Universität Hamburg
Department Informatik
Knowledge Technology, WTM

Comparing Methods of Multimodal Fusion for Wearable Stress and Affect Detection

Seminar Paper

Bio-Inspired Artificial Intelligence

Jakob Ambsdorf & Tassilo Hahm
Matr.Nr. 6919840 & 6917023
6ambsdor@informatik.uni-hamburg.de
6hahm@informatik.uni-hamburg.de

28.01.2021

Abstract

Using the WESAD Dataset for wearable stress and affect detection, we compare different methods of multi-modal fusion in neural networks. Results employing the Gated Multimodal Fusion model, linear sum and concatenation are presented and compared against AdaBoost as a baseline. In our experiment, the neural fusion methods achieved similar performance, while being substantially better than the AdaBoost baseline. When introducing random noise on one modality at inference time, performance of all methods dropped, but the models using the Gated Fusion module deteriorated most notably, resulting in below chance-level accuracy.

Contents

1 Introduction

Multi-modal fusion for neural networks has been a subject of interest in recent times. As humans we usually do not rely on a single sensory modality to make decisions[8]. Looking at fruit for example, to make the decision if it is edible we will look at its color, feel if it is soft to the touch and smell if it is ripe. We then combine these different inputs to come to a conclusion. Next to the basic fusion of information from different modalities, what this example also shows is the varying importance of modalities depending on context[8] (or in this case the specific fruit). A banana, when edible, changes in color from green to yellow, becomes softer and smells sweeter while an avocado barely smells, goes from a brighter green to a darker shade, but the main indicator of ripeness being its softness. Or another example: A tomato changes in color from green to red, basically does not change its smell and only becomes slightly softer when edible. But what do we do when the banana is faintly yellow, still hard and does not smell yet? What would be our decision if the avocado is dark green but still pretty hard? And how would we assess a very soft but green tomato? So it seems sometimes we do not want to just fuse sensory modalities, we also want to do so in some weighted way, giving a higher importance to certain modalities but also being able to override those modalities when others clearly indicate something else.

In this paper we concern ourselves with a new building block for neural network models that aims to achieve exactly this: The Gated Multimodal Unit (GMU)[1]. Previously existing methods of multi-modal fusion work by either averaging over results from each modality, adding the respective linearized outputs or simply concatenating inputs altogether. The GMU however ads a layer of complexity that lets a model make weighted decisions based on the learned importance of modalities for certain tasks. Here we want to compare how the GMU compares to the simpler methods on a given dataset.

We apply our model to the Multimodal Dataset for Wearable Stress and Affect Detection (WESAD)[11], a real world collection of multi-modal (physiological and motion) inputs to detect stress and affect, and compare it to the performance of the previously described methods of multi-modal fusion and to the baseline of basic machine learning methods described in[1]. Early stress detection by wearable devices could play an important role in preventing related health problems [17] and has been an active research topic in recent years.

1.1 Stress and Affect Detection

With the advent of IoT devices and smart watches, physiological data has become more readily available [15]. Interpretation of such data in order to detect stress and other affects, with the goal of preventive measures against stress related health problems has been an active topic of research in recent years. While many approaches employed traditional machine learning models, feature engineering and feature selection and/or statistical methods [14][17], neural networks and deep learning have shown promising solutions to predicting stress [12]. However,

real world scenarios are still difficult, due to a multitude of reasons, such as context dependence and interpersonal differences, mainly pointing to the need of larger scale, longitudinal studies [13].

1.2 Multimodal Fusion

Multimodal fusion can be broadly divided into two categories[2]: *early* and *late* fusion.

Early fusion tries to combine features or input modalities at an early level to output a single vector that is then further analyzed. All of the presented methods here are early fusion as we take the different modalities and combine them in different ways to present them to a classifier.

Late fusion utilizes decision units that make decisions based on single a modality and then fuses the decisions in some way. It can therefore be thought of as a way of combining unimodal decision models.

There are also some researchers that try to combine both in what is called *hybrid fusion* (e.g. [7] [18]).

1.3 Related Work

There has been some sparse research into applying neural network methods to the WESAD dataset: Reiss et al.[9] used convolutional neural networks and Di Martino et al.[4] used LSTMs. Very recently Sarkar et al.[10] employed self-supervised learning with apparently great success. To the best of our knowledge however no one has applied the novel GMU cell or for that matter any kind of dedicated neural network multimodal fusion method to WESAD.

