Multimodal Data Fusion based on the Global Workspace Theory

Cong Bao [1 2] **Zafeirios Fountas** [2 3] **Temitayo Olugbade** [1] **Nadia Bianchi-Berthouze** [1]

Abstract

We propose a novel neural network architecture, named the Global Workspace Network (GWN), that addresses the challenge of dynamic uncertainties in multimodal data fusion. The GWN is inspired by the well-established Global Workspace Theory from cognitive science. We implement it as a model of attention, between multiple modalities, that evolves through time. The GWN achieved F1 score of 0.92, averaged over two classes, for the discrimination between patient and healthy participants, based on the multimodal EmoPain dataset captured from people with chronic pain and healthy people performing different types of exercise movements in unconstrained settings. In this task, the GWN significantly outperformed a vanilla architecture. It additionally outperformed the vanilla model in further classification of three pain levels for a patient (average F1 score = 0.75) based on the EmoPain dataset. We further provide extensive analysis of the behaviour of GWN and its ability to deal with uncertainty in multimodal data.

1. Introduction

Reasoning about and interpreting multimodal data is an important task in machine learning research because life involves streaming of data from multiple modalities (Baltrusaitis et al., 2017). Multimodal data fusion, which leverages the combination of multiple modalities, is a valuable strategy (Atrey et al., 2010; Calhoun & Sui, 2016; Hori et al., 2017; Liu et al., 2018). Its benefits including complementarity of information, higher prediction performance, and robustness (Baltrusaitis et al., 2017). However, multimodal fusion comes with challenges; Lahat et al. (2015) specifies them under two categories: (1) challenges at the data observation and acquisition level, and (2) challenges due to uncertainty in the data (such as noise, missing values, conflicting information). Challenges at the observation level can be managed by data pre-processing, e.g. with data resampling, to deal with different temporal resolutions across modalities) (Aung et al., 2016). However, challenges due to uncertainty require the design of models that can exploit complementarity or discrepancy across modalities (Lahat et al., 2015), an area which is less explored. Findings in previous work on multimodal fusion have highlighted the effectiveness of weighting different modalities based on some “importance” metric (Wilderjans et al., 2011; imekli et al., 2013; Liberman et al., 2014; Kumar et al., 2007; Acar et al., 2011), which is the basis of the use of attention mechanisms in machine learning (Bahdanau et al., 2015). Despite the fact that uncertainty evolves through time in multimodal, sequential data (Lahat et al., 2015), relevant studies have not sufficiently explored mechanisms for both cross-modal cum temporal attention. For example, the architecture proposed by Beard et al. (2018) captures variations in importance along the time axis separately for the different modalities in the data. A drawback of their approach is that these variations are not simultaneously fused over the modalities.

To address this gap in multimodal data fusion, we propose the Global Workspace Network (GWN) which integrates variations in importance simultaneous through time and across modalities. Our GWN is inspired by the Global Workspace Theory (GWT) (Baars, 1997; 2002), which is a well-developed framework in cognitive science, and originally proposed as a model of human consciousness (Baars, 1988). The GWT states that concomitant cognitive processes *compete* for the opportunity to *broadcast* their current state (to peers) (Fountas et al., 2011). At each iteration, the winner (a single process or a coalition of processes) earns the privilege of contributing current information in a *global workspace* which can be accessed by all processes (including the winner) (Shanahan, 2008). This competition and broadcast cycle is believed to be ubiquitous in the perceptual regions of the brain (Baars, 1988). Although the literature contains architectures of biologically-realistic spiking neural networks based on GWT (Shanahan, 2008; Fountas et al., 2011), so far, to our knowledge, there has been no relevant

[1] Department of Computer Science, University College London, London, United Kingdom [2] Emotech Labs, London, United Kingdom [3] Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London, London, United Kingdom. Correspondence to: Cong Bao <cong.bao.18@ucl.ac.uk>, Zafeirios Fountas <f@emotech.co>, Temitayo Olugbade <temitayo.olugbade.13@ucl.ac.uk>, Nadia Berthouze <nadia.berthouze@ucl.ac.uk>.

arXiv preprint, 2020.

implementation in machine learning. Mechanistically, the main concept of this theory can be implemented as the combination of a compete-and-broadcast procedure and an external memory structure. In contrast to the global workspace, which can be seen as a communication module, external memory here is used as the means to store information for later application (Shanahan, 2006). Taking the processing of each modality in multimodal data as analogous to specialised processes in the brain, the similarity between the compete-and-broadcast cycle and typical cross-modality attention mechanism is obvious. The repetitiveness of the cycle allows the pattern of attention to evolve over time and, given the external memory module, be used in the primary prediction task of the network.

In order to simulate the two main components of the GWT (compete-and-broadcast and external memory), we employed two widely-tested algorithms, the transformer (Vaswani et al., 2017) and the Long Short-Term Memory (LSTM) neural network (Hochreiter & Schmidhuber, 1997; Gers et al., 1999) respectively. There are 3 key elements of the transformer that we leverage in the GWN. First is its self-attention mechanism (Cheng et al., 2016; Paulus et al., 2017) that we use as the GWN's compete-and-broadcast procedure where each modality independently scores all modalities and integrates the data from them based on the resulting weights. The second valuable component of the transformer is its memory-based attention mechanism (Weston et al., 2015; Sukhbaatar et al., 2015). Drawing from its traditional application in Natural Language Processing (NLP) question answering tasks (Sukhbaatar et al., 2015; Miller et al., 2016), this unit further maps the feature vector into query, key and value spaces to increase the weighting depth and robustness (Hu, 2018). This additionally enables distributed competition versus broadcasting computations. In essence, the query and key forms can be used for the competition, while broadcast is performed on value form, which can have more expressive information that is not valuable for the competition. The third merit of the transformer is its bagging approach, i.e. the use of multiple heads in which multiple attention patterns are learnt in parallel, which has the advantage of increasing robustness. We used the LSTM as the basis for the complementary external memory because it has been shown as effective for learning long-term dependencies (Lipton, 2015).

The contributions of this paper are as follows:

- The GWN architecture, a novel approach to fusion of sequential data from multiple modalities.

  We evaluate the proposed GWN architecture on the EmoPain dataset (Aung et al., 2016), which consists of motion capture and electromyography (EMG) data collected from patients with chronic lower back pain and healthy control participants while they performed exercise movements. This dataset is representative of real-life data with continuously-streamed multiple modalities, each with varying degrees of uncertainty.

- Analysis of the GWN's outputs demonstrating its effectiveness in handling uncertainty in data.

