Accommodating Missing Modalities in Time-Continuous Multimodal Emotion Recognition

Juan Vazquez-Rodriguez, Grégoire Lefebvre, Julien Cumin, James L. Crowley

ACCOMMODATING MISSING MODALITIES IN TIME-CONTINUOUS MULTIMODAL EMOTION RECOGNITION

AUTHOR VERSION

Juan Vazquez-Rodriguez[1,2], **Grégoire Lefebvre**[1], **Julien Cumin**[1], **James L. Crowley**[2]
[1] Orange Innovation, Grenoble, France
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

ABSTRACT

Decades of research indicate that emotion recognition is more effective when drawing information from multiple modalities. But what if some modalities are sometimes missing? To address this problem, we propose a novel Transformer-based architecture for recognizing valence and arousal in a time-continuous manner even with missing input modalities. We use a coupling of cross-attention and self-attention mechanisms to emphasize relationships between modalities during time and enhance the learning process on weak salient inputs. Experimental results on the Ulm-TSST dataset show that our model exhibits an improvement of the concordance correlation coefficient evaluation of 37% when predicting arousal values and 30% when predicting valence values, compared to a late-fusion baseline approach.

***K*eywords** Affective Computing, Multimodal Emotion Recognition, Machine Learning, Transformers.

1 Introduction

Technologies for automatic emotion recognition have been shown valuable for interpersonal communications [1], health and wellness concerns [1] and stress management [2], for example. People express emotions in both verbal and non-verbal manners. Facial expressions, pitch intensity or cardiac rhythms are examples of non-verbal communication.

Using multiple modalities for emotion recognition is advantageous since modalities may be complementary, and should thus improve the performance of the model when used together [3]. However, in real-world scenarios, there might be cases where a modality might not be available. If for example, the modalities consist of video, audio, and physiological signals, the camera field of view might be obstructed, the microphone can be too far away, or the physiological sensor might be on a wearable device that is not currently worn. Therefore, we need a model capable of handling missing modalities.

In this work, we extend a multimodal Transformer [4] as an encoder to obtain representations from the different modalities and a Transformer decoder [5] to process those representations and make predictions. A Transformer-based approach will continue to work in the case of missing modalities, although the performance often decreases [6]. We investigated a learning strategy to improve performance by eliminating the most important modalities during part of the training, so that the model is forced to learn from the less informative ones, that nevertheless may carry valuable features. This has two desirable effects. First, the model improves its performance, by training the model to draw information from all modalities rather than focusing on the most important ones. Second, the model becomes less sensitive to missing modalities, as it learns to handle the case where a modality is not present.

A critical aspect of multimodal emotion recognition is modeling the complementarity of information from different modalities. In other words, the model should be capable of weighing the different modalities according to their importance. We then present a novel approach of using the encoder-decoder attention (cross-attention) of the Transformer decoder to weigh the representations generated by the encoder, making this weighting scheme focus on choosing between modalities rather than paying attention to information from different time-steps. In addition, our Transformer decoder is auto-regressive, meaning that it takes into account past predicted values when doing the current inference, which is important when performing time-continuous predictions.

The main contributions of this research are: 1. we extend a multimodal Transformer-based architecture to perform time–continuous value-continuous multimodal emotion recognition, 2. we present a novel approach using cross-attention from the Transformer decoder to weigh the importance of different modalities, 3. and we develop a learning strategy to improve the performance of the model when a modality is missing.

2 Related Work

2.1 Time-Continous Multimodal Emotion Recognition

Several works address the problem of time-continuous multimodal emotion recognition. Traditionally, Long Short Term Memory - Recurrent Neural Networks (LSTM-RNN) have been employed to model the temporal relations of the inputs and to consider past predictions when predicting the current

time-step [7]. Recently, employing Transformer-based approaches [8–10] has gained popularity in addressing this task. Some works rely on LSTM-RNNs to complement the attention mechanisms from the Transformer to model the temporal information better [8], while other works use a pure attention-based approach [9, 10].

To fuse information from different modalities, some authors use late-fusion combining the outputs of single modality models [11], while others employ early-fusion by combining the input features before feeding them into the model [12]. Some approaches that use Transformer-based techniques have presented more elaborate solutions to aggregate multimodal information. Cross-modal attention [13] can be used to incorporate information from different modalities [8, 12]. Different from this, Zhang et al. [9] group the query vectors from each modality to form a single query vector and do the same for the key and value vectors. Then, they employ the grouped vectors to perform a modified version of the scaled dot-product attention described in the original Transformer paper [5]. Chen et al. [10] and He et al. [14] model temporal information using a standard Transformer approach and they model intermodal information through a multimodal attention mechanism.

A disadvantage of these approaches is that, at some point, the features coming from the different modalities are concatenated. This requires that all the modalities need to be present, thus breaking the approach if a modality is missing. Some authors have worked on addressing this situation, and we review some of these works below.

2.2 Handling Missing Modalities

There are three main types of approaches to handle missing modalities [15]: 1. learning a joint representation from the different modalities, so only one modality could be used at test time, 2. generating the missing modalities from the available ones, and 3. hiding some modalities during training.