The GMU has been used in a few different contexts, eg. action recognition[6], conversational systems[19], or event localization[16]. Again to the best of our knowledge it has never been used for emotion recognition and ours is a novel approach. In this context it needs to be said that the original implementation provided by Arevalo et al.[1] only offers the fusion of two modalities. Because of the way it is structured, making use of a convex combination of the two modalities, it cannot simply be extended, meaning researchers using the model either limited themselves to two modalities or had to write their own implementation that is often not public. With this in mind, performance comparisons in different contexts may be of varying quality, depending on the respective implementation chosen or written by the authors.

1.4 Gated Multimodal Unit

The GMU is a unit for multimodal fusion that is, similar to GRUs[3] or LSTMs[5], using gates to process its inputs. However it is not a recurrent cell and therefore not aimed at working on sequential or temporal data. It rather combines the feature vectors stemming from different modalities in a weighted manner.

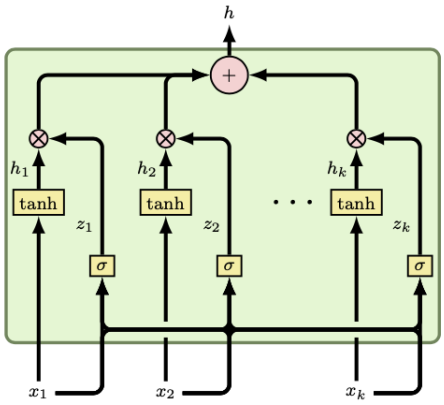


Figure 1: GMU[1]

A GMU is completely described by the following set of equations:

$$
\begin{aligned}
h_i &= tanh(W_i x_i) \\
z_i &= \sigma(W_{z_i}[x_1, ..., x_n]) \\
h &= z_1 h_1 + ... + z_n h_n
\end{aligned}
$$

Each feature vector x_i from the respective modality is fed into one node h_i, gated with a respective z_i function that receives the input from all modalities concatenated and then all gated hidden features are summed for the output activation h. This way, inputs can be weighted by their importance and contribute to the output in different amounts, analogous to a summation of the inputs that is weighted by the corresponding gates. The authors of the GMU also point out, that by gating the input modalities, noisy or corrupt channels can be effectively switched off, possibly making the GMU an effective fusion method in such situations.

2 Approach

2.1 WESAD Dataset

The WESAD Dataset[11] is a collection of data from 15 participants, employing two avenues of measurement: A wrist- and a chest-worn device, both respectively measuring a multitude of modalities at a rate of 700 Hz. The researchers exposed the subjects to different stimuli and elicited three states: neutral (or baseline), stress and amusement. Additionally, there is a rest condition after the stress condition, two meditation conditions and breaks between the conditions where the participants are filling out self-reports.

For this study, we have focused only on the physiological data of the chest-device, consisting of electrocardiogram (ECG), skin conductance (EDA), electromyography (EMG), respiratory rate (RESP), and body temperetaure (TEMP).

2.2 Data Preprocessing

In the dataset, the associated condition is labeled per-measurement with 0-7. Additionally to the labeling of the conditions baseline, stress and amusement (1-3), there are also labels for the meditation condition (4) and an "undefined" condition (0). The remaining labels (5-7) have no further explanation and are stated to be ignored. In order to clean up the dataset, conditions 0 and 5-7 have been removed. Unfortunately, we were unable to find any indication as to how the authors of the dataset treated the meditation condition in their experiments[11]. Furthermore we expected the condition to contain data resembling both the excitation conditions and the baseline condition, making differentiation of conditions harder, which is why we decided to also removed it.

2.3 Model

In Figure 2, a schematic overview of the model architecture is presented for the example of using the GMU for modality fusion.

For each modality, a feature extraction module is be created (labeled "Input Extractor in 2. It consists of three linear layers, with ReLU activation functions in between. In our concrete example, we have five modalities and, therefore, five input extractors with a vector length of 42000 in the first layer, according to the size of the sliding window. After that, layer sizes of 512 and 256 are used to extract relevant features from the large input window.

After the feature extraction, the modalities are fused. Here, the different methods that we compare are used (GMU, concatenation and linear sum). Concatenation and summation by itself are only arithmetic operations on the feature vectors without introducing any trainable parameters, contrary to the GMU, which added about two million parameters in our example. As a consequence, to allow for a fair comparison, we inserted an additional linear layer after the modality fusion and also adjusted the number of parameters used in the classification module afterwards.

The final classification happens in three stages. For the example of the GMU, two Linear layers with a size of 256 have been used with a ReLU activation in between, followed by another linear layer with an output size of three, for a one-hot encoding of the three conditions (neutral, stressed, amused).

