Inhaltsverzeichnis

Sitzung ABC

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

Katharina Weitz[1,2], Teena Hassan[1], Ute Schmid[2], and Jens Garbas[1]

[1] Fraunhofer IIS, Intelligent Systems Group,
Am Wolfsmantel 33, 91058 Erlangen
[2] University of Bamberg, Cognitive Systems Group,
An der Weberei 5, 96047 Bamberg

Abstract Deep learning networks are successfully used for object and face recognition in images and videos. In order to be able to apply such networks in practice, for example in hospitals as a pain recognition tool, the current procedures are only suitable to a limited extent. The advantage of deep learning methods is that they can learn complex non-linear relationships between raw data and target classes without limiting themselves to a set of hand-crafted features provided by humans. However, the disadvantage is that due to the complexity of these networks, it is not possible to interpret the knowledge that is stored inside the network. It is a black-box learning procedure. Explainable Artificial Intelligence (AI) approaches mitigate this problem by extracting explanations for decisions and representing them in a human-interpretable form. The aim of this paper is to investigate the explainable AI method Layer-wise Relevance Propagation (LRP) and apply it to explain how a deep learning network distinguishes facial expressions of pain from facial expressions of emotions such as happiness and disgust.

Keywords explainable artificial intelligence, deep learning, emotion recognition, pain recognition

1 Introduction

Facial expressions are one of the most important human nonverbal signals in interacting with other people and thus contribute to the emer-

gence and maintenance of social relationships [1]. One of the tasks of facial expressions is to communicate emotions [2, 3]. This is of particular importance when people are unable to express themselves verbally (e. g., because of illness, accidents or congenital disabilities). For this reason, nursing staff in clinics and care facilities are required to observe patients closely in order to be able to read their emotions and take action, if necessary. Due to the already significantly high number of patients, especially in nursing homes, and the prognosis that more and more people will be cared for in such facilities in the future, it would be beneficial to deploy a system to support the nursing staff to monitor a patient's facial expressions and alert them when a pain episode is detected. Additionally, humans often have problems in differentiating between pain and other facial expressions [4, 5]. Therefore, in addition to the (classical) exploration of emotions in a psychological context, research into a technical solution for distinguishing emotions and pain has gained greater importance in the last decade. A system, which uses explainable AI methods to describe how pain differs from other emotions, can be used to train nursing staff to improve their ability to recognise pain correctly.

Towards this goal, in this paper we examine and apply the LRP method [6] to explain the decisions made by a deep Convolutional Neural Network (CNN) that is trained to distinguish facial expressions of pain, happiness, and disgust.

2 Related Work

One of the deep learning architectures that has been successfully applied to image processing applications is CNN [7], which processes images in a hierarchical manner [8]. Compared to approaches based on explicit facial activity descriptors [9], the features of deep learning do not have to be handcrafted. Instead, the system learns the features by itself by projecting information from bitmaps into so-called convolutional layers. They can learn non-linear relationships to model dependencies among the features [10]. One disadvantage is that deep learning approaches require a lot of sample data to extract features [11]. This problem can be reduced by data augmentation methods such as flipping or rotation. The other disadvantage is that, due to its complexity, it is no

longer comprehensible for humans, what the network has learned and what it bases its predictions on. In a practical application, it could be shown that these systems were not going to be accepted, because people do not blindly trust a system which they do not understand [12]. Therefore, techniques that make the black-box learning comprehensible to humans are necessary. One of these techniques for explaining the black-box deep learning is called Layer-wise Relevance Propagation (LRP), which explains the network's decisions by pixel-wise decomposition [6]. In facial image analysis, LRP can be used to explain which pixels were important for the decision of the network. For this, LRP uses the model's prediction strategies in the given pre-processing conditions and visualises the results of this computation as heatmaps [13]. When applying LRP, different parameters can be set to improve the resulting heatmap. In some cases, the resulting relevance scores for each pixel generated by LRP can take on unbounded values [6]. To adjust and stabilize the relevance scores, an ε value can be used. Additionally, α and β values can be applied for stabilization. Besides the stabilizing effect, α and β values can be used to visualize positive and negative activations of pixels [6]. With different values for α and β, the strength of the influence of positive (α) and negative (β) portions can be controlled [6,14]. Besides these parameters, [15] showed that a 'preset' variant of the LRP algorithm achieves optimal results in the calculation of relevance maps. Using the preset approach, the relevance scores for all neurons of the lowest (first) layer are uniformly distributed to the input neuron instead of using the α and β values [13]. To control the resolution of the heatmaps generated by LRP, [16] describes an approach for 'mapping influence cut-off point'. This point describes the moment from which the forward mapping function of the classifier no longer influences relevance propagation, since only the receptive field of the classifier is relevant. The cut-off at this point is called the 'flat' rule. The reference of a receptive field is adapted from neuroscience [17]. In a CNN, the convolutional and pooling layers are inspired by the biological receptive field [11].