The paper is organized as follows. We discuss the state of the art in attention approaches in Section 2. We then describe the proposed GWN architecture in Section 3 and present both validation and analysis of the network in Section 4. Section 5 concludes the paper.

2. Related Work

Attention in the temporal dimension In the literature on neural networks for multimodal data, attention performed on the time axis is usually separated by modality, and the resulting context vectors from each modality are fused as non-temporal features. A representative case of this approach is the Recursive Recurrent Neural Network (RRNN) architecture proposed by Beard et al. (2018). In their work, different modalities (video, audio, and subtitles) extracted from a subtitled audiovisual dataset were divided into segments of uttered sentences and each segment was used an input to the network. For each modality in a segment, a bi-directional LSTM layer was used to extract features. At a given time step, attention computation is performed for each modality separately and the outputs are concatenated over all modalities together with the current state of a shared memory, which the authors implemented with a GRU cell (Cho et al., 2014). The outcome is then used to update the state of shared memory. An advantage of this work is that since each modality was encoded separately, they do not have to follow a common time axis, which allows each modality to optimally exploit its inherent temporal properties. However, as this method cannot account for attention between modalities, different modalities (some potentially more noisy than others) affect the final prediction equally and thus the problem of cross-modal uncertainty variation remains unsolved.

Attention across modalities Several related studies have considered the relation between modalities in fusing them. The typical approach (Wilderjans et al., 2011; imekli et al., 2013; Liberman et al., 2014) is the use of modality weighting although not particularly based on attention mechanisms (Bahdanau et al., 2015). One study that does explicitly use an attention mechanism is the work of Hori et al. (2017) on the modelling of video description. Their approach leverages attention between different modalities using an encoder-decoder architecture (Bahdanau et al., 2015) with separate encoders for each modality and a single decoder. Features of each modality are encoded separately and the decoder weights them to generate a context vector as an output. A similar study Caglayan et al. (2016) applies

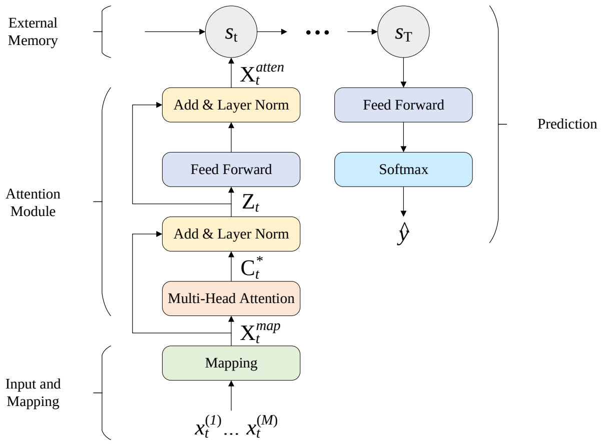

Figure 1. The architecture of the GWN. Here the intermediate matrices $\mathbf{X}_t^{map}$, $\mathbf{C}_\mathbf{t}^*$, $\mathbf{Z}_t$, and $\mathbf{X}_t^{atten}$ have the same dimensionality of $M \times H$.

multimodal attention in neural machine translation where images are leveraged in translating the description texts from one language to another. The image and text modalities were first encoded using pre-trained ResNet-50 (He et al., 2015) and bi-directional GRU (Cho et al., 2014) respectively. Then, attention scores were computed for these encodings. The common approach of encoding the temporal data before computing attention is appropriate for obtaining modality-specific feature representation, however, it does not allow in-depth capture of the complex interactions between modalities through time. In addition, it is not suitable for online process of live-streamed data.

The GWN architecture we propose addresses these limitations by considering both the interaction of multiple modalities and the temporal variations in this interaction. It is indeed a more intuitive approach to processing a stream of multimodal data, by weighting multiple modalities at each timestep.

3. Global Workspace Network (GWN)

The architecture of the GWN is shown in Figure 1. The network consists of five components: an input unit, a mapping block, an attention module, an external memory, and a prediction block. These components are described in detail below.

3.1. Mapping Inputs to a Common Feature Space

Consider M modalities that they have an identical sampling rate, i.e. for each data instance, each modality $m \in M$ in that instance can be written as $\{\boldsymbol{x}_1^{(m)}, \dots, \boldsymbol{x}_T^{(m)}\}$, where T denotes the common temporal length (common across modalities) of the data instance. The dimensionality at a given time t may nevertheless be different across these modalities, i.e. $\boldsymbol{x}_t^{(m)} \in \mathbb{R}^{d_m}$. The attention mechanism of the GWN requires identical dimension across modalities and so, it is necessary to have a module for mapping the modalities into the same dimensions.

Inspired by the work of Akbari et al. (2018) and Bollegala & Bao (2018), we take the approach of using multiple autoencoders (Vincent et al., 2008) that each learn a common feature space for multiple modalities. Assuming that the common feature space $\boldsymbol{c}$ has a dimensionality of H, the mapping function in the encoder for each autoencoder $E^{(m)}$ outputs a vector with dimensionality of H. This function can be designed as a feed forward network with one hidden layer which is activated with the rectified linear unit (ReLU) (Nair & Hinton, 2010) non-linearity, i.e.

$$E^{(m)}\left(\boldsymbol{x}_t^{(m)}\right) = \max\left(0, (\boldsymbol{x}_t^{(m)}\mathbf{W}_1 + \boldsymbol{b}_1)\right)\mathbf{W}_2 + \boldsymbol{b}_2 \tag{1}$$

where $\boldsymbol{x}_t^{(m)} \in \mathbb{R}^{d_m}$ is the data instance $\boldsymbol{x}$ sampled at modality m and time t; and $\mathbf{W}_1$, $\mathbf{W}_2$, $\boldsymbol{b}_1$, and $\boldsymbol{b}_2$ are trainable parameters of function. The findings of Cybenko (1989) suggest that such encoding should be capable of mapping different modalities into a common feature space. $\boldsymbol{c}$ can then be obtained by summing the outputs across the encoders

$$\boldsymbol{c} = \sum_m E^{(m)}\left(\boldsymbol{x}_t^{(m)}\right) \tag{2}$$

This is based on previous work in Bollegala & Bao (2018). The decoders have the same form as the encoders, i.e.

