Deep learning Model for Detection of Pain Intensity from Facial Expression

Jeffrey Soar[1], Ghazal Bargshady[1], Xujuan Zhou[1], and Frank Whittaker[2],

[1] University of Southern Queensland, Queensland, Australia
[2] Nexus eCare, Melbourne, Australia

Abstract. Many people who are suffering from a chronic pain face periods of acute pain and resulting problems during their illness and adequate reporting of symptoms is necessary for treatment. Some patients have difficulties in adequately alerting caregivers to their pain or describing the intensity which can impact on effective treatment. Pain and its intensity can be noticeable in ones face. Movements in facial muscles can depict ones current emotional state. Machine learning algorithms can detect pain intensity from facial expressions. The algorithm can extract and classify facial expression of pain among patients. In this paper, we propose a new deep learning model for detection of pain intensity from facial expressions. This automatic pain detection system may help clinicians to detect pain and its intensity in patients and by doing this healthcare organizations may have access to more complete and more regular information of patients regarding their pain.

1 Introduction

Painful conditions are one of the most common reasons that patients seek health care [1]; some patients may have difficulty providing accurate recall of pain or may be unwilling to disclose their pain. Assessment can be invasive and incon- venient [2].

The needs of the large number of people in chronic pain overwhelms health- care systems, and continues to worsen [3]. Healthcare providers may be less able to help patients who have chronic pain if there is not a complete understanding of symptoms and some patients such as the very young, or those with partic- ular disabilities or conditions, may not be able to articulate their experiences. Automatic pain management systems have been demonstrated to better help clinicians to detect the level of pain of patients [2].

Faces have evolved to convey rich information for social interaction, includ- ing the expression of emotions and pain [4]. Researchers have applied machine learning to the task of automatic pain detection in a real-world clinical setting involving patients suffering pain [5].

In this study, we proposed a new deep learning model for detection of pain intensity from facial expressions. This automatic pain detection model can be

used as a smart technology to detect pain level and can help patients and their careers to monitor and manage chronic pain.

2 Related Work

2.1 Deep Learning

In the last few years, deep neural networks have become the classifier of choice for many machine learning tasks. Simply put, deep neural networks (DNNs) are a group of models that perform nonlinear function approximation. Suppose there is a function f that relates input x to some label y, a neural network learns a function f ($y = f(x; \theta)$) that approximates f, the true mapping from x to y, using parameters θ [6]. A deep neural network is usually represented as the composition of multiple nonlinear functions. Therefore, $f(x)$ can be expressed in the following manner:

$$f(x) = f^{(N)}((f^{(3)}(f^{(2)}(f^{(1)}(x))))) \quad (1)$$

Eq. 1 defines a feed-forward network or multilayer perceptron (MLP). It is called feed-forward because the output of each function $f^{(i)}$ is passed as input to the next function $f^{(i+1)}$. Each function $f^{(i)}$ represents a layer in the neural network and is typically composed of an affine operation followed by an element-wise nonlinearity: sigmoid, hyperbolic tangent (tanh), or rectified linear unit (ReLU). Consider some input $x \varepsilon R^N$ and its corresponding output $z \varepsilon R^M$, a simple exam- ple of a neural network layer is given in Eq. 2, where $W \varepsilon R^{MxN}$ and $b \varepsilon R^M$ are the parameters to be learned and h is a user-selected nonlinear function. Each element of x is called an input unit and each element of z is an output unit. The depth of a neural network is defined as the number of layers. A neural network is usually classified as deep if it has a depth of three or more layers [6]:

$$z = h(W_x + b) \quad (2)$$

Convolutional Neural Networks (CNNs) One common modification to MLPs, which makes the network more suitable for tasks specific to computer vision, is to make the affine transformation in each layer a convolution operation. These networks are called Convolutional Neural Networks (CNNs) [7]. In an MLP, each output unit in z is dependent on all input units in x. This design, however, does not allow the network to model the local structures commonly found in images. With CNNs, each output unit in z is instead connected to a reduced number of input units, specifically a small contiguous region in the input. The reduction in the number of connections also means that CNNs have considerably fewer parameters to learn than MLPs [7].

Recurrent Neural Networks (RNNs) Despite being powerful tools for func- tion approximation, MLPs and CNNs have one main drawback. They are both have difficulty modelling sequential data. This can be

particularly problematic when dealing with speech or video data. In these instances, it would be advan- tageous to have a model whose feature representation can capture information from all the previous time steps but can also update its representation with what it sees in the future. Recurrent Neural Networks (RNNs) [8] present one way to create such a model. Consider a sequence of inputs of length

$T(x_1, \ldots, x_t)$. RNNs

have a state at each time point t, h_t, which captures all information of previous inputs $(x_1, \ldots, x_t)$. Then, when considering the input at the next time point, x_{t+1}, a new state, h_{t+1}, is computed using the new input x_{t+1} and the previous state h_t. At each time point t, the hidden state h_t can be used to compute an output O_t, typically a class label for classification or a continuous number for regression [9].

