End-to-end multimodal affect recognition in real-world environments

Panagiotis Tzirakis [*,1], Jiaxin Chen[1], Stefanos Zafeiriou, Björn Schuller

GLAM — Group on Language, Audio & Music, Imperial College London, UK

1. Introduction

Automatic affect recognition is a fundamental component towards a complete interaction between human and machine. Currently, intelligent systems, such as robots and virtual humans, try to use emotion recognition models to make the interaction with humans more *natural*. To this end, such systems should automatically sense and adapt their responses according to the human behavior. One application can be found in an automatic tutoring system, where the system adjusts the level of the tutorial depending on the user's affective state, such as excitement or boredom [1].

The task of recognizing human affect is very challenging as the duration of human emotions vary significantly and depends on the person and the situation. In addition, emotions are expressed differently among different individuals [2] and cultures. The difficulty of affect recognition is further escalated in real-world environments where uncontrolled conditions are entailed. Although most of the current research on emotion recognition is focused on exploiting a unimodal stream, it is important to consider multiple channels as complementary information exists among them [3].

In this paper, we study the automatic continuous affect recognition task using text, audio and visual information in an *end-to-end* manner. We develop a transformer-based [4] architecture to extract information from the text channel, use an audio network for audio feature extraction, and develop a variant of a high resolution network (HRNet) architecture [5] with three stages to extract information from the visual channel. To fuse the unimodal features to a single unified representation, we propose novel attention-based methods. The fused feature vector is then fed into a Long Short-Term Memory (LSTM) [6] before the prediction of the affective state of individuals. The training and evaluation of our models are based on the *concordance correlation coefficient* (ρ_c). To test the effectiveness of our model, we utilize the Sentiment Analysis in the Wild (SEWA) dataset, which was used in the Audio/Visual Emotion Challenge (AVEC) in 2017 [7] and compare with the three best papers of the competition.

The contributions of our paper can be summarized as follows:

- develop a transformer-based architecture in our text model and capture the semantics of a sentence by extracting context-aware features. To the best of our knowledge, this is the first time a transformer mechanism has been used in the sentiment analysis domain.
- propose three novel attention-based fusion strategies to combine the features extracted from the raw speech, text and visual modalities.

* Corresponding author.
E-mail address: panagiotis.tzirakis12@imperial.ac.uk (P. Tzirakis).

[1] Equal contribution.

- visualize (i) the attention mechanism in our text model and indicate the words that contribute the most to the final prediction, and (ii) the attention-based fusion strategies that can indicate which modality has the highest contribution to each frame.
- provide state-of-the-art results for the visual, text and multimodal modalities, and the second highest performing model in the audio modality, when comparing with the winning papers from the challenge that use several hand-crafted and deep features. On top of that, we perform an extensive experimentation of current state-of-the-art unimodal architectures in visual, and audio modalities.

The rest of the paper is organized as follows. Section 2 presents related studies on unimodal and multimodal emotion recognition. After defining the notation in Section 3, Section 4 describes the unimodal networks, the attention mechanism methods used to fuse the unimodal features, and the multimodal network. Finally, experimental studies and results are presented and discussed in Section 5, along with visualisation of the proposed attention methods.

2. Related work

Deep neural networks have been widely used by the speech community [8–10], and several affect recognition models have been proposed. For example, Neumann et al. [11] proposed an attentive convolutional neural network (ACNN) that combines CNNs with attention. Xu et al. [12] proposed head fusion, a multi-head self-attention method. Other studies use the raw waveform to model affect. For example, Tzirakis et al. [13] proposed a deeper architecture with a longer input size along with a strategy to compute kernel size and max-pooling size. In a more recent study, Li et al. [14] proposed the use of self-attention and global windowing in the transformer model for the task at hand.

Visual information has also been exploited to predict the emotional state of individuals. For example, Yang et al. [15] proposed De-expression Residue Learning (DeRL) where a generative model is used to extract the neutral face image, and then the method learns the residue (emotion) that remains in the generative model. In another study, [16] propose a deep learning approach based on attentional convolutional network. In a similar study, [17] uses a feature extractor based on VGG-Face [18] and a 2-layer Recurrent Neural Network (RNN) to account for the temporal information in the data, before getting the final prediction. For more affect recognition studies that exploit the visual information, the interested reader is referenced to Li et al. [19].

Affect recognition via textual information has been a very popular research area in natural language processing. Kim et al. [20] proposed a character-aware neural language model that first learns word representations at character-level using a CNN, and then models language tasks at word-level utilizing LSTM. In another example, Angelidis et al. [21] proposed a multiple instance learning method to utilize segment-level sentiment predictions to formulate document-level analysis. Sarma et al. [22] proposed a domain adapted word embedding model to facilitate the performance of sentiment recognition in specific domains/fields. In a more recent study, Park et al. [23] proposed emotion embedding model to classify story text emotions. A comprehensive review for the text classification task has been published by Minaee et al. [24].

Emotion recognition systems can be benefited by exploiting multiple modalities such as audio, visual and text [25–36]. Some approaches utilize multitask learning [27], others sentic blending [25], and few have been proposed for real-time analysis [34,37]. In a recent study, Tsai et al. [30], proposed an end-to-end multimodal transformer model to align variable sampling rates modalities. In another study, Mai et al. [38] proposed "divide, conquer and combine" multimodal fusion strategy that investigates local and global interactions between unimodal features in an hierarchical manner. Chaturvedi et al. [33] proposed convolutional fuzzy sentiment classifier that combines deep learning models with fuzzy logic classifier to predict the emotional state of individuals.

The importance of predicting emotion in real-world environments led the AVEC 2017 emotion sub-challenge [7] to use the SEWA dataset which is comprised by three modalities: text, audio, and visual. The baseline model [7] used hand-crafted features in all of the modalities and a Support Vector Regressor (SVR) for the final prediction. The winning model of the challenge was proposed by Chen et al. [39] and used a multi-task learning method exploiting both deep neural network, and hand-crafted, features. Their final model comprised of passing all the unimodal extracted features to a LSTM network before the final prediction. In a similar study, Huang et al. [40] utilized pre-train deep models and hand-crafted features on all modalities. Another interesting study was submitted by Dang et al. [41] where they proposed a fusion strategy using probabilistic predictions by utilizing hand-crafted features in all the modalities.