$$\begin{aligned}\hat{\boldsymbol{x}}_t^{(m)} &= D^{(m)}\left(\boldsymbol{x}_t^{(m)}\right) \\ &= \max\left(0, (\boldsymbol{c}\mathbf{W}_1' + \boldsymbol{b}_1')\right)\mathbf{W}_2' + \boldsymbol{b}_2'\end{aligned} \tag{3}$$

where $\hat{\boldsymbol{x}}_t^{(m)} \in \mathbb{R}^{d_m}$ is the reconstruction of data instance $\boldsymbol{x}$ sampled at modality m and time t; and $\mathbf{W}_1'$, $\mathbf{W}_2'$, $\boldsymbol{b}_1'$, and $\boldsymbol{b}_2'$ are trainable parameters of decoder. A sum $\mathcal{L}\left(E^{(m)}, D^{(m)}\right)$ of the mean squared error loss for each autoencoder can be used to train the full mapping module.

$$\mathcal{L}\left(E^{(m)}, D^{(m)}\right) = \sum_m \left|\left|\hat{\boldsymbol{x}}_t^{(m)} - \boldsymbol{x}_t^{(m)}\right|\right|^2 \tag{4}$$

Figure 2 provides an illustration with an example of two modalities mapped into a common feature space and then reconstructed, based on two autoencoders. After pre-training the autoencoders, the encoders are used directly as the mapping function in the GWN. The pre-trained parameters in the encoders then serve as initial values for the mapping block in the GWN. Though this approach introduces more learnable parameters, the findings of Hinton et al. (2006) suggest that unsupervised pre-training on shallow layers can improve the performance of a deep network.

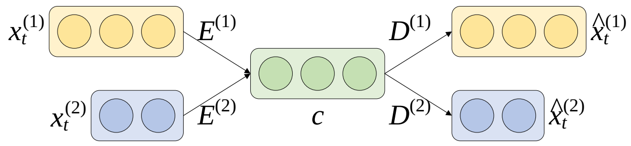

Figure 2. An illustration of the mapping module with two different modalities and two autoencoders.

For the subsequent attention module, the output vector from each modality's mapping are merged by stacking, to form a matrix $\mathbf{X}_t^{map} \in \mathbb{R}^{M \times H}$.

3.2. The Attention Module

The attention module is a single layer of the transformer encoder described in (Vaswani et al., 2017) with the difference that, in the GWN, the input is a set of different modalities for a number of data instances at a specific time t, rather than data sequences (i.e. multiple time steps and instances) based on a single modality. Since the input $\mathbf{X}_t^{map} \in \mathbb{R}^{M \times H}$ is already in matrix form, the following multi-head attention calculation can be performed:

$$\mathbf{C}_t^{*} = \text{concat}\left(\mathbf{C}_t^{1}, \ldots, \mathbf{C}_t^{K}\right) \mathbf{W}^{\text{O}} \tag{5}$$

where K is a set of heads and $\mathbf{W}^{\text{O}} \in \mathbb{R}^{KH \times H}$ is a trainable matrix. Each context matrix $\mathbf{C}_t^k \in \mathbb{R}^{M \times H}$ for a specific head $k \in K$ is calculated as

$$\mathbf{C}_t^{k} = \text{softmax}\left(\frac{\mathbf{Q}_t^{k} \mathbf{K}_t^{k\top}}{\sqrt{H}}\right) \mathbf{V}_t^{k} \tag{6}$$

The query, key, and value matrices of a specific head k at time t are calculated as:

$$\mathbf{Q}_t^{k} = \mathbf{X}_t^{map} \mathbf{W}_k^{\text{Q}} \tag{7}$$

$$\mathbf{K}_t^{k} = \mathbf{X}_t^{map} \mathbf{W}_k^{\text{K}} \tag{8}$$

$$\mathbf{V}_t^{k} = \mathbf{X}_t^{map} \mathbf{W}_k^{\text{V}} \tag{9}$$

Here, the query, key, and value are variations of the input $\mathbf{X}_t^{map}$, based on the idea of memory-based attention mechanism (Miller et al., 2016). Note that the trainable matrices $\mathbf{W}_k^{\text{Q}} \in \mathbb{R}^{H \times H}$, $\mathbf{W}_k^{\text{K}} \in \mathbb{R}^{H \times H}$, and $\mathbf{W}_k^{\text{V}} \in \mathbb{R}^{H \times H}$ are reused on different time steps t but are independent for different heads k.

As shown in Figure 1, there are two residual connections (He et al., 2015) in the attention module. Each of the residual connection is followed by a layer normalisation (Lei Ba et al., 2016). The first residual connection can be represented as:

$$\mathbf{Z}_t = \text{layernorm}\left(\mathbf{C}_t^{*} + \mathbf{X}_t^{map}\right) \tag{10}$$

Here, the assumption of identical dimensionality for residual connection is satisfied as $\mathbf{C}_t^{*} \in \mathbb{R}^{M \times H}$ and $\mathbf{X}_t^{map} \in \mathbb{R}^{M \times H}$. The subsequent feed forward layer and the final output of the attention module, respectively, are:

$$\text{FFN}\left(\mathbf{Z}_t\right) = \max\left(0, \left(\mathbf{Z}_t \mathbf{W}_1 + \boldsymbol{b}_1\right)\right) \mathbf{W}_2 + \boldsymbol{b}_2 \tag{11}$$

$$\mathbf{X}_t^{atten} = \text{layernorm}\left(\text{FFN}\left(\mathbf{Z}_t\right) + \mathbf{Z}_t\right) \tag{12}$$

both $\in \mathbb{R}^{M \times H}$.

3.3. External Memory

The external memory is implemented as an LSTM cell (Hochreiter & Schmidhuber, 1997) with updates:

$$\boldsymbol{f}_t = \sigma\left(\left[\boldsymbol{x}_t^{atten}; \boldsymbol{h}_{t-1}\right] \mathbf{W}^{f} + \boldsymbol{b}^{f}\right) \tag{13}$$

$$\boldsymbol{i}_t = \sigma\left(\left[\boldsymbol{x}_t^{atten}; \boldsymbol{h}_{t-1}\right] \mathbf{W}^{i} + \boldsymbol{b}^{i}\right) \tag{14}$$

$$\boldsymbol{o}_t = \sigma\left(\left[\boldsymbol{x}_t^{atten}; \boldsymbol{h}_{t-1}\right] \mathbf{W}^{o} + \boldsymbol{b}^{o}\right) \tag{15}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tanh\left(\left[\boldsymbol{x}_t^{atten}; \boldsymbol{h}_{t-1}\right] \mathbf{W}^{c} + \boldsymbol{b}^{c}\right) \tag{16}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh\left(\boldsymbol{c}_t\right) \tag{17}$$

