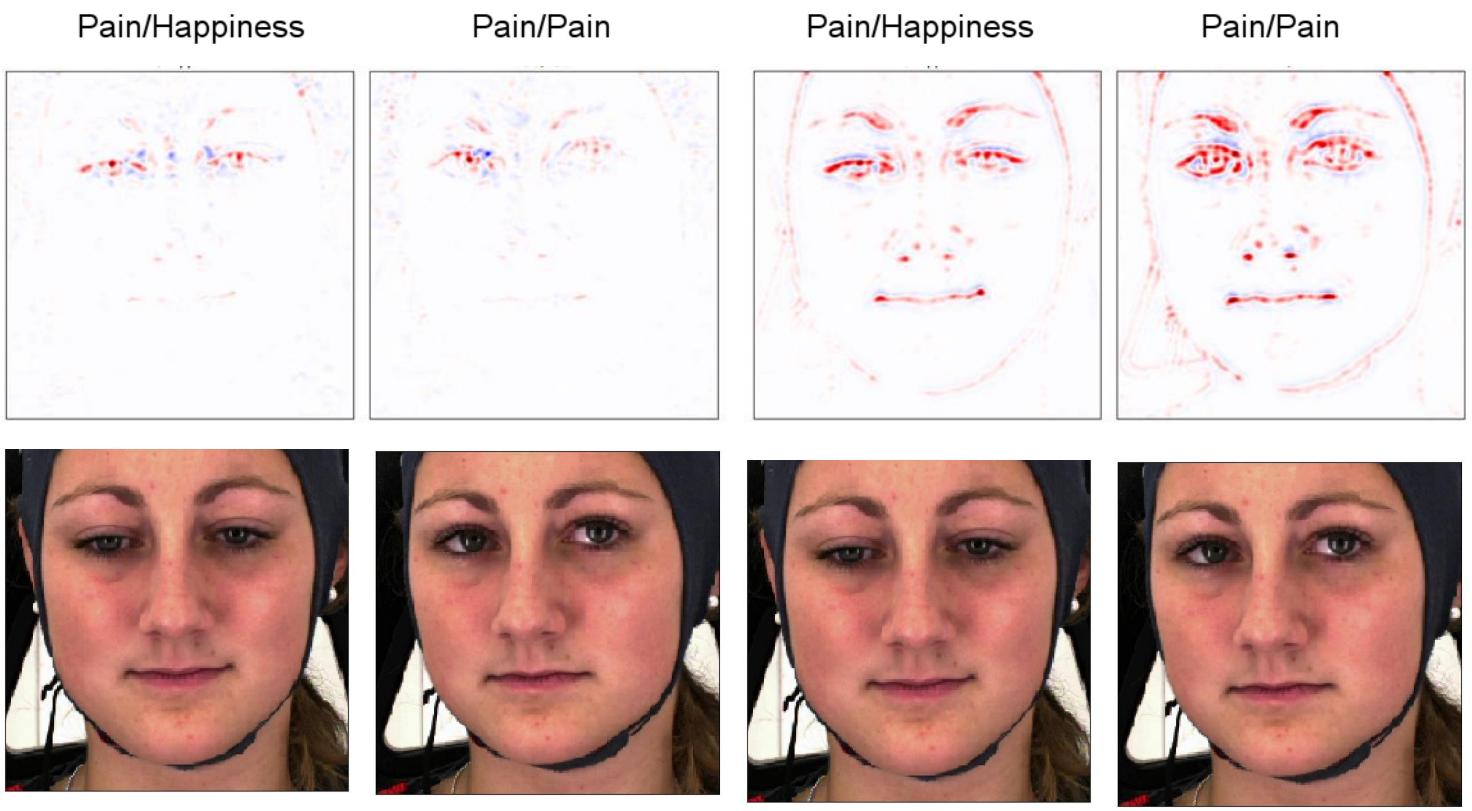
Cognitive Systems Group
University of Bamberg · www.uni-bamberg.de/cogsys

