Time-Continuous Audiovisual Fusion with Recurrence vs Attention for In-The-Wild Affect Recognition

Vincent Karas*
University of Augsburg
Augsburg, Germany

Mani Kumar Tellamekala
University of Nottingham
Nottingham, UK

vincent.karas@informatik.uni-augsburg.de

Adria Mallol-Ragolta
University of Augsburg
Augsburg, Germany

Michel Valstar
University of Nottingham
Nottingham, UK

Björn W. Schuller
University of Augsburg
Augsburg, Germany

Abstract

This paper presents our contribution to the 3rd Affective Behavior Analysis in-the-Wild (ABAW) challenge. Exploiting the complementarity among multimodal data streams is of vital importance to recognise dimensional affect from in-the-wild audiovisual data, as the contribution affect-wise of the involved modalities might change over time. Recurrence and attention are two of the most widely used modelling mechanisms in the literature for capturing the temporal dependencies of audiovisual data sequences. To clearly understand the performance differences between recurrent and attention models in audiovisual affect recognition, we present a comprehensive evaluation of fusion models based on LSTM-RNNs, self-attention, and cross-modal attention, trained for valence and arousal estimation. Particularly, we study the impact of some key design choices: the modelling complexity of CNN backbones that provide features to temporal models, with and without end-to-end learning. We train the audiovisual affect recognition models on the in-the-wild Aff-wild2 corpus by systematically tuning the hyper-parameters involved in the network architecture design and training optimisation. Our extensive evaluation of the audiovisual fusion models indicate that under various experimental settings, compared to RNNs, attention models may not necessarily be the optimal choice for time-continuous multimodal fusion for emotion recognition.

1. Introduction

The growing market penetration of *smart* devices is radically increasing the number of scenarios where we interact with machines. Nowadays, such interactions take place in a wide range of environments, including the workplace, at home, or even inside our vehicles. If technology is going to accompany us in all aspect of our lives, powering machines with affective capabilities is a requirement to humanise technology towards a more natural *Human-Computer Interaction* (HCI). Creating more human-like technology is one of the objectives of Affective Computing [31].

This paper focuses on the automatic recognition of valence and arousal with the aim to develop Emotional Artificial Intelligence solutions that could allow machines to adapt to the users' affective states. For instance, in the vehicle environment, if the car detects that the driver has been showing high levels of arousal and negative levels of valence, it could interpret this as anger. Alternatively, low levels of arousal and negative valence may indicate sadness or fatigue [9]. In these cases, the car could suggest playing calm music or even pulling over to take some rest for the safety of the driver and others. Mood improvement and relaxation systems already exist on the market for some premium brands, but knowing when to suggest them and adapting them based on the detected emotions could greatly enhance the user experience [1].

However, automatically detecting emotions and moods in a setting as described above, or any scenario in an uncontrolled environment, remains an open problem. It is commonly referred to as emotion recognition in-the-wild, and presents several challenges: Data is often noisy, for the visual modality, a person's face may be partially occluded, or there may be rapid changes in the illumination. For the acoustic modality, a person' voice may be less intelligible because of background noise. Another issue lies in cross-cultural emotion recognition [34] *i.e.*, automatic affect recognition systems needing to perform reliably for people of very diverse backgrounds and origins, who may express their feelings differently. In order to tackle this difficult problem, various methods have been proposed. These

frequently involve fusing multiple modalities in order to better judge the emotional state from complementary information [32]. Another common strategy is using temporal information, since the emotional state fluctuates over time.

For the purpose of processing time series, Recurrent Neural Networks (RNNs) continue to be popular. RNNs look at each element of the input sequentially and update their hidden states. Recent works in emotion recognition have also made use of networks based on self-attention and cross-modal attention [4, 14]. While self-attention finds relations between the elements of one sequence, cross-modal attention relates two sequences from different modalities to each other [41]. Compared to RNNs, attention-based network architectures have the advantage of allowing parallel computations. However, adding recurrence may still improve attention-based networks [14].

Although recurrent and attention models have been widely applied to the multimodal fusion for affect recognition and sentiment analysis [2, 28, 38, 39], it is not very clear how the quality of the input feature embeddings affects performance when modelling complex interactions between the modalities, particularly in end-to-end learning approaches. Specifically, to the best of our knowledge, not much attention is paid to comprehensively analysing the performance of RNNs and attention models based on the underlying CNN backbones' characteristics. To this end, we consider two commonly used CNN backbone models of two different complexity levels for extracting face image features: FaceNet, based on the InceptionResNetV1 architecture, and MobileFaceNet, based on the MobileNetV2 architecture. Using the visual features extracted using these two CNN backbones and systematically tuning the hyperparameters of network design and optimisation, we comprehensively evaluate the performance of LSTM-RNNs, self-attention and cross-modal attention models on the task of audiovisual affect recognition.

