2022 26th International Conference on Pattern Recognition (ICPR)
August 21-25, 2022, Montréal, Québec, Canada

Transformer-Based Self-Supervised Learning for Emotion Recognition

Juan Vazquez-Rodriguez*†, Grégoire Lefebvre*, Julien Cumin* and James L. Crowley†
*Orange Labs, Grenoble, France
Email: {juan.vazquezrodriguez, gregoire.lefebvre, julien1.cumin}@orange.com
†Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, Grenoble, France
Email: james.crowley@inria.fr

2022 26th International Conference on Pattern Recognition (ICPR) | 978-1-6654-9062-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICPR56361.2022.9956027

***Abstract*—In order to exploit representations of time-series signals, such as physiological signals, it is essential that these representations capture relevant information from the whole signal. In this work, we propose to use a Transformer-based model to process electrocardiograms (ECG) for emotion recognition. Attention mechanisms of the Transformer can be used to build contextualized representations for a signal, giving more importance to relevant parts. These representations may then be processed with a fully-connected network to predict emotions.**

To overcome the relatively small size of datasets with emotional labels, we employ self-supervised learning. We gathered several ECG datasets with no labels of emotion to pre-train our model, which we then fine-tuned for emotion recognition on the AMIGOS dataset. We show that our approach reaches state-of-the-art performances for emotion recognition using ECG signals on AMIGOS. More generally, our experiments show that transformers and pre-training are promising strategies for emotion recognition with physiological signals.

I. Introduction

When processing time-series signals with deep learning approaches, it is useful to be able to aggregate information from the whole signal, including long-range information, in a way that the most relevant parts are given more importance. One way of doing this is by employing an attention mechanism [1] that uses attention weights to limit processing to relevant contextual information, independent of distance.

Arguably, the Transformer [2] is one of the most successful attention-based approaches. Developed for Natural Language Processing (NLP), the Transformer uses attention mechanisms to interpret sequences of words, and is suitable for use in other tasks requiring interpretation of sequences, such as time series forecasting, [3], analysis of medical physiological signals [4], [5], and recognition of human activity from motion [6].

Physiological signal analysis can be seen as a form of time-series analysis and are thus amenable to processing with Transformers. Moreover, these signals can be used to predict emotions [7], and sensors for these types of signals can be incorporated into wearable devices, as a non-invasive means for monitoring the emotional reaction of users. Several works in this direction have emerged using signals like electrocardiograms (ECG) [8], [9], electroencephalograms (EEG) [10], [11], electrodermal activity (EDA) [12], and other types of physiological signals [13], [14].

Established approaches for deep learning with Convolutions and Recurrent networks require large datasets of labeled training data. However, providing ground truth emotion labels for physiological data is a difficult and expensive process, limiting the availability of data for training [15], [16], [17]. Pre-training models with self-supervised learning can help to overcome this lack of labeled training data. With such an approach, during pre-training the model learns general data representations using large volumes of unlabeled data. The model is then fine tuned for a specific task using labeled data. This approach has been successfully used in other domains including NLP [18], [19] and Computer Vision [20], [21]. It has also been successfully used in affective computing, in tasks like emotion recognition from physiological signals [9], [22] and from speech [23], personality recognition [24], and facial expression recognition [25], [26], [27].

In this paper, we address the problem of predicting emotions from ECG signals. We are interested in obtaining contextualized representations from these signals using a Transformer-based architecture, and then using these representations to predict low/high levels of arousal and valence. We believe that the contextualized representations obtained with the Transformer should capture relevant information from the whole signal, which the performance of the downstream task of emotion recognition should benefit from. Our main contributions are: 1) We show that it is feasible to use a Transformer-based architecture for emotion prediction from ECG signals. 2) We show that using a self-supervised technique to pre-train the model is useful for ECG signals, achieving superior performance in emotion recognition than a fully-supervised approach. 3) We show that our pre-trained Transformer-based model reaches state-of-the-art performances on a dataset of the literature.

II. Related Work

Traditional techniques for emotion recognition from physiological signals include Gaussian naive Bayes, Support Vector Machines, k-Nearest Neighbours, and Random Forests. [16], [17], [28], [29], [30], [31]. These approaches typically use manually-selected time and frequency features derived from intuition and domain knowledge. Shukla et al. [12] show that commonly used features for arousal and valence prediction are not necessarily the most discriminant. This illustrates the difficulty of selecting good hand-crafted features.

To overcome this, researchers have increasingly used deep learning techniques to extract features from physiological

978-1-6654-9062-7/22/$31.00 ©2022 IEEE

signals for emotion recognition. A common approach, described by Santamaria et al. [8], is to use a 1D Convolutional Neural Network (CNN) to extract the features (also called representations), followed by a fully-connected network (FCN) used as classifier to predict emotions. As an alternative, Harper and Southern [32] use a Long Short-Term Memory (LSTM) network concurrently with a 1D-CNN. Siddharth et al. [33], first convert signals into an image using spectrograms [34], and then use a 2D-CNN for feature extraction, followed by an extreme learning machine [35] for classification.