For the first type, an example is the work of Pham et al. [16], where a joint representation is learned by encoding the text into a representation (the joint representation) and generating the other modalities from this representation. At test time, only the text input is needed. For the second type, we have the work of Mittal et al. [17], where the model generates replacement features using a learned linear transformation that converts features from the available modalities into features of the missing one. For the third type, an example is the work of Neverova et al. [18], where a carefully designed network is designed so it can still work even with missing modalities. Then, at train time, some modalities are dropped randomly to make the model robust to missing modalities.

Although these approaches make the model robust to missing modalities, a disadvantage of the first type of approach is that it cannot take advantage of using all modalities if they are present at test time. For the second type of approach, a drawback is that there is no guarantee that the generated representation accurately resembles the missing one. And to implement the third type of approach, the architecture should be capable of working with missing modalities. On the contrary, if a Transformer-based approach is used, there is no need to generate the missing modality representations, or do modifications to the architecture so it can work with modalities absent. In this case, the attention mechanisms simply do not attend to the missing modalities, and it is capable of attending to all of them if they are present.

For the reasons stated in the previous paragraph, many works that use the third type of approach to handle missing modalities use a Transformer-based model. Some examples include the work of Goncalves and Busso [19], and the work of Parthasarathy and Sundaram [20], where they use a cross-modality Transformer to combine audio and visual modalities, improving the robustness of the model to missing modalities by eliminating a modality during training. A disadvantage of using a cross-modality Transformer is that expanding the approach to use more modalities is not straightforward. To overcome this problem, a Multimodal Transformer [4] can be employed, like in the work of Ma et al. [6], where robustness to missing modalities is increased using a multitasking approach.

The Transformer-based models are well suited to model long and short temporal relations of the inputs and to model the cross-modality dependencies. Nevertheless, in the reviewed Transformer-based approaches, the attention layers have to model the temporal and the intermodal dependencies at the same time. We argue that it can be advantageous to attend only to the cross-modal dependencies when aggregating the multimodal information. In addition, the reviewed approaches do not explicitly consider past predictions when making the current prediction, which we believe is beneficial.

Our approach is a Transformer-based approach that uses a Multimodal Transformer as encoder, making it suitable for any number of modalities. We also use the novel idea of using the cross-attention from a Transformer decoder [5] to weigh the information from the different modalities. The decoder uses only the information of each modality at the current prediction time-step, relieving it from modeling the temporal information, which is done by the encoder. In addition, we explicitly use past predictions to make the current one by employing an auto-regressive approach. To handle missing modalities, we use the approach of hiding some modalities at train time, but different from the state of the art, we employ a technique to find and then hide the important modalities.

3 Approach

In this section, we provide a detailed explanation of our approach to perform multimodal time-continuous value-continuous emotion recognition. Our objective is to predict values of arousal and valence. We start this section by explaining our encoder that generates multimodal representations. Then we explain our decoder that predicts the values of arousal and valence from those representations. Finally, we describe the loss that we use to train our model.

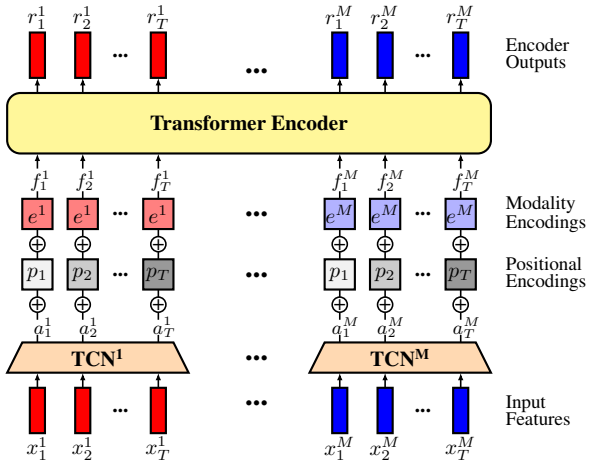


Figure 1: MultiModal Transformer Encoder (MMTE).

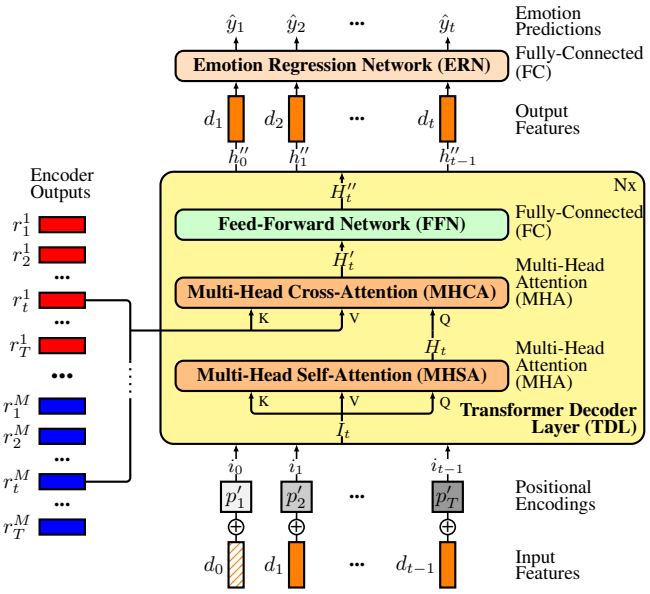


Figure 2: Auto-regressive MultiModal Transformer Decoder (AMMTD). Decoder when predicting the emotion value at time t. Nx means that there are N stacked TDL layers.