Herein, we present this comparative analysis of RNNs, self-attention and cross-modal attention as part of our contribution to the 3rd Affective Behavior in-the-Wild (ABAW) challenge. While similar comparisons have been performed, we focus our analysis specifically on the task of continuous emotion recognition in the wild. The Affwild2 dataset used in the challenge is the largest in-the-wild corpus annotated in terms of valence and arousal [20]. Its data presents many of the difficulties listed above, including a high diversity of subjects, varying illumination, and occlusions, and frequently noisy audio. We believe that it is beneficial to benchmark the algorithms on such a dataset. Our main contributions are:

1. We investigate the impact of CNN backbones with different complexities on the performance of LSTM-RNNs for audiovisual affect recognition in the wild, and show the effectiveness of end-to-end-learning.

2. We contrast the performance of LSTM-RNNs with self-attention and cross-modal attention and show that recurrent models can outperform attention models in combination with low-complexity CNN backbones.

The rest of the paper is structured as follows: We present our methodology in Sec. 3 and describe our experimental settings and results in Sec. 4. A discussion of the results follows in Sec. 5, and we suggest future works in Sec. 6. Finally, Sec. 7 concludes this paper.

2. Related Work

Recurrence vs Attention for Sequence Modelling. To model the underlying temporal dynamics embedded in the time-continuous data, recurrent [13] and attention [43] mechanisms have been widely used. While the recurrence models rely on gated sequential propagation of temporal dynamics encoded into a latent state, the attention models bypass the sequential propagation of information and directly attend to the past inputs. Thus, the attention models can capture long-term temporal contingencies by circumventing the problem of vanishing gradients. Although LSTM-RNNs [13] are designed to capture the long-term dependencies by controlling the information flow, they still fall short in practice due to their fixed dimensional latent state to hold the past information, unlike in the attention models.

Building on the success of attention mechanism [43] in sequence data modelling in recent years, cross-modal attention-based audiovisual fusion has been widely applied to the emotion recognition tasks [10, 23, 33, 46, 47]. However, this advantage with attention models comes at the cost of poor (quadratic) scalability with sequence length, which is not the case with RNNs. Furthermore, attention models can operate only within a fixed temporal context window whereas the RNNs can handle unbounded context [11].

Some recent works [15, 30] made systematic efforts to understand the trade-offs between the recurrence and attention mechanisms. However, in the case of time-continuous multimodal fusion, which requires modelling complex interactions among different modalities, not much is known about how their performance is influenced by some key design choices; for instance, the CNN backbone modelling complexity and the resultant input features quality. This observation motivates our attempt to study the impact of CNN backbones on the performance of LSTMs, self-attention, and cross-modal attention models, by systematically tuning the hyper-parameters involved in the network architecture design and training optimisation.

In-The-Wild Audiovisual Affect Recognition. The first affect in the wild challenge based on the Aff-wild dataset was introduced at CVPR 2017 [44]. In [19], the dataset and the challenge are described. Aff-wild contains 298 videos sourced from YouTube. Shown in it are subjects reacting

to a variety of stimuli; *e.g.*, film trailers. Subsequently, the corpus was extended with additional videos, and renamed to Aff-wild2 dataset [20]. Aff-wild2 is composed of 548 videos, with a total of about $2,78\,M$ frames. The total number of subjects is 455, of whom 277 are male. The dataset is annotated with three sets of labels: continuous affect (valence and arousal), basic expressions (six emotions and neutral), and Facial Action Units (FAUs). 545 videos have annotations for valence and arousal.

The first ABAW challenge was held as a workshop at FG2020 [17]. It consisted of three sub-challenges for estimating valence-arousal (VA track), classifying facial expressions (EXPR track), and detecting 8 FAUs (AU track). The winning team of the VA track [6] relied on a multi-task learning approach. To deal with the problem of incomplete labels in Aff-wild2 data used for the first ABAW competition, *i.e.* not all samples being annotated for each task, Deng *et al.* [6] proposed a teacher-student framework. A deep ensemble model was trained as the teacher, predicting the missing labels to guide the student.

In 2021, the second ABAW challenge took place in conjunction with ICCV 2021 [22]. Compared to the previous year, the database had been extended with more annotations. The challenge tracks were identical, but the AU track now included 12 FAUs. The winner of the VA track [7] was the same team as in the previous year, again utilising a multi-task teacher-student framework. The approach also included the prediction uncertainty of an ensemble of student models to further improve the performance.

Several multi-task learning models [6, 8, 24] effectively leveraged the availability of Aff-wild2 data jointly annotated with the labels of dimensional affect, categorical expressions, and FAUs. A holistic multi-task, multi-domain network for facial emotion analysis named FaceBehaviorNet was developed on Aff-wild2 and validated in a cross-corpus setting in [18, 19, 21].

3. Methodology

Since we want to compare fusion methods for time-continuous emotion recognition, our method is based on deep neural networks operating on sequences of features extracted from the visual and audio modalities. We use the cropped and aligned faces from the videos as visual inputs and fixed-length clips as audio inputs. Our approach is illustrated in Fig. 1.

