Multimodal Data Fusion based on the Global Workspace Theory

Cong Bao
cong.bao.18@ucl.ac.uk
Department of Computer Science, University College London
London, United Kingdom

Zafeirios Fountas
f@emotech.co
Emotech Labs
Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London
London, United Kingdom

Temitayo Olugbade
temitayo.olugbade.13@ucl.ac.uk
University College London (UCL) Interaction Centre, UCL
London, United Kingdom

Nadia Bianchi-Berthouze
nadia.berthouze@ucl.ac.uk
University College London (UCL) Interaction Centre, UCL
London, United Kingdom

ABSTRACT

We propose a novel neural network architecture, named the Global Workspace Network (GWN), which addresses the challenge of dynamic and unspecified uncertainties in multimodal data fusion. Our GWN is a model of attention across modalities and evolving through time, and is inspired by the well-established Global Workspace Theory from the field of cognitive science. The GWN achieved average F1 score of 0.92 for discrimination between pain patients and healthy participants and average F1 score = 0.75 for further classification of three pain levels for a patient, both 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 these tasks, the GWN significantly outperforms the typical fusion approach of merging by concatenation. We further provide extensive analysis of the behaviour of the GWN and its ability to address uncertainties (hidden noise) in multimodal data.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**; **Neural networks**.

KEYWORDS

Machine Learning; Multimodal Fusion; Global Workspace Theory

ACM Reference Format:
Cong Bao, Zafeirios Fountas, Temitayo Olugbade, and Nadia Bianchi-Berthouze. 2020. Multimodal Data Fusion based on the Global Workspace Theory. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3382507.3418849

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418849

1 INTRODUCTION

Reasoning about and interpreting multiple sources of information concurrently is an important task in machine learning research as life involves streaming of data from multiple modalities [9]. Multimodal data fusion, which leverages the combination of multiple modalities, is a valuable strategy [3, 14, 26, 38]. Its benefits include complementarity of information, higher prediction performance, and robustness [9]. However, multimodal fusion comes with challenges; [32] specifies them under two categories: (1) challenges of multimodal data acquisition, and (2) uncertainties (such as noisy modalities, missing values, conflicting information) in multimodal data. The former type of challenges could be managed with later pre-processing, e.g. resampling to reconcile different temporal resolutions across modalities [4]. However, addressing uncertainties in multimodal data requires specialised design of models that can exploit complementarity or discrepancy across modalities [32]. While there have been approaches such as [58] that address the particular problem of missing modalities, fusion of multimodal data with varying types or levels of uncertainty (e.g. noise) which are not known apriori has been less investigated. Findings of the efficacy of automatic learning of weights (e.g. some "importance" or "confidence" metric) for individual input features [1, 31, 36, 57, 60], the basis of attention mechanisms in machine learning [8], suggests that this may be a more relevant approach to factoring uncertainties into multimodal data fusion. However, while uncertainty also evolves through time [32], the typical attention approach has been unidimensional, i.e. attention across modalities alone or attention over time within individual modalities, e.g. in [11]. Few studies have explored the propagation of attention across modalities through time. The memory fusion network of [59] which is based on a cross-modality attention module with a memory is one of such rare cases.

To address this gap in multimodal data fusion, we propose the Global Workspace Network (GWN) which, like [59], propagates cross-modality attention through time. However, unlike previous work, the GWN further addresses the problem of differences in feature dimensionalities of the modalities via a common feature space, based on pre-trained autoencoders. In addition, different from [59], our approach is bio-inspired (grounded in the Global Workspace Theory [6, 7]) and we implement the GWN's cross-modality attention using the widely-tested transformer architecture [52].