3.1 MultiModal Tranformer Encoder (MMTE)

We depict our MultiModal Transformer Encoder (MMTE) in Figure 1. Our MMTE is based on the work by Gabeur et al. [4]. The inputs for our encoder are features extracted from raw data from the different modalities. We discuss the features we use in Section 6.

The first step in our MMTE architecture is to process each modality individually using a Temporal Convolutional Network (TCN) [21] to model local temporal information, similarly to [10]. Our model learns a different TCN for each modality. We define $x_t^m \in \mathbb{R}^{d_{\text{modality}}}$ as the feature corresponding to modality m at time-step t. If we denote $[x_1^m, \ldots, x_T^m]$ as the sequence with length T of features corresponding to modality m, then during this step we have:

$$[a_1^m, \ldots, a_T^m] = \text{TCN}^m([x_1^m, \ldots, x_T^m]), \tag{1}$$

where $a_t^m \in \mathbb{R}^{d_{\text{model}}}$. For all modalities, the TCN output will have a common size d_{model}.

The next step is to add positional encodings that allow the Transformer to take into account the actual order of the sequence [5]. If the sequence of positional encodings is $P = [p_1, \ldots, p_T]$, with $p_t \in \mathbb{R}^{d_{\text{model}}}$, then the output of this step is

$$[a_1^m + p_1, \ldots, a_T^m + p_T]. \tag{2}$$

The elements of P are parameters that are learned during the training of the whole architecture.

The Transformer also needs to differentiate each modality to process cross-modality information. To do this, we follow the original multimodal Transformer from Gabeur et al., and add modality encodings. Similar to positional encodings, these modality encodings are learned during training. For each modality m, an encoding $e^m \in \mathbb{R}^{d_{\text{model}}}$ is added to the input. The output of this step is then

$$[a_1^m + p_1 + e^m, \ldots, a_T^m + p_T + e^m]. \tag{3}$$

We then concatenate the sequences from all modalities to have a single sequence. If we define the input corresponding to modality m at time-step t as $f_t^m = a_t^m + p_t, +e^m$, and if we have M modalities, the concatenated input sequence is then

$$[f_1^1, \ldots, f_T^1, \ldots, f_1^M, \ldots, f_T^M]. \tag{4}$$

We process this sequence using a Transformer encoder. The output representations r_t^m of the Transformer are given by

$$\begin{aligned}&[r_1^1, \ldots, r_T^1, \ldots, r_1^M, \ldots, r_T^M] = \\ &\text{Transformer Encoder}([f_1^1, \ldots, f_T^1, \ldots, f_1^M, \ldots, f_T^M]).\end{aligned} \tag{5}$$

Following [10], we employ a bidirectional attention mask in the Transformer encoder. When processing a specific time-step, this mask *hides* the inputs that are farther than *mask_length* positions in the future and in the past. This allows the model to concentrate on recent information, and not to worry about information too far in time that probably does not influence the current emotional state. Note that we are not hiding complete modalities, therefore this technique is not intended to make the model robust to missing modalities.

3.2 Auto-regressive MultiModal Transformer Decoder (AMMTD)

One of the contributions of this paper is to develop a decoder that predicts emotion from the multimodal representations given by the encoder. To do this, we design an Auto-regressive MultiModal Transformer Decoder (AMMTD). This decoder has two important characteristics: first, it takes previous predictions into account to determine the current emotion; second, it aggregates the representations of the different modalities, giving more weight to the more important ones.

A Transformer Decoder Layer (TDL) [5] is composed of a Multi-Head Self-Attention module (MHSA), followed by a

Multi-Head Cross-Attention module (MHCA), and followed by a fully-connected Feed-Forward Network (FFN). Residual connections are used around each of these three components. The Multi-Head Attention (MHA) mechanism in MHSA and MHCA projects a query vector q from a given position to a key vector k from another position to determine the attention (i.e. the weight) given to a value v associated with the position of k. The final value is the weighted sum of the v vectors from the different positions. We denote the MHA mechanism as

$$\text{MHA}(Q, K, V), \tag{6}$$

where the three parameters Q, K, and V indicate the sequence used as query, key, and value, respectively. More details about MHA can be found in the original Transformer paper [5].

Our AMMTD architecture is depicted in Figure 2. It is composed of a stack of TDL followed by an Emotion Regression Network (ERN). The MHSA module in the TDL uses self-attention to attend to previous predictions. To do this, we use auto-regression, meaning that the previously generated outputs are used as inputs to the decoder. Note that we cannot use the output of the ERN, i.e. the predicted emotion values $\hat{y}$, since we are predicting continuous outputs. Instead, we use the features generated by the top TDL. When predicting the emotion value at time-step t, the TDL stack should have generated a sequence $[d_1, \ldots, d_{t-1}]$ with $d_i \in \mathbb{R}^{d_{\text{model}}}$. Then, the decoder input is

$$[d_0, d_1, \ldots, d_{t-1}], \tag{7}$$

where $d_0 \in \mathbb{R}^{d_{\text{model}}}$ is a randomly initialized vector.