One drawback of these CNN-based approaches is that they do not take context into account: after training, kernel weights of the CNN are static, no matter the input. For this reason, attention-based architectures such as the Transformer [2], capable of incorporating contextual information, have started to be used for emotion prediction. Transformers have been successfully used to recognize emotions with multimodal inputs composed of text, visual, audio and physiological signals [36], [37], [38], [39], [40]. In addition, Transformers have been used to process time-series in general [3], [41], and also to process uni-modal physiological signals in particular, with the aim of recognizing emotions. Arjun et al. [42] employ a variation of the Transformer, the Vision Transformer [43] to process EEG signals for emotion recognition, converting the EEG signals into images using continuous wavelet transform. Behinaein et al. [44] propose to detect stress from ECG signals, by using a 1D-CNN followed by a Transformer and a FCN as classifier.

Most of the approaches for measuring emotions, including those using multimodal physiological data, have relied on supervised learning, and thus are limited by the availability of labeled training data. Using self-supervised pre-training can improve performances of a model [45], as it allows to learn more general representations, thus avoiding overfitting in the downstream task. This is especially important for tasks with limited labeled data. Sarkar and Etemad [9] pre-train a 1D-CNN using a self-supervised task to learn representations from ECG signals. Their self-supervised task consists in first transforming the signal, with operations such as scaling or adding noise, and then using the network to predict which transformation has been applied. Ross et al. [22] learn representations from ECG signals using auto-encoders based on 1D-CNN. In both approaches, once the representations have been learned, they are used to predict emotions.

In contrast with the two previously mentioned approaches, we propose to take into account contextual information during pre-training by using a Transformer-based model. Such an approach has been used for pre-training Transformers from visual, speech and textual modalities [23], [46], [47], [48], [49]. Haresamudram et al. use this approach to pre-train a Transformer for human activity recognition using accelerometer and gyroscope data [6]. Zerveas et al. [50] develop a framework for multivariate time-series representation learning, by pre-training a Transformer-based architecture. However, none of these works deal with uni-modal physiological signals. In this work, we have extended this approach for use with ECG signals. Specifically, we investigate the effectiveness of pre-training a Transformer for ECG emotion recognition, which to the best of our knowledge has not been done before.

III. Our approach

Our framework for using deep learning for emotion recognition is based on the following two steps: first, we need to obtain contextualized representations from time-series signals using a deep model; then, we use those representations to perform the targeted downstream task. In this paper, the considered physiological time-series are raw ECG signals, and the downstream task is binary emotion recognition: predicting high/low levels of arousal, and high/low levels of valence.

For the first step (see Figure 1.a), we developed a signal encoder based on deep neural networks and attention, to obtain contextualized representations from ECG signals. The main component of the signal encoder is a Transformer [2]. This signal encoder is pre-trained with a self-supervised task, using unlabeled ECG data. For the second step (see Figure 1.b), we fine-tune the whole model (the signal encoder and the fully-connected classifier) for our downstream task of binary emotion recognition, using labeled ECG data.

In the following subsections, we describe in detail the different components of our approach.

A. Learning Contextualized Representations

At the heart of our signal encoder is a Transformer encoder [2], which we use to learn contextualized representations of ECG signals. In Transformers, contextual information is obtained through an attention mechanism, with the attention function considered as a mapping of a query vector along with a group of key-value vector pairs to an output. In the case of the Transformer encoder, each position in the output pays attention to all positions in the input. Several attention modules (also called *heads*) are used, creating various representation subspaces and improving the ability of the model to be attentive to different positions. The Transformer encoder is constructed by stacking several layers containing a multi-head attention module followed by a fully-connected network applied to each position, with residual connections. Since our implementation of the Transformer is almost identical to the one described in [2], we refer the readers to this paper for further details.

In Figure 2, we present our signal encoder, which we describe in the remainder of this subsection.

Input Encoder: to process an ECG signal with the Transformer, we first encode it into s feature vectors of dimension d_{model} that represent each one of the s values of the ECG signal. We use 1D Convolutional Neural Networks (1D-CNN) to perform this encoding, like in [6], [36], [51]. Thus, for a raw input signal $X = \{x_1, ..., x_s\}$ where x_i is a single value, after encoding X with the input encoder we obtain features $F = \{f_1, ..., f_s\}$ where $f_i \in \mathbb{R}^{d_{\text{model}}}$.

CLS token: given that our downstream task is a classification task, we need to obtain a single representation of the whole processed signal at the output of our signal encoder. Similar to what is done in BERT [19], we append a special classification

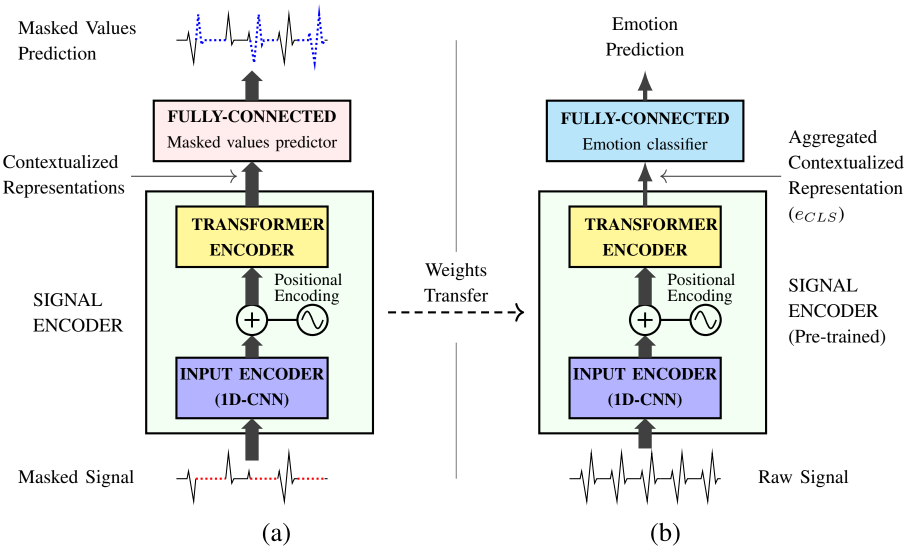

Fig. 1. Our approach with self supervised learning based on a Transformer (a) and fine-tuning strategy for learning the final emotion predictor (b).

