Attention-Based Multi-layer Perceptron to Categorize Affective Videos from Viewer's Physiological Signals

Lazib Sharar Shaiok[1], Ishtiaqul Hoque[2], Md Rakibul Hasan[1,4], Shreya Ghosh[4], Tom Gedeon[3,4,5], and Md Zakir Hossain[4(✉)]

[1] BRAC University, Dhaka, Bangladesh
[2] Islamic University of Technology, Dhaka, Bangladesh
lazib.sharar.shaiok@g.bracu.ac.bd
[3] Australian National University, Canberra, Australia
ishtiaqulhoque@iut-dhaka.edu
[4] Curtin University, Perth, Australia
[5] Obuda University, Budapest, Hungary
{Rakibul.Hasan,Shreya.Ghosh,Tom.Gedeon,Zakir.Hossain1}@curtin.edu.au

Abstract. The rapid growth of online video content has led to an increasing demand for effective video categorization methods. Current methods employed by video platforms include ratings from moderators, creators, and viewers. However, such a self-rating categorization method might not be the most efficient or insightful way to categorize videos. If physiological signals were taken into account, that would make the categorization more robust and could provide content creators, advertisers, and researchers with a better understanding of the viewers' emotional responses and preferences. In this paper, we develop a hybrid MLP architecture called "ATT-MLP" that utilizes self-attention in its layers and then test its performance on the AVDOS (Affective Video Dataset Online Study) dataset – a database where viewers' physiological signals were measured whilst they watched pre-classified videos. ATT-MLP outperformed MLP and traditional ML algorithms (Gaussian Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Linear Ridge, and Random Forrest) across all five data modalities (HRV, IMU, EMG-A, EMG-C, and ALL) of the AVDOS dataset. Accuracy and F1 were used as performance metrics, and the hybrid MLP architecture recorded the highest accuracy and F1 score, 93.8% and 93.1%, when the EMG-A data modality of the AVDOS dataset was used. This study shows that the MLP employing self-attention mechanisms within its hidden layers can be a powerful tool in the classification tasks of affective datasets. The code for the aforementioned model is publicly available on Github: https://github.com/IshtiaqHoque/ATT-MLP.

Keywords: affective computing · machine learning · deep learning · self-attention mechanisms

L.S. Shaiok and I. Hoque—Equal contribution.

N. T. Nguyen et al. (Eds.): ACIIDS 2024, CCIS 2145, pp. 25–34, 2024.
https://doi.org/10.1007/978-981-97-5934-7_3

1 Introduction

Stimuli from external multimedia elicit a diverse affective experience in people, driven by physiological signal changes. Moreover, watching a video can trigger fluctuations in multimodal physiological responses, including Electroencephalogram (EEG), Electrocardiogram (ECG), Heart Rate Variability (HRV) signals, and Electromyography (EMG) [1]. Online video platforms and other stakeholders, such as advertisers, are keen to explore the relationship between such physiological changes in viewers and video content. The recognition of affective states in viewers whilst watching videos is a popular branch of study in affective computing and has been achieved by the utilization of physiological features in various computational models. Affective states refer to individual emotional experiences, which can be classified into two main dimensions: valence and arousal. These components collectively depict the manner in which individuals engage in the processing and integration of their emotions within their conscious experiences [2]. Another branch of study is video agent emotion recognition, where the emotion of the agent or actor in the video is classified using audio and visual features [1].

In this study, we conduct an affective video content analysis using the Affective Video Database Online Study (AVDOS) database [3,4]. Subsequently, we develop a hybrid Multi-layer Perceptron MLP model utilizing self-attention [5] in its hidden layer to classify the affective videos into three categories: positive, negative, and neutral. We further compared our hybrid model with traditional MLP and traditional machine learning algorithms like KNN to substantiate the validity of our approach.

The primary contributions of this paper can be summarized in two key aspects:

1. A novel hybrid MLP model with self-attention was developed for comprehensive video categorization utilizing the AVDOS dataset;
2. Performed a comparative analysis between traditional MLP and standard machine learning algorithms, confirming that our hybrid approach is more effective and dependable.

This paper is organized as follows: Sect. 2 highlights the related studies based on affective video categorization and the use of MLP in previous studies; Sect. 3 describes our proposed model, establishing the novelty of our work by incorporating MLP and a self-attention mechanism for video categorization; Sect. 4 presents the discussion of our results and a brief description of the AVDOS dataset; Sect. 5 provides the conclusion and prospects of this study.