As for our encoder, we learn positional encodings $p'_t \in \mathbb{R}^{d_{\text{model}}}$ for our decoder and add them to the inputs before feeding them to the TDL stack. Thus, when performing the prediction at time-step t, the input sequence $I_t = [i_0, i_1, \ldots, i_{t-1}]$ with $I_t \in \mathbb{R}^{t \times d_{\text{model}}}$ becomes

$$I_t = [d_0 + p'_0, d_1 + p'_1, \ldots, d_{t-1} + p'_{t-1}]. \tag{8}$$

Inside the TDL, the features are first processed by the MHSA module. This module uses self-attention to integrate information from its own inputs. This means that the query, key, and value for the MHSA all come from the input sequence. To preserve the auto-regressive property, we make sure that a given input at a certain time-step can only attend inputs from past time-steps. Using Expression 6, the sequence of features $H_t = [h_0, h_1, \ldots, h_{t-1}]$ with $H_t \in \mathbb{R}^{t \times d_{\text{model}}}$ at the output of the MHSA module is

$$H_t = \text{MHA}(I_t, I_t, I_t). \tag{9}$$

The sequence of features H_t is then processed with the MHCA module, which is used to incorporate information from the input modalities. Specifically, the MHCA module attends to the outputs of the encoder. This means that the query comes from the output sequence of the MHSA, and the key and value come from the output of the encoder. If we are predicting the emotion value at time-step t, the MHCA attends only to the outputs of each modality corresponding to this time step. The output sequence of the MHCA, $H'_t \in \mathbb{R}^{t \times d_{\text{model}}}$, using the output of the encoder from Equation 5 and the output of the MHSA from Equation 9, is

$$H'_t = \text{MHA}([h_0, h_1, \ldots, h_{t-1}], [r_t^1, \ldots, r_t^M], [r_t^1, \ldots, r_t^M]), \tag{10}$$

Note that we force the model to only attend to the encoder outputs at time t instead of attending to all encoder outputs (or other outputs around t) because we want that the MHCA focuses only on finding the best weighting between the different modalities. We want to avoid the MHCA having to weigh which other time-steps in the different modalities might be important. Moreover, this restricts the information flow between modalities, which has been demonstrated to be beneficial [22], because it forces the shared representation to condense the most significant information.

The final step in the TDL is processing each feature of the sequence H'_t through a fully connected feed-forward network (FFN), applied independently to each position:

$$H''_t = \text{FFN}(H'_t). \tag{11}$$

If the TDL stack has more than one layer, the sequence from Equation 11 becomes the input of the next layer. Concretely, the new layer implements Equations 9, 10, and 11 using as input $I_t = H''_t$.

For the last TLD layer, the sequence in Equation 11 is the newly generated sequence $[d_1, \ldots, d_t]$ that will be used as input for the decoder to predict the emotion value for the next time-step, thus if $H''_t = [h''_0, h''_1, \ldots, h''_{t-1}]$, then $d_i = h''_{i-1}$ with $i \in [1, t]$.

Once the complete output sequence $D = [d_1, \ldots, d_T]$ has been generated, the final step is to process it with the Emotion Regression Network (ERN). As shown in Figure 2, our ERN is a Fully Connected (FC) layer that independently processes each element of the sequence D to predict the emotion values for each time-step, producing the predicted sequence $[\hat{y}_1, \ldots, \hat{y}_T]$.

3.3 Loss Function

As suggested in previous works that address the problem of recognizing emotion in a time-continuous manner [7, 11, 14], we use the concordance correlation coefficient (CCC) [23] as the loss to train our model. Specifically, the loss is

$$\begin{aligned} \mathcal{L} &= 1 - \text{CCC} \\ \text{CCC} &= \frac{2\rho\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2}, \end{aligned} \tag{12}$$

where ρ is the Pearson correlation coefficient between the predicted values $\hat{y}$ and the ground-truth values y. σ denotes the standard deviation and μ denotes the average of either the predicted or the ground-truth values.

4 Handling Missing Modalities

We now describe another contribution of our paper, which is our approach to deal with missing modalities. In our case, a

missing modality means that the modality is completely absent, thus the input sequence defined in Expression 4 is built with the remaining modalities. By construction, our model will not break in the case a modality is missing. The attention mechanisms in our Transformer-based approach can accommodate missing modalities as explained below. In the case of the MMTE, if a modality is not present, the Tranformer encoder will simply attend to the remaining modalities. Similarly, in the AMMTD, the cross-modal attention will be able to attend to the remaining representations without the need to generate a replacement for the missing ones.

Even if our approach is capable to continue working in the case of missing modalities, its performance may be degraded. To increase the robustness of the model to missing modalities, we perform an *optimized training* that we describe below.

First, we identify the most important modalities. To do this, we first train our model in a standard way, then test it without one modality at a time. We can then identify in which cases the performance is reduced more, meaning that the missing modalities in those cases should be important. Next, we retrain the model, without the important modalities a part of the time. Specifically, for each batch, we randomly select to eliminate the important modality i with probability $\rho^i_{\text{eliminate}}$, and to keep all modalities with probability $\rho_{\text{none}} = 1 - \sum_{i=1}^{n} \rho^i_{\text{eliminate}}$, where n is the number of important modalities.

Our reasoning behind this training strategy is that by hiding the important modalities, the model is forced to learn from the remaining ones, thus making the model more robust when those important modalities are missing. Moreover, we believe that this training strategy should lead to better results in general, even without missing modalities, as more information will be incorporated from all the modalities, rather than just relying on the important ones.

