PREPROCESSING IS NOT NEEDED: AN END-TO-END SOLUTION FOR PHYSIOLOGICAL SIGNALS BASED EMOTION RECOGNITION

Ziqing Yang & Houwei Cao
Department of Computer Science, New York Institute of Technology
{zyang23, hcao02}@nyit.edu

ABSTRACT

While the recent advances in automatic emotion recognition with human physical signals such as audio, visual and textual inputs have been remarkable, research on emotion recognition with internal physiological signals has received considerable attention only in recent years, and most of the studies focused on feature engineering and traditional machine learning algorithms. In this study, we propose an advanced domain alignment transformer (DATransformer) framework that addresses the major challenges of physiological signals based emotion recognition–the domain inconsistency and sample rate difference between the multivariate physiological signals and emotional states. Our proposed DATransformer framework does not require any preprocessing on the raw physiological signal inputs, but can obtain comparable or even better emotion recognition performance than the preprocessed signals. We evaluate the proposed DATransformer on the Continuously Annotated Signals of Emotion (CASE) dataset and achieve the state-of-the-art (SOTA) performance.

1 INTRODUCTION

Emotions are essential to human life. They directly influence human perception and behaviors, and have big impacts on our daily tasks, such as learning, rational decision-making, and social interaction. Positive emotions can improve human health and lead to greater job satisfaction and productivity, while negative emotions may cause health problems and reduce the general life satisfaction. Automatic emotion recognition has found applications in many domains, including smart healthcare and human-computer interaction. For instance, a smart health system can sense the affective states based on multimodal and multidimensional data collected with mobile and wearable devices. Such a system can then track the change of affect over time and create emotion evolution profiles, which can be used to detect the negative feelings and monitor the mental health condition.

The emotional states of people usually change accompanied by external physical changes including voice, facial expressions, body gestures, etc., as well as the internal changes relevant to human organs and tissues, which usually can be captured by physiological signals such as electroencephalogram (EEG), temperature (T), electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR), respiration (RSP), etc. (Zhang et al., 2023; Kim & Lee, 2023; Ahmad & Khan, 2022). While the recent advances in automatic emotion recognition with human physical signals such as audio, visual and textual inputs have been remarkable, research on emotion recognition with internal physiological signals has received considerable attention only in recent years. Most of the studies still focused on feature engineering and traditional machine learning algorithms, which did not benefit from the recent development of the deep-learning and large machine learning models.

There are several challenges in the recognition of emotion from physiological signals: 1. Due to the complexities of the physiological signals, the conventional solution requires handcrafted features that are heavily dependent on the level of experience for each signal. Lack of domain knowledge can result in inappropriate features. (Han et al., 2023) 2. Due to different collection conditions among different datasets, the preprocessing setting need to be optimized for each individual dataset. 3. The domain inconsistency and sample rate difference between the physiological signals and

emotional states make the physiological signal based emotion recognition more challenging than the traditional time series prediction task.

To address these challenges, we proposed an advanced end-to-end domain alignment transformer (DATransformer) framework. We first normalize both the physiological signals and emotional state signals to a similar distribution via instance normalization (Liu et al., 2022; Kim et al., 2021). A constant standard deviation and mean will be calculated for both signals separately, and the final output will be denormalized back to the emotional signal distribution based on the calculated standard deviation and mean. This normalization and denormalization process effectively solves the domain inconsistency between the physiological signal and emotional state. To tackle the sample rate difference between the physiological signal and the emotional state, we introduce the patch embedding (Zhang & Yan, 2022; Nie et al., 2022), which embeds the signals as fix length windows sliced from the complete time steps. It is easy to align the physiological signals with the emotional states by setting the patch size to match with their correspondent sample rate difference. In addition, patch embedding further allows the model to utilize channel information via transposing the embedding dimension from channel to time step. The model dynamically assigns the weights to each channel based on their decoder's cross attention score, and sum up all the channel representations to generate the final prediction. Moreover, the experimental results show that our solution doesn't require preprocessing on the physiological signals. The raw physiological signals can generate comparable or even better results than the preprocessed signal inputs. To summarize, our contributions lie in three folds:

- We propose a generic solution that tackle the domain inconsistency and sample rate difference of the input and target signal, which are the two major challenges of the physiological-based emotion recognition.
- Experimentally, our proposed end-to-end solution does not require any preprocessing on the input physiological signals.
- Our proposed DATransformer framework achieve the state-of-the-art (SOTA) results on the real-world physiological signal emotion recognition dataset.