To avoid overfitting and allow for better generalization of the model, two dropout layers where introduced, one after the very first layer of each input extractor and one before the the last layer in the classifier.

2.4 Experiment

We trained the model on the WESAD Dataset using leave-one-subject-out cross-validation (as described in [11]), thereby using the data of one subject as the validation set, training 15 models, one for subjects in the dataset. We then took the

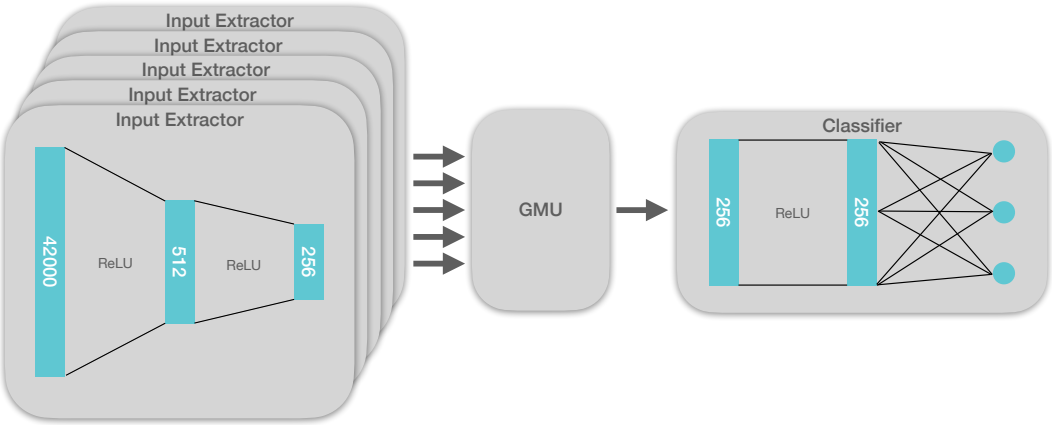


Figure 2: Overview of the model architecture, using the GMU for multimodal fusion.

mean value for the accuracy and F_1-score and calculated the standard deviation. This procedure simulates how the model would generalize to a novel subject.

We used only the physiological data from the chest-worn device, consisting of electrocardiogram (ECG), skin conductance (EDA), electromyography (EMG), respiratory rate (RESP), and body temperetaure (TEMP). With this data, the original authors achieved the best results using an AdaBoost classifier, which we use as a baseline for our results[11]. Additionally, this data is already synchronized and all modalities have the same length for their input sequences, which simplifies data loading and model architecture for our test scenario.

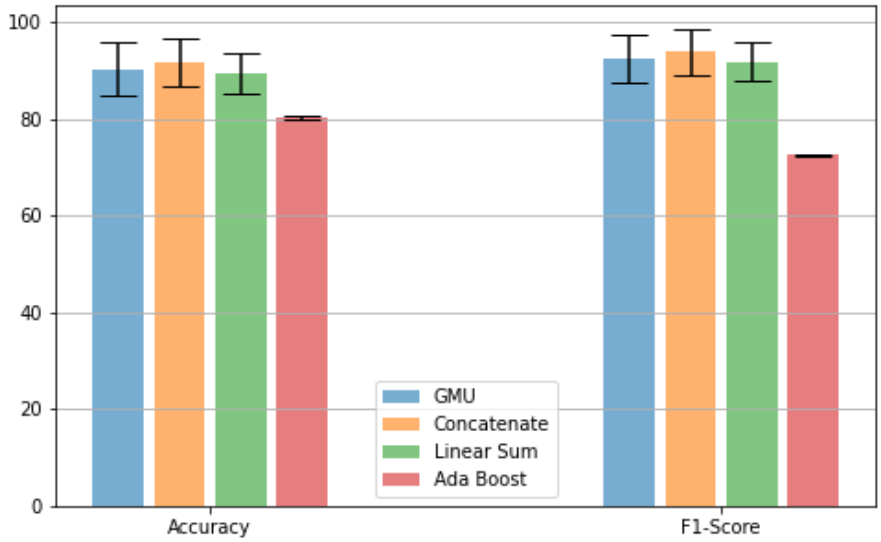
Each model was trained for eight epochs, as we observed overfitting when trained beyond this point in initial experiments. Stochastic gradient descent was employed as an optimizer, using a learning rate of 0.001 and a momentum of 0.9. These hyperparameters, as well as the specific model configuration have been selected after an initial, exploratory search.

3 Results

We present the results of our model using the GMU, linear sum and concatenation and compare the performance of the three methods on the WESAD data. We also include the original authors results as a baseline. Accuracy on the validation set, as well as F1-Scores will be presented.