3 Research Questions

Answers must be found for the following research questions for automatic pain classification becoming applicable in real-life settings:

- **Predictive performance**: How well can facial expressions of pain be automatically distinguished from those of disgust and happiness using self-learned spatial features?
- **Decision interpretation**: How can the decisions made by the model be presented to people in a comprehensible and transparent way?
- **Feature explanation**: How do the self-learned features differ for the facial expressions of pain and those of disgust and happiness?

This paper[3] would like to provide answers of the questions above. For this, a pre-trained VGG-Face model [18] implemented with the Keras framework was finetuned to distinguish pain from happiness and disgust. For the fine-tuning images of the BioVid dataset[4] [19] were used. Then, the Keras implementation of the LRP approach[5] [20] was used to generate heatmaps at pixel level to illustrate which pixels were relevant for the classification by the VGG-Face model.

4 Material & Procedure

The procedure of this study consisted of the following steps: First, data preparation was done on the BioVid dataset [19]. After that, the VGG-Face architecture was used to fine-tune the CNN for the three-class problem of distinguishing pain from happiness and disgust.

The BioVid dataset [19] was used for finetuning the VGG-Face model. Frames were extracted from the video sequences of pain, happiness, and disgust. For the class pain, part A (pain stimulation without facial EMG) was used. For the class disgust and happiness, frames from

[3] This paper is based on the master's thesis of the first author submitted on August, 31, 2018 to the University of Bamberg. Online link: `https://www.uni-bamberg.de/en/cogsys/research/theses/advised-theses/`

[4] `http://www.iikt.ovgu.de/BioVid.html`

[5] `https://github.com/albermax/innvestigate`

video sequences in part D (posed pain & basic emotions) were used. The video sequences for the pain condition are each 5 seconds long (24,012 frames), the video sequences for the emotions are each 1 minute long (114,076 frames for disgust and 112,575 frames for happiness). The dataset was balanced by manually selecting 3×107 frames from each of the happiness and disgust sequences. One subject in the condition 'disgust' turned away from the camera and talked to the study leaders and showed no disgust expression. This subject was removed from the dataset. In Table 1.1, the amount of frames for each class after the data cleaning steps is provided. The extracted BioVid dataset was then used

Table 1.1: Extracted BioVid data after data cleaning steps.

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

to fine-tune VGG-Face. For the implementation, Tensorflow (version 1.8) and Keras (version 2.2.0) were used. After that, the explainable AI method LRP was applied. For LRP, the Keras implementation from [20] was adapted.

5 Results

The VGG-Face CNN was fine-tuned and tested using 5-fold cross-validation. Here, 4 folds were used for training the model and the remaining fold was used to test the model. The best performing fold had an accuracy of 0.67 on the test data, and was used for generating explanations using the LRP method. In Table 1.2, the class-wise performance of the best fold is presented. When looking at the confusion matrix (see Figure 1.1), it becomes clear that the CNN had problems to to classify happy faces as happy. 28% of the happy images were classified as pain. To take a closer look at this problem, the LRP approach was used. It was used to get an insight of the pixel-related areas of the image which were important for decisions of pain and happiness. To gain this insight,

Table 1.2: Class-wise results of the best performing fold.