where the input vector $\boldsymbol{x}_t^{atten} \in \mathbb{R}^{MH}$ is the flattened form of $\mathbf{X}_t^{atten} \in \mathbb{R}^{M \times H}$, $\sigma\left(\cdot\right)$ is the sigmoid function:

$$\sigma\left(x\right) = \frac{1}{1 + \exp\left(-x\right)} \tag{18}$$

$\tanh\left(\cdot\right)$ is the hyperbolic tangent function

$$\tanh\left(x\right) = \frac{\exp\left(x\right) - \exp\left(-x\right)}{\exp\left(x\right) + \exp\left(-x\right)} \tag{19}$$

and $\odot$ denotes the Hadamard product (i.e. element-wise product). $\boldsymbol{s}_t \in \mathbb{R}^{2G}$ is the recurrent state at time step t, and consists of a memory cell $\boldsymbol{c}_t \in \mathbb{R}^{G}$ and the output $\boldsymbol{h}_t \in \mathbb{R}^{G}$ at that time step, with G as an hyperparameter that indicates the size of the external memory. The initial state $\boldsymbol{s}_0 = [\boldsymbol{c}_0; \boldsymbol{h}_0]$ is set with zeros. $\boldsymbol{f}_t$, $\boldsymbol{i}_t$, and $\boldsymbol{o}_t$ represent forget, input, and output gates respectively (Hochreiter & Schmidhuber, 1997; Gers et al., 1999). All the gates have the same dimensionality G. The output vector $\boldsymbol{h}_T \in \mathbb{R}^{G}$ in the last recurrent state $\boldsymbol{s}_T$ is used by the final prediction component.

3.4. Prediction

The final prediction module consists of a feed forward layer with one hidden layer activated with a ReLU followed by a softmax function. The layer serves as a simple non-linear transformation from the external memory and can be applied at any time step, making it suitable for online prediction with streaming data. The equations are given as

$$\boldsymbol{r} = \max\left(0, \boldsymbol{h}_T \mathbf{W}_1 + \boldsymbol{b}_1\right) \mathbf{W}_2 + \boldsymbol{b}_2 \tag{20}$$

$$\hat{\boldsymbol{y}} = \text{softmax}\left(\boldsymbol{r}\right) \tag{21}$$

i.e.

$$\hat{y}_i = \frac{\exp(r_i)}{\sum_j \exp(r_j)} \tag{22}$$

where $\boldsymbol{r}$ is the prediction result mapped into the distribution $\hat{\boldsymbol{y}}$. Both $\boldsymbol{r}$ and $\hat{\boldsymbol{y}}$ have the same dimensionality, the size of label L.

4. Experiments

To evaluate the proposed GWN architecture, we conducted experiments on the multimodal EmoPain dataset (Aung et al., 2016). In Section 4.1, the dataset, data preprocessing, and experiment tasks are introduced. Section 4.2 describes the methods and metrics used for evaluation against a baseline model. Finally, Section 4.3 presents the performance and empirical analyses of the GWN.

4.1. Data

4.1.1. The EmoPain Dataset

The EmoPain dataset (Aung et al., 2016) is suitable given that it consists of sequential data from multiple modalities and in unconstrained settings where there are bound to be uncertainties (e.g. in form of sensor noise) in the data, and in varying degrees over time. The data was collected from 22 patients with chronic low back pain and 28 healthy control participants and includes motion capture (MC) and muscle activity data based on surface electromyography (EMG). The data for each participant was acquired while they performed physical exercises that put demands on the lower back. For each exercise, there were two levels of difficulty. There is the normal trial, for 7 types of exercise ((1) balancing on preferred leg, (2) sitting still, (3) reaching forward, (4) standing still, (5) sitting to standing and standing to sitting, (6) bending down, and (7) walking). There is additionally the difficult trial, where four of these exercise types were modified to increase the level of physical demand, i.e. (8) balancing on each leg, (9) reaching forward while standing holding a 2 kg dumbbell, (10) sitting to standing and return to sitting initiated upon instruction, (11) walking with 2 kg weight in each hand, starting by bending down to pick up the weights, and exercises (2) and (4) repeated without modification. The data was acquired so as to build automatic detection models for pain and related cognitive and affective states, and so after each exercise type, patients self-reported the level of pain they experienced, on a scale of 0 to 10 (0 for no pain and 10 for extreme pain) (Jensen & Karoly, 1992). In this paper, we used the subset of the EmoPain dataset with the self-reported pain labels available and where consent was given for further use of the data. This subset consists of 14 patients with chronic pain and 8 healthy control participants resulting in a total of 200 exercise instances.

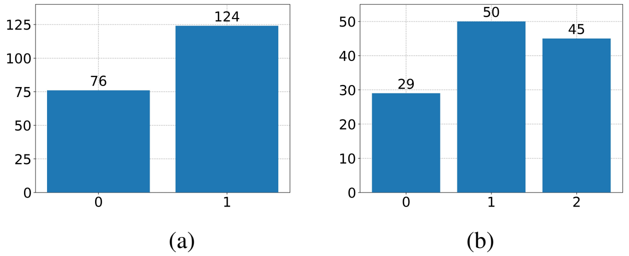

Figure 3. Number of exercise instances per each classes for (a) Healthy-vs-Patient Discrimination Task and (b) Pain Level Detection Task.

4.1.2. Evaluation Experiment Tasks

The proposed GWN architecture was evaluated on two classification tasks based on the multimodal EmoPain dataset:

Pain Level Detection Task The aim of this task is to detect the level of a person with chronic pain. The motivation for creating such system is to endow technology with the capability for supporting physical rehabilitation by providing timely feedback or prompts, and personalised recommendations tailored to the pain level of a person with chronic pain. For example, a person with low level pain may be reminded to take breaks at appropriate times and not overdo, whilst a person with high pain may be reminded to breath to reduce tension which may further increase pain levels (Olugbade et al., 2019).

A formal description of the task is as follows. Given M and E, denoting MC and EMG data, for an unseen subject known to have chronic pain (i.e. the event $cp = 1$), infer the probability $p\,(l|cp = 1, M, E)$ that the data corresponds to one of three levels of pain. A random variable l represents the level of chronic pain and is $\in \{0, 1, 2\}$. In this paper, 0 represents zero level pain, i.e. pain self-report = 0, 1 represents low level pain, i.e $0 <$ pain self-report ≤ 5), and 2 represents high level pain, i.e pain self-report > 5).