5 Experimental Setup

We test our model on the task of recognizing time-continuous values of arousal and valence. In this section, we describe the dataset, features, and parameters of our model for these experiments.

5.1 Dataset

To evaluate our model, we use the Ulm-Trier Social Stress Test dataset (Ulm-TSST), which was presented for the Muse 2021 Challenge [24, 25]. This is a multimodal dataset, where participants were recorded in a stressful situation emulating a job interview, following the TSST protocol [26]. Each participant gave a 5 minutes speech supervised by two interviewers, who did not intervene during that time. Besides audio and video, the following physiological signals are collected: Electrocardiogram (ECG), Respiration (RESP), and heart rate (BPM). A transcription of the speech is also provided.

The data set was annotated by 3 raters giving continuous values of arousal and valence in the range $[-1, 1]$. The annotations are done in a time-continuous fashion every 0.5s. To aggregate the valence annotations from the 3 raters, Rater Aligned Annotation Weighting (RAAW) [27] is used. For arousal, the annotations corresponding to the lowest inter-rater agreement are discarded, and replaced by the subject's Electrodermal Activity (EDA) signal recorded during the session. The authors of the dataset do this because EDA has been demonstrated to be a good indicator of arousal [28]. Like it was done for valence, RAAW is used to aggregate the annotations from the two remaining raters and the EDA signal.

The dataset includes 69 samples, each being a 5-minute presentation given by a subject. In the original dataset, 41 samples are used as train set, 14 as validation set, and 14 as test set. Since annotations are not provided for the test set, we randomly pick 4 samples from the validation set and 6 from the train set to form a new test set with 10 samples. In summary, we have 35 samples in the train set, 10 in the validation set, and 10 in the test set.

5.2 Input Features

We use audio, video, and physiological signals as input modalities, with features directly provided in Ulm-TSST. All features are aligned with annotations; that is, they are sampled at a rate of 2 Hz. For audio, we use extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features. For video, we use Facial Action Units (FAU) intensity. For physiological signals, the features are the concatenation of the values of ECG, RESP, and BPM. We selected these features by running some experiments in the baseline model provided by the authors of the Ulm-TSST dataset and selecting the features that lead to good performance.

5.3 Model Hyperparameters and Training

We optimize the hyperparameters of our model using the Ray Tune Framework [29] based on the validation set. Our model is parameterized as follows: we use a TCN with 6 layers and a kernel of size 9, with ReLU activation function. We have a model dimension of $d_{\text{model}} = 64$. In the Transformer encoder and the TDL, we use the GELU activation function. The size of the FFN inside the Transformers is $d_{\text{model}} \times 4 = 256$. We use a Transformer encoder with 2 attention heads and 2 layers. Our decoder is composed of a single TDL with one head. The FC layer in the ERN has a single hidden layer of $d_{\text{model}}/2 = 32$, with ReLU activation function. The bidirectional attention mask for the Transformer encoder has a *mask_length* of 50 seconds (100 time-steps).

During training, we segment each 5-minute sample into smaller samples, as suggested by Christ et al. [7]. Searching across different options, we found that segments of 125 seconds (250 time-steps) with a hop size of 25 seconds (50 time-steps) work well in our experimental protocol.

We train our model with a batch size of 64 for a maximum of 100 epochs. We start with a learning rate of 0.0001, and halve it if the metric does not improve for 5 epochs on the validation set, and we early-stop the training if there is not improvement for 15 epochs. We use Adam optimizer with $B_1 = 0.9$ and

$B_2 = 0.999$. We use a dropout rate of 0.2 throughout all the model.

6 Experiments

This section presents and discusses the experimental results. For each experiment, we obtain 30 results by training the model with 30 different initialization seeds, reporting the average of those results. We use the Holm-Bonferroni method to assert statistically significant difference in the comparisons in Section 6.1. For other results, we do a t-test using a threshold of p-value < 0.05 to assert statistical significance, using a one-sided t-test to state that a result is significantly better than another, and a two-sided t-test to check for statistical difference. We use as metric the Root-Mean-Square Error (RMSE) and the Concordance Correlation Coefficient (CCC) [23] between ground truths and predicted values.

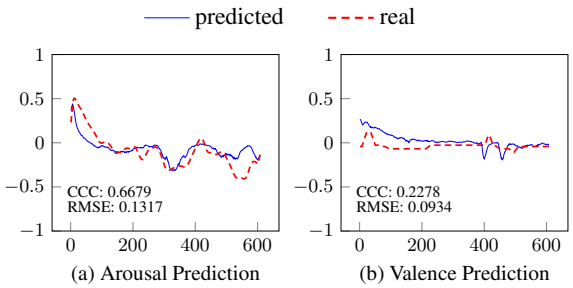
6.1 Performance with all Modalities Present

We present in Table 1 the performance of our model, along with baseline approaches, when all modalities are present. Approach Nº1 corresponds to the baseline model developed for the Muse 2022 Challenge [7], where the Ulm-TSST dataset was presented. We use the provided code[1] and the original hyperparameters to evaluate this model with the same features we employ, using our partition of the Ulm-TSST dataset. This approach is based on Long Short-Term Memory (LSTM) networks and uses late-fusion to aggregate the different modalities. Approach Nº2 corresponds to a model where instead of using our AMMTD to process the representations from the MMTE, it uses directly the ERN. Recall that the ERN is a FC network that performs regression of the emotion values. Similarly, approach Nº3 uses an LSTM to process the outputs of the MMTE. The LSTM has 4 layers, with a hidden dimension of 32. We used a grid search to tune this LSTM. For both approaches Nº2 and 3, the input is the concatenation of all modalities per time-step. The last two entries in Table 1 correspond to the approach presented in this paper. Approach Nº4 is our model trained in a standard way, i.e. with all modalities present during training. Approach Nº5 is our model trained with our optimized training strategy as presented in Sections 4 and 6.2, i.e. hiding some modalities during training.