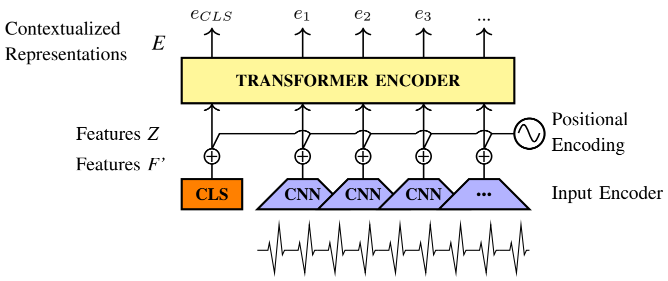

Fig. 2. Our Transformer-based signal encoder that produces contextualized representations. The aggregated representation e_{CLS} is used for classification.

token (CLS) at the start of the feature sequence F, resulting in the sequence $F' = \{CLS, f_1, ..., f_s\}$. We use a trainable vector of dimension d_{model} as CLS token. At the output of the Transformer, we obtain an embedding of the CLS token (e_{CLS}), along with the rest of the representations of the signal (see Figure 2 and Equation 2). Through the attention mechanisms of the Transformer, e_{CLS} is capable of aggregating information from the entire input signal and its contextualized representations. For this reason, at classification time, e_{CLS} can be used as input for the classifier network.

Positional Encoding: positional information of each input is required so that the Transformer can take into account the actual ordering of time-steps in the input sequence. As in [2], we use fixed sinusoidal positional embeddings. We sum the positional embeddings with the features F':

$$Z = \{CLS + pe_0, f_1 + pe_1, ..., f_s + pe_s\}, \qquad (1)$$

where $pe_i \in \mathbb{R}^{d_{\text{model}}}$ is the positional embedding for time-step i. We then apply layer normalization [52] to Z. Please refer to [2] for details on how to obtain the positional embeddings.

Transformer Encoder: we obtain contextualized representations E using a Transformer encoder with h heads and l layers on the sequence Z:

$$E = \{e_{CLS}, e_1, ..., e_s\} = \text{Transformer}_{h,l}(Z). \qquad (2)$$

We then use the representations E for emotion recognition, as is described in Section III-C

B. Pre-training Task

To pre-train our signal encoder, we employ a self-supervised approach inspired in BERT [19]. We mask random segments of a certain length by replacing them with zeros, and then we train our model to predict the masked values, as shown in Figure 1a. Labeled data is not needed for this step.

Similar to [51], a proportion p of points is randomly selected from the input signal as starting points for masked segments, and then for each starting point the subsequent M points are masked. The masked segments may overlap.

To predict masked points, we use a fully-connected network (FCN) on top of the signal encoder, as shown in Figure 1a. We only predict values of masked inputs, as opposed to reconstructing the whole signal. We use the mean square error between predicted and real values as the reconstruction loss $\mathcal{L}_r$ during pre-training:

$$\mathcal{L}_r = \frac{1}{N_m}\sum_{j=1}^{N_m}(\hat{x}_j - x_{p(j)})^2, \qquad (3)$$

where N_m is the number of masked values, $\hat{x}_j$ is the prediction corresponding to the j^{th} masked value, and $x_{p(j)}$ is the original input value selected to be the j^{th} masked value, whose position is $p(j)$ in the input signal.

C. Fine-tuning

We fine-tune our model to perform binary emotion prediction, as shown in Figure 1b. This step is supervised, using labeled data. To make the prediction, a FCN is added on top of the signal encoder, using e_{CLS} as input. We initialize the signal encoder with the weights obtained after pre-training, while the FCN is randomly initialized. We then fine-tune all the parameters of the model, including the pre-trained weights. For this task, we minimize the binary cross-entropy loss $\mathcal{L}_{ft}$:

$$\mathcal{L}_{ft} = -w_p y \log[\sigma(out)] - (1-y)\log[1-\sigma(out)] \qquad (4)$$

where y is an indicator variable with value 1 if the class of the ground truth is positive and 0 if it is negative, out is the output of the classifier, σ is the sigmoid function, and w_p is the ratio of negative to positive training samples, used to compensate unbalances that may be present in the dataset.

IV. Experimental Setup

In this section, we describe the experimental choices taken to evaluate our approach for a downstream task of binary emotion recognition (high/low levels of arousal and valence), on ECG signals. We present the datasets used, the pre-processes employed, and the parametrization of our two steps of pre-training and fine-tuning.