2 Related Work

Affective videos are used as stimuli in experiments to provoke certain emotions in viewers. The classification of these affective videos can be done using audio and visual features from the video itself [6]. For instance, Kang et al. used visual

features to classify videos into three categories: fearful, sad, and happy, with an accuracy of 79% [7]. With an accuracy of 74.69%, Wang et al. did an experiment with 2,040 video clips and classified them into seven emotion categories [8]. Classification of affective videos using the viewer's physiological signals is also an option among researchers. Lin [3] used EEG signals to detect music video emotion recognition. Soleymani et al. utilized physiological features like ECG, GSR, EEG, and respiration amplitude to classify arousal and valence levels induced by affective videos [9]. In other studies, such as the ones conducted by [1,10], affective video classification tasks were done by using multi-modal datasets utilizing both video and physiological features [10].

In addition, the utilization of Multilayer Perceptron (MLP) in conjunction with multimodal datasets has also been the subject of substantial research in previous studies. In a research conducted by Xing et al., six machine-learning algorithms were employed for the purpose of emotion recognition. These algorithms included Liblinear, REPTree, XGBoost, MultilayerPerceptron, RandomTree, and RBFNetwork. The authors conducted a comparative analysis of these algorithms in order to determine the optimal model for video emotion recognition using a multi-modal dataset. The results showed that MLP achieved the best performance where the classification accuracy for arousal was 97.79%, and the classification accuracy for valence was 96.79%. Additionally, the authors noted that Liblinear and REPTree had better performance in arousal recognition, while MLP achieved a balanced classification result for both arousal and valence. Regarding the dataset used in the research, it is worth noting that the existing EEG dataset is limited due to the complexity of EEG signal processing and the relatively longer time required for EEG feature collection. Not only this, the dataset used in the research had a limited variety of physiological signals [1].

3 Proposed Model

Our proposed model, ATT-MLP, is designed to capture complex patterns and dependencies within multi-modal data. It consists of an integrated structure combining a Multilayer Perceptron (MLP) and a self-attention layer. The self-attention mechanism enhances the MLP architecture to promote the sequential and variable data handling capacity of the architecture compared to the MLP and ML models. As the AVDOS dataset includes diverse and interconnected multi-modal information regarding physiological signals, using our hybrid approach becomes particularly useful in predicting the video category.

The different physiological signals do not have fixed lengths. The model essentially takes the input and passes it to the hidden layers. After the MLP hidden layers, the output is further enhanced by a hybrid hidden layer, which incorporates an MLP hidden layer and a self-attention layer. Figure 1 clearly illustrates the workflow of our proposed model. This combination makes use of the MLP's feature extraction skills and self-attention context awareness. As a result, the output exhibits a fused representation that includes both learned features and contextual data.

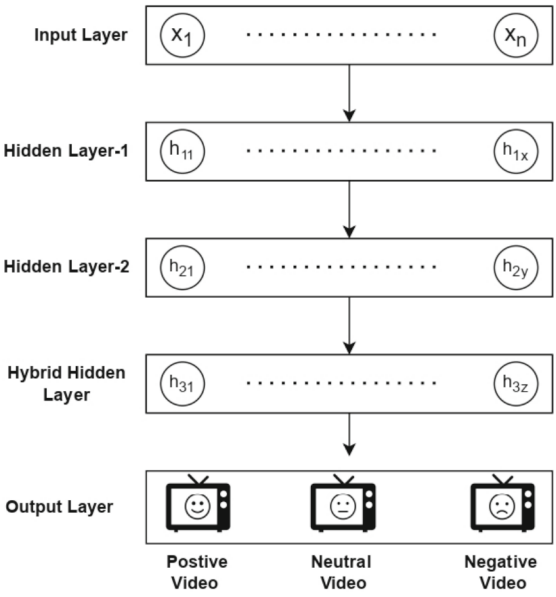

Fig. 1. Modified MLP Architecture (ATT-MLP)

3.1 Multilayer Perceptron (MLP)

Multilayer perceptron (MLP) is a feed-forward neural network [11]. It consists of an input layer, hidden layers, and an output layer. The presence of multiple layers distinguishes MLP from a linear perceptron. Video classification through recording multi-channel physiological responses is a complex nonlinear system [12]. Compared with other traditional machine learning approaches such as SVM, KNNs, and Decision Trees, the MLP model has a greater nonlinear mapping ability. [13].