3. Notation

Before describing our model we define our notation. Matrices are defined as uppercase bold letters as $\mathbf{X}$, vectors as lowercase bold letters as $\mathbf{z}$, and scalars as non-bold letters. The ith row of a matrix is defined as $\mathbf{X}[i,:]$, the jth column as $\mathbf{X}[:,j]$, and the element at position $[i,j]$ as $\mathbf{X}[i,j]$. The kth element of a vector is specified as $\mathbf{z}[k]$. The row-wise concatenation of the vectors $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m\}$ is denoted as $[\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m]$.

4. Proposed model

We introduce a multimodal system that is comprised of three DNNs utilized to extract features from text, audio and visual modalities. The features are combined to form a universal representation using attention mechanism. Temporal information is vital in our task and as such we utilize LSTM networks to process the extracted features before the final prediction.

4.1. Unimodal networks

4.1.1. Text network

Our proposed text network is depicted in Fig. 1 and it is comprised of six different parts, which are summarized below, along with the input representation.

Input. For our word-level representation, we adopt the pre-trained XLNet [42] of 768 dimensions with enriched subword information such that a sentence is represented as the matrix $\mathbf{S} \in \mathbb{R}^{768\times L}$, where L is the maximum words in a sentence of our corpus.

Position-wise N-Grams. We use a CNN to capture position-wise N-grams of the input matrix $\mathbf{S}$ for each frame. In particular, each word vector $\mathbf{S}[:,l]$ is processed by several convolutional layers and depending on their filter width, i.e. $N = 2, 3, 4, \dots, T$, they are used to capture N-grams producing a feature map as follows:

$$\mathbf{f}[l] = g(\tilde{\mathbf{S}}[:, l : l + N - 1] \circ \mathbf{H} + b) \tag{1}$$

where $\tilde{\mathbf{S}}$ is the matrix $\mathbf{S}$ padded, $g(\cdot)$ is an activation function (in our case $tanh$), N is the width of the filter $\mathbf{H} \in \mathbb{R}^{768\times N}$, and $\circ$ denotes the Hadamard multiplication. We concatenate the original feature matrix $\mathbf{S}$ with the N_k feature maps extracted from the convolution layers to form a new matrix $\mathbf{C} = [\mathbf{S}, \mathbf{f}^1, \dots, \mathbf{f}^{N_k}] \in \mathbb{R}^{D_e\times L}$ where $D_e = 768 + N_k$.

Multihead Linear Projection. To improve the diversity of the input features and inspired by [4], we propose to use an attention mechanism to enable each frame to discover multiple representative features in different context spaces. More particularly, given the input matrix $\mathbf{C}$, we first apply different linear projections on $\mathbf{C}$ to explore different perspective of context for the current frame, i.e. $\mathbf{V}_p = \mathbf{W}_p\mathbf{C}$, where $\mathbf{W}_p \in \mathbb{R}^{D_p\times D_e}$ is the p-th linear projection matrix and D_p is the dimension of the p-th context space.

Recurrent Context Generator. To capture semantic information in the different context spaces we utilize a 2-layer Bi-LSTM with hidden state $\mathbf{h}_p^l$ at position l.

Selector. The semantic-aware hidden states $\mathbf{h}_p^l$ of the Bi-LSTM are processed by an attention mechanism $\mathbf{s}_p$ to enable the model to *select* the most useful context-aware features in each context space p.

Context Feature Concatenation. In this step we concatenate the context-aware features obtained in each projection space to form a single feature representation, i.e. $\mathbf{z} = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_p] \in \mathbb{R}^{pD_l}$.

Highway Networks. Highway networks [43] are proven to have a strong empirical performance especially for deep networks. To this end we adopt highway networks by feeding them our frame-level feature $\mathbf{z}$, i.e.,

$$\tilde{\mathbf{z}} = \mathbf{o} \circ g(\mathbf{W}^z \mathbf{z} + b^z) + (\mathbf{1} - \mathbf{o}) \circ \mathbf{z} \tag{2}$$

where $\mathbf{o} = \sigma(\mathbf{W}^o \mathbf{z} + b^o)$ acts as a transformation gate which allows shortcut from input to output directly, and $g(\cdot)$ is an activation function (in our case *tanh*).

4.1.2. Audio network

Extracting features from the audio signal is an important step in the field of paralinguistics. For our purposes we utilise the network architecture proposed by Tzirakis et al. [13]. The network architecture is comprised of 3 blocks, each containing one convolution and one max-pooling layer. The convolution layers have 50, 125, and 250 number of channels with kernel sizes $8, 6, 6$ with stride 1, respectively, and the max-pooling layers $10, 5, 5$ kernel size and same stride size, respectively.

The input to the network is the raw waveform, segmented into 160 s long sequences. At a sampling rate of 22050 Hz this corresponds to a $3,528,000$-dimensional vector. In addition, we utilize the interlocutor information provided with the SEWA dataset by using an additional vector in each frame so that we can distinguish the target from the chatting partner in the contained dyadic conversations. More particularly, half of the entries of this vector are zeros depending on the speaker turn, and the other half is filled with the extracted features. The final input to the network is a $(2 * 3,528,000)$-dimensional vector.

4.1.3. Visual network

Our visual feature extractor is a High-Resolution network (HRNet) [44] of 3 stages. HRNets maintain high-resolution representations of the input by performing multiscale and fusion across parallel convolution. They have shown superior results in a number of computer vision problems, such as semantic segmentation and object detection.