A. Datasets

For pre-training, we only require datasets that contain ECG signals, regardless of why they were actually collected or which labeling they have, if any. The datasets that we use in our experiments are: ASCERTAIN [16], DREAMER [53], PsPM-FR [54], PsPM-HRM5 [55], PsPM-RRM1-2 [56], and PsPM-VIS [57]. We also employ the AMIGOS dataset [17], taking care of not using the same data for pre-training and evaluating our model, as this dataset is also used for the downstream task. To gather as much data as possible, we use all the ECG channels available in the datasets. For ASCERTAIN, we discard some signals according to the quality evaluation provided in the dataset: if a signal has a quality level of 3 or worse in the provided scale, it is discarded. In total, there are around 230 hours of ECG data for pre-training.

To fine-tune our model to predict emotions, we use the AMIGOS dataset [17]. In this dataset, 40 subjects watched videos specially selected to evoke an emotion. After watching each video, a self-assessment of their emotional state is conducted. In this assessment, subjects rated their levels of arousal and valence on a scale of 1 to 9. Of the 40 subjects, 37 watched a total of 20 videos, while the other 3 subjects watched only 16 videos. During each trial, ECG data were recorded on both left and right arms. We use data only from the left arm to fine-tune our model. AMIGOS includes a pre-processed version of the data, that was down-sampled to 128Hz and filtered with a low-pass filter with 60Hz cut-off frequency. We use these pre-processed data for our experiments, including the pre-training phase. The ECG data that we use for fine-tuning amounts to around 65 hours of recordings.

B. Signal Pre-processing

We first filter signals with an 8$^{\text{th}}$ order Butterworth band-pass filter, having a low-cut-off frequency of 0.8Hz and a high-cut-off frequency of 50Hz. We then down-sample the signals to 128 Hz, except for AMIGOS which already has that sampling rate. Signals are normalized so they have zero-mean and unit-variance, for each subject independently. Signals are finally divided into 10-second segments (we also report results for segments of 20 seconds and 40 seconds).

C. Pre-training

As stated previously, we use ASCERTAIN, DREAMER, PsPM-FR, PsPM-RRM1-2, PsPM-VIS, and AMIGOS for pre-training. Since we also use AMIGOS for fine-tuning, we need to avoid using the same segments both for pre-training and for evaluating the model. To do this, we pre-train two models, one using half of the data from AMIGOS, and the second using the other half. When testing our model with certain segments from AMIGOS, we fine-tune the model that was pre-trained with the half of AMIGOS that do not contain those segments. More details are given in Section IV-D. In total, both of our models are pre-trained with 83401 10-second segments.

We select a proportion of $p = 0.0325$ points from each input segment to be the starting point of a masked span of length $M = 20$, resulting in around 47% of the input values masked.

The input encoder is built with 3 layers of 1D-CNN with ReLU activation function. We use layer normalization [52] on the first layer, and at the output of the encoder. Kernel sizes are (65, 33, 17), the numbers of channels are (64, 128, 256) and the stride for all layers is 1. This results in a receptive field of 113 input values or 0.88s. We selected this receptive field size because it is comparable with the typical interval between peaks on an ECG signal, which is between 0.6s and 1s, including when experiencing emotions [58].

The Transformer in our signal encoder has a model dimension $d_{\text{model}} = 256$, 2 layers and 2 attention heads, with its FCN size of $d_{model} \cdot 4 = 1024$. The FCN used to predict the masked values consists of a single linear layer of size $d_{\text{model}}/2 = 128$ followed by a ReLU activation function. An additional linear layer is used to project the output vector to a single value, which corresponds to the predicted value of a masked point.

We pre-train the two models for 500 epochs, warming up the learning rate over the first 30 epochs up to a value of 0.001 and using linear decay after that. We employ Adam optimization, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and L_2 weight decay of 0.005. We use dropout of 0.1 at the end of the input encoder, after the positional encoding, and inside the Transformer.

We tuned the number of layers and heads in the Transformer, the learning rate, and the warm-up duration using the Ray Tune framework [59] with BOHB optimization [60].

D. Fine-Tuning

We fine-tune our model (both the signal encoder and FCN classifier) for emotion recognition with the AMIGOS dataset, using each of the 10-second segments as a sample. As labels, we use the emotional self-assessments given in the dataset. Since these assessments provide values of arousal and valence on a scale 1 to 9, we use the average arousal and the average valence as threshold value to determine a low or a high level.

We use 10-fold cross-validation to evaluate our approach. Recall that we pre-train two signal encoders. After dividing