3.1. Visual Features

Visual features are extracted with the help of 2D-CNNs. We use pre-trained networks trained on facial recognition tasks. Specifically, we use FaceNet [37], based on the Inception-Resnetv1 architecture and trained on VGGFace2 [3]. We also employ MobileFaceNet [5], a lightweight architecture designed for facial recognition in embedded devices. MobileFaceNet is built upon residual blocks used in the MobileNetv2 network [36]. Its usefulness as a feature extractor for emotion recognition was demonstrated in [7]. Both CNNs return 512-dimensional feature embeddings. The FaceNet has approximately $27\,M$ parameters, while the MobileFaceNet has $0.99\,M$ parameters.

3.2. Audio Features

For audio feature extraction, we choose a 1D CNN network based on the architecture proposed by Zhao *et al.* [48]. The CNN encoder has 4 local feature learning blocks consisting of 1D convolutions and maxpooling layers. The kernel sizes and output channels are [3,3,3,3] and [64, 64, 128, 128]. The choice of this architecture is motivated by its low parameter count (about $88\,k$) and proven effectiveness for speech emotion recognition on a number of corpora.

We use the RECOLA dataset [35], a corpus of spontaneous affective interactions between French speakers, to pre-train the audio network. For this purpose, we combine the 1D-CNN with a 2-layer LSTM and a fully connected output layer and train the model end-to-end using the End2You toolkit [1] [42]. Then, the LSTM and output layers were removed to obtain the convolutional feature extractor. We then added a global average pooling layer at the end so the network returns 128-dimensional embeddings.

3.3. Sequence Modelling

Recurrence Models (RNN) are widely used for sequential data modelling, whose fundamental strength lies in their ability to learn the underlying temporal context in the form of a hidden state; *i.e.*, $h_t = f(h_{t-1}, ..)$. This approach based on maintaining the hidden states is a natural solution to model the sequential data that is irregularly sampled from an underlying time-continuous series phenomenon [11] such as dimensional affect recognition. However, the limitations of recurrence models in terms of capturing cross-modal interactions in multimodal temporal data, which is critical for audiovisual emotion recognition, is not very clear. In this work, we consider the canonical Long-Short Term Memory (LSTM) RNNs [13], using both unidirectional and bidirectional models, for a comprehensive evaluation on audiovisual affect estimation.

Self-Attention (SA). Second, we use networks based on the Transformer architecture [43]. Specifically, we use multi-headed scaled dot-product attention blocks with feed-forward networks as employed in the transformer encoder. The scaled dot-product attention is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (1)$$

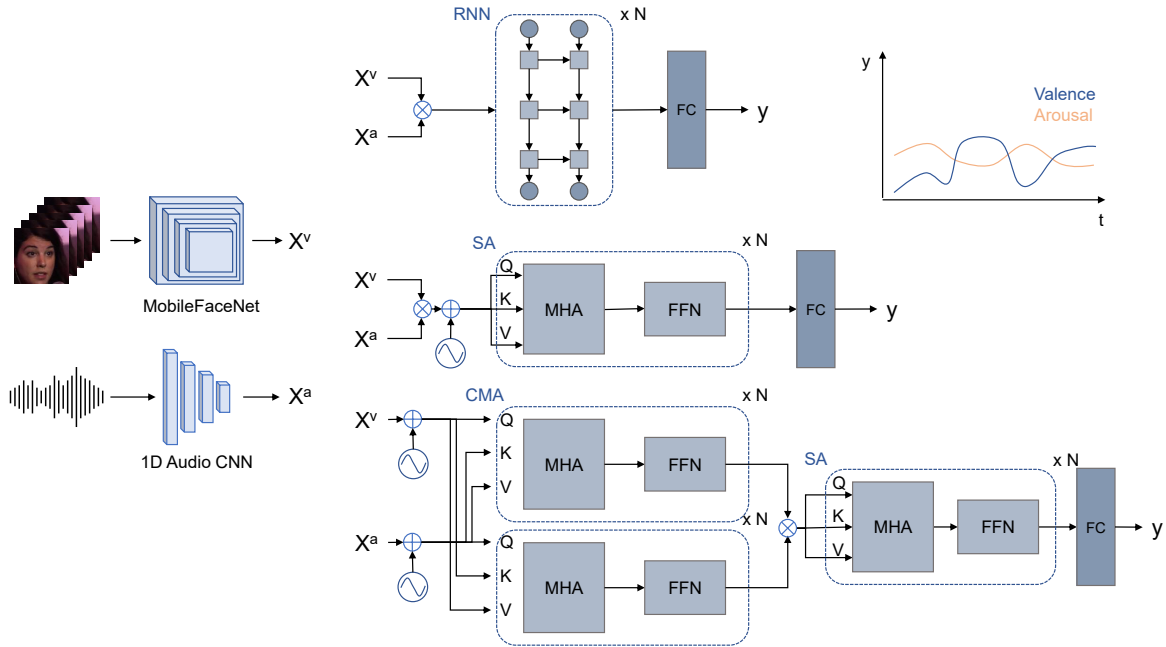[1] https://github.com/end2you/end2you

Figure 1. Our proposed approach. We use pre-trained CNNs as feature extractors from sequences of faces, and raw audio clips (left). Then, we process them with three different architectures: Recurrent Neural Networks with LSTM cells (RNN, top), Self-Attention (SA, middle), and Cross-Modal Attention (CMA, bottom). Each model predicts a sequence of scores for valence and arousal for each time step of the input (right).