Results: LRP on BioVid (1/2)



Cognitive Systems Group
University of Bamberg · www.uni-bamberg.de/cogsys

Results: LRP on BioVid (2/2)

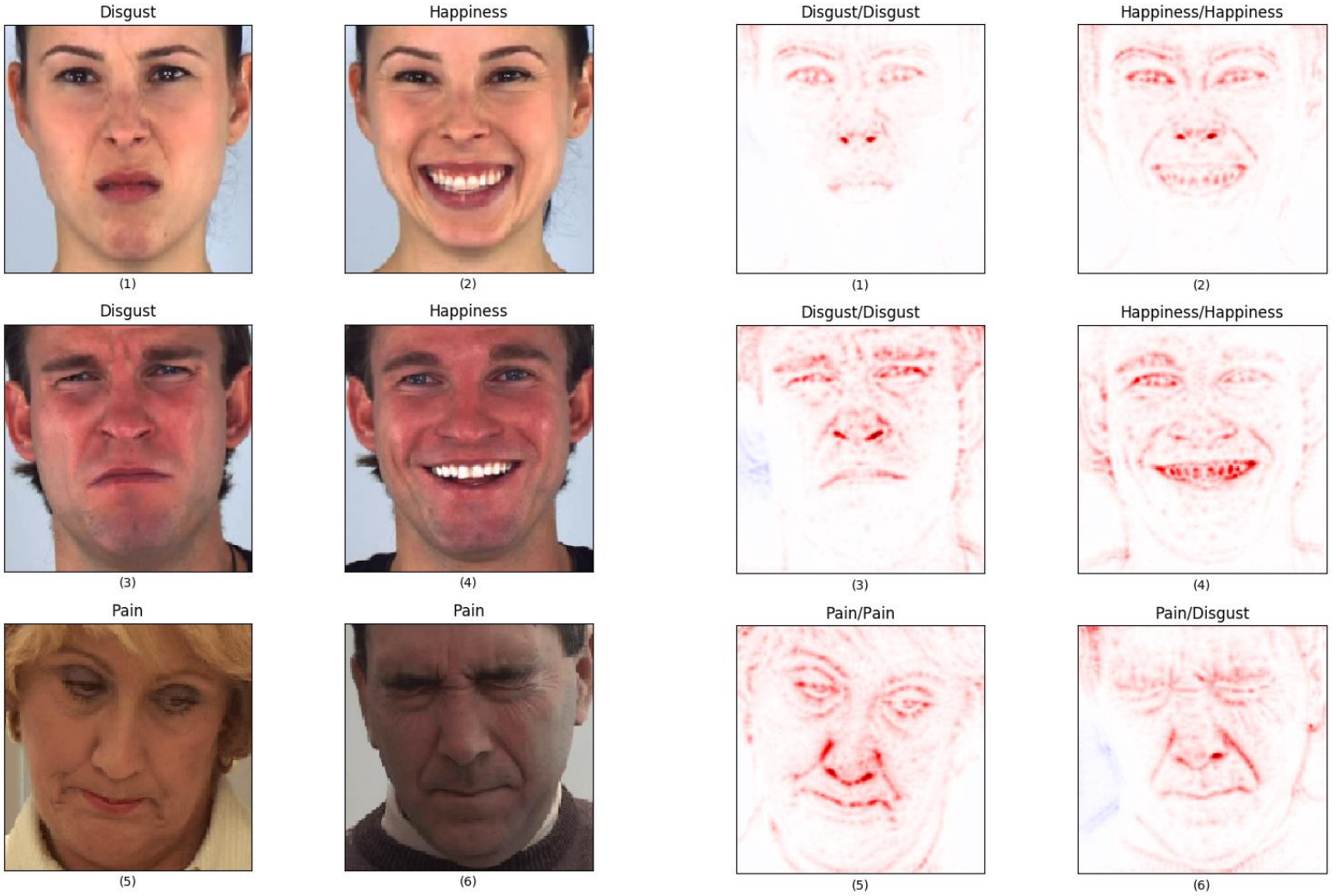


Cognitive Systems Group
University of Bamberg · www.uni-bamberg.de/cogsys

Results: LRP on UNBC & Actorstudy

Lucey et al. (2011), Fraunhofer IIS/Intelligent Systems Group

LRP PresetAFlat



COGSYS
Cognitive Systems Group
University of Bamberg · www.uni-bamberg.de/cogsys

Summary

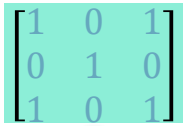
Predictive performance

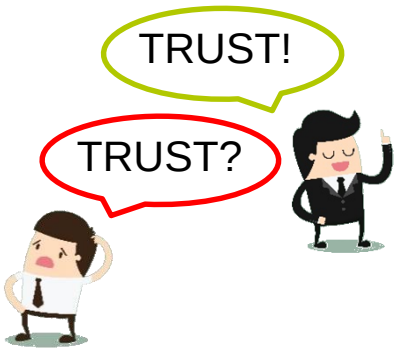
Disgust

Pain

Happiness

Decision interpretation





Feature explanation

Future Work

Linguistic information

"In this image, the eyes are important for the classification of happiness"

Uncertainty information

"Pixel activations for pain have an uncertainty value of 90 out of 100"

References

Pictures

- https://www.destatis.de/DE/Publikationen/Thematisch/Bevoelkerung/VorausberechnungBevoelkerung/BevoelkerungDeutschland2060Presse5124204159004.pdf?__blob=publicationFile
- http://www.heatmapping.org/
- https://www.kdnuggets.com/2016/08/introduction-local-interpretable-model-agnostic-explanations-lime.html
- http://iphome.hhi.de/samek/pdf/GCPR2017.pdf
- www.colourbox.de

References

Literature

- Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., ... & Kindermans, P. J. (2018). iNNvestigate neural networks!. arXiv preprint arXiv:1808.04260
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, *10*(7), e0130140.
- Bach, S., Binder, A., Müller, K. R., & Samek, W. (2016). Controlling explanatory heatmap resolution and semantics via decomposition depth. arXiv preprint arXiv:1603.06463
- Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating visual explanations. In *European Conference on Computer Vision* (pp. 3-19). Springer, Cham.
- Kohlbrenner, M. H. (2017). On the stability of neural network explanations. Bachelor's Thesis.
- Lapuschkin, S., Binder, A., Müller, K.-R., & Samek, W. (2017). Understanding and comparing deep neural networks for age and gender classification. In Proceedings of the international conference on computer vision (pp. 1629–1638)
- Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., & Matthews, I. (2011). Painful data: the unbc-mcmaster shoulder pain expression archive database. In Proceedings of the international conference on automatic face & gesture recognition and workshops (pp. 57–64). IEEE. doi:10.1109/FG.2011.5771462

References

Literature

- Montavon, G., Samek, W., & Müller, K.-R. (2017). Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, 1–15. doi:10.1016/j.dsp.2017.10.011
- Rodriguez, P., Cucurull, G., Gonzàlez, J., Gonfaus, J. M., Nasrollahi, K., Moeslund, T. B., & Roca, F. X. (2017). Deep pain: Exploiting long short-term memory networks for facial expression classification. *IEEE transactions on cybernetics*.
- Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: the all convolutional net. arXiv preprint arXiv:1412.6806, 1–14.
- Walter, S., Gruss, S., Ehleiter, H., Tan, J., Traue, H. C., Werner, P., ... & da Silva, G. M. (2013). The BioVid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In *Cybernetics (CYBCONF), 2013 IEEE International Conference on* (pp. 128-131). IEEE.

GitHub

- Lapuschkin, S., Alber, M., Hägele, M., Schütt, K., & Binder, A. (2018). Lrp. https://github.com/albermax/innvestigate. GitHub.

SourceCode

- Weitz, K. (2018). Explainable AI. https://git01.iis.fhg.de/grp-ils-deeplearning/PainDetection