The Global Workspace Theory (GWT) is a well-developed framework (originally proposed as a model of human consciousness [5]) in cognitive science. The GWT states that concomitant cognitive processes *compete* for the opportunity to *broadcast* their current state (to peer processes) [21]. 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) [49]. This competition and broadcast cycle is believed to be ubiquitous in the perceptual regions of the brain [5]. Although the literature on GWT includes architectures of biologically-realistic spiking neural networks [21, 49], to our knowledge, there has been no direct implementation in machine learning. For such implementation, the GWT can be conceptualised 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, the external memory stores information for later use [48]. By considering 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 becomes clear. 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 our implementation of the GWN, the transformer [52] was leveraged to simulate the compete-and-broadcast component of the GWT, and the Long Short-Term Memory (LSTM) neural network [22, 25] as its external memory. There are 3 key elements of transformers that illustrate their advantage and relevance to the current task. First is a self-attention mechanism [15, 47] 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. A second merit is the transformer's bagging approach, where multiple attention patterns are learnt in parallel, with the advantage of increased robustness. Finally, a third valuable attribute is its memory-based structure [51, 55]. Drawing from traditional applications in Natural Language Processing question answering tasks [43, 51], this unit further maps the feature vector into query, key, and value spaces to increase the weighting depth and robustness [27]. 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. As for the external memory module, in contrast to the use of a custom two-gated recurrent network in [59], we used the well-established LSTM which has two additional gates [37]. Finally, unlike [59], we provide extensive analysis of the behaviour of the GWN in the presence of varying degrees of uncertainties across modalities and over time.

The contribution of this paper is the GWN architecture which we propose as an approach to fusion of sequential data from multiple modalities. We evaluate the architecture on the EmoPain dataset [4], 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. While the EMG has four feature dimensions, the motion capture data comprises 78 dimensions. Further, we provide 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-based machine learning in Section 2. We then describe in Section 3 the proposed GWN architecture that builds on these and present both validation and analysis of the network in Section 4. Section 5 concludes the paper.

2 RELATED WORK

As earlier-stated, there have been different approaches to multimodal fusion. For example, [33] simply concatenated vectors from individual encoders for each modality. The architecture of [58], which was mainly tested on non-sequential inputs, learns both individual encodings as well as a common encoding for the different modalities. For the joint encoding in [58], the individual encodings are merged by multiplication. Rather than cover the literature on multimodal data fusion, we refer the reader to [10] for a comprehensive review and focus our discussion here on attention-based approaches to multimodal data fusion.

Attention over time in multimodal fusion. In the literature on neural networks for multimodal data, attention performed on the time axis is usually done separately for each modality, and the resulting context vectors from each modality are then fused as non-temporal features. A representative case of this approach is the Recursive Recurrent Neural Network (RRNN) architecture proposed by [11]. 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 Gated Recurrent Unit (GRU) cell [16]. The outcome is then used to update the state of the 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 affect the final prediction equally despite the fact that some modalities could be more noisy than others. Thus, the challenge of the dynamics of uncertainty across modalities remains unsolved.

Attention across multiple modalities. Several studies have modelled the relation between modalities in multimodal fusion. The typical approach [36, 57, 60] is the use of modality weighting although not particularly based on attention mechanisms [8]. One study that does explicitly use the attention mechanism is the work of [26] on automatic video description. Their approach leverages attention between different modalities using an encoder-decoder architecture [8] 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 [13] applies 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 [23] and bi-directional GRU neural networks [16] respectively. Then, attention scores were computed for these encodings. More recently, authors of [41] place an attention layer on top of several modality-specific feature encoding layers to model the importance of different modalities in book genre prediction. There are many other works [20, 35, 39, 40] that leverage this technique, i.e. encoding sequential/temporal data for each modality before computing attention weighting and fusing encoded modality-specific features. While it is appropriate for obtaining modality-specific feature representation, it does not allow in-depth quantification of the complex interactions between modalities through time.