Our model features an input layer whose dimensions vary based on the number of input features, ensuring that it can effectively process data from different sources. This is followed by an MLP consisting of one to three hidden layers. The number of hidden layers and their respective sizes are adjustable to capture complex patterns in the data. Specifically, the hidden layer sizes range from 32 to 256 neurons, offering a balance between model expressiveness and computational efficiency. Two activation functions have been employed, ReLU (Rectified Linear Unit) and Tanh (Hyperbolic Tangent), to introduce non-linearity into the model. L2 regularization, which varies between $1e^{-5}$ to $1e^{-1}$ on a logarithmic scale, is applied to the weights of the hidden layers, promoting generalization and preventing over-fitting. Lastly, the Output Layer consists of a Dense layer with a 'softmax' activation function. The number of units in the Output Layer matches the number of classes, which are the three outputs (Positive, Neutral, Negative).

3.2 Self-Attention Mechanism

The self-attention mechanism employs the scaled dot-product attention approach as shown in Fig. 2b. This allows the model to assign importance scores to different parts of the input sequence, capturing dependencies and relationships between elements. As illustrated in Fig. 2a, the output of the last hidden layer is concatenated with the output of the self-attention layer, resulting in a hybrid hidden layer that encompasses both the acquired features and contextual insights.

The hybrid layer consists of a hidden layer with an activation function given by Eq. 2. The output vector of the hidden layer, h_3, is passed into the self-attention layer and undergoes linear transformation to become vector h_{3a} using a weight matrix w_w. Subsequently, using scaled dot-product attention, query (q), key (k), and value (v) vectors are calculated, as shown by Eqs. 4, 5 and 6, respectively. w_q, w_k and w_v are weight matrices and b_q, b_k and b_v are bias vectors. The sizes of the q, k and v vectors are determined by the choice of attention units (16, 32 or 64). Following the q, k, and v calculation, the attention vector α is calculated using a 'softmax' function. As indicated by Eq. 7, the query vector and transpose of the key vector are multiplied and scaled between 0 and 1 to calculate the value of α. Afterwards, the h_3' vector is calculated by multiplying the attention vector (α) with the value vector as shown in Eq. 8. The vector h_3' is the output of the self-attention mechanism embedded within a particular hidden layer. Lastly, h_3' is concatenated with the output vector of the hidden layer 3 as shown in Fig. 2a and described by Eq. 9. The concatenated output is the final contextualized vector that is passed into the output layer.

$$h_2 = \text{Output of hidden layer 2}$$

$$a = \sum w_k h_{2k} + b \tag{1}$$

$$h_3 = y(a) \tag{2}$$

$$h_{3a} = h_3 w_w + b_1 \tag{3}$$

$$q = h_{3a} w_q + b_q \tag{4}$$

$$k = h_{3a} w_k + b_k \tag{5}$$

$$v = h_{3a} w_v + b_v \tag{6}$$

$$\alpha = \text{softmax}(\frac{q^* k^T}{\sqrt{d_k}}) \tag{7}$$

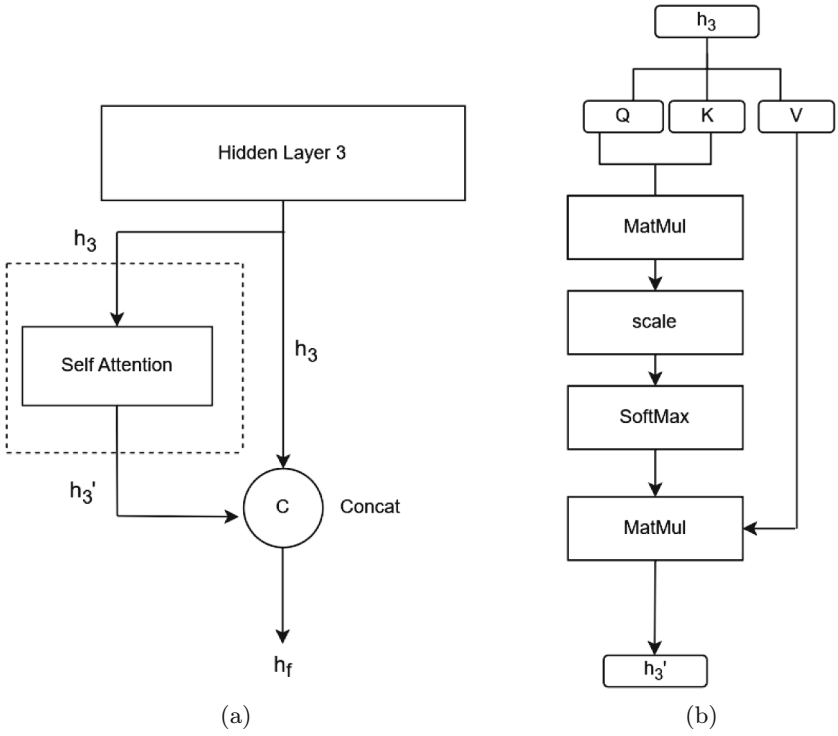

Fig. 2. (a) Hybrid Hidden Layer and (b) The Scaled Dot Product Self Attention Mechanism Applied To The Last Layer.