Studies in the literature have shown the beneficial properties of using an attentional pooling layer [45] instead of the last average one in deep convolution networks. To this end, we adopt an hierarchical attention scheme, where we replace the last average pooling layer with an attention mechanism following the low-rank second-order pooling (top-down attention) [45]. On top of that, we use another attention between the feature maps. Utilizing this attention scheme speeds up the training of the network and increases its performance on the development set.

4.2. Attentional fusion strategies

Before feeding the features extracted from text ($\mathbf{x}^t \in \mathbb{R}^{D_t}$), audio ($\mathbf{x}^a \in \mathbb{R}^{D_a}$) and visual ($\mathbf{x}^v \in \mathbb{R}^{D_v}$) modality to the recurrent network, we consider five different strategies to fuse them together.

Concatenation. The first approach is a concatenation of the unimodal features, i. e., $\mathbf{x}^{fusion} = [\mathbf{x}^t, \mathbf{x}^a, \mathbf{x}^v]$. This is a standard feature-level fusion approach that has been used extensively in the literature.

Hierarchical Attention. We propose to perform hierarchical attention to the unimodal features so as to maximize the relevant information that is propagated forward to the network. More particularly, we first perform attention to all paired unimodal features, i.e. audio-text, audio-visual, and visual-text, before using a higher level attention to the paired outputs. More formally, given the features extracted from text $\mathbf{x}^t$, audio $\mathbf{x}^a$ and visual $\mathbf{x}^v$ modalities, we first perform a linear projection to each of these features such that they are projected to the same vector space (with the same number of dimensions D_u), namely,

$$\begin{aligned} \tilde{\mathbf{x}}^t &= \mathbf{W}^t \mathbf{x}^t + b^t \\ \tilde{\mathbf{x}}^a &= \mathbf{W}^a \mathbf{x}^a + b^a \\ \tilde{\mathbf{x}}^v &= \mathbf{W}^v \mathbf{x}^v + b^v \end{aligned} \tag{3}$$

where $\mathbf{W}^t \in \mathbb{R}^{D_u \times D_t}$, $\mathbf{W}^a \in \mathbb{R}^{D_u \times D_a}$, $\mathbf{W}^v \in \mathbb{R}^{D_u \times D_v}$ are projection matrices for text, audio and visual modalities, respectively, with dimension D_u. The projected features are combined pair-wise, and then are passed through a selective attention mechanism to obtain the final

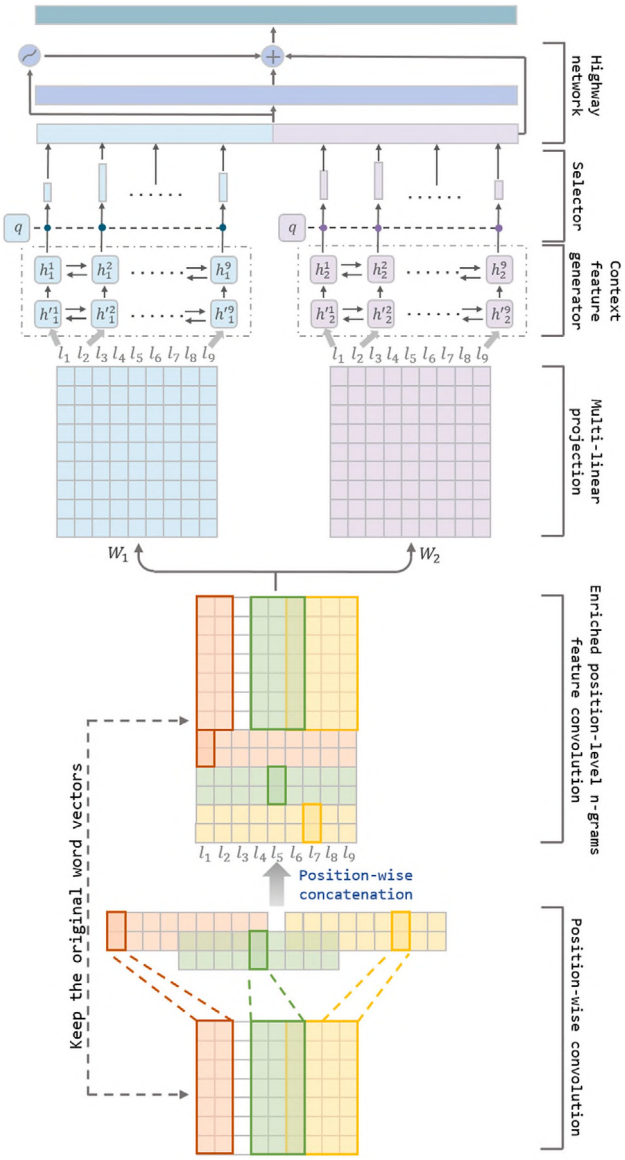

Fig. 1. Text network architecture. The input is a word-embedding matrix that is passed to a N-gram feature extractor (padding is not shown), before being concatenated with the N-gram features. Then linear projections are applied such that different contexts are extracted. Each context-matrix is processed by a 2-layer Bi-LSTM to get the semantics in the sentence before the selector that weights the outputs. Finally, the outputs of each recurrent network is concatenated and passed by a highway network to get the semantic-aware feature vector. (Best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

fusion feature, i.e.,

$$\alpha^i = softmax(\frac{\tilde{\mathbf{x}}^i \mathbf{q}^i}{\sqrt{D_u}})$$
$$Attention(\mathbf{x}^1 \cdots \mathbf{x}^M) = \sum_{i=1}^{M} \alpha^i \tilde{\mathbf{x}}^i \quad (4)$$

where $\mathbf{q}^i \in \mathbb{R}^{D_u}$ is a learnable vector that attends to different modality features, and $i \in [1, M]$ is an index denoting the $M-th$ modality. We obtain final attentive fusion feature $\mathbf{x}^{fusion}$ as follows:

$$\mathbf{x}^{vt} = Attention(\tilde{\mathbf{x}}^v, \tilde{\mathbf{x}}^t)$$
$$\mathbf{x}^{va} = Attention(\tilde{\mathbf{x}}^v, \tilde{\mathbf{x}}^a)$$
$$\mathbf{x}^{ta} = Attention(\tilde{\mathbf{x}}^t, \tilde{\mathbf{x}}^a) \quad (5)$$
$$\mathbf{x}^{fusion} = Attention(\mathbf{x}^{vt}, \mathbf{x}^{va}, \mathbf{x}^{ta})$$

Self-Attention. We also propose to apply self-attention fusion strategy by enabling the extracted features to attend to each other. To do so we first apply a linear projection on the unimodal features $\mathbf{x}^t$, $\mathbf{x}^a$ and $\mathbf{x}^v$ like in Eq. (3).