Multi-head attention linearly projects the query, key and value pairs into different sub-spaces and performs attention on them in parallel, before recombining and projecting into the output dimension. It is defined as:

$$\text{MHA}(Q, K, V) = \text{Concat}\left(\text{head}_1, ..., \text{head}_n\right) W^O,$$
$$\text{where head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right). \quad (2)$$

In order to fuse modalities within our models, we either use a simple concatenation of our feature embeddings, or a cross-modal fusion architecture.

Cross-Modal Attention (CMA) Fusion is proposed in Tsai et al. [41] to implement the Multimodal Transformer network in which pair-wise attention modelling across different modalities is performed. On the task of discrete emotion recognition from multimodal signals, CMA demonstrates superior generalisation performance compared to LSTM-RNNs [41]. However, when it comes to the continuous emotion recognition from multimodal data, the performance gains that CMA can achieve over the canonical RNNs is unclear. To delineate the trade-offs between the CMA and the other aforementioned sequence models, we herein evaluate different CMA-based audiovisual fusion models. We implement the audiovisual CMA models by tai-

loring the multimodal transformer architecture[2] which was originally designed for text, audio, and visual modalities.

Our cross-modal architecture is based on the cross-modal attention blocks introduced by [41]. In the self-attention mechanism used in the transformer encoder, Q, K, and V are identical. In the cross-modal attention however, the queries and the key-value pairs come from two different modalities, where Q is denoted as the target and K,V as the source, respectively. It is similar to the transformer decoder, but it does not involve self-attention. At each layer, the target modality is reinforced with the low-level information of the source modality [41].

When employing concatenation of feature vectors, we pass the result through either a stack of recurrent layers or a self-attention stack. When using cross-modal fusion, we pass the features through two cross-modal blocks in parallel, one of them using the audio features to attend to the visual features, and the other vice-versa. We then concatenate the outputs of the cross-modal blocks before passing them to a self-attention stack. We use fully connected and 1D convolutional layers to reduce the dimensionality of the features returned by our extractor networks before passing them to our sequence models. When using 1D-CNNs with kernel size larger than 1, this also serves to encode the local temporal context. For the transformer networks, we add

---

[2]https://github.com/yaohungt/Multimodal-Transformer

additional position embedding layers with fixed sinusoidal patterns to help them distinguish the sequence order [43].

3.4. Loss Functions

We use fully connected layers to return the outputs of our models. Each model has two output heads. The first head has size 2 and is used for prediction of valence and arousal scores.

We use two losses for the regression head. The first is based on the Concordance Correlation Coefficient (CCC) [26], which is defined as in Eq. (3). It measures the correlation between two sequences and ranges between -1 and 1, where -1 means perfect anti-correlation, 0 means no correlation, and 1 means perfect correlation. The loss is calculated as $1 - CCC$ for both valence and arousal, then averaged. We also compute the Mean Square Error (MSE), which is defined as Eq. (4).

$$CCC(x,y) = \frac{2 * cov(x,y)}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{3}$$
$$\text{where } cov(x,y) = \sum (x - \mu_x) * (y - \mu_y).$$

$$MSE(x,y) = \sum (x - y)^2. \tag{4}$$

In addition to regressing the scores, we also add a classification head that predicts the category the scores belong to. Jointly estimating continuous and categorical emotions from faces has been shown to be effective for facial affect analysis in the wild [40]. While the Affwild2 dataset is annotated in terms of both continuous and categorical emotions, the rules of the ABAW challenge do not allow using multiple annotations for the valence-arousal track. Therefore, we discretise the labels, by dividing the two-dimensional affect space into 24 sections. These are derived by plotting valence and arousal in polar coordinates, with 3 equidistant radial subdivisions and 8 angular subdivisions.

The Cross-entropy loss is used as the loss function for the classification head. Since the Affwild2 dataset is imbalanced towards positive arousal and valence, we weigh the logits to emphasise minority classes.

Our total loss is thus composed of three terms. We add weights to the MSE and cross-entropy losses to adjust their contribution, leading to our loss function Eq. (5)

$$\mathcal{L} = \mathcal{L}_{ccc} + \lambda_{mse} * \mathcal{L}_{mse} + \lambda_{ce} * \mathcal{L}_{ce}. \tag{5}$$

4. Experiments and Results

We describe our experimental settings and the obtained results on the validation set of the challenge.