It can be easily deduced from Figure 3 that the neural multimodal fusion methods perform substantially better than the classical ML methods, while the different variations of the network achieve approximately the same performance. The "Con-

Figure 3: Mean accuracy and F_1-score results using leave-one-subject-out cross validation, error bars are indicating standard deviation.



Measure	GMU	Concatenate	Linear Sum	Baseline (AdaBoost)
Accuracy	90.33 ± 5.56	**91.77** ± 4.89	89.49 ± 4.10	80.34 ± 0.43
F_1-Score	92.40 ± 4.89	**93.79** ± 4.67	91.85 ± 4.16	72.51 ± 0.17

Table 1: Mean accuracy and F_1-score results using leave-one-subject-out cross validation, including standard deviation.

catenate" variant produced the best result in our testing with an average accuracy of 91.77 percent and an F_1-score of 93.79, however, as indicated by the standard deviation of the results, this is not a clear and general outcome. Detailed measurements are included in Table 1.

3.1 Evaluation with Noise

The noise was applied to a single, random modality for each sliding window. It was produced by generating a random sequence of the same length as the input measurements, with a uniform distribution between -5 and +5. This sequence was subsequently multiplied element-wise with the randomly selected modality.

To our surprise, the GMU performed substantially worse with noise than the other two methods, as depicted in Figure 3.1. We expected the gating mechanism of the GMU cell to be able to respond to noise and therefore increase performance. The opposite seemed to be the case, as the performance of the GMU models deteriorated to less than chance level. Although it has to be noted that the models

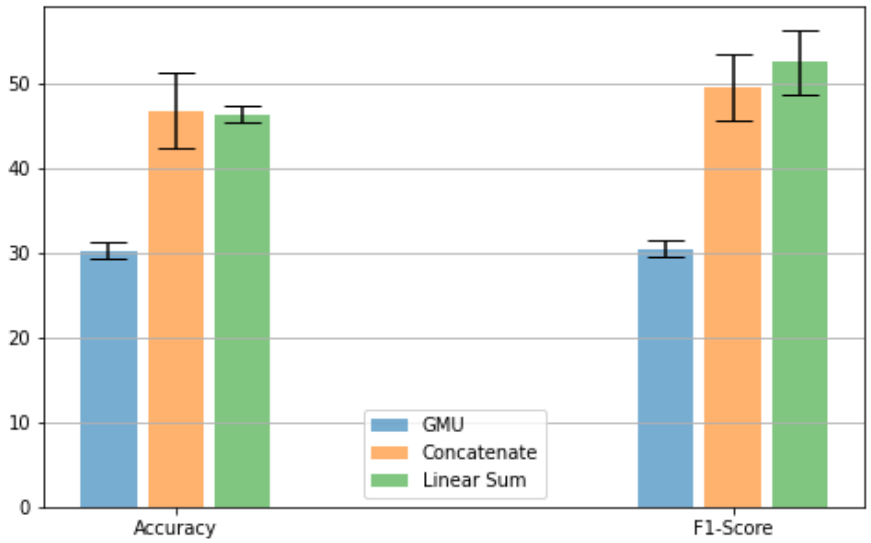


Figure 4: Mean accuracy and F_1-score results using leave-one-subject-out cross validation in the noise condition, error bars are indicating standard deviation.

were not trained on a noisy dataset. Concatenate and linear sum performed considerably worse as well, but still performed above chance with above 46 percent mean accuracy.

Measure	GMU	Concatenate	Linear Sum
Accuracy	30.26 ± 1.02	**46.79** ± 4.50	46.31 ± 1.02
F_1-Score	30.44 ± 0.94	49.48 ± 3.83	**52.46** ± 3.75

Table 2: Evaluation with artificial noise added to a random modality. Mean accuracy and F_1-score results using leave-one-subject-out cross validation, including standard deviation.

4 Discussion

The main takeaway seems to be that a GMU does not increase performance on a small, clean dataset over far simpler methods and therefor in this context does not justify its increase in complexity. Although slightly surprising, this was not an unexpected result.

Far more jarring however was the fact that the GMU did not just *not* perform better on noisy data but substantially worse. We had hoped the GMU to perform better in a more real-world setting, something that we followed from [1] where

the authors proposed noisy data as a prime application of the GMU. There are however some caveats to our methods that might have influenced the performance of the GMU:

- We did not train the GMU on a noisy dataset. We trained it on a very clean dataset and then tried to feed that model artificially noisy data. It might be the case that our model was so "overfitted" to the clean data that the gating mechanisms might have made performance even worse than no gating at all. It remains a question however if the GMU could then generalize from noisy data if it is presented different noise from the one it was trained on.

