An Interpretable Framework for Pain Evaluation

Run Wang[1], **Zifan Jiang**[2], **Gari Clifford**[2]
[1] Fudan University [2] Emory & Gatech

Introduction

Pain is defined as a distressing experience complicated with tissue damage and cognitive suffering. A valid and reliable pain assessment is necessary for choosing adequate treatment. However, the predominant pain assessment guideline, self-report, not only is not objective but also fails for patients without cognitive ability. Furthermore, prior studies on pain classification fail to achieve reliable accuracy and lack explainability for clinicians to trust. Explainable Machine Learning has the potential to fix these problems. Our work includes two aspects of interpretable pain monitoring. The first aspect is for physiological signal: we explored which part of the signal contributes to the network's final prediction and suspected the existence of the end effect, i.e. the end part of the signal accounts for much more than the average. On the other hand, for video signals, we utilized surveys on Amazon MTurk to evaluate the utility of the proposed interpretable framework on video signals. More specifically, we evaluated the consistency difference for every video's mean confidence and mean accuracy and how much improvement was observed on the human performance with the assistance of the interpretable framework.

Data

Biovid Heat Pain Database

The Biovid database (*1*) was used for following experiments. Five different levels of heat-induced pain was applied on 87 subjects for a total of 100 times, with 20 times for each level of pain. It includes video signals and physiological signals, where the physiological signals include galvanic skin response (GSR), electrocardiogram (ECG) and electromyography (EMG), each cut into 5.5s. To avoid the influences from the EMG sensors presented in the other parts of the Biovid, only part A was utilized in the analysis. Besides, based on the findings of a previous study in (*2*), only GSR signals were sufficient to yield reliable accuracy thus only GSR signal was used for following experiment. For video signals, based on the suggestion from (*3*), 45 subjects were excluded due to their low level of visual responses to pain. The criterion to select subjects was based on the strength and variance of the calcuated facial feature scores.

Amazon Mturk Participants

Participants from Amazon Mturk were recruited for video experiments. A qualification task to classify 10 videos was implemented in advance to screen qualified workers. 15 workers with high performance in qualification test were recruited for video experiment 1. 16 workers with high performance in qualification test were recruited for video experiment 2 and another 16 workers with low performance in qualification test but great history record (history pass rate larger than 95%) was recruited for video experiment 2. All recruited workers have a great history record.

Method

Data Processing

Though the experiment of physiological signals and video signals were conducted separately, they share the same classification neural network structure in (*2*). The details about the network are omitted in this poster for compactness. Classification network's performances were estimated with leave-one-subject-out cross-validation.

Physiological signals

Three processing steps were applied to physiological signal: filtering, classification, and interpretation. A third-order low-pass Butterworth filter with a cut-off frequency of 3 Hz was applied on the GSR signals. After the filter, the signals were averaged to 0. Based on the results in article(*2*), the classification result is comparable to the state-of-art using only GSR, so only GSR was used. To point out, only the highest pain level data (PA4) and no pain level (BL1) data were used, as data with other levels of pain is too difficult to be classified accurately. All 87 subjects' data was included.

Video

Fig.1. shows the flow of data for videos. Videos were first passed through Openface (*4*) to obtain 38 AU features and some eye related features(Openface feature 1-100). Then, those features were passed through a pain classification network. Finally those features calculated at the first step were assigned with significance scores by interpreters. A subset of subjects were excluded in this research based on the sum of variance of all AU's intensity and presence, which measure the level of pain visible from the videos.

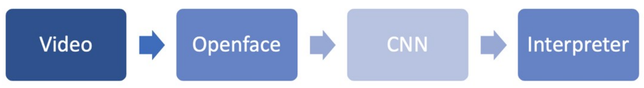

Fig.1. Data Flow for Video Signal

Interpretation Method

The interpreter and the evaluation of its effectiveness are the focus of this work and will be presented in this section.

Interpreter

Interpretation methods explain the prediction of a model for a input data sample. For a given target, contribution values to this target are assigned to each input feature, explaining how the model evaluates the input data regarding the target. For the sake of computation cost, **DeepLift**(*5*), **LRP**(*6*), **Saliency**(*7*) and **IG**(*5*) were adopted in this work as interpreter. **Occlusion**(*8*) was used as the baseline for interpreters. The sliding window shape and the stride are 100. The target for all interpreters is pain.

Interpretation Method Evaluation

Physiological Signal Evaluation

During the initial exploration for physiological signal interpretation, we found that interpreters paid more attention to the end of the signal.To investigate this phenomenon, we truncated the physiological signal by 100 points each time from original length, 2816 points, to 2000 points and repeat the procedure of classification and interpretation. The ratio of the average of last 100 points' significance value and the average of all points' significance value were calculated for each truncation.

Video Evaluation: Interpreter Comparison

Fig. 2. Survey Questions

15 workers with high qualification test performance were invited to take this survey. They were shown 42 videos, each of which were selected randomly from a subject's PA4 video collection. They were asked to answer which time ranges and face regions affected their decision (multi-selection). The example survey question is exhibited in Fig.2.
The significance values given by interpreters were compared with results from Amazon Mturk survey.

Video Evaluation: Human Performance Change

Human performances change was designed to analyze if human can make better classification about pain video given interpreter hint.