Attention across modalities and through time. As discussed in the introduction, [59] addresses the limitation of attention over time alone or across modality only by considering both the interaction of multiple modalities and the temporal variations in this interaction. Their architecture is based on separate time encoding of individual modalities. A cross-modality attention is then computed and applied for each time slice. Instead of a single time step per slice, each slice consists of successive time steps t and $t-1$. The weighted multimodal encodings for a given time slice are then fed into a memory module with retain and update gates which are based on neural networks that have the encodings as input. A recurrent update is done using the gate outputs, the previous memory state, and the proposed memory state which is also the output of a neural network computation on the encodings. The findings of [59] in a set of ablation studies suggest that propagation of attention through time improves prediction performance. The GWN architecture that we propose makes further advance with implementation of the cross-modality attention module based on the self-attending, multi-head attention transformer architecture [52]. The GWN additionally addresses the confounding challenge of different feature and/or temporal dimensionalities across the modalities to be fused. While [59] evaluate their model on data with such characteristic, they do not clarify how their architecture deals with this. In the GWN, we take the approach of learning a common dimensionality across modalities. Based on further controlled experiments, we also contribute analysis of the effect of noise in one of the modalities.

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 $\{\boldsymbol{x}_1^{(m)}, \ldots, \boldsymbol{x}_T^{(m)}\}$, 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.

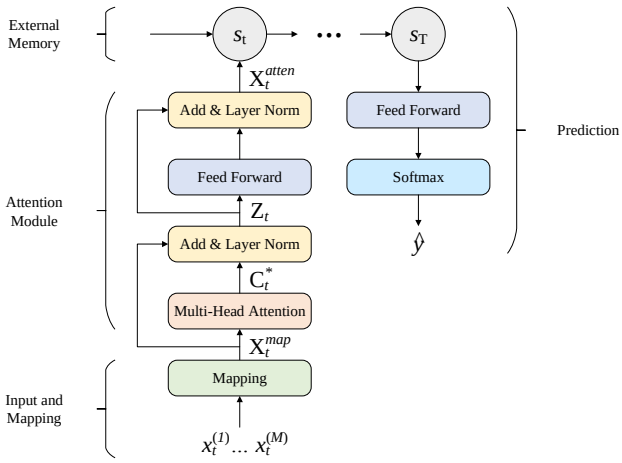


Figure 1: The architecture of the GWN. Here the intermediate matrices $\mathbf{X}_t^{map}$, $\mathbf{C}_t^*$, $\mathbf{Z}_t$, and $\mathbf{X}_t^{atten}$ have the same dimensionality of $M \times H$.

Inspired by the work of [2] and [12], we take the approach of using multiple autoencoders [53] 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) [44] 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 [17] 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 [12]. 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.

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

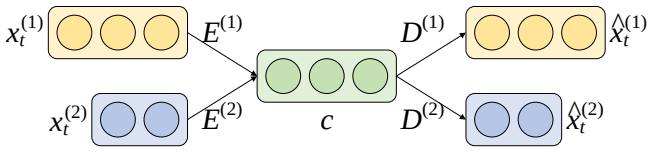


Figure 2: An illustration of the mapping module with two modalities.

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 [24] suggest that unsupervised pre-training on shallow layers can improve the performance of a deep network.

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 [52] 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 [43]. 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 [23] in the attention module. Each of the residual connection is followed by a layer normalisation [34]. 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 [25] 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 $\mathbf{X}_t^{atten} \in \mathbb{R}^{M \times H}$; $\sigma(\cdot)$, $\tanh(\cdot)$, and $\odot$ are sigmoid, hyperbolic tangent, and Hadamard product (i.e. element-wise product) functions respectively. $\boldsymbol{s}_t \in \mathbb{R}^{2G}$ 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 [22, 25]. 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

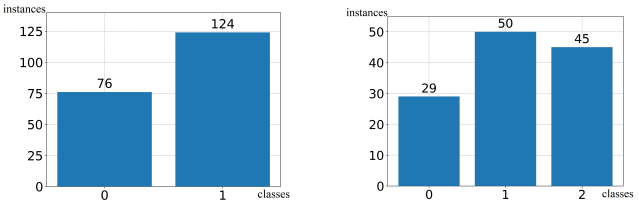
$$\boldsymbol{r} = \max\left(0, \boldsymbol{h}_T\mathbf{W}_1 + \boldsymbol{b}_1\right)\mathbf{W}_2 + \boldsymbol{b}_2 \tag{18}$$

$$\hat{\boldsymbol{y}} = \text{softmax}\left(\boldsymbol{r}\right) \tag{19}$$

i.e.

$$\hat{y}_i = \frac{\exp\left(r_i\right)}{\sum_j \exp\left(r_j\right)} \tag{20}$$

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.