Healthy-vs-Patient Discrimination Task The healthy control participants were assumed to have no pain. However, patients with chronic pain who reported pain as 0 were not considered to be in the same class as these participants. Hence, a separate model may be needed to first distinguish a person with chronic pain from healthy participants.

The formal definition of the task is as follows. Given M and E, infer the probability $p\,(cp|M, E)$ that the data belongs to a person with chronic pain. A random variable cp represents the event that an unseen subject has chronic pain, and $cp \in \{0, 1\}$ with 0 for healthy and 1 for chronic pain person.

Figure 3 shows the number of exercise instances for each class, for the Healthy-vs-Patient Discrimination Task and Pain Level Detection Task respectively.

4.1.3. Data Preprocessing

Here, we describe the preprocessing performed to prepare the data for the evaluation experiments.

Dealing with A High Sampling Rate The EMG data of the EmoPain dataset had been downsampled from 1000Hz to 60Hz for consistency with the MC data. However, 60Hz results in high dimensionality whereas preliminary experiments suggest that 10Hz may be sufficient for the Healthy-vs-Patient Discrimination Task. Thus, we downsampling both MC and EMG data further, to 10 Hz to be suitable for the Healthy-vs-Patient Discrimination Task. The original 60Hz was found to be more appropriate for the Pain Level Detection Task.

Padding for Uniform Sequence Lengths Based on the findings in Dwarampudi & Reddy (2019); Wang et al. (2019), we used pre-padding rather than post-padding to obtain uniform time sequence lengths for different data instances. Further, we used zero padding, which is the common approach used in modelling when assuming no prior knowledge about the input data (Shi et al., 2015).

Dealing with Imbalanced Data As can be seen in Figure 3, the class distribution of the data is skewed for both pain classification tasks. To reduce bias toward the majority class, we randomly over-sampled data instances of the minority class (Kotsiantis et al., 2005).

Data Augmentation The total number of exercise instances available for training and evaluation was 200, which is a limited amount for training a neural network. We employed data augmentation, particularly creating new instances from the original by rotating them, to address this problem. Preliminary experiments that we performed show that rotation about y-axis, which is along the cranial-caudal, outperforms the mirror reflection augmentation used in Olugbade et al. (2018). This augmentation approach used four angles, 0°, 90°, 180°, and 270°, and resulted in four times the original data size. For each newly created instance, only the original MC data was changed by the rotation; for these instances, the original EMG data was used unchanged as they are not affected by the orientations.

4.2. Evaluation Methods

4.2.1. Baseline Model

A simple concatenation (CONCATN) architecture, which is representative of the traditional multimodal data fusion approach, was used as the baseline network against which we evaluated our GWN architecture. This baseline allows evaluation of the contribution of the GWN's mapping and attention components to its performance. The CONCATN has identical external memory and prediction units. Hence, it can be seen as a network that does not pay particular attention to different modalities over time, but rather treats them equally through time.

In the CONCATN, multiple modalities are concatenated along the feature axis and fed into a LSTM network. The feed forward equations are

$$\boldsymbol{x}_t^* = \text{concat}\left(\boldsymbol{x}_t^{(1)}, \ldots, \boldsymbol{x}_t^{(M)}\right) \tag{23}$$

$$\boldsymbol{c}_t, \boldsymbol{h}_t = \text{lstm}\left(\boldsymbol{x}_t^*, \boldsymbol{c}_{t-1}, \boldsymbol{h}_{t-1}\right) \tag{24}$$

where M is the number of modalities, $\boldsymbol{c}_t$ is a memory cell and $\boldsymbol{h}_t$ is the hidden state. Initial states $\boldsymbol{c}_0$ and $\boldsymbol{h}_0$ have values of zero. Assuming the dimensionality of each modality input at a specific time t is d_m, the dimensionality of the concatenated vector $\boldsymbol{x}_t^*$ is $\sum_m^M d_m$. The dimensionalities of $\boldsymbol{c}_t$ and $\boldsymbol{h}_t$ have the same values as in the GWN model. The prediction module is also identical to the GWN model, i.e. the last LSTM output $\boldsymbol{h}_T$ is fed into a feed forward network with one hidden layer activated with ReLU (Nair & Hinton, 2010) non-linearity.

4.2.2. Validation Technique

In the experiments carried out, we used the leave-one-subject-out cross-validation (LOSOCV), where the data for a single subject is left out for testing in each fold as is the standard approach for evaluating the generalisation capability of a model to unseen subjects. However, for statistical tests to compare the proposed GWN with the baseline CONCATN, the LOSOCV has the limitation of overlapping training sets across folds that has the risk of high Type I error (Dietterich, 1998). Thus, in this work, we additionally perform 5×2 CV (i.e. 5 random replications of 2-fold CV) which has a lower risk of Type I errors (Dietterich, 1998) for the purpose of model comparison. The advantage of the 2-fold CV is that there are no overlap between training sets.

For both LOSOCV and 5×2 CV, we perform Wilcoxon signed-rank test (Wilcoxon, 1945) to compare the proposed GWN and the baseline CONCATN.

4.3. Results and Discussion

4.3.1. Comparison with the Baseline

Both the GWN and the CONCATN baseline model are trained with optimisation algorithm (Adam (Kingma & Ba, 2014)), learning rate (0.001), and batch size (32), which were chosen by grid search. The dimensionality of LSTM cell, which is the shared hyperparameter of the two models, are also kept the same, i.e. 64. The performance of the GWN can be seen in Table 1 comparison with the CONCATN baseline model, based on accuracy (ACC), Matthews Correlation Coefficient (MCC) (Matthews, 1975), and F1

Task	Validation	Model	ACC	MCC	F_1 (0)	F_1 (1)	F_1 (2)	F_1 (avg)	r	p
Healthy-vs-Patient Discrimination Task	LOSOCV	CONCATN	0.765	0.489	0.662	0.820	-	0.745	0.628	0.003
		GWN*	0.920	0.831	0.887	0.938	-	0.915		
	5 × 2 CV	CONCATN	0.587	0.110	0.434	0.675	-	0.555	0.768	0.015
		GWN*	0.648	0.225	0.482	0.733	-	0.613		
Pain Level Detection Task	LOSOCV	CONCATN	0.653	0.465	0.464	0.667	0.756	0.629	0.487	0.068
		GWN	0.766	0.645	0.581	0.800	0.857	0.748		
	5 × 2 CV	CONCATN	0.395	0.075	0.249	0.438	0.441	0.379	0.596	0.059
		GWN†	0.448	0.151	0.309	0.474	0.503	0.430		

Table 1. Evaluation Experiment Results Comparing the GWN with the Baseline CONCATN. $*$ indicates that a Wilcoxon Signed-Rank test showed that the model performance is significantly (significance level p = 0.05) higher. † indicates that the model accuracy is marginally significantly higher.

scores.