- The dataset was comparably small and of limited diversity. We only trained our model on data from 15 participants that were respectively exposed to three conditions one single time. Even though we had many data points, a lack of diversity within participants and their respective reactions might influence how well a model could generalize and also maybe respond to noisy data. If there is no diversity among the participant's reactions it might be the case that it is easy to overfit, especially with cross-subject validation.

- As described in Chapter 2 we reduced the amount of data significantly, filtering out only what we deemed useful. There might be other approaches that could have lead to better performance of a GMU compared to the other methods.

- We set the number of parameters of both the linear sum and concatenation model variants to try to match the complexity of the GMU. This was done to avoid having performance gains solely based on an increased number of parameters. However, it could also be that this has impacted the performance of the other two more than we anticipated while possibly making them inefficient on a larger scale or there being some ceiling after which a GMU might perform better.

- It might be the case that the added complexity of the GMU lead to overfitting on such a small dataset compared to the other two, simpler methods, even with our considerations for parameter number in mind.

5 Conclusion

We set out to apply a relatively novel method of multimodal fusion to a dataset that provided an opportunity, however limited, for emotion recognition. Even though we significantly increased performance from the classical ML methods presented in [11], we could not show that the GMU provided any form of benefit over linear sum or concatenate.

6 Outlook

There are a multitude of ways that one may wish to deviate from or build upon our work here. We will suggest a few possible avenues:

- Training a GMU with (different kinds of) noise might might increase performance significantly, especially in the noise condition
- Training the model on a larger/more diverse dataset
- Training the model on the WESAD dataset but with differing preprocessing
  - Training the model on an unadulterated WESAD dataset
  - Training the model with the meditation condition included
  - Training the model with a fourth output as a container for everything that is not stress, baseline or amusement
- Use alternative methods of validation
- Vary the level at which the GMU is employed
- Vary the number of GMUs employed
- Integrate it into a more complex model where the input is already processed by non-linear components

There is still much to explore when it comes to this novel cell in relation to emotion recognition. We still believe the GMU has a certain potential but it has yet to prove it is worth the increased complexity.

References

[1] John Arevalo, Thamar Solorio, Manuel Montes-y Gomez, and Fabio A González. Gated multimodal networks. *Neural Computing and Applications*, pages 1–20, 2020.

[2] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[3] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[4] Flavio Di Martino and Franca Delmastro. High-resolution physiological stress prediction models based on ensemble learning and recurrent neural networks. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE, 2020.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[6] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019.

[7] Jianjun Ni, Xiaoping Ma, Lizhong Xu, and Jianying Wang. An image recognition method based on multiple bp neural networks fusion. In *International Conference on Information Acquisition, 2004. Proceedings.*, pages 323–326. IEEE, 2004.

[8] Maria Nilsson. *Mind the gap: human decision making and information fusion.* PhD thesis, Örebro University, 2008.

[9] Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep ppg: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14):3079, 2019.

[10] Pritam Sarkar and Ali Etemad. Self-supervised ecg representation learning for emotion recognition. *IEEE Transactions on Affective Computing*, 2020.

[11] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 400–408, 2018.

[12] Rajiv Ranjan Singh, Sailesh Conjeti, and Rahul Banerjee. A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals. *Biomedical Signal Processing and Control*, 8(6):740–754, 2013.

[13] Elena Smets, Walter De Raedt, and Chris Van Hoof. Into the wild: the challenges of physiological stress detection in laboratory and ambulatory settings. *IEEE journal of biomedical and health informatics*, 23(2):463–473, 2018.

[14] Feng-Tso Sun, Cynthia Kuo, Heng-Tze Cheng, Senaka Buthpitiya, Patricia Collins, and Martin Griss. Activity-aware mental stress detection using physiological sensors. In Martin Gris and Guang Yang, editors, *Mobile Computing, Applications, and Services*, pages 282–301, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[15] H. Thapliyal, V. Khalus, and C. Labrado. Stress detection and management: A survey of wearable smart health devices. *IEEE Consumer Electronics Magazine*, 6(4):64–69, 2017.

[16] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018.

[17] J. Wijsman, B. Grundlehner, H. Liu, H. Hermens, and J. Penders. Towards mental stress detection using wearable physiological sensors. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1798–1801, 2011.

[18] Huaxin Xu and Tat-Seng Chua. Fusion of av features and external information sources for event detection in team sports video. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2(1):44–67, 2006.

[19] Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. Towards implicit content-introducing for generative short-text conversation systems. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2190–2199, 2017.