$$h_3' = \alpha v \tag{8}$$

$$h_f = \text{concat}(h_3\prime, h_3) \tag{9}$$

3.3 Further Enhancements

In order to enhance the performance and generalization capability of the model, the process of hyperparameter tuning is carried out utilizing Optuna, a robust framework for hyperparameter optimization. The hyperparameters that require adjustment include the number of hidden layers, the dimensions of each hidden layer, the selection of activation function, the strength of L2 regularisation, and the number of attention units. The Optuna framework utilizes an exhaustive search strategy to determine the optimal combination of hyperparameters that maximizes a preset goal function. This process guarantees that the model is adjusted to optimize its performance for the particular job and AVDOS-VR dataset. The hyperparameters and the subsequent search space used for ATT-MLP can be found in Table 1.

Table 1. Hyperparameter Search Space for Optimization

Classifier	Optimization Framework	Hyperparameter	Search Space
ATT-MLP	Optuna	hidden layer size	32, 64, 128
		activation function	relu, tanh
		alpha	1e−5, 1e−4, 1e−3, 1e−2
		attention units	16, 32, 64
MLP	GridSearch	optimizer learning rate	0.05, 0.001
		hidden layer sizes	(100,), (50, 50,)
		model dropout	0, 0.5
Linear Ridge	GridSearch	alpha	10^{-5} to 10^{5}
Gaussian SVM	GridSearch	C	1, 10, 100, 1000
		gamma	0.1, 0.01, 0.001
Random Forrest	GridSearch	n estimators	10, 50, 100
		max depth	5, 10, 20
K-Nearest Neighbour	GridSearch	no of neighbours	1, 5, 11, 15

4 Experiment

Table 2. AVDOS Dataset: Data Modality by Number of Features

Data Modality	Number of Features
HRV	42
IMU	108
EMG-A	84
EMG-C	84
All	318

4.1 Dataset

The dataset gathered from an open-source GitHub repository is part of a wider study titled AVDOS – Affective Video Database Online Study conducted by researchers from Bournemouth University, Aegean University, and Emteq Labs [3]. In this dataset, 37 participants' physiological signals were measured when they watched videos of three categories: positive, neutral, and negative. The

physiological signals were further divided into five different modalities: HRV (heart rate variability), IMU (inertial measurement unit), EMG-A (electromyography amplitude), EMG-C (electromyography contact), and ALL (all the physiological signals) (Table 2). In this study, LOSO-cross validation was used wherein, for each classifier, one participant was made to be the test dataset, whereas the other thirty-six participants were used to train the model. This process was repeated for each participant, ensuring each participant was the test data at least once, and then the average accuracy and F1 score were calculated.

4.2 Results

Table 3. Different Classifier's Performance by Data Modality. Best performance is highlighted in **boldface** font.

Classifier	Data Modality									
	HRV		IMU		EMG-A		EMG-C		All	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
MLP	0.389	0.362	0.506	0.482	0.752	0.733	0.439	0.408	0.754	0.736
Gaussian SVM	0.373	0.328	0.523	0.497	0.793	0.786	0.454	0.402	0.769	0.752
KNN	0.379	0.358	0.412	0.391	0.746	0.727	0.419	0.395	0.640	0.626
Linear Ridge	0.396	0.344	0.502	0.471	0.798	0.792	0.441	0.387	0.764	0.750
RF	0.353	0.326	0.443	0.405	0.788	0.780	0.427	0.384	0.780	0.771
ATT-MLP (Ours)	**0.757**	**0.736**	**0.853**	**0.736**	**0.938**	**0.931**	**0.849**	**0.832**	**0.933**	**0.927**

In this study, ATT-MLP reported the highest accuracy and F1 score compared to the rest of the five classifiers (MLP, Gaussian SVM, KNN, Linear Ridge, Random Forrest) across all five modalities (HRV, IMU, EMG-A, EMG-C, and All) as shown in Table 3. The hyperparameters and the subsequent search space for the six classifiers are reported in Table 1. Accuracy and F1 scores were the primary performance parameters used to distinguish the performance of the different classifiers. Accuracy and F1 scores were calculated according to the following expressions, respectively.