Our results show that the GWN significantly outperforms the baseline for the Health-vs-Patient Discrimination task (significance level p = 0.05) with F1 score of 0.913 based on LOSOCV, averaged over the two classes. The effect size is r=0.768 for the 5×2 CV and r=0.628 for the LOSOCV. As expected, due to smaller training data size in the 5×2 CV, it gives lower performance estimation than the LOSOCV for both the baseline CONCATN and the GWN. Although only marginally significant in this case, the GWN also outperforms the baseline CONCATN in the Pain Level Detection Task, effect size r=0.596, for the 5×2 CV.

4.3.2. Attention Patterns

An additional advantage of the proposed GWN model is that the attention patterns obtained in modelling can provide insight into the relevance of each modality through time. In our experiments, we found 5 attention patterns (see Figure 4 for further specification of each pattern):

Favours-Itself-Always (FIA) The given modality always pays attention to itself and never switches attention to the other modality.

Favours-Other-Sometimes (FOS) The given modality mostly pays attention to itself but sometimes switches its attention to the other modality.

Favours-Itself-and-Other-in-Balance (FIOB) The given modality pays balanced attention to itself and the other modality.

Favours-Itself-Sometimes (FIS) The given modality mostly pays attention to the other modality but sometimes switches attention to itself.

Favours-Other-Always (FOA) The given modality always pays attention to the other modality and never to itself.

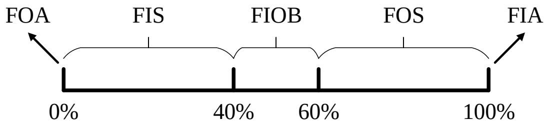

Figure 4. The percentage a modality pays attention to itself in the five patterns. The threshold 40% and 60% used in this definition were chosen heuristically as a ±10% interval around 50%.

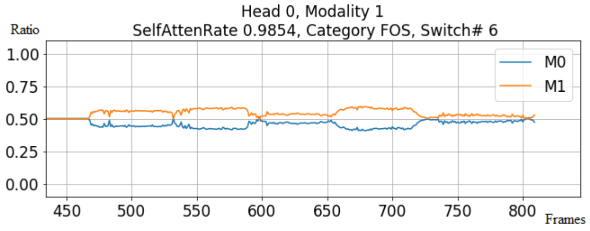

Figure 5. An example of attention distribution of one exercise instance. Head 0 means the first attention head. Modality 0 (M0) represents MC and modality 1 (M1) represents EMG.

Figure 5 gives an example of the FOS pattern. In this case, modality 1 (EMG) pays attention to itself most of the time (98.54%), with a few switches (6 times) to modality 0 (MC).

The frequency of occurrence of each of the five attention cases are shown in Table 2 (row 3). It can be seen that MC tends to always pay attention to either only itself or mostly to the EMG (higher FIA and FOA frequencies), whereas the EMG balances its attention (higher FOS, FIOB and FIS frequencies). One possible explanation is that, since the dimensionality of EMG (4) is much lower than the dimensionality of MC data (78), EMG is always trying to balance the difference in information. In contrast, the modality of MC is rich in information, and so can afford to pay 100 percent attention to itself.

1	Noise	FIA		FOS		FIOB		FIS		FOA		mean of switch #		std. of switch #	
2		MC	EMG	MC	EMG	MC	EMG	MC	EMG	MC	EMG	MC	EMG	MC	EMG
3	None	0.51	0.40	0.04	0.29	0.03	0.05	0.05	0.15	0.37	0.11	0.40	14.3	1.32	30.9
4	In MC	0.31	0.43	0.08	0.36	0.02	0.05	0.11	0.10	0.48	0.07	6.92	14.6	25.0	30.4
5	In EMG	0.50	0.46	0.02	0.27	0.02	0.05	0.06	0.09	0.41	0.13	0.35	12.6	1.52	30.7

Table 2. Relative frequency of the five attention patterns for the Pain Level Detection Task, with or without noise added in the data.

Noise	ACC	MCC	F_1 (0)	F_1 (1)	F_1 (2)	F_1 (avg)
None	0.766	0.645	0.581	0.800	0.857	0.748
In MC	0.734	0.594	0.557	0.763	0.822	0.715
In EMG	0.734	0.599	0.590	0.747	0.813	0.721

Table 3. Results of Pain Level Detection Task with or without noise in each MC and EMG.

4.3.3. Evaluating How The GWN Deals with Uncertainty in Data

In order to further examine the behaviour of the GWN model with respect to uncertainties in the data, noise was added to one modality at a time. We experimented with different levels of noise. We expected that if the GWN manages uncertainty in data, the modality without added noise would pay less attention to the noisy modality.

The noise was sampled from a Gaussian distribution with zero mean and standard deviation σ_{noise}, equal to 10% of the standard deviation in the original data for this modality. For instance, as the standard deviation of MC in the Pain Level Detection Task is 105.4, in this case, $\sigma_{\text{noise}} = 10$ (round as integral ten digits). Similarly, in the case of the EMG recordings of the same dataset, $\sigma_{\text{noise}} = 0.001$.

Table 3 presents the result of adding noise. A Wilcoxon Signed-Rank test showed no significant (significance level of p = 0.05) difference between the accuracy of the GWN model with and without noise in the MC data, based on the LOSOCV ($r = 0.492, p = 0.066$) or with and without noise in the EMG also based on the LOSOCV ($r = 0.045, p = 0.866$). This suggests that the proposed GWN may be tolerant to this level of noise.

Table 2 shows the GWN's behaviour with the noisy input (row 4 for noisy MC and row 5 for noisy EMG), separated based on the detected attention patterns. Compared with frequencies of the 5 attention cases without added noise, with the noisy MC data, the frequency of FIA for the MC decreases while its frequencies of FOS, FIS, and FOA increase. This indicates that the MC modality is able to recognise noise in itself and rely more on the other modality (EMG). This is also evident in the increase in mean switch frequency.