After we concatenate the three vectors together, i. e., $\mathbf{X}^c = [\tilde{\mathbf{x}}^t, \tilde{\mathbf{x}}^a, \tilde{\mathbf{x}}^v]^T \in \mathbb{R}^{3 \times D_u}$, we apply a multi-head 3-way linear projection of the matrix [4], i. e., $\mathbf{Z}_i^j = \mathbf{X}^c \mathbf{W}_i^j \in \mathbb{R}^{3 \times D_s}$ for $j = 1, 2, 3$, where $\mathbf{W}_i^j \in \mathbb{R}^{D_u \times D_s}$, D_s is the dimension of each projection space, and $i \in \mathbb{N}^+$ is an index for operations in different projection space. Then, we utilize self-attention to explore complementary relationships among the different modalities. The attention is applied as follows:

$$\boldsymbol{A}_i = softmax(\frac{\mathbf{Z}_i^1 (\mathbf{Z}_i^2)^T}{\sqrt{D_s}}) \mathbf{Z}_i^3$$
$$\boldsymbol{\alpha}_i = \boldsymbol{A}_i \mathbf{q}_i \quad (6)$$
$$\mathbf{v}_i = \sum_{k=1}^{3} \boldsymbol{\alpha}_i[k] \boldsymbol{A}_i[k, :]$$

where $\mathbf{q}_i \in D_s$ here is again a learnable variable for self-attentive features and the subscript i denotes the ith linear projection. The final fusion feature vector is the concatenation of all the projected features, i.e., $\mathbf{x}^{fusion} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ where p is the number of projections applied.

Residual Self-Attention. Inspired by the work on non-local neural networks [46] we extend the self-attention fusion strategy by incorporating a residual connection. To this end, the residual self-attention can be defined as follows:

$$\mathbf{z}_i = \mathbf{W}_z \mathbf{v}_i + \mathbf{X}^c, \quad (7)$$

where i represents the i-th linear projection (see Eq. (6)), "$+\mathbf{X}^c$" denotes the residual connection, and $\mathbf{v}_i$ is defined in Eq. (6). The final fusion representation is computed by concatenating the projected features $\mathbf{z}_i$, i. e., $\mathbf{x}^{fusion} = [\mathbf{z}_1, \ldots, \mathbf{z}_p]$ where p is the number of projections applied.

Cross-modal Hierarchical Self-Attention. We also propose a cross-modal hierarchical self-attention fusion, where each modality is combined with the rest using self-attention, and then the features per modality are fused in an attentional manner. Mathematically, we define the cross-modal attention as follows:

$$\mathbf{a}^{m,c} = \tilde{\mathbf{x}_1}^m softmax(\frac{\tilde{\mathbf{x}_2}^m (\tilde{\mathbf{x}}_1^c)^T}{\sqrt{D_s}}), \quad (8)$$

where the modality m is fused with modality c, $\tilde{\mathbf{x}}_i^m$ is the i-th linear projection with dimensions D_s of modality $\mathbf{x}_i^m$, as defined in Eq. (3). The cross-modal attentional representation of modality m with the rest modalities c_i is defined as $\mathbf{v}_f^m = Attention(\mathbf{a}^{m,c_1}, \mathbf{a}^{m,c_2}, \ldots \mathbf{a}^{m,c_i})$. Finally, we fuse the M (one per modality) cross-modal representations using another attention, i. e.,

$$x^{fusion} = Attention(\mathbf{v}_f^{m_1}, \mathbf{v}_f^{m_2}, \ldots, \mathbf{v}_f^{m_M}). \quad (9)$$

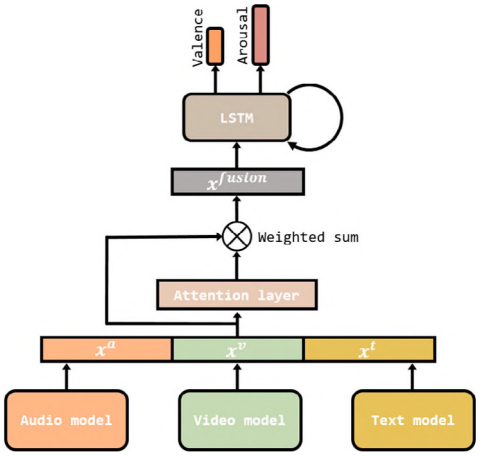

Fig. 2. The multimodal network is comprised of the features extracted from the text ($\mathbf{x}^t$), audio ($\mathbf{x}^a$), and visual ($\mathbf{x}^v$) modalities, the attentional layer that fuses the extracted features ($\mathbf{x}^{fusion}$) using a weighted vector, and 1-layer LSTM that capture the contextual dynamics in the data before the final prediction. (Best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.3. Multimodal model

Our multimodal model is comprised of all the unimodal networks, the attention layer that performs our fusion strategies, and 1-layer LSTM network. Fig. 2 depicts the overall architecture.