TABLE I
COMPARISON OF DIFFERENT STRATEGIES OF OUR APPROACH ON AMIGOS DATASET

		Arousal Acc.	**Arousal F1**	**Valence Acc.**	**Valence F1**
Aggregation Method	Max-Pooling 1	$0.85\pm6.6\mathrm{e}^{-3}$	$0.84\pm6.4\mathrm{e}^{-3}$	$0.78\pm6.5\mathrm{e}^{-3}$	$0.78\pm6.6\mathrm{e}^{-3}$
	Max-Pooling 2	$0.86\pm7.4\mathrm{e}^{-3}$	$0.84\pm7.3\mathrm{e}^{-3}$	$0.8\pm6.3\mathrm{e}^{-3}$	$0.8\pm5.9\mathrm{e}^{-3}$
	Average-Pooling 1	$0.87\pm8.3\mathrm{e}^{-3}$	$\mathbf{0.87}\pm7.3\mathrm{e}^{-3}$	$0.82\pm6.2\mathrm{e}^{-3}$	$0.82\pm6.7\mathrm{e}^{-3}$
	Average-Pooling 2	$\mathbf{0.88}\pm4.4\mathrm{e}^{-3}$	$\mathbf{0.87}\pm4.6\mathrm{e}^{-3}$	$\mathbf{0.83}\pm6.4\mathrm{e}^{-3}$	$\mathbf{0.83}\pm6.6\mathrm{e}^{-3}$
	Last Representation	$0.85\pm1.3\mathrm{e}^{-2}$	$0.84\pm1.2\mathrm{e}^{-2}$	$0.8\pm7.6\mathrm{e}^{-3}$	$0.8\pm8.0\mathrm{e}^{-3}$
Segment Length	40 seconds	$0.86\pm1.2\mathrm{e}^{-2}$	$0.85\pm1.1\mathrm{e}^{-2}$	$0.82\pm1.0\mathrm{e}^{-2}$	$0.81\pm9.9\mathrm{e}^{-3}$
	20 seconds	$0.87\pm5.6\mathrm{e}^{-3}$	$0.86\pm6.4\mathrm{e}^{-3}$	$0.82\pm7.8\mathrm{e}^{-3}$	$0.82\pm8.1\mathrm{e}^{-3}$
Our Best Approach	CLS with 10s segment	$\mathbf{0.88}\pm5.4\mathrm{e}^{-3}$	$\mathbf{0.87}\pm5.4\mathrm{e}^{-3}$	$\mathbf{0.83}\pm7.8\mathrm{e}^{-3}$	$\mathbf{0.83}\pm7.4\mathrm{e}^{-3}$

TABLE II
NO PRE-TRAINING VS PRE-TRAINED MODEL

Pre-train	**Arousal Acc.**	**Arousal F1**	**Valence Acc.**	**Valence F1**
No	$0.85\pm5.6\mathrm{e}^{-3}$	$0.84\pm5.8\mathrm{e}^{-3}$	$0.8\pm6.5\mathrm{e}^{-3}$	$0.8\pm6.4\mathrm{e}^{-3}$
Yes	$\mathbf{0.88}\pm5.4\mathrm{e}^{-3}$	$\mathbf{0.87}\pm5.4\mathrm{e}^{-3}$	$\mathbf{0.83}\pm7.8\mathrm{e}^{-3}$	$\mathbf{0.83}\pm7.4\mathrm{e}^{-3}$

AMIGOS into 10 folds, we use folds 1 to 5 to pre-train one signal encoder (SE_1), and folds 6 to 10 to pre-train the second one (SE_2) (and all data from the other datasets, for both). Then, when we fine-tune the models to be tested with folds 1 to 5, we use the weights from SE_2 to initialize the signal encoder parameters. In a similar fashion, we use SE_1 as initialization point of the signal encoder when we fine-tune the models to be tested with folds 6 to 10. This method allows us to pre-train, fine-tune and test our model in a more efficient way than pre-training 10 different models, one for each fold, while retaining complete separations between training and testing data.

The FCN classifier used to predict emotions has two hidden layers of sizes [1024, 512] with ReLU activation functions, and an output layer that projects the output to a single value. We fine-tune one model to predict arousal and another to predict valence. For each task, we fine-tune our model for 100 epochs using Adam optimization, with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and L_2 weight decay of 0.00001. We start with a learning rate of 0.0001, and decrease it every 45 epochs by a factor of 0.65. We keep using a dropout of 0.1 at the end of the input encoder, after the positional encoding, and inside the Transformer. We use dropout of 0.3 in the FCN classifier.

We used Ray Tune with BOHB, as we did on pre-training, to tune the learning rate, the learning rate schedule, the shape and dropout of the FCN classifier, and the L_2 weight decay.

V. RESULTS

In our results, we use as metrics the mean accuracy and mean F1-score between positive and negative classes. We report the mean and confidence intervals of the metrics across our 10 folds of cross-validation. The confidence intervals are calculated using a t-distribution with 9 degrees of freedom, for a two-sided 95% confidence.

A. Comparing Aggregation Methods and Segment Lengths

We report in Table I the performances of our approach for different strategical choices. Firstly, we compare different aggregation approaches to combine the contextualized representations at the output of the signal encoder, given to the FCN classifier. Secondly, we compare performances for different segment lengths used to divide the input signals.

Aggregation Method: we compared 4 strategies for aggregating representations, to be given as input to the FCN: max-pooling, average-pooling, using only the last representation e_s, and using only the embedding of the CLS token e_{CLS} (we call this strategy CLS). Max-pooling 1 and average-pooling 1 are the result of max-pooling and average-pooling across all representations, to obtain a single representation of size $d_{model} = 256$. Max-pooling 2 was optimized on the validation set: representations are reduced to a size of 64, divided into two groups, then max-pooling was applied on each group and the results concatenated to obtain a single representation of size 128. Average-pooling 2 was optimized on the validation set: representations are divided into 4 groups, average-pooling is applied on each group and the results concatenated to obtain a single representation of size 1024.