In contrast, having a noisy EMG (see row 5 in Table 2) does not result in the same behaviour. Compared with the frequencies of the 5 attention cases (see row 3), the frequency of the EMG's FIA with noisy EMG unexpectedly increases. The frequencies of FOS and FIS also do not increase. Only the FOA frequencies shows expected albeit slight increase. In addition, the mean of switch frequency shows no increment. These results suggest that the EMG modality is less sensitive to its noisiness. One explanation is that the amount of noise added to EMG is not sufficient enough to influence the feature representation. Another possible reason is that the system is sensitive to precise amount of information being lost per modality. Since the dimensionalities of MC and EMG are different, 78 and 4 respectively, the 10% noise added to MC corrupts more information than when added to the EMG, leading to a more sensitive MC in the case of the former.

5. Conclusion

Here we proposed the GWN, a novel neural network architecture for multimodal fusion in temporal data. At each time step, multiple modalities compete for broadcasting information, and each broadcast is accumulated over time. We find that the GWN outperforms baseline multimodal fusion by concatenation, for pain level detection based on the EmoPain dataset. Our analysis further highlights the selectivity of the different modalities in this dataset. Moreover, modality-specific noise manipulations revealed the ability of GWN to deal with changes in uncertainty over time. We believe that our system presents a promising direction for future research in multimodal neural networks, while promoting a close connection with cognitive neuroscience. Such interdisciplinary links can be fruitful for both communities and help to propel each other forward.

References

Acar, E., M. Dunlavy, D., G. Kolda, T., and Mrup, M. Scalable tensor factorizations for incomplete data. ***Chemometrics and Intelligent Laboratory Systems***, 106:41–56, 03 2011. doi: 10.1016/j.chemolab.2010.08.004.

Akbari, H., Karaman, S., Bhargava, S., Chen, B., Vondrick, C., and Chang, S. Multi-level multimodal common semantic space for image-phrase grounding. ***CoRR***, abs/1811.11683, 2018. URL http://arxiv.org/abs/1811.11683.

Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. Multimodal fusion for multimedia analysis: A survey. ***Multimedia Syst.***, 16(6):345–379, November 2010. ISSN 0942-4962. doi: 10.1007/s00530-010-0182-0. URL http://dx.doi.org/10.1007/s00530-010-0182-0.

Aung, M. S. H., Kaltwang, S., Romera-Paredes, B., Martinez, B., Singh, A., Cella, M., Valstar, M., Meng, H., Kemp, A., Shafizadeh, M., Elkins, A. C., Kanakam, N., de Rothschild, A., Tyler, N., Watson, P. J., Williams, A. C. d. C., Pantic, M., and Bianchi-Berthouze, N. The automatic detection of chronic pain-related expression: Requirements, challenges and the multimodal emopain dataset. ***IEEE Trans. Affect. Comput.***, 7 (4):435–451, October 2016. ISSN 1949-3045. doi: 10.1109/TAFFC.2015.2462830. URL https://doi.org/10.1109/TAFFC.2015.2462830.

Baars, B. J. ***A Cognitive Theory of Consciousness***. Cambridge University Press, Cambridge, MA, 1988.

Baars, B. J. ***In the Theater of Consciousness***. Oxford University Press, New York, NY, 1997.

Baars, B. J. The conscious access hypothesis: Origins and recent evidence. ***Trends in Cognitive Sciences***, 6(1):47–52, 2002. doi: 10.1016/s1364-6613(00)01819-2.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In ***3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings***, 2015. URL http://arxiv.org/abs/1409.0473.

Baltrusaitis, T., Ahuja, C., and Morency, L. Multimodal machine learning: A survey and taxonomy. ***CoRR***, abs/1705.09406, 2017. URL http://arxiv.org/abs/1705.09406.

Beard, R., Das, R., Ng, R. W. M., Gopalakrishnan, P. G. K., Eerens, L., Swietojanski, P., and Miksik, O. Multimodal sequence fusion via recursive attention for emotion recognition. In ***Proceedings of the 22nd Conference on Computational Natural Language Learning***, pp. 251–259, Brussels, Belgium, October 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/K18-1025.

Bollegala, D. and Bao, C. Learning word meta-embeddings by autoencoding. In ***Proceedings of the 27th International Conference on Computational Linguistics***, pp. 1650–1661, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1140.

Caglayan, O., Barrault, L., and Bougares, F. Multimodal attention for neural machine translation. ***CoRR***, abs/1609.03976, 2016. URL http://arxiv.org/abs/1609.03976.

Calhoun, V. D. and Sui, J. Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. ***Biological Psychiatry: Cognitive Neuroscience and Neuroimaging***, 1(3):230 – 244, 2016. ISSN 2451-9022. doi: https://doi.org/10.1016/j.bpsc.2015.12.005. URL http://www.sciencedirect.com/science/article/pii/S2451902216000598. Brain Connectivity in Psychopathology.

Cheng, J., Dong, L., and Lapata, M. Long short-term memory-networks for machine reading. ***CoRR***, abs/1601.06733, 2016. URL http://arxiv.org/abs/1601.06733.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. ***CoRR***, abs/1409.1259, 2014. URL http://arxiv.org/abs/1409.1259.

Cybenko, G. Approximation by superpositions of a sigmoidal function. ***Mathematics of Control, Signals and Systems***, 2:303–314, 1989.

Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. ***Neural Computation***, 10(7):1895–1923, Oct 1998. ISSN 0899-7667. doi: 10.1162/089976698300017197.

Dwarampudi, M. and Reddy, N. V. S. Effects of padding on lstms and cnns. ***ArXiv***, abs/1903.07288, 2019.

Fountas, Z., Gamez, D., and Fidjeland, A. K. A neuronal global workspace for human-like control of a computer game character. In ***2011 IEEE Conference on Computational Intelligence and Games (CIG'11)***, pp. 350–357, Aug 2011. doi: 10.1109/CIG.2011.6032027.

Gers, F. A., Schmidhuber, J., and Cummins, F. Learning to forget: continual prediction with lstm. In ***1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)***, volume 2, pp. 850–855 vol.2, Sep. 1999. doi: 10.1049/cp:19991218.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. ***CoRR***, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Hinton, G. E., Osindero, S., and Teh, Y.-W. A fast learning algorithm for deep belief nets. ***Neural Comput.***, 18(7): 1527–1554, July 2006. ISSN 0899-7667. doi: 10.1162/neco.2006.18.7.1527. URL `http://dx.doi.org/10.1162/neco.2006.18.7.1527`.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. ***Neural Comput.***, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL `http://dx.doi.org/10.1162/neco.1997.9.8.1735`.