$$\text{ACC} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{10}$$

$$\text{F1} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{11}$$

Meanwhile, Gaussian SVM was the second best-performing classifier as it reported performance metrics higher than other classifiers, excluding ATT-MLP, in three data modalities (IMU, EMG-A and ALL). KNN seems to be the worst-performing classifier as it reported the lowest score in three data modalities (IMU, EMG-A, and ALL). Contrary to expectations, the highest accuracy and F1 score reported by any classifier came from EMG-A data modality than ALL

data modality. ATT-MLP scored 0.938 in accuracy and 0.931 in F1 score when it came to classifying EMG-A data, whilst for ALL data, it scored 0.933 and 0.927, respectively. In fact, all six classifiers performed better when EMG-A data was used compared to ALL data, though the difference was quite minute. This might indicate that EMG-A has the most weight in terms of classifying emotional state compared to all other physiological signals though further data analysis is needed before this claim can be verified. The lowest accuracy and F1 score were reported by the Random Forrest classifier at 0.353 and 0.326, respectively for HRV data modality. The accuracy and F1 score variations across different modalities are illustrated by Table 3.

5 Conclusion

Video is currently the most popular multimedia stimulus capable of conveying complex emotional meanings through visual and aural cues [1]. As such, the automatic classification of videos by using features from the video and the viewer is a growing field of study in affective research. In this study, we presented a novel hybrid MLP architecture called ATT-MLP that utilizes self-attention mechanisms in its layers and then tested its performance on the AVDOS dataset. The study showed the effectiveness of this model in classifying the videos based on the viewer's physiological signals, outperforming several popular ML algorithms and MLP without any self-attention mechanism it's hidden layers. One of the key limitations of the proposed model is that it has been specifically designed for the AVDOS dataset; thus, the future of this research consists of testing the developed model with other affective datasets with physiological features.

References

1. Xing, B., et al.: Exploiting EEG signals and audiovisual feature fusion for video emotion recognition. IEEE Access **7**, 59844–59861 (2019)
2. Santamaria-Granados, L., Munoz-Organero, M., Ramirez-Gonzalez, G., Abdulhay, E., Arunkumar, N.: Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos). IEEE Access **7**, 57–67 (2018)
3. Gnacek, M., et al.: Avdos-affective video database online study video database for affective research emotionally validated through an online survey. In: 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1–8. IEEE (2022)
4. Michalgnacek: Github - michalgnacek/AVDOS-VR: scripts repository for analysis of DRAP database. https://github.com/michalgnacek/AVDOS-VR
5. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
6. Fonnegra, R.D., Díaz, G.M.: Deep learning based video spatio-temporal modeling for emotion recognition. In: Kurosu, M. (ed.) HCI 2018. LNCS, vol. 10901, pp. 397–408. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91238-7_32
7. Kang, H.B.: Affective content detection using HMMs. In: Proceedings of the eleventh ACM International Conference on Multimedia, pp. 259–262 (2003)

8. Wang, H.L., Cheong, L.F.: Affective understanding in film. IEEE Trans. Circuits Syst. Video Technol. **16**(6), 689–704 (2006)
9. Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. IEEE Trans. Affect. Comput. **3**(1), 42–55 (2011)
10. Duan, L., Ge, H., Yang, Z., Chen, J.: Multimodal fusion using kernel-based ELM for video emotion recognition. In: Cao, J., Mao, K., Wu, J., Lendasse, A. (eds.) Proceedings of ELM-2015 Volume 1. PALO, vol. 6, pp. 371–381. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-28397-5_29
11. Li, D., Huang, F., Yan, L., Cao, Z., Chen, J., Ye, Z.: Landslide susceptibility prediction using particle-swarm-optimized multilayer perceptron: comparisons with multilayer-perceptron-only, BP neural network, and information value models. Appl. Sci. **9**(18), 3664 (2019)
12. Zhang, X., Xu, C., Xue, W., Hu, J., He, Y., Gao, M.: Emotion recognition based on multichannel physiological signals with comprehensive nonlinear processing. Sensors **18**(11), 3886 (2018)
13. Amendolia, S.R., Cossu, G., Ganadu, M., Golosio, B., Masala, G.L., Mura, G.M.: A comparative study of k-nearest neighbour, support vector machine and multilayer perceptron for thalassemia screening. Chemom. Intell. Lab. Syst. **69**(1–2), 13–20 (2003)