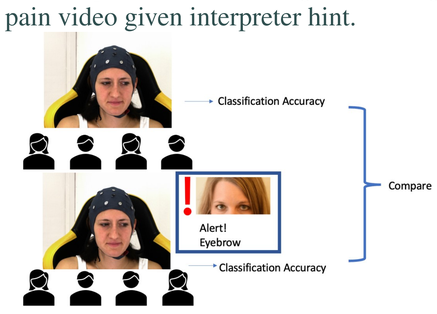

Fig. 3

16 workers with high performance and another 16 workers with low performance in qualification test were recruited. Each of these two group was separated into group A and group B. For group A, videos with odd index were shown with hints from interpreters, while hints were given for videos with even indices for group B.
The interpreter hint is a one second video whose average interpreter score was the highest among all the one second snippets within the video. And the video zooms in to the face region with highest significance score during that time range. During the experiment, videos with and without hint were shown alternatively. Workers are asked to **classify** the video first and use two sliders ranging from 1 to 10 to evaluate their **confidence** towards their decision and the **pain level** in the video.

Result

Baseline Network

Physiologicl Signal	84.3%
Video Signal	82.1%

Table 1: Baseline Network Result

For physiological signal, the classifier achieved an accuracy of 84.3%, which is similar to the results from (*2*). For video signal, the accuracy is 82.1% within the 42 subjects included in the analysis.

Physiological Signal

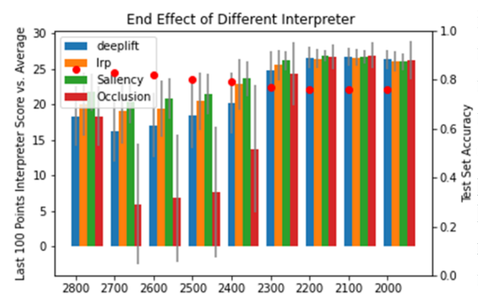

Fig. 4. End Effect of the Interpretation

Fig.4. shows the end part of GSR signals has higher significance score given by four different interpreters. In Fig.4., the x-axis is the length of the truncated signal ranging from the original length to 2000. The left y-axis is the ratio between the average of the last 100 points and the average of all points. As the figure shows, no matter which interpreter, no matter the lengh of the signal, the ratio is much higher than one. The right y-axis shows the LOSO accuracy for that truncated signals which drops gradually by tiny margin.

Video Signal

Exp1. Interpreter Comparison

Interpreter	MSE	SNR
Saliency	0.169	9.024
Deeplift	0.370	1.070
LRP	0.228	6.324
IG	0.220	6.801

Table 2: Interpreter Comparison

Suppose Human selection v.s. interpreter score are regarded as signal before communication channel and after it. And suppose the sum of the workers results for time range selection is regarded as original signal and the sum of interpreter's score for extant AU features as signal with noise. Therefore, the MSE and SNR, which are used to evaluate the discrepancy between original signal and signal with noise, are adopted as evaluation metric for the reliability of interpreter. The average of MSE and SNR for all videos are listed in table 2. Saliency has the lowest MSE and highest SNR, which indicates the time range selected by Saliency has the highest congruence with human's decision. Therefore, Saliency is adopted for the following experiment 2.

Exp2. Human Performance Change

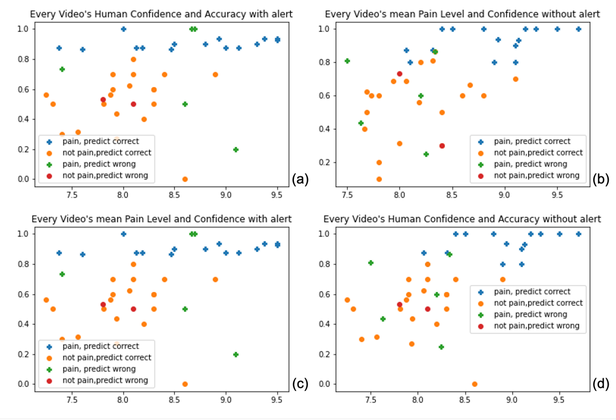

Fig. 5. Video Exp.2

In Experiment 2, we did not observe a change in the overall accuracy of the workers when given the interpreter scores. Specifically, when given the interpreter score, people were more inclined to suspect that the video was "pain" because the mechanism of the interpreter is to pick out the features that influence the algorithm to make a "pain" judgment. This tendency then led to more "pain" videos being selected correctly, and equally "non-pain" videos being selected incorrectly more often, resulting in little change in overall accuracy.
However, we found out the difference in consistency between every video's mean confidence v.s. mean accuracy and every video's mean pain level v.s. mean accuracy. As shown in Fig.5., for (a) and (c)(given interpreter), points are more consistent than points in (b) and (d)(without interpreter). This phenomenon is more obvious within groups with different qualification performance.

Discussion

There are three main works in the future. Firstly we will explore the reasons for the interpreter's end effect in physiological signals. Secondly, we will add one more experiment in video experiment 1. Finally, we will explore and explain the reason for the consistency difference in the video experiment 2.

References

1. S. Walter *et al.*, en, presented at the 2013 IEEE International Conference on Cybernetics (CYBCO), pp. 128–131, ISBN: 978-1-4673-6469-0.
2. P. Thiam, P. Bellmann, H. A. Kestler, F. Schwenker, *Sensors* **19**, 4503 (2019).
3. P. Werner, A. Al-Hamadi, S. Walter, presented at the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp. 176–180.
4. T. Baltrusaitis, A. Zadeh, Y. C. Lim, L.-P. Morency, presented at the 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), pp. 59–66.
5. A. Shrikumar, P. Greenside, A. Kundaje, presented at the International Conference on Machine Learning, pp. 3145–3153.
6. S. Bach *et al.*, *PloS one* **10**, e0130140 (2015).
7. K. Simonyan, A. Vedaldi, A. Zisserman, *arXiv preprint arXiv:1312.6034* (2013).
8. M. D. Zeiler, R. Fergus, presented at the European conference on computer vision, pp. 818–833.