(a) (b)

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

4 EXPERIMENTS

To evaluate the proposed GWN architecture, we conducted experiments on the multimodal EmoPain dataset [4]. The dataset, data preprocessing, and experiment tasks are introduced in Section 4.1. Section 4.2 describes the baseline model used for comparison and the methods and metrics of this evaluation. 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 [4] is suitable for exploring the GWN architecture 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 at preferred pace, (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) holding a 2 kg dumbbell while reaching forward, (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) [28]. 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.

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 [46].

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 downsampled 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 [19, 54], 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 [50].

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 oversampled data instances of the minority class [30].

Data Augmentation. The total number of exercise instances available for training and evaluation was 200, which is a limited amount

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 our 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.

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 [45]. 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

$$\boldsymbol{x}_t^* = \text{concat}\left(\boldsymbol{x}_t^{(1)}, \ldots, \boldsymbol{x}_t^{(M)}\right) \tag{21}$$

$$\boldsymbol{c}_t, \boldsymbol{h}_t = \text{lstm}\left(\boldsymbol{x}_t^*, \boldsymbol{c}_{t-1}, \boldsymbol{h}_{t-1}\right) \tag{22}$$

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 [44] non-linearity.

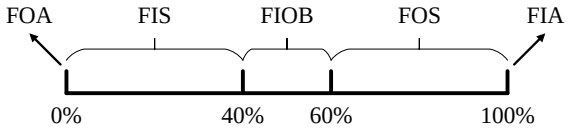


Figure 4: The percentage of itself that a modality pays attention to in the five different attention patterns we found. The thresholds 40% and 60% used in this definition were chosen heuristically as a $\pm 10\%$ interval around 50%.

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 lack of independence between folds (due to overlapping training sets across folds) that has higher risk of Type I error [18]. 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 [18] for the purpose of model comparison. The advantage of the 2-fold CV is that there is no overlap between training sets.

For both LOSOCV and 5×2 CV, we perform Wilcoxon signed-rank test [56] 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 Adam optimisation algorithm [29], learning rate = 0.001, and batch size = 32, which were chosen by grid search. The dimensionality of each LSTM cell are also kept the same, i.e. 64, for the two models. The performance of the GWN can be seen in Table 1 showing comparison with the CONCATN baseline model, based on accuracy (ACC), Matthews Correlation Coefficient (MCC) [42], and F1 scores.

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.

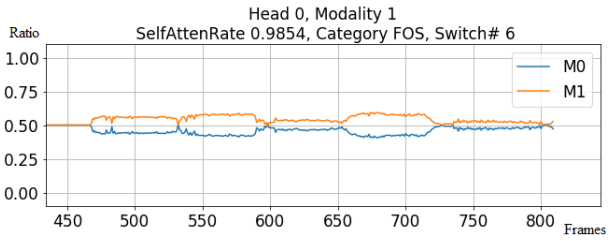


Figure 5: An example of the attention distribution of one exercise instance. Head 0 means the first attention head. Modality 0 (M0) represents MC and modality 1 (M1) represents EMG.

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 patterns of its attention scores

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

$\forall t$ ($\mathbf{a}_t^k$ is one of the terms in equation 6 and $\mathbf{a}_t^k \in \mathbb{R}^{M\times M}$) 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.

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.

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 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.