5. Experiments

5.1. Dataset

To test the effectiveness of our model on time-continuous predictions in real-world environments, we utilize the Sentiment Analysis in the Wild (SEWA) dataset that was used in the AVEC 2017 challenge. It consists of audio-visual recordings that were collected from web-cameras and microphones, and captured spontaneous and natural emotions (arousal and valence). In total, 64 subjects (age from 18 to 60 years old) are paired (i. e., 32 pairs) to watch a 90 s commercial video, and their task was to discuss the content with their partner for maximum 3 min. The dataset provides three modalities, namely, text, audio and visual, and it is split into 3 partitions: training (17 pairs), development (7 pairs), and test (8 pairs). In total, the training, development and test sets contain 54,813, 22,556, and 27,950 audiovisual frames, and 1,329, 604, and 629 sentences, respectively. Finally, 6 German-speaking annotators (3 female, 3 male) were used to annotate the dataset in continuous arousal and valence.

5.2. Objective function

As our objective function we utilize the Concordance Correlation Coefficient (ρ_c) that was also used in the AVEC 2017 challenge. ρ_c evaluates the agreement level between the predictions and the gold standard by scaling their correlation coefficient with their mean square difference. More particularly, we define the concordance loss J_c as follows:

$$J_c = 1 - \rho_c = 1 - \frac{2\sigma_{xy}^2}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \quad (10)$$

where $\mu_x = \mathbb{E}(\mathbf{x})$, $\mu_y = \mathbb{E}(\mathbf{y})$, $\sigma_x^2 = \text{var}(\mathbf{x})$, $\sigma_y^2 = \text{var}(\mathbf{y})$ and $\sigma_{xy}^2 = \text{cov}(\mathbf{x}, \mathbf{y})$.

For our purposes we train our networks to predict both arousal and valence, and as such we define the loss function as: $J = (J_c^a + J_c^v)/2$, where J_c^a and J_c^v are the concordance loss of the arousal and valence, respectively.

Table 1
Range of values for hyper-parameter tuning.

Hyper-parameter	Range
Learning rate	$[10^{-3} - 10^{-6}]$
Batch size (Text/audio)	[10, 20, 30]
Batch size (Visual/multimodal)	[2, 3]
Dropout	[0.1 − 0.6]
Sequence length	[75, 100, 200]
Hidden units (LSTM)	[64, 128, 256, 512]
Num layers (LSTM)	[1, 2, 3]

5.3. Training process

The first step in our training process is to train individually each unimodal network. To this end, we stack on top of each network a 1-layer LSTM of 256 hidden units before the predictions of the arousal and valence.

The weights of each trained unimodal network are used to initialize the weights in the multimodal network. As in the unimodal cases, we use 1-layer LSTM of 256 hidden units to capture the temporal dynamics in the data, which was initialized by the Glorot based method [47].

5.4. Experimental setup

Our networks have a number of parameters that need to be tuned. For our purposes, we consider tuning the following hyper-parameters: learning rate, batch size, dropout, and number of hidden units and number layers for the LSTM. The range of the values for each hyper-parameter is shown in Table 1. We should point out that due to memory constrains the batch size for the visual and multimodal models is in the range [2, 3], whereas for the text and audio modalities was set in the range [10, 20, 30].

Due to the high number of configurations, we adopt the Hyperband algorithm [48] to tune the hyper-parameters. This is an iterative algorithm that starts by randomly sampling a number of hyper-parameter configurations in the search space, and discards some of them (in our case a 25%) during the training of the model.

For training the models, we use Adam optimizer [49], with initial learning rate of 5×10^{-4} throughout all experiments. The length of a sequence in a batch is set to 100. The batch size for training the visual and multimodal networks was set to 3, while for the text and audio models it was set to 20. Additionally, we apply dropout [50] to the text and audio models so that our model would not overfit on the training data. In particular, for the audio model dropout is performed with a probability of 0.5 after each convolution layer. For the text model, we apply dropout of 0.1 for original word representations, 0.3 for the output hidden states of context feature generator, and 0.2 for the output of the multi-linear projection matrices. Finally, each LSTM network we use in the training phase is trained with a dropout of 0.5.

5.5. Ablation study

We experiment with different state-of-the-art networks for each modality in order to choose the best performing one per modality. All models were trained using 1-layer LSTM on top of the extracted features such that the temporal dynamics in the data are captured. Using the visual modality we experiment with VGG [18], ResNet-50 [5], DenseNet [51], MobileNet [52], and HRNet [44]. Using the text modality we experiment with Fasttext [53], BERT [54], AlBert [55], XLNet [42], RoBerta [56], ULMFit [57] as our word embedding representation. Finally, we experiment with three end-to-end models using the audio modality, i. e, Trigeorgis et al. [58], Tzirakis et al. [13], and Li et al. [14]. Table 2 depicts the results for all modalities. From the results we conclude that on average the highest performing model for the visual modality is HRNet, for the text modality is the XLNet, and for the audio is Tzirakis et al. model. We utilize these models in our multimodal network.

Table 2
SEWA dataset development results (in terms of ρ_c) for the visual, text, and audio modalities, for the prediction of arousal and valence.

Modality	Network	Arousal	Valence	Avg
Visual	Resnet-50 [5]	.641	.689	.665
	VGG-16 [18]	.624	.668	.646
	DenseNet [51]	.617	.672	.645
	MobileNet [52]	.621	.689	.655
	HRNet [44]	**.647**	**.695**	**.671**
Text	Fasttext [53]	.508	.554	.531
	BERT [54]	.537	.578	.558
	AlBert [55]	.532	.577	.555
	RoBerta [56]	.523	.561	.542
	ULMFit [57]	.527	.559	.543
	XLNet [42]	**.544**	**.581**	**.563**
Audio	Trigeorgis et al. [58]	.541	.488	.513
	Li et al. [14]	.552	.534	.543
	Tzirakis et al. [13]	**.563**	**.532**	**.545**

Table 3
SEWA dataset test results (in terms of ρ_c) for the text modality, and for the prediction of arousal and valence. In parenthesis are the performances obtained on the development set.