	Precision	Recall	F1-score	#Images
Pain	0.62	0.69	0.66	4692
Disgust	0.70	0.73	0.71	4815
Happiness	0.67	0.57	0.62	4815
Average/Total	0.67	0.66	0.66	14322

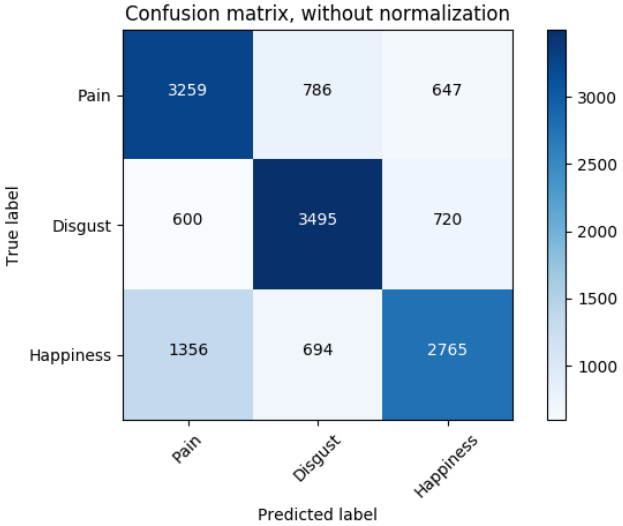

Figure 1.1: Confusion matrix without normalization for the best fold of the 5-fold cross-validation.

two test images from the pain category were selected from the BioVid test fold. In Figure 1.2, the first image displays the subject experiencing pain intensity 3, and in the second image the same person experiencing pain intensity 4. The first label above the image refers to the true class, and the second label to the predicted class.

In Figure 1.3, the visualization generated using LRP with different parameters are shown. In LRP-Z, the basic LRP approach without sta-

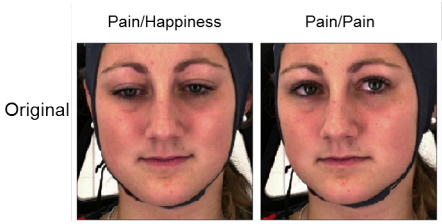

Figure 1.2: Original image from the test fold of the BioVid dataset. First label indicates the true class, second label indicates the predicted class.

bilizers is applied. Here, a face is roughly recognizable. The noise due to the absence of stabilizers is present. In LRP-Epsilon, an ε value of 1e-7 is used. In comparison to the basic LRP approach (LRP-Z), much less noise is represented here. For the LRP-PresetAFlat, an α value of 1, a β value of 0, and an ε value of 1e-1 are used. The flat rule is also applied. For LRP-PresetBFlat, an α value of 2, a β value of 1, and an ε value of 1e-1 are used. Again, the flat rule was applied. In both visualizations, red pixel indicate a positive contribution to the predicted class, and blue pixel indicate a negative contribution. In comparison to the LRP-Z heatmap, the visualization of preset-flat variants are much more detailed and clearer. In the two preset variants, it can be observed that highly positive pixel values are more important for the decision of the CNN. It becomes apparent that with the increase of the α value (LRP-PresetAFlat), the positive pixel values become more prominent. With the increase of the β value (LRP-PresetBFlat), the negative pixel values are highlighted more clearly. When looking at the LRP PresetAFlat visualization, it can be seen that mostly the same areas in the face, namely the eyes, the nose and the mouth contribute to the classification of happiness and pain. This could be an indication why the accuracy of the CNN is not very high. When looking at the LRP PresetBFlat, slight differences in the contribution of negative pixels for the classification are visible. For pain, more negative pixels around the nostrils and on the lower side of the eyebrows are detectable on the heatmaps.

Besides test images from the BioVid dataset, images from the UNBC-McMaster shoulder pain expression archive database [21] for pain were used for visualization using LRP methods (see left part of Figure