Hori, C., Hori, T., Lee, T., Sumi, K., Hershey, J. R., and Marks, T. K. Attention-based multimodal fusion for video description. ***CoRR***, abs/1701.03126, 2017. URL `http://arxiv.org/abs/1701.03126`.

Hu, D. An introductory survey on attention mechanisms in NLP problems. ***CoRR***, abs/1811.05544, 2018. URL `http://arxiv.org/abs/1811.05544`.

Jensen, M. P. and Karoly, P. Self-report scales and procedures for assessing pain in adults. ***Handbook of pain assessment***, pp. 135–151, 1992.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization, 2014. URL `http://arxiv.org/abs/1412.6980`. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. Handling imbalanced datasets: A review. ***GESTS International Transactions on Computer Science and Engineering***, 30:25–36, 11 2005.

Kumar, M., Garg, D. P., and Zachery, R. A. A method for judicious fusion of inconsistent multiple sensor data. ***IEEE Sensors Journal***, 7(5):723–733, May 2007. ISSN 1530-437X. doi: 10.1109/JSEN.2007.894905.

Lahat, D., Adali, T., and Jutten, C. Multimodal data fusion: An overview of methods, challenges, and prospects. ***Proceedings of the IEEE***, 103(9):1449–1477, Sep. 2015. ISSN 0018-9219. doi: 10.1109/JPROC.2015.2460697.

Lei Ba, J., Ryan Kiros, J., and E. Hinton, G. Layer normalization. ***arXiv***, abs/1607.06450, 07 2016. URL `https://arxiv.org/abs/1607.06450`.

Liberman, Y., Samuels, R., Alpert, P., and Messer, H. New algorithm for integration between wireless microwave sensor network and radar for improved rainfall measurement and mapping. ***Atmospheric Measurement Techniques***, 7(10):3549–3563, 2014. doi: 10.5194/amt-7-3549-2014. URL `https://www.atmos-meas-tech.net/7/3549/2014/`.

Lipton, Z. C. A critical review of recurrent neural networks for sequence learning. ***CoRR***, abs/1506.00019, 2015. URL `http://arxiv.org/abs/1506.00019`.

Liu, H., Wu, Y., Sun, F., Fang, B., and Guo, D. Weakly paired multimodal fusion for object recognition. ***IEEE Transactions on Automation Science and Engineering***, 15(2):784–795, April 2018. doi: 10.1109/TASE.2017.2692271.

Matthews, B. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. ***Biochimica et Biophysica Acta (BBA) - Protein Structure***, 405(2):442 – 451, 1975. ISSN 0005-2795. doi: https://doi.org/10.1016/0005-2795(75)90109-9. URL `http://www.sciencedirect.com/science/article/pii/0005279575901099`.

Miller, A. H., Fisch, A., Dodge, J., Karimi, A., Bordes, A., and Weston, J. Key-value memory networks for directly reading documents. ***CoRR***, abs/1606.03126, 2016. URL `http://arxiv.org/abs/1606.03126`.

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In ***Proceedings of the 27th International Conference on International Conference on Machine Learning***, ICML'10, pp. 807–814, USA, 2010. Omnipress. ISBN 978-1-60558-907-7. URL `http://dl.acm.org/citation.cfm?id=3104322.3104425`.

Olugbade, T. A., Newbold, J. W., Johnson, R. M. G., Volta, E., Alborno, P., Niewiadomski, R., Dillon, M., Volpe, G., and Bianchi-Berthouze, N. Automatic detection of reflective thinking in mathematical problem solving based on unconstrained bodily exploration. ***CoRR***, abs/1812.07941, 2018. URL `http://arxiv.org/abs/1812.07941`.

Olugbade, T. A., Singh, A., Bianchi-Berthouze, N., Marquardt, N., Aung, M. S. H., and Williams, A. C. D. C. How can affect be detected and represented in technological support for physical rehabilitation&#x003f;. ***ACM Trans. Comput.-Hum. Interact.***, 26(1):1:1–1:29, January 2019. ISSN 1073-0516. doi: 10.1145/3299095. URL `http://doi.acm.org/10.1145/3299095`.

Paulus, R., Xiong, C., and Socher, R. A deep reinforced model for abstractive summarization. ***CoRR***, abs/1705.04304, 2017. URL `http://arxiv.org/abs/1705.04304`.

Shanahan, M. A cognitive architecture that combines internal simulation with a global workspace. ***Consciousness and Cognition***, 15(2):433–449, 2006. doi: 10.1016/j.concog.2005.11.005.

Shanahan, M. A spiking neuron model of cortical broadcast and competition. ***Consciousness and Cognition***, 17(1):288 – 303, 2008. ISSN 1053-8100. doi: https://doi.org/10.1016/j.concog.2006.12.005. URL http://www.sciencedirect.com/science/article/pii/S1053810007000050.

Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., and Woo, W. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. ***CoRR***, abs/1506.04214, 2015. URL http://arxiv.org/abs/1506.04214.

Sukhbaatar, S., Szlam, A., Weston, J., and Fergus, R. Weakly supervised memory networks. ***CoRR***, abs/1503.08895, 2015. URL http://arxiv.org/abs/1503.08895.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In ***Advances in neural information processing systems***, pp. 5998–6008, 2017. URL https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In ***Proceedings of the 25th International Conference on Machine Learning***, ICML '08, pp. 1096–1103, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390294. URL http://doi.acm.org/10.1145/1390156.1390294.

Wang, C., Olugbade, T. A., Mathur, A., de C. Williams, A. C., Lane, N. D., and Bianchi-Berthouze, N. Automatic detection of protective behavior in chronic pain physical rehabilitation: A recurrent neural network approach. ***CoRR***, abs/1902.08990, 2019. URL http://arxiv.org/abs/1902.08990.

Weston, J., Chopra, S., and Bordes, A. Memory networks. ***CoRR***, abs/1410.3916, 2015.

Wilcoxon, F. Individual comparisons by ranking methods. ***Biometrics Bulletin***, 1(6):80–83, 1945. ISSN 00994987. URL http://www.jstor.org/stable/3001968.

Wilderjans, T., Ceulemans, E., Van Mechelen, I., and van den Berg, R. Simultaneous analysis of coupled data matrices subject to different amounts of noise. ***The British journal of mathematical and statistical psychology***, 64:277–90, 05 2011. doi: 10.1348/000711010X513263.

imekli, U., Ermi, B., Cemgil, A. T., and Acar, E. Optimal weight learning for coupled tensor factorization with mixed divergences. In ***21st European Signal Processing Conference (EUSIPCO 2013)***, pp. 1–5, Sep. 2013.