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

Katharina Weitz, Teena Hassan, Ute Schmid & Jens Garbas
Motivation

Aviezer et al. (2012)

Brahnam et al. (2007)
Motivation

www.destatis.de
Research Questions

**Predictive performance**

- Disgust
- Pain
- Happiness

**Decision interpretation**

```
1 0 1
0 1 0
1 0 1
```

TRUST!

TRUST?

**Feature explanation**

Disgust/Disgust
Research Questions

**Deep Learning**

- Image size: 224x224
- Image
  - conv-64
  - conv-64
- Image size: 112x112
  - max pooling
  - conv-128
  - conv-128
- Image size: 56x56
  - max pooling
  - conv-256
  - conv-256
- Image size: 28x28
  - max pooling
  - conv-512
  - conv-512
  - conv-512
- Image size: 14x14
  - max pooling
  - conv-512
  - conv-512
  - conv-512
- Image size: 7x7
  - max pooling
  - fc-4096
  - fc-4096
  - fc-2622
  - softmax
  - Parkhi et al. (2016)

**VGG-Face**

**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)
XAI: Idea

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

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

Montavon et al. (2017)

Class discrimination

Good resolution

LRP
XAI: Layerwise Relevance Propagation (LRP)

Variations: $\varepsilon$, $\alpha$, $\beta$, preset, flat-rule

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

<table>
<thead>
<tr>
<th>Part</th>
<th>Name</th>
<th>Subjects</th>
<th>Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part A</td>
<td>Pain intensity 3</td>
<td>87</td>
<td>12,006</td>
</tr>
<tr>
<td></td>
<td>Pain intensity 4</td>
<td>87</td>
<td>12,006</td>
</tr>
<tr>
<td>Part D</td>
<td>Disgust</td>
<td>76</td>
<td>114,076</td>
</tr>
<tr>
<td></td>
<td>Happiness</td>
<td>75</td>
<td>112,575</td>
</tr>
</tbody>
</table>

### After data cleaning

<table>
<thead>
<tr>
<th>Part</th>
<th>Name</th>
<th>Subjects</th>
<th>Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part A</td>
<td>Pain intensity 3</td>
<td>87</td>
<td>12,006</td>
</tr>
<tr>
<td></td>
<td>Pain intensity 4</td>
<td>87</td>
<td>12,006</td>
</tr>
<tr>
<td>Part D</td>
<td>Disgust</td>
<td>75</td>
<td>24,075</td>
</tr>
<tr>
<td></td>
<td>Happiness</td>
<td>75</td>
<td>24,075</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Fold</th>
<th>Training</th>
<th>Validation</th>
<th>Testing</th>
<th>Epochs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Loss</td>
<td>Accuracy</td>
<td>Loss</td>
<td>Accuracy</td>
</tr>
<tr>
<td>1</td>
<td>0.86</td>
<td>0.998</td>
<td>2.08</td>
<td>0.623</td>
</tr>
<tr>
<td>2</td>
<td>0.85</td>
<td>0.998</td>
<td>2.11</td>
<td>0.608</td>
</tr>
<tr>
<td>3</td>
<td>0.85</td>
<td>0.998</td>
<td>1.99</td>
<td>0.637</td>
</tr>
<tr>
<td>4</td>
<td>0.86</td>
<td>0.997</td>
<td>2.18</td>
<td>0.600</td>
</tr>
<tr>
<td>5</td>
<td>0.55</td>
<td>0.999</td>
<td>1.81</td>
<td>0.625</td>
</tr>
</tbody>
</table>

Average  2.24  0.593
Results: CNN – Best Fold

- True label: Pain
- Predicted label: Happiness

- True label: Pain
- Predicted label: Pain
Results: LRP on BioVid (1/2)
Results: LRP on BioVid (2/2)

LRP- Epsilon

LRP- PresetBFlat
Results: LRP on UNBC & Actorstudy

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

LRP PresetAFlat

Disgust  Happiness  Disgust/Disgust  Happiness/Happiness

(1)  (2)  (1)  (2)

Disgust  Happiness  Disgust/Disgust  Happiness/Happiness

(3)  (4)  (3)  (4)

Pain  Pain  Pain/Pain  Pain/Disgust

(5)  (6)  (5)  (6)
Summary

Predictive performance

Disgust

Pain

Happiness

Feature explanation

Disgust/Disgust

Decision interpretation

TRUST?!

TRUST?

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/
- www.colourbox.de
References

Literature


References

Literature


GitHub


SourceCode