Predictor	Arousal	Valence
Baseline [7]	.375 (.373)	.425 (.390)
Dang et al. [41]	.320 (.441)	.394 (.499)
Huang et al. [40]	.483 (.489)	.520 (.523)
Chen et al. [39]	.463 (.478)	.515 (.532)
Proposed	**.532** (.544)	**.568** (.581)

Table 4
SEWA dataset test results (in terms of ρ_c) for the audio modality, and for the prediction of arousal and valence. In parenthesis are the performances obtained on the development set.

Predictor	Arousal	Valence
Baseline [7]	.225 (.344)	.244 (.351)
Dang et al. [41]	.344 (.494)	.346 (.507)
Huang et al. [40]	**.583** (.584)	**.487** (.585)
Chen et al. [39]	.422 (.524)	.405 (.504)
Proposed	.456 (.563)	.428 (.532)

5.6. Unimodal results

We compare each of our unimodal networks to the baseline paper of the AVEC 2017 challenge along with the results of the best performing models of that year.

5.6.1. Text modality

Most of the studies in the AVEC 2017 challenge use BoTW to extract features from the text. Only Chen et al. [39] uses word2vec features. Table 3 depicts the results in the test and development sets. Our model outperforms all the other methods in predicting both the arousal and valence dimensions with high margin.

5.6.2. Audio modality

The audio modality is more effective at predicting the arousal dimension. Table 4 shows the results of our model and the rest of the AVEC challenge papers. Our model has the second highest performance behind the Huang et al.'s [40] method. We should note, however, that the network they use was pretrained on 300 hours of a spontaneous English speech recognition corpus before fine-tuning it to the SEWA dataset. In addition to the features of the network, they also utilize several hand-engineered features.

5.6.3. Visual modality

The visual information can efficiently predict the valence dimension rather than the arousal. Table 5 depicts the results. Our model

Table 5
SEWA dataset test results (in terms of ρ_c) for the visual modality, and for the prediction of arousal and valence. In parenthesis are the performances obtained on the development set. A dash is inserted if the results could not be obtained.

Predictor	Arousal	Valence
Baseline [7]	.308 (.466)	.455 (.400)
Dang et al. [41]	.390 (.518)	.496 (.583)
Huang et al. [40]	.531 (.682)	.670 (.720)
Chen et al. [39]	— (.675)	— (.693)
Proposed	**.608** (.647)	**.675** (.695)

Table 6
Results (in terms of ρ_c) on the development set of the SEWA dataset for the prediction of arousal and valence using five different fusion strategies, namely, concatenation, hierarchical attention, self-attention, residual self-attention, and cross-modal self-attention.

Method	Arousal	Valence
Concatenation	.698	.737
Self-attention	.729	.751
Hierarchical attention	.785	.808
Residual self-attention	.748	.767
Cross-modal hierarchical self-attention	**.797**	**.818**

outperforms all the other models. In addition, looking at the development results we can conclude that the generalization capability of our model is higher than the other models. We believe this is due to the hierarchical attentional pooling layer we used.

5.7. Fusion strategy results

In this subsection we show the results of our different fusion strategies, namely, concatenation, hierarchical attention, self-attention, residual self-attention, and cross-modal hierarchical self-attention. Table 6 shows the results on the development set. As expected the simple concatenation method produces the worst results. The hierarchical attention outperforms, in both arousal and valence, the self-attention and residual self-attention, but not the cross-modal hierarchical self-Attention. We believe that the bottom-up approach of the hierarchical attention provides the network the capability to more efficiently attend to one of the modalities before combining them together.

5.8. Multimodal results

To have a fair comparison with the models used in the AVEC 2017 challenge, and the Singh et al. [59]. and Khorram et al. [35] studies, our model was trained utilizing all the modalities of the dataset, namely, text, audio, and visual, for the final predictions. The results are shown in Table 7. Our model is the highest performing model compared with the other models. We should note that Chen et al.'s [39] model utilizes several hand-engineered and deep features, while our model operates on the raw signal. For example, the visual features they use are extracted from the DenseNet [51] and the VGG [18], while for the audio the features are the IS10 [60] and the features extracted from the SoundNet [61]. Finally, we run the Wilcox statistical test and found that the results are statistically significant with level of significance 0.05.

5.9. Hyper-parameter optimization

Each hyper-parameter has a different effect on the final multimodal network. In Table 8 we show, in terms of the average ρ_c, the effect of each hyper-parameter on the development set of the SEWA database. Results indicate that dropout has the highest effect on the model's performance with the highest value of standard deviation on both arousal and valence. On the other hand, the batch size has the lowest effect with 0.01 standard deviation.

Table 7
SEWA dataset results (in terms of ρ_c) for our multimodal network, and for the prediction of arousal and valence. In parenthesis are the performances obtained on the development set.

Predictor	Arousal	Valence
Baseline [7]	.375 (.525)	.466 (.507)
Singh et al. [59]	.276 (.294)	.365 (.346)
Khorram et al. [35]	.412 (.530)	.379 (.542)
Dang et al. [41]	.523 (.657)	.540 (.602)
Huang et al. [40]	.599 (.721)	.721 (.728)
Chen et al. [39]	.672 (.823)	.756 (.796)
Proposed	**.690** (.797)	**.783** (.818)

Table 8
SEWA dataset results (in terms of ρ_c) for the multimodal model on the development set for different hyper-parameters of the model.