We see in Table I that the best results were obtained with average-pooling strategies and with CLS, with accuracies up to 0.88 for arousal, for example. In the following experiments, we will thus use CLS as our aggregation method. Indeed, although results are practically identical for CLS and average-pooling 2 (e.g. $0.88\pm5.4\mathrm{e}^{-3}$ compared to $0.88\pm4.4\mathrm{e}^{-3}$ accuracies for arousal), CLS has the advantage of being a commonly-used strategy for Transformers, which does not require any kind of tuning on validation data, contrary to average-pooling 2.

Segment length: we compare 3 different segment lengths for dividing ECG signals into input instances: 10, 20, and 40 second segments. We can see in Table I that shorter segments lead to better results on average, both for arousal and valence. For example for arousal, 10-second segments lead to an accuracy of $0.88\pm5.4\mathrm{e}^{-3}$, compared to $0.87\pm5.6\mathrm{e}^{-3}$ for 20-second segments, and $0.86\pm1.2\mathrm{e}^{-2}$ for 40-second segments.

Two explanations emerge for this observation: firstly, since emotions are relatively volatile states, longer segmentation might cover fluctuating emotional states, thus making it harder to characterize emotion; secondly, longer segments should

TABLE III
COMPARISON OF DIFFERENT METHODS ON AMIGOS DATASET

	Model	Subj. Ind.	Input Seg. Size	Arousal Acc.	Arousal F1	Valence Acc.	Valence F1
Various experiment protocols	Gaussian Naive Bayes [17]	Yes	20 seconds	-	0.551	-	0.545
	1D-CNN [8]	No	200 peaks	0.81	0.76	0.71	0.68
	2D-CNN [33]	Yes	Not segmented	0.83	0.76	0.82	0.80
	1D-CNN with LSTM [32]	Yes	Not segmented	-	-	0.81	0.80
	Convolutional autoencoder [22]	No	10 seconds	0.85	0.89	-	-
Our protocol	Pre-trained CNN [9]	No	10 seconds	$0.85 \pm 5.4\text{e}^{-3}$	$0.84 \pm 5.3\text{e}^{-3}$	$0.77 \pm 5.5\text{e}^{-3}$	$0.77 \pm 5.1\text{e}^{-3}$
	Pre-trained Transformer (ours)	No	10 seconds	**0.88**$\pm 5.4\text{e}^{-3}$	**0.87**$\pm 5.4\text{e}^{-3}$	**0.83**$\pm 7.8\text{e}^{-3}$	**0.83**$\pm 7.4\text{e}^{-3}$

require more complex models (i.e. bigger Transformer and FCN), which are harder to train with the relatively restricted amount of labeled data in AMIGOS. Moreover, shorter segments are faster to process, allowing a high number of training epochs and smaller learning rates. In the following experiments, we will thus use 10-second segments.

B. Effectiveness of Pre-training

To demonstrate the effectiveness of our pre-training approach, we tested our architecture by fine-tuning our model on AMIGOS with all parameters randomly initialized, instead of using a pre-trained signal encoder (thus skipping step (a) of our process in Figure 1). As reported in Table II, the pre-trained model is on average significantly better than the model with no pre-training, for both accuracy and F1-score. For example, for arousal, the pre-trained model reaches an average accuracy of $0.88 \pm 5.4\text{e}^{-3}$, compared to $0.85 \pm 5.6\text{e}^{-3}$ for the model with no pre-training. These results illustrate the benefits of pre-training Transformers for our task. Moreover, during our experiments, we observed that the model with no pre-training had a tendency to overfit quickly, which was not the case for the pre-trained model. Pre-training the model on many different datasets should increase its robustness to overfitting when fine-tuning on a specific dataset.

C. Comparisons With Other Approaches

We report in Table III various state-of-the-art results for emotion recognition from ECG signals on the AMIGOS dataset. The first section of the table contains results from works which all use different experiment protocols, such as different segment sizes, different separations of data into training and test sets, subject dependent and independent evaluations, etc. These results are therefore not directly comparable with one another, nor are they directly comparable with ours. Nevertheless, we report them to showcase the variety of state-of-the-art approaches published for this task, and give a relative idea of achieved performances on AMIGOS.

To compare our approach with another state-of-the-art approach as fairly as possible, it is required that both use exactly the same experiment protocol. For this, we fully retrained and tested the pre-trained CNN approach proposed by Sarkar and Etemad [9], with the experiment protocol we presented. To this end, we use the implementation provided by the authors[1]. To ensure fair comparisons, the exact same data was used to pre-train, fine-tune, and test both our approach and also Sarkar and Etemad's approach, for each fold of cross-validation.

We see in Table III that our approach achieves better performance on average than Sarkar and Etemad's approach with the same experiment protocol, for both arousal and valence. For example, our approach achieves an F1-score of $0.83 \pm 7.4\text{e}^{-3}$ for valence, compared to $0.77 \pm 5.1\text{e}^{-3}$ for the pre-trained CNN. These results are statistically significant with $p < 0.01$ following a t-test.

This final set of results shows that our approach, and more generally self-supervised Transformer-based approaches, can be successfully applied to obtain contextualized representations from ECG signals for emotion recognition tasks.

VI. CONCLUSIONS AND PERSPECTIVES

In this paper, we investigate the use of transformers for recognizing arousal and valence from ECG signals. This approach used self-supervised learning for pre-training from unlabeled data, followed by fine-tuning with labeled data. Our experiments indicate that the model builds robust features for predicting arousal and valence on the AMIGOS dataset, and provides very promising results in comparison to recent state-of-the-art methods. This work showcases that self-supervision and attention-based models such as Transformers can be successfully used for research in affective computing.