4.1. Dataset

We use the subset of Affwild2 annotated for the Valence-Arousal (VA) Estimation task. The training set consists of 341 videos, the validation set consists of 71 videos and the test set consists of 152 videos. Several videos have more than one person in them, those videos are annotated separately for each person and are considered like multiple videos. Frames are annotated in terms of valence and arousal belonging to the range [-1, 1].

We use the cropped and aligned faces from the videos provided by the challenge organisers. After discarding frames marked as invalid, we create sequences from the remainder. The frame rate of the Affwild2 dataset is 30 fps for the majority of videos. Thus, consecutive frames are very similar. In order to provide our model with more temporal information, one option would be to increase the sequence length, at the cost of additional computational resources. We choose instead an approach similar to [24] and use dilated sampling; *i.e.*, we select only 1 in N frames. With sequence length T, this gives a temporal context t of:

$$t = \frac{N}{30} * T. \tag{6}$$

In order not to reduce the size of the training set, we also apply an interleaved sampling method to select the remaining frames. We do not apply this dilated sampling method to the validation set. While this introduces some discrepancy with the training, it maintains equal conditions to the test set.

We use a fixed sequence length of 16 frames for our experiments since we hypothesise that enough information is contained in this time span, and to limit the memory cost of our models. Audio clips are extracted at a fixed window length of 0.5 s, centred at the frame timestamps. We convert the audio of the entire dataset to 16 kHz mono, 16 bit PCM. The images are resized to the shape required by the CNN feature extractor. We use randomly affine transformations and changes in saturation, brightness, and contrast as data augmentation on the images. We also apply Gaussian noise to the audio frames.

4.2. Training

We implement our models using the PyTorch framework and train them on servers with Nvidia RTX3090 and A40 GPUs. Per model training, we allocate 40 CPUs and 40 GB RAM in order to accelerate the loading of batches. The batch size is 64. Models are trained using the AdamW optimiser [27]. We apply cosine annealing with warm restarts as learning rate scheduling, setting it to restart after 200 steps.

In order to find the best configurations for our models, we perform extensive hyperparameter optimisation. We train our models in groups, choosing first the feature extractors and the general architecture (recurrent or transformer),

Hyper-parameter	Value Range
General parameters	
n_{layers}	[1, 5]
d_{model}	64, 128, 256
activation	GELU, SELU
dropout	[0.1, 0.6]
learning rate	$[10^{-5}, 10^{-2}]$
weight decay	$[10^{-3}, 10^{-1}]$
λ_{mse}	[0.0, 1.0]
λ_{ce}	[0.0, 1.0]
Attention Models	
$d_{feedforward}$	64, 128, 256
n_{heads}	2, 4, 8
Cross-modal Attention	
$n_{layers}^{V \rightarrow A}$	[1, 5]
$n_{layers}^{A \rightarrow V}$	[1, 5]
Recurrent Models	
Context aggregation	{unidirectional, bidirectional}
n_{layers}	[1, 5]
d_{hidden}	64, 128, 256

Table 1. Search space of hyper-parameters used for training.

Method	Visual CNN	Vale.	Arou.	Avg.
Baseline [16]	ResNet50	0.310	0.170	0.240
Recurrent Models (RNNs)				
Aud-RNN	-	0.094	0.233	0.163
Vis-RNN	Inception	0.277	0.188	0.233
Vis-RNN	Mobile	0.285	0.357	0.321
AV-RNN	Inception	0.339	0.486	**0.413**
AV-RNN	Mobile	0.319	0.436	0.378
Self-Attention (SA) Models				
Aud-SA	-	0.076	0.317	0.197
Vis-SA	Inception	0.318	0.203	0.261
Vis-SA	Mobile	0.324	0.414	0.369
AV-SA	Inception	0.344	0.404	0.374
AV-SA	Mobile	0.248	**0.529**	0.389
Cross-Modal Attention (CMA) Models				
AV-CMA	Inception	**0.393**	0.363	0.378
AV-CMA	Mobile	0.324	0.460	0.392

Table 2. Validation results in CCC ↑ (valence, arousal and average), evaluated on the validation set of the Aff-wild2 corpus as partitioned in the ABAW 2022 challenge for unimodal and multimodal models with frozen feature extractors. Results are given for each type of architecture investigated - RNNs, SA, and CMA.

then varying the architecture's parameters, as well as the learning rate for our optimiser and the contributions of our losses. The list of hyperparameters used is shown in Tab. 1.

Since the potential number of hyperparameter combinations is very large, making a simple grid search inefficient, we use a tuning algorithm instead to cover a larger number of choices. For this, we choose Ray Tune [3], a flexible tuning toolkit that supports parallel training on multiple GPUs. We use the ASHA scheduling algorithm [25] to quickly discover suitable configurations and stop trials early if they are not performing well.

Our preliminary experiments showed that adding the MSE loss term led to more stable training than using CCC loss alone. Adding the cross-entropy loss on the discretised labels further improved performance, which agrees with the findings of [40].