Hyper-parameter	Arousal	Valence
Learning rate	$.765 \pm .03$	$.798 \pm .02$
Batch size	$.785 \pm .01$	$.805 \pm .01$
Dropout	$.742 \pm .05$	$.766 \pm .04$
Sequence length	$.783 \pm .03$	$.796 \pm .03$
Hidden units (LSTM)	$.783 \pm .02$	$.801 \pm .02$
Num layers (LSTM)	$.759 \pm .03$	$.792 \pm .03$

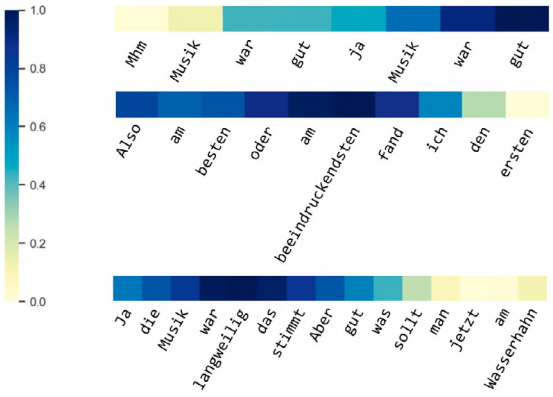

Fig. 3. Visualization of 3 sentences of the attention mechanism (*Selector*) in the text model. In the first sentence the highest weight is assigned to the word *gut* which means *good*, in the second sentence to the word *beeindruckendsten* which translates to *the most impressive*, and in the last sentence to the word *langweilig* which means *boring*. (Best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.10. Attention visualization

One of the benefits of using attention mechanism is the interpretability of its results. To this end, we visualize the attentions used in the text and multimodal networks.

5.10.1. Text visualization

The text network utilizes an attention mechanism (*Selector*), which we use to visualize the weights the network assigns to different words in the sentence. Fig. 3 depicts 3 sentences from the development set of the SEWA along with the weights the attention assigns to. As we can observe the most attentive words of the network are the ones that are the most meaningful. More particularly, in the first sentence the highest weight is assigned to the word *gut* which means *good*, in the second sentence to the word *beeindruckendsten* which translates to *the most impressive*, and in the last sentence to the word *langweilig* which means *boring*.

5.10.2. Multimodal visualization

We visualize the results of the hierarchical attention mechanism that is utilized to fuse the unimodal features. Fig. 4 shows the weights (color

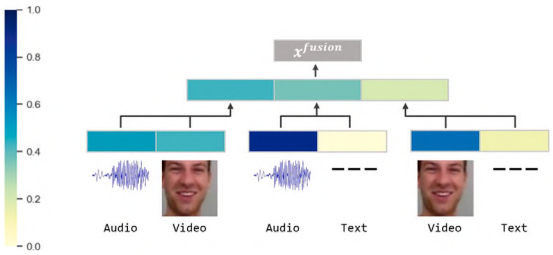

Fig. 4. Visualizing the hierarchical attention fusion strategy in a frame where the participant is laughing. The model provides the highest scores in the audiovisual stream in both the pair-wise attention (first layer) and in the higher layer. (Best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

encoded) of a frame where the participant is laughing and there is no text transcriptions. As expected the network attends mostly to the audio and visual channels in the first layer, while the highest score in the second layer is provided by the audiovisual pair of the first layer.

6. Conclusions

In this paper, we proposed a transformer-based text architecture along with attention-based fusion strategies to combine the different modality features to achieve better performance for affect recognition task. Our text model utilizes multi-linear projection and context-aware feature generator that can capture the semantics of a sentence. Furthermore, the proposed fusion strategies can balance the relationship among different modalities better than a simple concatenation to achieve a higher recognition performance on the SEWA dataset. Our text, visual and multimodal models outperform the state-of-the-art methods, while our audio network achieve the second best performance when compared with models that utilize several hand-crafted and deep features. Finally, our model can provide a series of visualizations for the text and the multimodal networks.

CRediT authorship contribution statement

Panagiotis Tzirakis: Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing - original draft, Writing - review & editing. **Jiaxin Chen:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing - original draft, Writing - review & editing. **Stefanos Zafeiriou:** Conception and design of study, Writing - original draft, Writing - review & editing. **Björn Schuller:** Conception and design of study, Writing - original draft, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The support of the EPSRC, UK Center for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1) is gratefully acknowledged. All authors approved the version of the manuscript to be published.

References

[1] N. Banda, P. Robinson, Multimodal affect recognition in intelligent tutoring systems, in: Affective Computing and Intelligent Interaction, Springer, 2011, pp. 200–207.
[2] C. Anagnostopoulos, T. Iliou, I. Giannoukos, Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, Artif. Intell. Rev. 43 (2) (2015) 155–177.
[3] P. Tzirakis, S. Zafeiriou, B. Schuller, Real-world automatic continuous affect recognition from audiovisual signals, in: Multimodal Behavior Analysis in the Wild, vol. 1, Academic Press, 2019, pp. 387–406, (Chapter 18).
[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017, pp. 5998–6008.
[5] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.
[6] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[7] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, M. Pantic, AVEC 2017: Real-life depression, and affect recognition workshop and challenge, in: The Workshop on AVEC, 2017, pp. 3–9.
[8] J. Pons, X. Serra, Randomly weighted CNNs for (music) audio classification, in: ICASSP, 2019, pp. 336–340.
[9] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, et al., The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring, in: Computational Paralinguistics Challenge (ComParE), Interspeech 2017, 2017, pp. 3442–3446.
[10] B.W. Schuller, S. Steidl, A. Batliner, P.B. Marschik, H. Baumeister, F. Dong, S. Hantke, F.B. Pokorny, E.-M. Rathner, K.D. Bartl-Pokorny, et al., The Interspeech 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats.
[11] M. Neumann, N. Vu, Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech, 2017, arXiv preprint arXiv:1706.00612.
[12] M. Xu, F. Zhang, S.U. Khan, Improve accuracy of speech emotion recognition with attention head fusion, in: CCWC, 2020, pp. 1058–1064.
[13] P. Tzirakis, J. Zhang, B. Schuller, End-to-end speech emotion recognition using deep neural networks, in: ICASSP, 2018, pp. 5200–5204.
[14] Y. Li, T. Zhao, T. Kawahara, Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning, in: Interspeech 2019, 2019, pp. 2803–2807.
[15] H. Yang, U. Ciftci, L. Yin, Facial expression recognition by de-expression residue learning, in: CVPR, 2018, pp. 2168–2177.
[16] S. Minaee, A. Abdolrashidi, Deep-emotion: Facial expression recognition using attentional convolutional network, 2019, arXiv preprint arXiv:1902.01019.
[17] D. Kollias, M.A. Nicolaou, I. Kotsia, G. Zhao, S. Zafeiriou, Recognition of affect in the wild using deep neural networks, in: CVPR Workshops, 2017, pp. 1972–1979.
[18] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: BMVC, 2015, p. 6.
[19] S. Li, W. Deng, Deep facial expression recognition: A survey, TAFFC (2020).
[20] Y. Kim, Y. Jernite, D. Sontag, A. Rush, Character-aware neural language models, in: AAAI, 2016, pp. 2741–2749.
[21] S. Angelidis, M. Lapata, Multiple instance learning networks for fine-grained sentiment analysis, Trans. Assoc. Comput. Linguist. 6 (2018) 17–31.
[22] K. Sarma, Y. Liang, A. Sethares, Domain adapted word embeddings for improved sentiment classification, in: ACL, 2018, pp. 37–42.
[23] S. Park, B. Bae, Y. Cheong, Emotion recognition from text stories using an emotion embedding model, in: BigComp, pp. 579–583.
[24] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep learning based text classification: A comprehensive review, 2020, arXiv preprint arXiv:2004.03705.
[25] E. Cambria, N. Howard, J. Hsu, A. Hussain, Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics, in: CIHLI, IEEE, 2013, pp. 108–117.
[26] P. Tzirakis, G. Trigeorgis, M. Nicolaou, B. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, IEEE J. Sel. Top. Sign. Proces. 11 (8) (2017) 1301–1309.
[27] B.T. Atmaja, M. Akagi, Multitask learning and multistage fusion for dimensional audiovisual emotion recognition, in: ICASSP, 2020, pp. 4482–4486.
[28] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, I. Marsic, Multimodal affective analysis using hierarchical attention strategy with word-level alignment, in: ACL, 2018, pp. 1–11.
[29] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, ICON: interactive conversational memory network for multimodal emotion detection, in: EMNLP, 2018, pp. 2594–2604.
[30] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: ACL, 2019, pp. 6558–6569.
[31] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, A. Hussain, Multimodal sentiment analysis: Addressing key issues and setting up the baselines, IEEE Intell. Syst. 33 (6) (2018) 17–25.