Multiple perspectives emerge from our work. New pre-training tasks can be investigated: other methods such as contrastive loss or triplet loss might be more efficient with regards to the specificities of ECG signals, compared to masked points prediction which we used in this work. Extending our work to other input modalities (EEC, GSR, and even non-physiological inputs such as ambient sensors) and, in general, to process multimodal situations could prove useful for improving performances of emotion recognition. Finally, larger scale experiments, with new datasets captured in varied situations, will allow for a better understanding of the behaviour of our approach.

Acknowledgements: This work has been partially supported by the MIAI Multidisciplinary AI Institute at the Univ. Grenoble Alpes: (MIAI@Grenoble Alpes - ANR-19-P3IA-0003).

[1] https://code.engineering.queensu.ca/pritam/SSL-ECG

References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "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*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[3] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, and X. Yan, "Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[4] G. Yan, S. Liang, Y. Zhang, and F. Liu, "Fusing Transformer Model with Temporal Features for ECG Heartbeat Classification," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2019, pp. 898–905.

[5] D. Ahmedt-Aristizabal, M. A. Armin, S. Denman, C. Fookes, and L. Petersson, "Attention Networks for Multi-Task Signal Analysis," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, Jul. 2020, pp. 184–187.

[6] H. Haresamudram, A. Beedu, V. Agrawal, P. L. Grady, I. Essa, J. Hoffman, and T. Plötz, "Masked reconstruction based self-supervision for human activity recognition," in *Proceedings of the 2020 International Symposium on Wearable Computers*. New York, NY, USA: Association for Computing Machinery, Sep. 2020, pp. 45–49.

[7] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A Review of Emotion Recognition Using Physiological Signals," *Sensors*, vol. 18, no. 7, p. 2074, Jul. 2018.

[8] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-González, E. Abdulhay, and N. Arunkumar, "Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS)," *IEEE Access*, vol. 7, pp. 57–67, 2019.

[9] P. Sarkar and A. Etemad, "Self-Supervised Learning for ECG-Based Emotion Recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 3217–3221.

[10] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532–541, Jul. 2020.

[11] H. Becker, J. Fleureau, P. Guillotel, F. Wendling, I. Merlet, and L. Albera, "Emotion Recognition Based on High-Resolution EEG Recordings and Reconstructed Brain Sources," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 244–257, Apr. 2020.

[12] J. Shukla, M. Barreda-Angeles, J. Oliver, G. C. Nandi, and D. Puig, "Feature Extraction and Selection for Emotion Recognition from Electrodermal Activity," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[13] S. Chen, K. Jiang, H. Hu, H. Kuang, J. Yang, J. Luo, X. Chen, and Y. Li, "Emotion Recognition Based on Skin Potential Signals with a Portable Wireless Device," *Sensors*, vol. 21, no. 3, p. 1018, Jan. 2021.

[14] M. R. Kose, M. K. Ahirwal, and A. Kumar, "A new approach for emotions recognition through EOG and EMG signals," *Signal, Image and Video Processing*, vol. 15, no. 8, pp. 1863–1871, Nov. 2021.

[15] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A Database for Emotion Analysis ;Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[16] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, Apr. 2018.

[17] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[18] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep Contextualized Word Representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Nov. 2020, pp. 1597–1607.

[21] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual Transformers: Token-based Image Representation and Processing for Computer Vision," *arXiv:2006.03677 [cs, eess]*, Nov. 2020.

[22] K. Ross, P. Hungler, and A. Etemad, "Unsupervised multi-modal representation learning for affective computing with multi-corpus wearable data," *Journal of Ambient Intelligence and Humanized Computing*, Oct. 2021.

[23] M. Macary, M. Tahon, Y. Estève, and A. Rousseau, "On the Use of Self-Supervised Pre-Trained Acoustic and Linguistic Features for Continuous Speech Emotion Recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2021, pp. 373–380.

[24] S. Song, S. Jaiswal, E. Sanchez, G. Tzimiropoulos, L. Shen, and M. Valstar, "Self-supervised Learning of Person-specific Facial Dynamics for Automatic Personality Recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[25] O. Wiles, A. S. Koepke, and A. Zisserman, "Self-supervised learning of a facial attribute embedding from video," *arXiv:1808.06882 [cs]*, Aug. 2018.

[26] Y. Li, J. Zeng, S. Shan, and X. Chen, "Self-Supervised Representation Learning From Videos for Facial Action Unit Detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 10 916–10 925.

[27] S. Roy and A. Etemad, "Self-supervised Contrastive Learning of Multi-view Facial Expressions," in *Proceedings of the 2021 International Conference on Multimodal Interaction*. Montréal QC Canada: ACM, Oct. 2021, pp. 253–257.

[28] M. Gjoreski, B. Mitrevski, M. Luštrek, and M. Gams, "An Inter-domain Study for Arousal Recognition from Physiological Signals," *Informatica*, vol. 42, no. 1, Mar. 2018.

[29] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion Recognition from Multimodal Physiological Signals for Emotion Aware Healthcare Systems," *Journal of Medical and Biological Engineering*, vol. 40, no. 2, pp. 149–157, Apr. 2020.