In the first round of experiments, we freeze the layer weights of the feature extraction networks to limit the number of trainable parameters. Then, we test end-to-end learning with the full set of parameters. For these experiments, we restrict the choice of the visual encoder network to MobileFaceNet to avoid overfitting.

4.3. Validation Results

The validation results for preliminary experiments on models with frozen feature extraction networks are presented in Tab. 2. We denote the three types of architectures employed as Audiovisual-[RNN, SA, CMA] for recurrent, self-attention, and cross-modal attention, respectively. The second column specifies the feature extraction network used for visual information, as Inception or Mobile for Inception-Resnetv1 and MobileFaceNet, respectively. For comparison, we also state the results of unimodal models trained with self-attention and RNN.

As Tab. 2 shows, our audiovisual models outperform the challenge baseline by a wide margin. We report the validation results of the best models per architecture, trained end-to-end, in Tab. 3. All models share the same feature encoders – i.e., MobileFaceNet and the 1D CNN pre-trained on RECOLA. The hyperparameter configurations of the best models are given in Tab. 4. In addition, we report the number of parameters for the best performing audiovisual models in Tab. 5, comparing their computational complexity.

Method	Valence	Arousal	Avg.
E2E-AV-RNN	0.361	**0.551**	**0.456**
E2E-AV-SA	0.380	0.520	0.450
E2E-AV-CMA	**0.388**	0.492	0.440

Table 3. Validation results in CCC ↑, evaluated on the validation set of the Aff-wild2 corpus as partitioned in the ABAW 2022 challenge. The reported results are for the best multimodal models trained end-to-end with MobileFaceNet as visual encoder and 1D CNN pre-trained on RECOLA as audio encoder, and using RNN, SA and CMA for sequence modelling.

Hyper-Parameter	E2E Models		
	AV-RNN	AV-SA	AV-CMA
n_{layers}	1	3	4
d_{model}	64	64	256
activation	SELU	GELU	GELU
dropout	0.5	0.5	0.6
learning rate	0.0002	0.002	0.0001
weight decay	0.023	0.008	0.06
λ_{mse}	0.84	0.78	0.18
λ_{ce}	0.88	0.27	0.76
$d_{feedforward}$	-	256	256
n_{heads}	-	8	4
$n_{layers}^{V \rightarrow A}$	-	-	3
$n_{layers}^{A \rightarrow V}$	-	-	1
Context aggregation	uni	-	-
d_{hidden}	64	-	-

Table 4. Hyperparameter configurations for the best performing models. The models are trained end-to-end with RNNs, SA, and CMA networks, respectively.

4.4. Test results

The results of our end-to-end models on the test set are given in Tab. 6. We also report the performance of an ensemble model obtained by averaging the individual model predictions, which achieved 5th place on the VA track of the ABAW 2022 challenge.

5. Discussion

We judge performance by the mean CCC for valence and arousal, which is the metric used in the VA Track of the ABAW 2022 challenge. We first discuss how the choice of the visual CNN impacts the RNN models, and the impact of end-to-end learning. We then compare the performance

Method	Visual Encoder	$P_{sequence}$	P_{total}
Recurrent Models (RNNs)			
AV-RNN	Inception	109 K	28.8 M
AV-RNN	Mobile	4.4 M	5.4 M
E2E-AV-RNN	Mobile	76 K	1.1 M
Self-Attention (SA) Models			
AV-SA	Inception	765 K	28.1 M
AV-SA	Mobile	482 K	1.51 M
E2E-AV-SA	Mobile	193 K	1.2 M
Cross-Modal Attention (CMA) Models			
AV-CMA	Inception	134 K	28.1 M
AV-CMA	Mobile	2.1 M	3.1 M
E2E-AV-CMA	Mobile	2.4 M	3.4 M

Table 5. Size of our models. We state the total number of parameters for the audiovisual models, grouped by architecture. For clarity, we report the number of parameters in the sequence models and the full number of parameters separately.

Model	Valence	Arousal	Avg.
Baseline [16]	0.180	0.170	0.175
E2E-AV-RNN	0.376	0.380	0.378
E2E-AV-SA	0.396	0.376	0.386
E2E-AV-CMA	0.327	0.359	0.343
Fused	**0.418**	**0.407**	**0.413**
Winner [29]	**0.606**	0.596	**0.601**
Runner-Up [45]	0.520	**0.602**	0.561

Table 6. Test results in terms of CCC↑ evaluated on the test set of the Aff-wild2 corpus as partitioned in the ABAW 2022 challenge. Reported are scores for our three models that performed best on the validation set, as well as for an ensemble obtained by averaging the predictions. Also shown are the scores of the two best performing teams and the challenge baseline.

of SA and CMA, before contrasting RNNs and attention models.

5.1. RNN performance

When using RNN as the sequence model, performance decreases considerably when replacing the FaceNet feature encoder with the less complex MobileFaceNet (0.413 to 0.378). At the same time, the number of parameters in the trainable part of the model increases sharply from 109 k to 4.4 M. We interpret this as the model having difficulties to learn valence and arousal effectively from the features returned by the smaller CNN.