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, $\sigma_{\text{noise}} = 10$ (rounded to the nearest one significant figure number). 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 the EMG data 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 and so since the dimensionalities of MC and EMG are 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

We propose the GWN, a novel neural network architecture for multimodal fusion of sequential, multimodal data. Drawing from the Global Workspace Theory, at each time step of the GWN, multiple modalities compete to broadcast information, and each broadcast is propagated through time. We find that this approach outperforms simply concatenating multiple modalities, for pain level detection based on the EmoPain dataset. Our analysis further highlights the modality selectivity that occurs in the GWN for this dataset. Moreover, controlled experiments with simulated noise suggest that the GWN addresses uncertainty and its variation over time. This could be a promising direction for future research in multimodal neural networks while promoting a close connection with cognitive neuroscience research. Such interdisciplinary links may be valuable in consolidating the myriad of advances in both communities.

ACKNOWLEDGMENTS

The project was partially supported by the Future and Emerging Technologies (FET) Proactive Programme H2020-EU.1.2.2 (Grant agreement 824160; EnTimeMent). It also received support from Emotech Ltd.

REFERENCES

[1] Evrim Acar, Daniel M. Dunlavy, Tamara G. Kolda, and Morten Mørup. 2011. Scalable Tensor Factorizations for Incomplete Data. *Chemometrics and Intelligent Laboratory Systems* 106 (03 2011), 41–56. https://doi.org/10.1016/j.chemolab.2010.08.004

[2] Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. 2018. Multi-level Multimodal Common Semantic Space for Image-Phrase Grounding. *CoRR* abs/1811.11683 (2018). arXiv:1811.11683 http://arxiv.org/abs/1811.11683

[3] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal Fusion for Multimedia Analysis: A Survey. *Multimedia Syst.* 16, 6 (Nov. 2010), 345–379. https://doi.org/10.1007/s00530-010-0182-0

[4] Min S. H. Aung, Sebastian Kaltwang, Bernardino Romera-Paredes, Brais Martinez, Aneesha Singh, Matteo Cella, Michel Valstar, Hongying Meng, Andrew Kemp, Moshen Shafizadeh, Aaron C. Elkins, Natalie Kanakam, Amschel de Rothschild, Nick Tyler, Paul J. Watson, Amanda C. de C. Williams, Maja Pantic, and Nadia Bianchi-Berthouze. 2016. The Automatic Detection of Chronic Pain-Related Expression: Requirements, Challenges and the Multimodal EmoPain Dataset. *IEEE Trans. Affect. Comput.* 7, 4 (Oct. 2016), 435–451. https://doi.org/10.1109/TAFFC.2015.2462830

[5] Bernard J. Baars. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press, Cambridge, MA.

[6] Bernard J. Baars. 1997. *In the Theater of Consciousness*. Oxford University Press, New York, NY.

[7] Bernard J. Baars. 2002. The Conscious Access Hypothesis: Origins and Recent Evidence. *Trends in Cognitive Sciences* 6, 1 (2002), 47–52. https://doi.org/10.1016/s1364-6613(00)01819-2

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. 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*. http://arxiv.org/abs/1409.0473

[9] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal Machine Learning: A Survey and Taxonomy. *CoRR* abs/1705.09406 (2017). arXiv:1705.09406 http://arxiv.org/abs/1705.09406

[10] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[11] Rory Beard, Ritwik Das, Raymond W. M. Ng, P. G. Keerthana Gopalakrishnan, Luka Eerens, Pawel Swietojanski, and Ondrej Miksik. 2018. Multi-Modal Sequence Fusion via Recursive Attention for Emotion Recognition. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Brussels, Belgium, 251–259. https://www.aclweb.org/anthology/K18-1025

[12] Danushka Bollegala and Cong Bao. 2018. Learning Word Meta-Embeddings by Autoencoding. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1650–1661. https://www.aclweb.org/anthology/C18-1140

[13] Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal Attention for Neural Machine Translation. *CoRR* abs/1609.03976 (2016). arXiv:1609.03976 http://arxiv.org/abs/1609.03976

[14] Vince D. Calhoun and Jing Sui. 2016. 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 (2016), 230 – 244. https://doi.org/10.1016/j.bpsc.2015.12.005 Brain Connectivity in Psychopathology.