In Figure 3, we present an example of the predictions of our model when using the optimized training strategy. For the same sample, we present the results when predicting arousal, Fig. 3(a), and valence, Fig. 3(b). As observed, the real valence values tend to be flat and have less variability than the arousal values, which we noted is a common occurrence in the dataset.

6.1.1 Comparison to the LSTM-based baseline model

If we compare our approach with standard training (approach Nº4 in Table 1) with the LSTM baseline (approach Nº1), we can see that in all metrics except for valence RMSE, our model performs better than the LSTM baseline, demonstrating that in general, our Tranformer-based approach is well suited for this task.



(a) Arousal Prediction (b) Valence Prediction

Figure 3: Example of an output of our model with optimized training compared with the ground-truth, when predicting arousal (a) and valence (b) for the same sample.

6.1.2 Comparison with other predictors

To test our idea of using cross-attention to weigh the input modalities and an auto-regressive approach to incorporate past predictions, we compare our approach with standard training (approach Nº4), with approaches Nº2 and Nº3 in Table 1, that use the ERN and an LSTM respectively instead of our AMMTD. The results show that our AMMTD module increases the performance in most of the metrics, demonstrating the effectiveness of our ideas of using cross-attention and auto-regression. The performance of our approach is statistically significantly better for all the metrics except for valence RSME, where although our approach outperforms both baselines, the improvement is not statistically significant.

In general, the baseline models perform well in terms of the RMSE metric when predicting valence, but our model performs better in terms of the CCC metric. We hypothesize that this behavior is produced because the simpler architecture of the baselines is good enough to predict flat sequences of valence values that are close enough to the flat ground-truth. On the other hand, those approaches fail to predict the small changes in the valence values, penalizing the CCC score.

6.1.3 Comparison with our optimized training strategy

The results presented in entries Nº4 and Nº5 in Table 1 show that our optimized training approach, designed to improve the handling of missing modalities, also has the desirable effect of increasing the performance of the model when all modalities are present. For example, arousal RMSE decreases from 0.2948 to 0.2869 and valence CCC increases from 0.1502 to 0.1656. As we expected, the model seems to learn to use more information from the *weak* modalities, improving the overall performance.

6.2 Accommodating Missing Modalities

We present in Table 2 the results of experiments we conducted when a modality is missing.

[1] https://github.com/EIHW/MuSe2022

Table 1: Comparison of our results with the baselines. The best result is in bold, the second best is underlined. The standard deviation is in parentheses. (•), (†), (‡) indicate that our results are statistically significantly different than the late-fusion, MMTE+FC, and MMTE+LSTM baselines, respectively. (↓) and (↑) indicate that a lower and a higher score is desirable respectively.

		Arousal		Valence	
Nº	**Approach**	**RMSE↓**	**CCC↑**	**RMSE↓**	**CCC↑**
1	Late-fusion (LSTM) [7]	0.3046 (0.0199)	0.2702 (0.0258)	**0.1585** (0.0156)	0.1273 (0.0528)
2	MMTE+FC (ERN)	0.3238 (0.0227)	0.1388 (0.0682)	0.1850 (0.0167)	0.1221 (0.0309)
3	MMTE+LSTM	0.3189 (0.0244)	0.1387 (0.0575)	0.1842 (0.0653)	0.0435 (0.0640)
4	MMTE+AMMTD (ours)	<u>0.2948</u>†‡ (0.0125)	<u>0.3578</u>•†‡ (0.0317)	0.1796 (0.0114)	<u>0.1502</u>†‡ (0.0272)
5	MMTE+AMMTD, optimized train (ours)	**0.2869**•†‡ (0.0120)	**0.3703**•†‡ (0.0351)	<u>0.1739</u> (0.0089)	**0.1656**•†‡ (0.0169)

Table 2: Summary of results when modalities are missing, for standard training and our optimized training strategy. We use bold font to indicate that the result is better than its counterpart trained in a different fashion, and if it is statistically significantly better we indicate this with the symbol (‡). We use a checkmark (✓) to indicate that a result obtained with a modality missing is not statistically significantly different than the result obtained with all the modalities.