2 Related Work

Physiological signals are the signals that carry physiological arousal information from the central nervous system (CNS) and autonomic nervous system (ANS) (Arya et al., 2021). Common physiological signals include electroencephalogram (EEG), electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR), skin temperature (SKT), respiration (RSP), etc. Previous studies have shown the effectiveness of physiological signals on prediction of discrete emotions, including basic emotions such as anger, fear, happy, sad, as well as the affective states like valence, arousal and dominance (Ahmad & Khan, 2022). Compared with conventional models relied on hand-crafted features, signal preprocessing and feature engineering, deep learning models can extract and select features automatically and dynamically. As a result, the deep learning methods are less dependent on experience and domain knowledge, which make them widely applied to many industrial applications (Han et al., 2023). Recently, several studies have investigated the transformer-based framework for physiological based emotion recognition (Vu et al., 2023; Vazquez-Rodriguez et al., 2022b;a; Yang et al., 2022). However, most of the works are focusing on the discrete emotion detection task. The more challenged continuous affective states prediction task still need to be further explored. In this paper, we aim to develop an advanced deep-learning framework that can predict continuous affective states based on raw physiological signals.

Transformer and its variants have dominated time series prediction. Previous studies have explored different model architecture, and attention selection.(Zeng et al., 2023; Zhou et al., 2022; Wu et al., 2021) Recently, several works emphasized the importance of the embedding design (Zhang & Yan, 2022; Nie et al., 2022; Liu et al., 2023) and the efficiency of normalization (Kim et al., 2021; Liu et al., 2022). In this paper, we focus on the embedding design and normalization technique and further investigate how they will benefit the physiological-based emotion recognition. Details of our model implementation will be discussed in the following section.

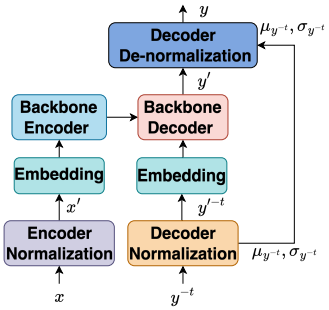

Figure 1: Overall Framework of the model

3 METHODOLOGY

To better reveal the contribution of the embedding design and normalization method, we used the conventional vanilla transformer as our model backbone in our proposed domain alignment transformer (DATransformer). The raw physiological signals are feed into the model as the input for the encoder, and the past time-step of emotional states are feed into the model as the input for the decoder. Both of the encoder and decoder inputs are normalized and embedded, and then further processed by the model. The final output will be denormalized back to the domain on the emotional states. The overall structure of the model is presented in Figure 1, where μ and σ denote mean and standard deviation of the series, y^{-t} denotes the past t time-step of the target signal where t is the label length, and x', y' denotes the normalized signals.

3.1 NORMALIZATION

Reversible instance normalization (RevIN) (Kim et al., 2021) proposed an instance normalization that ensure the series will follow a similar distribution. It has a symmetric structure that contains the normalization and denormalization process. It defines an unsupervised normalization based on statistics (mean and standard deviation) of the input and learnable affine parameters. It shows that the normalization applies on the I/O process is better than that on the intermediate process. Non-stationary Transformer (NSTransformer) (Liu et al., 2022) proposed series stationarization normalization, which drops the learnable affine parameters of RevIN and adapt it to a more straightforward implementation.

In this work, we adopt the NSTransformer implementation. Unlike the traditional time series prediction task in which model inputs and targets are in the same domain, the input and target signals are in totally different domains in the physiological-based emotion recognition. Therefore, we first normalize both of the input and target signals to stationary. The normalization process ensures that the input and target signals will follow a similar distribution, and the stationarization will further help the model to better align the input and target signals in the time domain. The final output of the model will be de-stationarized back to the target domain.

3.2 EMBEDDING

Early studies on time series prediction with transformer-based backbone usually adopt the token embedding, which is commonly used in natural language processing (NLP) and computer vision (CV), and many of them focused on studying the trade off between the positional and timestamp embedding (Zhou et al., 2021; 2022; Zeng et al., 2023). Recently, several works emphasized the difference between the time series prediction and other tasks in NLP and CV, and proposed that the embedding for time series signals should transpose the embedding axis from channel to time step (Zhang & Yan, 2022; Nie et al., 2022; Liu et al., 2023). Furthermore, the PatchTST (Nie et al., 2022) and the Crossformer (Zhang & Yan, 2022) utilized the patch segmentation, which segment the complete time steps into several fix-length windows and further embed these segmented windows. Both of them adopt positional embedding instead of timestamp embedding. On the other hand,

Table 1: Physiological-based emotion recognition performance on various machine learning and end-to-end deep learning models (RMSE)

Model	Valence	Arousal	Average
Vanilla Transformer	3.98	2.52	3.25
Multi-scale Transformer	1.50	1.64	1.57
AutoGluon	0.95	0.91	0.93
RF / XGBoost	0.87	0.85	0.86
DATransformer (ours)	**0.79**	**0.70**	**0.75**

iTransformer (Liu et al., 2023) proposed the embedding design without patching, which embeds the complete time steps. The embedding layer concatenates timestamp as an extra channel to embed with the input rather than embed timestamp separately and summing back to the value embedding.