[15] Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long Short-Term Memory-Networks for Machine Reading. *CoRR* abs/1601.06733 (2016). arXiv:1601.06733 http://arxiv.org/abs/1601.06733

[16] KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *CoRR* abs/1409.1259 (2014). arXiv:1409.1259 http://arxiv.org/abs/1409.1259

[17] George Cybenko. 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems* 2 (1989), 303–314.

[18] T. G. Dietterich. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation* 10, 7 (Oct 1998), 1895–1923. https://doi.org/10.1162/089976698300017197

[19] Mahidhar Dwarampudi and N. V. Subba Reddy. 2019. Effects of padding on LSTMs and CNNs. *ArXiv* abs/1903.07288 (2019).

[20] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous Memory Enhanced Multimodal Attention Model for Video Question Answering. *CoRR* abs/1904.04357 (2019). arXiv:1904.04357 http://arxiv.org/abs/1904.04357

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

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

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[24] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput.* 18, 7 (July 2006), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[26] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Kazuhiro Sumi, John R. Hershey, and Tim K. Marks. 2017. Attention-Based Multimodal Fusion for Video Description. *CoRR* abs/1701.03126 (2017). arXiv:1701.03126 http://arxiv.org/abs/1701.03126

[27] Dichao Hu. 2018. An Introductory Survey on Attention Mechanisms in NLP Problems. *CoRR* abs/1811.05544 (2018). arXiv:1811.05544 http://arxiv.org/abs/1811.05544

[28] Mark P. Jensen and Paul Karoly. 1992. Self-report scales and procedures for assessing pain in adults. *Handbook of pain assessment* (1992), 135–151.

[29] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. 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.

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

[31] M. Kumar, D. P. Garg, and R. A. Zachery. 2007. A Method for Judicious Fusion of Inconsistent Multiple Sensor Data. *IEEE Sensors Journal* 7, 5 (May 2007), 723–733. https://doi.org/10.1109/JSEN.2007.894905

[32] D. Lahat, T. Adali, and C. Jutten. 2015. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proc. IEEE* 103, 9 (Sep. 2015), 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697

[33] Michelle A. Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. 2018. Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks. *CoRR* abs/1810.10191 (2018). arXiv:1810.10191 http://arxiv.org/abs/1810.10191

[34] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. *arXiv* abs/1607.06450 (07 2016). https://arxiv.org/abs/1607.06450

[35] Aske Rasch Lejbølle, Benjamin Krogh, Kamal Nasrollahi, and Thomas B. Moeslund. 2018. Attention in Multimodal Neural Networks for Person Re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW))*. IEEE, United States, 292–300. https://doi.org/10.1109/CVPRW.2018.00055 IEEE Conference on Computer Vision and Pattern Recognition, 2018, IEEE CVPR 2018 ; Conference date: 18-06-2018 Through 22-06-2018.

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

[37] Zachary Chase Lipton. 2015. A Critical Review of Recurrent Neural Networks for Sequence Learning. *CoRR* abs/1506.00019 (2015). arXiv:1506.00019 http://arxiv.org/abs/1506.00019

[38] H. Liu, Y. Wu, F. Sun, B. Fang, and D. Guo. 2018. Weakly Paired Multimodal Fusion for Object Recognition. *IEEE Transactions on Automation Science and Engineering* 15, 2 (April 2018), 784–795. https://doi.org/10.1109/TASE.2017.2692271

[39] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual Attention Model for Name Tagging in Multimodal Social Media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1990–1999. https://doi.org/10.18653/v1/P18-1185

[40] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering. *CoRR* abs/1606.00061 (2016). arXiv:1606.00061 http://arxiv.org/abs/1606.00061

[41] Suraj Maharjan, Manuel Montes, Fabio A. González, and Thamar Solorio. 2018. A Genre-Aware Attention Model to Improve the Likability Prediction of Books. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3381–3391. https://doi.org/10.18653/v1/D18-1375