		All Modalities		Missing Audio		Missing Video		Missing Physio	
	Training Mode	**RMSE↓**	**CCC↑**	**RMSE↓**	**CCC↑**	**RMSE↓**	**CCC↑**	**RMSE↓**	**CCC↑**
AROUSAL	Standard	0.2948	0.3578	0.2926✓	0.3589✓	0.3252	0.2713	0.2920✓	0.3539✓
	Optimized	**0.2869**‡	**0.3703**	**0.2850**✓‡	**0.3644**✓	**0.3249**	**0.2984**‡	**0.2878**✓	**0.3571**✓
VALENCE	Standard	0.1796	0.1502	0.2533	0.0738	0.2170	0.1564✓	0.1808✓	0.1486✓
	Optimized	**0.1739**‡	**0.1656**‡	**0.2052**‡	**0.1170**‡	**0.1809**✓‡	**0.1676**✓‡	**0.1746**✓‡	**0.1637**✓‡

First, we analyze the case where we are predicting arousal with the model trained in a standard way. In Table 2, we see that there is no significant performance degradation when the audio or physiological modalities are missing. For example, when audio is missing, RMSE goes from 0.2948 to 0.2926, and CCC goes from 0.3578 to 0.3589. These differences are not statistically significant, as indicated by the checkmark (✓). We also see that performance drops significantly when the video modality is missing, with RMSE increasing from 0.2948 to 0.3252 and CCC decreasing from 0.3578 to 0.2713. These results confirm that our model continues to accurately predict arousal when a modality is missing, although performance is reduced in some cases.

For valence, we see that there is no significant performance degradation when physiological signals are missing, with RMSE going from 0.1796 to 0.1808 and CCC going from 0.1502 to 0.1486. On the contrary, the model performance drops significantly when the audio or video modalities are missing. For example, RMSE increases from 0.1796 to 0.2533 when the audio modality is missing and increases to 0.2170 when the video modality is missing. Much like for arousal, these results show that our model continues to estimate valence when a modality is missing, albeit with a drop in performance.

With these results, we can identify the most important modalities for our model, by identifying the modalities that when missing, induce a significant drop in performance. When predicting arousal, the most important modality is video, and for valence the most important modalities are audio and video.

Models tend to rely heavily on the important modalities, even though other modalities may carry useful features for recognizing emotions. To improve this, we apply our optimized training strategy of hiding important modalities during parts of the training, forcing the model to rely on other modalities. Specifically, we use the strategy described in Section 4, eliminating the video modality with probability $\rho^{\text{video}}_{\text{eliminate}} = 0.25$, and maintaining all modalities with probability $\rho_{\text{none}} = 0.75$ when training for arousal prediction. For valence, we use $\rho^{\text{audio}}_{\text{eliminate}} = 0.333$, $\rho^{\text{video}}_{\text{eliminate}} = 0.333$ and $\rho_{\text{none}} = 0.334$. These probabilities were found empirically by testing several configurations and keeping the best ones for the validation set.

We see in Table 2 that our optimized training strategy improves all results when any modality is missing. For example, when the physiological signals are missing, CCC improves from 0.3539 to 0.3571 when predicting arousal and from 0.1486 to 0.1637 when predicting valence. Notably, the improvement is statistically significant, as indicated by the double dagger (‡), in all cases when the *important* modalities are missing, except for RMSE when predicting arousal with the video modality missing. In addition, the performance of the model improves even when all modalities are present. For instance, RMSE decreases from 0.2948 to 2869 for arousal and from 0.1796 to 0.1739 for valence.

This shows that our optimized training strategy works well, making our model less reliant on the important modalities and using more information from the other ones. This strategy improves the performance of our approach not only when a modality is missing, but also when all modalities are present.

7 Conclusions and Perspectives

In this work, we presented a novel Transformer-based architecture with the coupling of self- and cross-attention mechanisms for emotion recognition from multimodal signals. We experimentally showed, using the Ulm-TSST dataset, that our proposal can competitively recognize emotional valence and arousal. In addition, we demonstrated that our optimized learning strategy improves performance. Consequently, our architecture is capable of reaching high performances for emotion recognition even with missing modalities.

Future works include investigating new ways to pre-train multimodal models with contrastive learning strategies. Metric learning between different signals, sharing the same semantic meaning, could positively influence the whole system to build strong correlations between input sequences produced at different moments in time. Moreover, learning similarities between signals even when some modalities are missing may help to better understand how pretraining on Transformer-based models behave for multimodal emotion recognition.

Ethical Impact Statement

We consider two main issues: privacy and harmful applications. First, regarding privacy, knowing the emotional state of a person leads to privacy concerns since it is evident that the emotional state is a private matter. Therefore, privacy should be guaranteed. There might be cases that people involved might agree to share this private information, as in the case of people participating in the dataset that we use. That is why we adhere to the dataset usage agreement. In the general case, to avoid privacy concerns, we think the emotional record of a person should be treated at the same level of privacy that medical records are treated.

The second issue has to do with potentially harmful applications. Although in this work we do not develop an application that uses the emotional state of a person as input, we are aware that bad actors may use emotion recognition techniques in unethical ways. For example, knowing the emotional state of a person can be used to manipulate their behavior. Direct control of this is out of our hands, and therefore, we think the affective computing community should pressure governmental entities for laws and ways to control these types of applications.

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).

References

[1] A. Zenonos, A. Khan, G. Kalogridis, S. Vatsikas, T. Lewis, and M. Sooriyabandara, "HealthyOffice: Mood recognition at work using smartphones and wearable sensors," in *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, Mar. 2016, pp. 1–6.

[2] E. Nunez, M. Hirokawa, M. Perusquia-Hernandez, and K. Suzuki, "Effect on Social Connectedness and Stress Levels by Using a Huggable Interface in Remote Communication," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Sept. 2019, pp. 1–7.