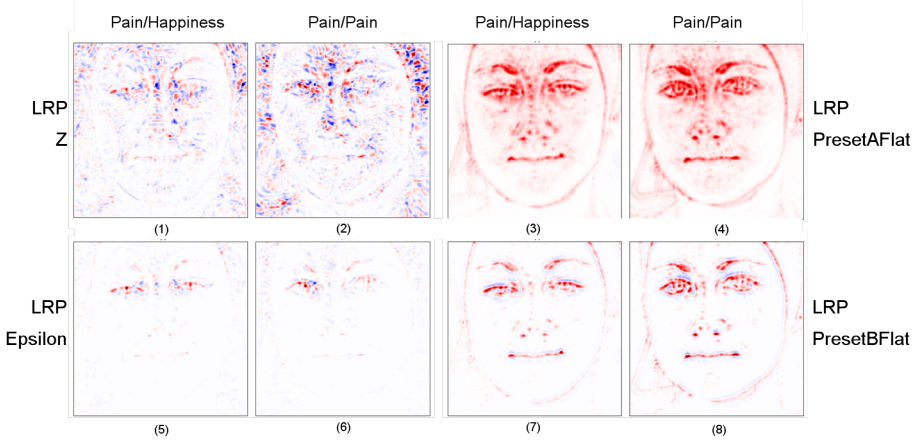

Figure 1.3: Visualizations for applying LRP method with different parameters on two pain images. First label indicates the true class, second label indicates the predicted class. The heatmap generated with the basic LRP approach (LRP-Z) is displayed in subimages (1) and (2). Subimages (3) and (4) display the heatmap generated with the LRP-PresetAFlat Variant. Subimages (5) and (6) display the heatmap generated with an ε stabilizer. Subimages (7) and (8) are the results of applying LRP-PresetBFlat. The visualizations display the predicted class.

1.4). For happiness and disgust, images from the Actorstudy dataset[6] were used. The visualizations using the LRP-PresetAFlat approach are shown on the right part in Figure 1.4. Here it can be seen that for happiness, the eyes and the mouth are important areas for the classification. For disgust, the focus lays on the nose and the eyes. This could be a reason that the pain image (subimage 6) was misclassified as disgust. For pain, the nostrils seems to be important.

6 Discussion

For the topic of **predictive performance**, this paper shows that the CNN could distinguish images of pain, disgust, and happiness only with an accuracy of 67%. Above all, happy faces were often misclassified as

[6] Unpublished dataset from Intelligent Systems Group, Fraunhofer IIS, Erlangen

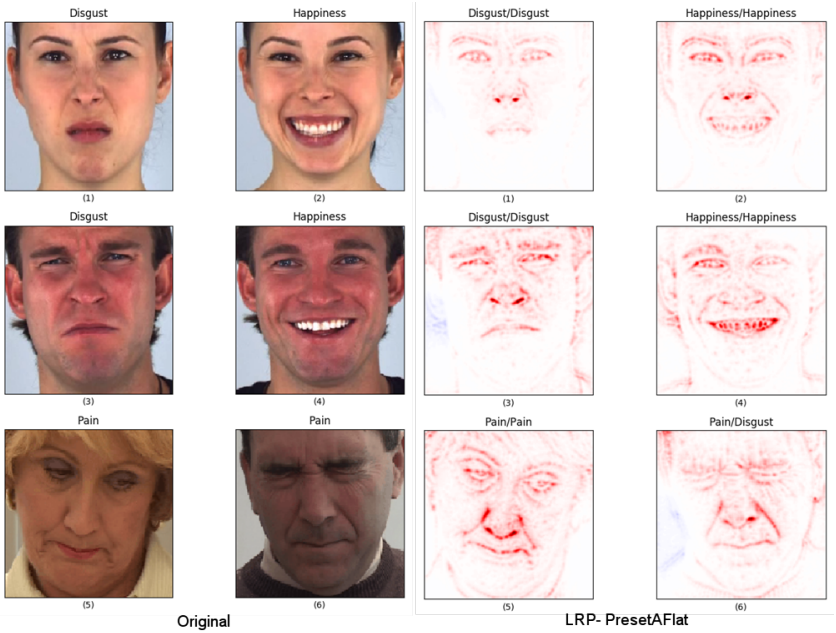

Figure 1.4: Left: Input images 1-4 from Actorstudy dataset (©Fraunhofer IIS) and images 5 & 6 from the UNBC-McMaster shoulder pain expression archive database (©Jeffrey Cohn) to visualize LRP approach. Right: Visualizations for applying LRP PresetAFlat method. The visualizations display the predicted class. First label refers to the true class, second label refers to the predicted class.

faces of pain.