[32] P. Tzirakis, S. Zafeiriou, B.W. Schuller, End2You–The imperial toolkit for multimodal profiling by end-to-end learning, 2018, arXiv preprint arXiv:1802.01115.
[33] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognit. Lett. 125 (2019) 264–270.
[34] H.-N. Tran, E. Cambria, Ensemble application of ELM and GPU for real-time multimodal sentiment analysis, Memetic Computing 10 (1) (2018) 3–13.
[35] S. Khorram, M. McInnis, E. Mower Provost, Jointly aligning and predicting continuous emotion annotations, IEEE Trans. Affect. Comput. (2019) 1–1.
[36] L. Stappen, A. Baird, G. Rizos, P. Tzirakis, X. Du, F. Hafner, L. Schumann, A. Mallol-Ragolta, B.W. Schuller, I. Lefter, et al., MuSe 2020–The first international multimodal sentiment analysis in real-life media challenge and workshop, 2020, arXiv preprint arXiv:2004.14858.
[37] K. Bahreini, R. Nadolski, W. Westera, Data fusion for real-time multimodal emotion recognition through webcams and microphones in e-learning, International Journal of Human–Computer Interaction 32 (5) (2016) 415–430.
[38] S. Mai, H. Hu, S. Xing, Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing, in: ACL, 2019, pp. 481–492.
[39] S. Chen, Q. Jin, J. Zhao, S. Wang, Multimodal multi-task learning for dimensional and continuous emotion recognition, in: Workshop on AVEC, 2017, pp. 19–26.
[40] J. Huang, Y. Li, J. Tao, Z. Lian, Z. Wen, M. Yang, J. Yi, Continuous multimodal emotion prediction based on long short term memory recurrent neural network, in: Workshop on AVEC, 2017, pp. 11–18.
[41] T. Dang, B. Stasak, Z. Huang, S. Jayawardena, M. Atcheson, M. Hayat, P. Le, V. Sethu, R. Goecke, J. Epps, Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in AVEC 2017, in: Workshop on AVEC, 2017, pp. 27–35.
[42] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R.R. Salakhutdinov, Q.V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: NIPS, 2019, pp. 5754–5764.
[43] K. Srivastava, K. Greff, J. Schmidhuber, Training very deep networks, in: NIPS, 2015, pp. 2377–2385.
[44] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, B. Xiao, Deep high-resolution representation learning for visual recognition, TPAMI.
[45] R. Girdhar, D. Ramanan, Attentional pooling for action recognition, in: NIPS, 2017, pp. 34–45.
[46] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: CVPR, 2018, pp. 7794–7803.
[47] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: AIStat, 2010, pp. 249–256.
[48] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, A. Talwalkar, Hyperband: A novel bandit-based approach to hyperparameter optimization, J. Mach. Learn. Res. (2017) 6765–6816.
[49] D. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[50] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
[51] G. Huang, Z. Liu, L. v. d. Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: CVPR, 2017, pp. 2261–2269.
[52] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.
[53] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. 5 (2017) 135–146.
[54] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[55] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, in: ICLR, 2019.
[56] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.
[57] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: ACL, 2018, pp. 328–339.
[58] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M.A. Nicolaou, B. Schuller, S. Zafeiriou, Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network, in: ICASSP, 2016, pp. 5200–5204.
[59] N. Singh, N. Singh, A. Dhall, Continuous multimodal emotion recognition approach for AVEC 2017, 2017, arXiv preprint arXiv:1709.05861.
[60] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, S. Narayanan, The Interspeech 2010 paralinguistic challenge, in: Interspeech 2010, 2010, pp. 2794–2797.
[61] Y. Aytar, C. Vondrick, A. Torralba, Soundnet: Learning sound representations from unlabeled video, in: NIPS, 2016.