In this work, we transpose the embedding axis from channel to time steps, whose effectiveness has been proven by previous studies (Liu et al., 2023; Zhang & Yan, 2022). To address the sample rate difference between the input and the target signals, instead of embedding the complete time step, we patch the input signals into windows with the length as the sample rate difference Δ, where $\Delta = (\text{Input sample rate})/(\text{Output sample rate})$. The target signals are embedded with window size as 1 to align the number of patches of input and target signals. For both embeddings, we concatenate the timestamp as an extra channel to be embedded. As the timestamp contains sequential information, we dropped the positional embedding. Meanwhile, via transposing the embedding axis, our proposed model can further explore the channel dependencies of the physiological signals and emotional signals. We utilize the cross attention score to generate the weight for each channel and sum them up to generate the final prediction.

4 Experiments & Results

4.1 Dataset

The evaluation dataset used in this study is the Continuously Annotated Signals of Emotion (CASE) dataset.(Sharma et al., 2019) It contains data from 30 participants, including 15 male and 15 female. The continuous valence and arousal scores are reported by the participants via a joystick while they are watching various videos. Eight raw physiological measurements are also collected simultaneously including electrocardiograph (ECG), blood volume pulse(BVP), electromyography(EMG) (3 channels), galvanic skin response (GSR) (or electrodermal (EDA)), respiration (RSP) and skin temperature (SKT) sensors. ECG reflects the electrical signal generated by the heart muscles during contraction. RSP reflects the expansion and contraction of the chest cavity. BVP is also known as photoplethysmography (PPG), which changes according to the blood flowing through the vessels, serving as a measure for the cardiac activity. GSR, also known as EDA, measures the variation in electrical conductance resulting from sweat released by the glands on the skin. EMG measures the surface voltage associated with muscle contractions. More details of how the physiological signals are collected can be found in (Sharma et al., 2019). The sample rate of the physiological signals is 1000Hz and the sample rate of the emotional states (valence and arousal) is 20Hz.

4.2 Experiment Setup

In our study, we follows the across-time scenarios experiment settings of the Emotion Physiology and Experience Collaboration (EPiC) challenge[1]. The 240 data files are divided into training data and testing data based on the timestamp. For each file the former part are selected as the training data, and the latter part become the testing data. The sequence length of the model input is 2250 (ms). The corresponding label length and prediction length are 15 and 30 (samples with 50 ms each). The corresponding input and output dimension will be $(batch, sequence\_length, 8)$ and $(batch, prediction\_length, 2)$. Eight and two are the number of physiological signals and emotional signals respectively.

[1] https://github.com/Emognition/EPiC-2023-competition

Table 2: Emotion recognition performance with and without the data preprocessing on the raw physiological signals (RMSE)

	Model	Valence	Arousal	Average
Preprocessing	Vanilla Transformer	2.15	1.89	2.02
	DATransformer (ours)	**0.788**	0.719	0.754
Non-Preprocessing	Vanilla Transformer	3.98	2.52	3.25
	DATransformer (ours)	0.792	**0.703**	**0.748**

Table 3: The ablation study on our proposed end-to-end domain alignment transformer (DATransformer) (RMSE)

Components					
Encoder Normalization	Decoder Normalization	Series Embedding	Valence	Arousal	Average
✓	✓	✓	0.792	**0.703**	**0.748**
✗	✓	✓	**0.786**	0.723	0.755
✓	✓	✗	0.815	0.711	0.763
✗	✓	✗	0.951	0.713	0.832
✗	✗	✓	1.228	1.523	1.376
✓	✗	✗	2.839	1.487	2.163
✗	✗	✗	3.98	2.52	3.25

4.3 RESULTS

Now we turn to discuss the physiological-based emotion recognition performance. We compared our proposed Domain Alignment Transformer(DATransformer) with the EPiC challenge winners and several other state-of-the-art transformer models, and the results are reported in Table 1. The AutoGluon solution, proposed by (Dollack et al., 2023), achieves the first place in EPiC challenge. The model focused on feature engineering and utilized the open source auto ML toolkit to explore the best ensemble combination of models. The traditional machine learning approach, RF/XGBoost with hand-crafted features, proposed by (D'Amelio et al., 2023), achieves the third place in the challenge and the best performance in the across-time scenario. The multi-scale Transformer solution, proposed by (Vu et al., 2023), focuses on fusing the representations extracted from multi-scale inputs, providing the model performance of transformer-based solution with a different model architecture. We report the results of the conventional Vanilla Transformer as well. It is worth noticing that all the end-to-end solution including Vanilla Transformer, multi-scale Transformer, and our proposed DATransformer do not apply any preprocessing to the raw physiological signals, while the EPiC challenge winners such as AutoGluon and the RF/XGBoost solution apply data preprocessing and feature engineering to the physiological signals inputs. From the results in Table 1, it's clearly shown that the proposed DATransformer significantly outperform the other two transformer-based models, as well as the two EPiC challenge winners, achieves the best prediction performance on both Arousal and Valence emotional states.