[42] B.W. Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405, 2 (1975), 442 – 451. https://doi.org/10.1016/0005-2795(75)90109-9

[43] Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-Value Memory Networks for Directly Reading Documents. *CoRR* abs/1606.03126 (2016). arXiv:1606.03126 http://arxiv.org/abs/1606.03126

[44] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning* (Haifa, Israel) *(ICML'10)*. Omnipress, USA, 807–814. http://dl.acm.org/citation.cfm?id=3104322.3104425

[45] Temitayo A. Olugbade, Joseph W. Newbold, Rose M. G. Johnson, Erica Volta, Paolo Alborno, Radoslaw Niewiadomski, Max Dillon, Gualtiero Volpe, and Nadia Bianchi-Berthouze. 2018. Automatic Detection of Reflective Thinking in Mathematical Problem Solving based on Unconstrained Bodily Exploration. *CoRR* abs/1812.07941 (2018). arXiv:1812.07941 http://arxiv.org/abs/1812.07941

[46] Temitayo A. Olugbade, Aneesha Singh, Nadia Bianchi-Berthouze, Nicolai Marquardt, Min S. H. Aung, and Amanda C. De C. Williams. 2019. How Can Affect Be Detected and Represented in Technological Support for Physical Rehabilitation?. *ACM Trans. Comput.-Hum. Interact.* 26, 1, Article 1 (Jan. 2019), 29 pages. https://doi.org/10.1145/3299095

[47] Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A Deep Reinforced Model for Abstractive Summarization. *CoRR* abs/1705.04304 (2017). arXiv:1705.04304 http://arxiv.org/abs/1705.04304

[48] Murray Shanahan. 2006. A Cognitive Architecture That Combines Internal Simulation with a Global Workspace. *Consciousness and Cognition* 15, 2 (2006), 433–449. https://doi.org/10.1016/j.concog.2005.11.005

[49] Murray Shanahan. 2008. A spiking neuron model of cortical broadcast and competition. *Consciousness and Cognition* 17, 1 (2008), 288 – 303. https://doi.org/10.1016/j.concog.2006.12.005

[50] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *CoRR* abs/1506.04214 (2015). arXiv:1506.04214 http://arxiv.org/abs/1506.04214

[51] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. Weakly Supervised Memory Networks. *CoRR* abs/1503.08895 (2015). arXiv:1503.08895 http://arxiv.org/abs/1503.08895

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008. https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[53] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and Composing Robust Features with Denoising Autoencoders. In *Proceedings of the 25th International Conference on Machine Learning* (Helsinki, Finland) *(ICML '08)*. ACM, New York, NY, USA, 1096–1103. https://doi.org/10.1145/1390156.1390294

[54] Chongyang Wang, Temitayo A. Olugbade, Akhil Mathur, Amanda C. de C. Williams, Nicholas D. Lane, and Nadia Bianchi-Berthouze. 2019. Automatic Detection of Protective Behavior in Chronic Pain Physical Rehabilitation: A Recurrent Neural Network Approach. *CoRR* abs/1902.08990 (2019). arXiv:1902.08990 http://arxiv.org/abs/1902.08990

[55] Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory Networks. *CoRR* abs/1410.3916 (2015).

[56] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83. http://www.jstor.org/stable/3001968

[57] Tom Wilderjans, Eva Ceulemans, Iven Van Mechelen, and Robert van den Berg. 2011. Simultaneous analysis of coupled data matrices subject to different amounts of noise. *The British journal of mathematical and statistical psychology* 64 (05 2011), 277–90. https://doi.org/10.1348/000711010X513263

[58] Mike Wu and Noah Goodman. 2018. Multimodal generative models for scalable weakly-supervised learning. In *32nd Conference on Neural Information Processing Systems*. 5575–5585.

[59] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[60] U. Şimşekli, B. Ermiş, A. T. Cemgil, and E. Acar. 2013. Optimal weight learning for Coupled Tensor Factorization with mixed divergences. In *21st European Signal Processing Conference (EUSIPCO 2013)*. 1–5.