[3] Yassine Ouzar, Frédéric Bousefsaf, Djamaleddine Djeldjli, and Choubeila Maaoui, "Video-Based Multimodal Spontaneous Emotion Recognition Using Facial Expressions and Physiological Signals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2460–2469.

[4] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid, "Multi-modal Transformer for Video Retrieval," in *Computer Vision – ECCV 2020*, Cham, 2020, vol. 12349, pp. 214–229, Springer International Publishing.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is All you Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[6] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng, "Are Multimodal Transformers Robust to Missing Modality?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18177–18186.

[7] Lukas Christ, Shahin Amiriparian, Alice Baird, Panagiotis Tzirakis, Alexander Kathan, Niklas Müller, Lukas Stappen, Eva-Maria Meßner, Andreas König, Alan Cowen, Erik Cambria, and Björn W. Schuller, "The MuSe 2022 Multimodal Sentiment Analysis Challenge: Humor, Emotional Reactions, and Stress," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, New York, NY, USA, Oct. 2022, pp. 5–14, Association for Computing Machinery.

[8] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue 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.

[9] Su Zhang, Ruyi An, Yi Ding, and Cuntai Guan, "Continuous Emotion Recognition using Visual-audio-linguistic Information: A Technical Report for ABAW3," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA, June 2022, pp. 2375–2380, IEEE.

[10] Haifeng Chen, Dongmei Jiang, and Hichem Sahli, "Transformer Encoder With Multi-Modal Multi-Head Attention for Continuous Affect Recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 4171–4183, 2021.

[11] Yiping Liu, Wei Sun, Xing Zhang, and Yebao Qin, "Improving Dimensional Emotion Recognition via Feature-wise Fusion," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, New York, NY, USA, Oct. 2022, pp. 55–60, Association for Computing Machinery.

[12] Vincent Karas, Mani Kumar Tellamekala, Adria Mallol-Ragolta, Michel Valstar, and Bjorn W. Schuller, "Time-Continuous Audiovisual Fusion with Recurrence vs Attention for In-The-Wild Affect Recognition," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA, June 2022, pp. 2381–2390, IEEE.

[13] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019, pp. 6558–6569, Association for Computational Linguistics.

[14] Yu He, Licai Sun, Zheng Lian, Bin Liu, Jianhua Tao, Meng Wang, and Yuan Cheng, "Multimodal Temporal Attention in Sentiment Analysis," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, New York, NY, USA, Oct. 2022, pp. 61–66, Association for Computing Machinery.

[15] Jinming Zhao, Ruichen Li, and Qin Jin, "Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Online, Aug. 2021, pp. 2608–2618, Association for Computational Linguistics.

[16] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos, "Found in Translation: Learning Robust Joint Representations by Cyclic Translations between Modalities," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 6892–6899, July 2019.

[17] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha, "M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 02, pp. 1359–1367, Apr. 2020.

[18] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout, "ModDrop: Adaptive Multi-Modal Gesture Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, Aug. 2016.

[19] Lucas Goncalves and Carlos Busso, "AuxFormer: Robust Approach to Audiovisual Emotion Recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 7357–7361.

[20] Srinivas Parthasarathy and Shiva Sundaram, "Training Strategies to Handle Missing Modalities for Audio-Visual Expression Recognition," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, New York, NY, USA, Dec. 2021, pp. 400–404, Association for Computing Machinery.

[21] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *arXiv:1803.01271*, Apr. 2018.

[22] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun, "Attention Bottlenecks for Multimodal Fusion," in *Advances in Neural Information Processing Systems*. 2021, vol. 34, pp. 14200–14213, Curran Associates, Inc.

[23] L. I. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, Mar. 1989.

[24] Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Meßner, Erik Cambria, Guoying Zhao, and Björn W. Schuller, "The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress," in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, Virtual Event China, Oct. 2021, pp. 5–14, ACM.

[25] Lukas Stappen, Eva-Maria Meßner, Erik Cambria, Guoying Zhao, and Björn W. Schuller, "MuSe 2021 Challenge: Multimodal Emotion, Sentiment, Physiological-Emotion, and Stress Detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, New York, NY, USA, Oct. 2021, pp. 5706–5707, Association for Computing Machinery.

[26] C. Kirschbaum, K. M. Pirke, and D. H. Hellhammer, "The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993.

[27] Lukas Stappen, Lea Schumann, Benjamin Sertolli, Alice Baird, Benjamin Weigell, Erik Cambria, and Björn W. Schuller, "MuSe-Toolbox: The Multimodal Sentiment Analysis Continuous Annotation Fusion and Discrete Class Transformation Toolbox," in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, New York, NY, USA, Oct. 2021, pp. 75–82, Association for Computing Machinery.

[28] Alice Baird, Lukas Stappen, Lukas Christ, Lea Schumann, Eva-Maria Messner, and Björn W. Schuller, "A Physiologically-Adapted Gold Standard for Arousal during Stress," in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, New York, NY, USA, Oct. 2021, pp. 69–73, Association for Computing Machinery.

[29] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica, "Tune: A Research Platform for Distributed Model Selection and Training," *arXiv:1807.05118 [cs, stat]*, July 2018.