However, using the lightweight architecture together

with end-to-end-learning presents a very different picture. When the feature extractors are fully trainable, the performance of the recurrent model is greatly increased, yielding an average CCC of 0.456. At the same time, the number of model parameters decreased to merely 76 k in the sequence part. Examining the hyperparameter configuration of this winning model shows that it has a single, unidirectional LSTM layer, with a hidden dimension $d = 64$. We conclude from this that the RNNs seem more efficient in learning representations of the emotional states if it is trained end-to-end in combination with shallow CNN encoders.

5.2. Comparing Self-attention and Cross-attention

When comparing the SA models with different visual CNNs, the model with MobileFaceNet outperforms the one with FaceNet, with average CCC scores of 0.389 and 0.374, respectively. At the same time, the size of the attention model is smaller for the architecture with MobileFaceNet (482 k parameters compared to 765 k parameters).

For the CMA models, replacing FaceNet with Mobile-FaceNet also increases performance, from 0.378 to 0.392. However, the transformer network becomes much larger, going from 134 k parameters to 2.1 M parameters. These scores show that the performances of SA and CMA appear to be similar if the feature extractors are frozen, with CMA performing slightly better.

When using end-to-end learning, performance increases significantly for the SA model, with an average validation CCC of 0.450. At the same time, the number of parameters in the sequence part of the network shrinks to 193 k. The CMA model also benefits greatly from end-to-end learning, achieving a score of 0.440. The number of parameters in the sequence part of the model is 2.4 M.

Comparing the two best end-to-end models shows that the SA model outperforms the CMA model, while being considerably smaller. We hypothesise that the lower complexity of the self-attention model helped discover a more efficient architecture during end-to-end training.

5.3. Comparing LSTM-RNNs and Attention

We now compare the performances of our RNN models and attention models directly based on the results on the validation set discussed in the two previous sections. In the case where the feature extractors are frozen, for the larger FaceNet, the RNN outperforms the attention models. If frozen MobileFaceNet is used, the attention models outperform RNN. With end-to-end learning, RNN beats both SA and CMA, while also having fewer parameters.

Evaluation on the test set has shown that SA outperforms RNN, with scores of 0.386 and 0.378 respectively, while both outperform CMA (0.343) by a wide margin. When all models are combined into an ensemble, the average CCC increases to 0.413.

We conclude from these observations that attention-based models do not consistently outperform RNNs for emotion recognition in the wild. When end-to-end learning is used in combination with shallow CNNs for feature encoding, RNNs perform on par or marginally better on the validation set than the SA models investigated in this paper, and on par or marginally worse on the test set, while being superior to CMA.

6. Outlook

We compared fusion performance using two CNNs of different sizes as visual feature extractors, while using a small 1D-CNN for extracting audio features. Another study could focus on choosing different audio networks, *e.g.* a larger model like VGGish [12], and comparing the effects.

The models used in this work had limited temporal context due to computational constraints. Future studies could extend towards longer sequence lengths to investigate how well the models capture long-term dependencies.

Our analysis focused on the average of valence and arousal as a metric, in order to judge the overall performance of the models. We leave the analysis of trade-offs between valence and arousal for future work.

7. Conclusion

On a wide range of sequence modelling tasks, attention models demonstrated superior generalisation performance than recurrent models in recent years. However, it is worth noting that the recurrent models have the natural ability to cope with the challenges in learning from time-continuous sequence data, by inferring the latent states with unbounded context, at least in principle. Therefore, in the case of time-continuous multimodal affect recognition, a recurrent neural network architecture may still be a natural choice to model the latent states of face and voice data and their interactions in a time-continuous manner. The extensive evaluation of LSTM-RNNs, self-attention, and cross-modal attention on in-the-wild audiovisual affect recognition suggests that attention models may not necessarily be the optimal choice to perform time-continuous multimodal fusion.

Acknowledgements

References

[1] Michael Braun, Jonas Schubert, Bastian Pfleging, and Florian Alt. Improving driver emotions with affective strategies. *Multimodal Technologies and Interaction*, 3(1), 2019. 1

[2] Cong Cai, Yu He, Licai Sun, Zheng Lian, Bin Liu, Jianhua Tao, Mingyu Xu, and Kexin Wang. Multimodal sentiment analysis based on recurrent neural network and multimodal attention. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pages 61–67. 2021. 2

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 67–74, 2018. 3

[4] H. Chen, D. Jiang, and H. Sahli. Transformer encoder with multi-modal multi-head attention for continuous affect recognition. *IEEE Transactions on Multimedia*, pages 1–1, 2020. 2

[5] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In Jie Zhou, Yunhong Wang, Zhenan Sun, Zhenhong Jia, Jianjiang Feng, Shiguang Shan, Kurban Ubul, and Zhenhua Guo, editors, *Biometric Recognition*, pages 428–438, Cham, 2018. Springer International Publishing. 3