[30] Y. Hsu, J. Wang, W. Chiang, and C. Hung, "Automatic ECG-Based Emotion Recognition in Music Listening," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 85–99, Jan. 2020.

[31] L. Shu, Y. Yu, W. Chen, H. Hua, Q. Li, J. Jin, and X. Xu, "Wearable Emotion Recognition Using Heart Rate Data from a Smart Bracelet," *Sensors*, vol. 20, no. 3, p. 718, Jan. 2020.

[32] R. Harper and J. Southern, "A Bayesian Deep Learning Framework for End-To-End Prediction of Emotion from Heartbeat," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[33] S. Siddharth, T. Jung, and T. J. Sejnowski, "Utilizing Deep Learning Towards Multi-modal Bio-sensing and Vision-based Affective Computing," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[34] S. A. Fulop and K. Fitz, "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," *The Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 360–371, Jan. 2006.

[35] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, Dec. 2006.

[36] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6558–6569.

[37] Z. Wu, X. Zhang, T. Zhi-Xuan, J. Zaki, and D. C. Ong, "Attending to Emotional Narratives," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Sep. 2019, pp. 648–654.

[38] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal Transformer Fusion for Continuous Emotion Recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 3507–3511.

[39] C. Cai, Y. He, L. Sun, Z. Lian, B. Liu, J. Tao, M. Xu, and K. Wang, "Multimodal Sentiment Analysis based on Recurrent Neural Network and Multimodal Attention," in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*. Virtual Event China: ACM, Oct. 2021, pp. 61–67.

[40] W.-S. Chien, H.-C. Chou, and C.-C. Lee, "Self-assessed Emotion Classification from Acoustic and Physiological Features within Small-group Conversation," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*. Montreal QC Canada: ACM, Oct. 2021, pp. 230–239.

[41] N. Wu, B. Green, X. Ben, and S. O'Banion, "Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case," *arXiv:2001.08317 [cs, stat]*, Jan. 2020.

[42] A. Arjun, A. S. Rajpoot, and M. Raveendranatha Panicker, "Introducing Attention Mechanism for EEG Signals: Emotion Recognition with Vision Transformers," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, Nov. 2021, pp. 5723–5726.

[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv:2010.11929 [cs]*, Oct. 2020.

[44] B. Behinaein, A. Bhatti, D. Rodenburg, P. Hungler, and A. Etemad, "A Transformer Architecture for Stress Detection from ECG," in *2021 International Symposium on Wearable Computers*. Virtual USA: ACM, Sep. 2021, pp. 132–134.

[45] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why Does Unsupervised Pre-training Help Deep Learning?" in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, Mar. 2010, pp. 201–208.

[46] A. Khare, S. Parthasarathy, and S. Sundaram, "Multi-Modal Embeddings Using Multi-Task Learning for Emotion Recognition," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 384–388.

[47] W. Rahman, M. K. Hasan, S. Lee, A. Bagher Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating Multimodal Information in Large Pretrained Transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 2359–2369.

[48] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion," *IEEE Access*, vol. 8, pp. 176 274–176 285, 2020.

[49] A. Khare, S. Parthasarathy, and S. Sundaram, "Self-Supervised Learning with Cross-Modal Transformers for Emotion Recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2021, pp. 381–388.

[50] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A Transformer-based Framework for Multivariate Time Series Representation Learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. Virtual Event Singapore: ACM, Aug. 2021, pp. 2114–2124.

[51] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[52] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," *arXiv:1607.06450 [cs, stat]*, Jul. 2016.

[53] S. Katsigiannis and N. Ramzan, "DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, Jan. 2018.

[54] A. Tzovara, D. R. Bach, G. Castegnetti, S. Gerster, N. Hofer, S. Khemka, C. W. Korn, P. C. Paulus, B. B. Quednow, and M. Staib, "PsPM-FR: SCR, ECG and respiration measurements in a delay fear conditioning task with visual CS and electrical US." Aug. 2018.

[55] P. C. Paulus, G. Castegnetti, and D. R. Bach, "PsPM-HRM5: SCR, ECG and respiration measurements in response to positive/negative IAPS pictures, and neutral/aversive sounds," Jun. 2020.

[56] D. R. Bach, S. Gerster, A. Tzovara, and G. Castegnetti, "PsPM-RRM1-2: SCR, ECG, respiration and eye tracker measurements in response to electric stimulation or visual targets," Sep. 2019.

[57] Y. Xia, F. Melinščak, and D. R. Bach, "PsPM-VIS: SCR, ECG, respiration and eyetracker measurements in a delay fear conditioning task with visual CS and electrical US," Jul. 2020.

[58] Y. Wu, R. Gu, Q. Yang, and Y.-j. Luo, "How Do Amusement, Anger and Fear Influence Heart Rate and Heart Rate Variability?" *Frontiers in Neuroscience*, vol. 13, p. 1131, 2019.

[59] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica, "Tune: A Research Platform for Distributed Model Selection and Training," *arXiv:1807.05118 [cs, stat]*, Jul. 2018.

[60] S. Falkner, A. Klein, and F. Hutter, "BOHB: Robust and Efficient Hyperparameter Optimization at Scale," in *Proceedings of the 35th International Conference on Machine Learning*. PMLR, Jul. 2018, pp. 1437–1446.