For the part of **decision interpretation**, LRP is a helpful tool to generate a fine-granular heatmap of relevant pixels. The usage of LRP with its various parameters allows a wide range of adjustments. The results presented here for the categorization of pain, disgust, and happiness represent only an initial step into the research of making decisions of black-box systems comprehensible for humans. Lapuschkin et al. [13] already investigated the application of LRP for the recognition of age and gender from images of faces. They could show that the visualizations of relevant pixels allow an interpretation of the relevant facial ar-

eas to classify age and gender. However, when looking at facial expressions of happiness, pain, and disgust it becomes clear that pixel activation alone cannot yet provide a clear difference between the predicted classes for the human eye. Therefore, for the topic of **feature explanation**, the relevant features for the classification are not easy detectable by humans. This is partly due to the classification capabilities of the VGG-Face CNN. The classification accuracy of 67% must be taken into account when looking at the visualizations.

Montavon et al. [14] describe some practical recommendations to improve the visualizations generated by the LRP method: using dropout as regularization technique, preferring sum pooling, instead of max pooling and not to use too many fully connected layers in the network (whereas no definition is given for what is meant by 'many'). Nevertheless, additional information is needed for a clearer interpretation of the results. Future research approaches may focus on the implementation of such additional information sources. Additional sources of information could, for example, take the form of linguistic information (e.g., 'In this image, the eyes are important for the classification of happiness') or the form of uncertainty formulations (e.g., pixel activations for happiness have an uncertainty value of 20 out of 100, while pixel activations for pain have an uncertainty value of 90 out of 100). Only when a result that is informative and interpretable for humans is achieved, a comprehensive application in real-life can be considered.

References

1. C. Frith, "Role of facial expressions in social interactions," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3453–3458, 2009.
2. N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychological Bulletin*, vol. 111, no. 2, pp. 256–274, 1992.
3. P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.
4. H. Aviezer, Y. Trope, and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," *Science*, vol. 338, no. 6111, pp. 1225–1229, 2012.

5. S. Brahnam, C.-F. Chuang, F. Y. Shih, and M. R. Slack, "Machine recognition and representation of neonatal facial displays of acute pain," *Artificial Intelligence in Medicine*, vol. 36, no. 3, pp. 211–222, 2006.
6. S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.
7. Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems*, 1990, pp. 396–404.
8. C. Ravat and S. A. Solanki, "Survey on different methods to improve accuracy of the facial expression recognition using artificial neural networks," in *Proceedings of the National Conference on Advanced Research Trends in Information and Computing Technologies*, vol. 4, no. 2, 2018.
9. P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, and H. C. Traue, "Automatic pain assessment with facial activity descriptors," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 286–299, 2017.
10. J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.
11. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
12. H. C. Lane, M. G. Core, M. Van Lent, S. Solomon, and D. Gomboc, "Explainable artificial intelligence for training and tutoring," University of Southern California Marina del Rey CA Institute for Creative Technologies, Tech. Rep., 2005.
13. S. Lapuschkin, A. Binder, K.-R. Müller, and W. Samek, "Understanding and comparing deep neural networks for age and gender classification," in *Proceedings of the International Conference on Computer Vision*, 2017, pp. 1629–1638.
14. G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2017.
15. M. H. Kohlbrenner, "On the stability of neural network explanations," Apr 2017, bachelor's Thesis.
16. S. Bach, A. Binder, K.-R. Müller, and W. Samek, "Controlling explanatory heatmap resolution and semantics via decomposition depth," in *Proceedings of the International Conference on Image Processing*. IEEE, 2016, pp. 2271–2275.

17. D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex." *Journal of Physiology*, vol. 160, no. 1, pp. 106–154, 1962.

18. O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, vol. 1, no. 3, 2015, pp. 1–12.

19. S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. C. Traue, P. Werner, A. Al-Hamadi, S. Crawcour, A. O. Andrade, and G. M. da Silva, "The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system," in *Proceedings of the International Conference on Cybernetics*. IEEE, 2013, pp. 128–131.

20. M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans, "innvestigate neural networks!" *arXiv preprint arXiv:1808.04260*, 2018.

21. P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database," in *Proceedings of the International Conference on Automatic Face & Gesture Recognition and Workshops*. IEEE, 2011, pp. 57–64.