2.2 CNN Models for Facial Expression Analysis

Deep models such as deep belief and deep convolutional networks have allowed us to have an insight into the effect on extracting robust and abstract features [10] and some deep models are used for facial expression [11]. Susskind et al. [11] learned deep belief nets without supervision for recognizing facial action units and they showed features extracted by learned belief nets could easily accom- modate different constraints in a real expressive environment. Differently, Xu et al. [12] used transfer features from deep convolutional networks to recognize fa- cial expression, and they showed deep convolutional networks are more suitable for classification than deep belief nets. They used a facial expression recognition model on transfer features from deep convolutional networks (ConvNets) for face identification.

Liu et al. [13] introduced an AU (Action Unit) aware receptive field layer in a deep network, designed to search subsets of the over- complete representa- tion, each of which aims at best simulating the combination of AUs. Its output is then passed through additional layers aimed at the expression classification, showing a large improvement over the traditional hand-crafted image features such as LBPs, SIFT and Gabors. Another example occurs where a CNN is jointly trained for detection and intensity estimation of multiple AUs. The authors pro- posed a network architecture composed of 3 convolutional and 1 max-pooling layers [14]. Zhao, et al [15] introduced an intermediate region layer that can learn region specific weights of CNNs. The region layer returns an importance map for each input image and the network is trained for joint AU detection. All these methods focus solely on feature extraction while the network output remains unstructured. Walecki et al. [16] introduced model structures and esti- mate complex feature representations simultaneously by combining conditional random field (CRF) encoded AU dependencies with deep learning. They de- signed a novel Copula CNN deep learning approach for modelling multivariate ordinal variables. Their model accounts for ordinal structure in output variables and their non-linear dependencies via copula functions modelled as cliques of a CRF.

3 Proposed Model

Based on the previous studies, the proposed model to detect patients pain inten- sity from facial expressions effectively was designed as BiLSTM-CNN-VSL-CRF framework. Fig. 1 illustrates the architecture of our network in detail.

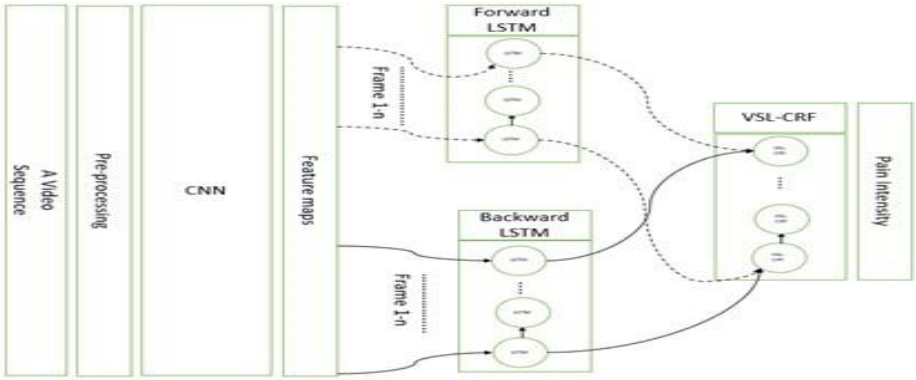

Fig. 1. The proposed model for pain intensity estimation from facial expression

3.1 Long Short-Term Memory Networks (LSTM)

LSTM is a type of RNN which is capable of learning long-term dependencies present on sequential data. Standard RNNs are theoretically capable of learning long term dependencies, but in practice, it is difficult to train them because the gradients tend to either explode or vanish. LSTM differs from standard RNN because it has a cell state controlled by three gates, which decide how much information should be let through. These gates are known as forget, input and output gates. The amount of information that is let through each gate is controlled by a point-wise multiplication and sigmoid function, as the sigmoid function output is between 0 and 1, indicating how much of the information should let through the gate. At each time-step, the input gate is computed depending on the input to the LSTM for that time-step and the previously hidden state.

3.2 Bidirectional Long Short-Term Memory (BiLSTM)

For many sequences labeling tasks it is beneficial to have access to both past (left) and future (right) frames. However, the LSTM's hidden state ht takes information only from past, knowing nothing about the future. An elegant solu- tion whose effectiveness has been proven by previous work is bidirectional LSTM (BiLSTM). The basic idea is to present each sequence forwards and backwards as two separate hidden states to capture past and future information, respectively. Then the two hidden states are concatenated to form the final output.

3.3 VSL-CRF (Variable-state Latent Conditional Random Field)

CRFs are a class of log-linear models that represent the conditional distribution

$P(h/x)$ as the Gibbs form clamped on the observation x [17]:

$$P(h/x,\theta) = \frac{1}{Z(x;\theta)} e^{s(x,h;\theta)} \quad (3)$$

Here, $Z(x;\theta) = \sum_{h \in H^e} e^{s(x,h;\theta)}$ is the normalizing partition function (H is a set of all possible output configurations), and θ are the parameters of the score function (or the negative energy) $s(x, h; \theta)$. Note that in this model, the states h is observed, and they represent the frame labels.