[6] D. Deng, Z. Chen, and B. E. Shi. Multitask Emotion Recognition with Incomplete Labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 592–599, 2020. 3

[7] Didan Deng, Liang Wu, and Bertram E. Shi. Iterative distillation for better uncertainty estimates in multitask emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3557–3566, October 2021. 3

[8] Nhu-Tai Do, Tram-Tran Nguyen-Quynh, and Soo-Hyung Kim. Affective expression analysis in-the-wild using multitask temporal statistical deep learning model. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 624–628. IEEE, 2020. 3

[9] Florian Eyben, Martin Wöllmer, Tony Poitschke, Björn Schuller, Christoph Blaschke, Berthold Färber, and Nhu Nguyen-Thien. Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car. *Advances in human-computer interaction*, 2010, 2010. 1

[10] Ziwang Fu, Feng Liu, Hanyang Wang, Jiayin Qi, Xiangling Fu, Aimin Zhou, and Zhibin Li. A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition. *arXiv preprint arXiv:2111.02172*, 2021. 2

[11] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 2, 3

[12] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, March 2017. 8

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2, 3

[14] 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)*, pages 3507–3511, May 2020. 2

[15] Giancarlo Kerg, Bhargav Kanuparthi, Anirudh Goyal, Kyle Goyette, Yoshua Bengio, and Guillaume Lajoie. Untangling tradeoffs between recurrence and self-attention in neural networks. *arXiv preprint arXiv:2006.09471*, 2020. 2

[16] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022. 6, 7

[17] D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 637–643, 2020. 3

[18] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 3

[19] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019. 2, 3

[20] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 2, 3

[21] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 3

[22] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 3

[23] DN Krishna and Ankita Patil. Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks. In *Interspeech*, pages 4243–4247, 2020. 2

[24] F. Kuhnke, L. Rumberg, and J. Ostermann. Two-stream aural-visual affect analysis in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 600–605, 2020. 3, 5

[25] Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-Tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A system for massively parallel hyperparameter tuning. *Proceedings of Machine Learning and Systems*, 2:230–246, 2020. 6

[26] Lawrence I.-Kuei Lin. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, 45(1):255–268, 1989. 5

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[28] Ziyu Ma, Fuyan Ma, Bin Sun, and Shutao Li. Hybrid muti-modal fusion for dimensional emotion recognition. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pages 29–36. 2021. 2

[29] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Yuan Cheng, Meng Wang, Chuanhe Liu, and Qin Jin. Multi-modal emotion estimation for in-the-wild videos, 2022. 7

[30] Danny Merkx and Stefan L Frank. Human sentence processing: Recurrence or attention? *arXiv preprint arXiv:2005.09471*, 2020. 2

[31] Rosalind W Picard. Affective computing: from laughter to ieee. *IEEE Transactions on Affective Computing*, 1(1):11–17, 2010. 1

[32] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017. 2

[33] Vandana Rajan, Alessio Brutti, and Andrea Cavallaro. Is cross-attention preferable to self-attention for multi-modal emotion recognition? *arXiv preprint arXiv:2202.09263*, 2022. 2

[34] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC '19, page 3–12, New York, NY, USA, 2019. Association for Computing Machinery. 1

[35] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013. 3

[36] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 3

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015. 3

[38] Licai Sun, Zheng Lian, Jianhua Tao, Bin Liu, and Mingyue Niu. Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, pages 27–34, 2020. 2

[39] Licai Sun, Mingyu Xu, Zheng Lian, Bin Liu, Jianhua Tao, Meng Wang, and Yuan Cheng. Multimodal emotion recognition and sentiment analysis via attention enhanced recurrent

model. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pages 15–20. 2021. 2

[40] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50, 2021. 5, 6

[41] 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*, pages 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. 2, 4

[42] Panagiotis Tzirakis, Stefanos Zafeiriou, and Bjorn W Schuller. End2you–the imperial toolkit for multimodal profiling by end-to-end learning. *arXiv preprint arXiv:1802.01115*, 2018. 3

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł. ukasz Kaiser, and Illia Polosukhin. Attention is all you need. pages 5998–6008. Curran Associates, Inc, 2017. 2, 3, 5

[44] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 2

[45] Su Zhang, Ruyi An, Yi Ding, and Cuntai Guan. Continuous emotion recognition using visual-audio-linguistic information: A technical report for abaw3, 2022. 7

[46] Su Zhang, Yi Ding, Ziquan Wei, and Cuntai Guan. Continuous emotion recognition with audio-visual leader-follower attentive fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3567–3574, 2021. 2

[47] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, and Shiguang Shan. M 3 f: Multi-modal continuous valence-arousal estimation in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 632–636. IEEE, 2020. 2

[48] Jianfeng Zhao, Xia Mao, and Lijiang Chen. Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical signal processing and control*, 47:312–323, 2019. 3