Next, we turn to discuss how the preprocessing of the physiological signal inputs would affect our model performance, and the results are reported in table 2. For the sake of comparison, we apply the same preprocessing techniques used in the RF/XGBoost solution from the EPiC challenge winner, and the details of the preprocessing could be found in (D'Amelio et al., 2023). We also report the results on Vanilla Transformer with the same setting in table 2 as well. From the results, it's clearly shown that our proposed DATransformer achieves comparable performance on Valence, and even better results on Arousal without the signal preprocessing. While for Vanilla Transformer, preprocessing is the crucial part, and it will substantially improve the model performance. This further prove the advantage of our proposed end-to-end solution, which could extract instructive representation directly from the raw physiological signals.

Finally, we perform the ablation study to further investigate the importance of different components i.e., encoder normalization, decoder normalization, and series embedding, for our proposed DATransformer. The results are shown in Table 3. We first noticed that each component contribute differently to the final emotion recognition performance, and the decoder normalization seems the

most powerful single component, which can decrease the average RMSE from 3.25 to 0.832. Moreover, the results show that all component are essential to our proposed DATransformer model, and we can achieve the best performance with average RMSE at 0.748 when the model includes all the three components together.

5 CONCLUSION

In this study, we proposed an end-to-end domain alignment transformer (DATransformer) framework that tackle the domain inconsistency and sample rate difference of the input and target signal, which are the two major challenges of the physiological-based emotion recognition. Our proposed solution does not require any preprocessing on the raw physiological signal inputs, but can obtain comparable or even better emotion recognition performance than the preprocessed signals. For future work, we will investigate more advanced backbones, and apply various attention designs to further improve the physiological-based emotion recognition.

REFERENCES

Zeeshan Ahmad and Naimul Khan. A survey on physiological signal-based emotion recognition. *Bioengineering*, 9(11):688, 2022.

Resham Arya, Jaiteg Singh, and Ashok Kumar. A survey of multidisciplinary domains contributing to affective computing. *Computer Science Review*, 40:100399, 2021.

Felix Dollack, Kiyoshi Kiyokawa, Huakun Liu, Monica Perusquia-Hernandez, Chirag Raman, Hideaki Uchiyama, and Xin Wei. Ensemble learning to assess dynamics of affective experience ratings and physiological change. *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2023.

Tomás A. D'Amelio, Nicolás M. Bruno, Leandro A. Bugnon, Federico Zamberlan, and Enzo Tagliazucchi. Affective computing as a tool for understanding emotion dynamics from physiology: A predictive modeling study of arousal and valence. *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2023.

Ean-Gyu Han, Tae-Koo Kang, and Myo-Taeg Lim. Physiological signal-based real-time emotion recognition based on exploiting mutual information with physiologically common features. *Electronics*, 12(13):2933, 2023.

Jeong-Yoon Kim and Seung-Ho Lee. Coordvit: A novel method of improve vision transformer-based speech emotion recognition using coordinate information concatenate. In *2023 International Conference on Electronics, Information, and Communication (ICEIC)*, pp. 1–4. IEEE, 2023.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2021.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Rethinking the stationarity in time series forecasting. *arXiv preprint arXiv:2205.14415*, 2022.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Karan Sharma, Claudio Castellini, Egon L van den Broek, Alin Albu-Schaeffer, and Friedhelm Schwenker. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific data*, 6(1):196, 2019.

Juan Vazquez-Rodriguez, Grégoire Lefebvre, Julien Cumin, and James L Crowley. Emotion recognition with pre-trained transformers using multimodal signals. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8. IEEE, 2022a.

Juan Vazquez-Rodriguez, Grégoire Lefebvre, Julien Cumin, and James L Crowley. Transformer-based self-supervised learning for emotion recognition. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 2605–2612. IEEE, 2022b.

Tu Vu, Van Thong Huynh, and Soo-Hyung Kim. Multi-scale transformer-based network for emotion recognition from multi physiological signals. *arXiv preprint arXiv:2305.00769*, 2023.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34:22419–22430, 2021.

Kangning Yang, Benjamin Tag, Yue Gu, Chaofan Wang, Tilman Dingler, Greg Wadley, and Jorge Goncalves. Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pp. 562–570, 2022.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Saining Zhang, Yuhang Zhang, Ye Zhang, Yufei Wang, and Zhigang Song. A dual-direction attention mixed feature network for facial expression recognition. *Electronics*, 12(17):3595, 2023.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2022.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pp. 27268–27286. PMLR, 2022.