In VSL-CRF $v = (v1, ..., vk)$ be a vector of symbolic states or labels encoding the nature of the latent states h^v of the $i-th$ sequence, $i = 1, ..., N_y$ from class $y = (1, ..., K)$, either as nominal ($v_y = 0$) or ordinal ($v_y = 1$). The score function for class y in the VSL-CRF model is then defined as:

$$s(y,x,h,v;\partial) = \begin{cases} \sum_{k=1}^{k} I(k=y) \cdot s(x,h;\theta^n), & \text{if } v_y = 0 \\ \sum_{k=1}^{k} I(k=y) \cdot s(x,h;\theta^o), & \text{if } v_y = 1 \end{cases} \quad (4)$$

where the nominal ($s(x, h; \theta^n)$) and ordinal ($s(x, h; \theta^o)$) score functions represent the sum of the node and edge potentials, respectively. Then, the full conditional probability of the VSL-CRF model is given by:

$$P(y/x) = \sum_{h,v} P(y,h,v\,x) = \frac{\sum_{h,v} exp(s(y,x,h,v))}{Z(x)} \quad (5)$$

where $Z(x) = \sum_{k,h,v} exp(s(k,x,h,v))$

The VSL-CRF also performs integration over the latent variable m, the state of which (ordinal or nominal) defines the type of the latent states for each se- quence of facial expressions [16]. The definition of the VSL-CRF in Eq. 5 allows it to simultaneously fit both ordinal and nominal latent states to each sequence, which may result in the model overfitting [16].

4 Conclusion and Future Work

There is interest in automatic pain intensity estimation in healthcare and med- ical fields. In the past decade, many approaches have been proposed for auto- matic pain intensity estimation. Early researchers

tended to focus on estimating whether the subject is in pain or not, and thus, conduct pain intensity estima- tion as a classification problem by using deep learning. While the effectiveness of these models has been demonstrated on many general vision problems, in the facial data domain obtaining accurate and comprehensive labels is typically difficult. The model proposed in this paper has advantages over previous work offering significant potential for pain sufferers, their families and their careers.

For future work, we will develop a prototype system based on our new model.

This automatic pain detection system will be tested using real work dataset.

References

1. Bicket, M. C. and Mao, J.: Chronic pain in older adults. Anesthesiology clinics, Elsevier, 2015, 33, 577-590
2. Kharghanian, R.; Peiravi, A. and Moradi, F.: Pain detection from facial images using unsupervised feature learning approach. Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the, 2016, 419-422
3. Thomas, D.; Frascella, J.; Hall, T.; Smith, W.; Compton, W.; Koroshetz, W.; Briggs, J.; Grady, P.; Somerman, M. and Volkow, N.: Reflections on the role of opioids in the treatment of chronic pain: a shared solution for prescription opioid abuse and pain. Journal of internal medicine, Wiley Online Library, 2015, 278, 92-94
4. Frank, M. G.; Ekman, P. and Friesen, W. V.: Behavioral markers and recognizability of the smile of enjoyment. Journal of personality and social psychology, US: American Psychological Association, 1993, 64, 83
5. Fasel, B. and Luettin, J.: Automatic facial expression analysis: a survey. Pattern recognition, Elsevier, 2003, 36, 259-275
6. LeCun, Y.; Bengio, Y. and Hinton, G.: Deep learning. nature, Nature Publishing Group, 2015, 521, 436
7. LeCun, Y.; Bottou, L.; Bengio, Y. and Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE, IEEE, 1998, 86, 2278-2324
8. Rumelhart, D. E.; Hinton, G. E. and Williams, R. J.: Learning representations by back-propagating errors. nature, Nature Publishing Group, 1986, 323, 533
9. Zhou, J.; Hong, X.; Su, F. and Zhao, G.: Recurrent convolutional neural network regression for continuous pain intensity estimation in video. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, 84-92
10. Krizhevsky, A.; Sutskever, I. and Hinton, G. E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012, 1097-1105
11. Susskind, J. M.; Hinton, G. E.; Movellan, J. R. and Anderson, A. K.: Generating facial expressions with deep belief nets. Affective Computing, InTech, 2008
12. Xu, M.; Cheng, W.; Zhao, Q.; Ma, L. and Xu, F.: Facial expression recognition based on transfer learning from deep convolutional networks. Natural Computation (ICNC), 2015 11th International Conference on, 2015, 702-708
13. Liu, M.; Li, S.; Shan, S. and Chen, X.: Au-aware deep networks for facial expression recognition. Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, 2013, 1-6
14. Gudi, A.; Tasli, H. E.; Den Uyl, T. M. and Maroulis, A.: Deep learning based facs action unit occurrence and intensity estimation. Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, 2015, 6, 1-5
15. Zhao, K.; Chu, W.-S. and Zhang, H.: Deep region and multi-label learning for facial action unit detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 3391-3399
16. Walecki, R.; Rudovic, O.; Pavlovic, V. and Pantic, M.: Variable-state Latent Conditional Random Field models for facial expression analysis. Image and Vision Computing, Elsevier, 2017, 58, 25-37
17. Lafferty, John and McCallum, Andrew and Pereira, Fernando CN.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceeding of the 18th International Conference on Machine Learning 2001 (ICML 2001), 2001, 282-289