electronics

MDPI

Article

Depression Detection in Speech Using Transformer and Parallel Convolutional Neural Networks

Faming Yin [1], Jing Du [2], Xinzhou Xu [3,*] and Li Zhao [2]

[1] School of Network and Communication, Nanjing Vocational College of Information Technology, Nanjing 210023, China

[2] School of Information Science and Engineering, Southeast University, Nanjing 210096, China

[3] School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

[*] Correspondence: xinzhou.xu@njupt.edu.cn

Abstract: As a common mental disorder, depression becomes a major threat to human health and may even heavily influence one's daily life. Considering this background, it is necessary to investigate strategies for automatically detecting depression, especially through the audio modality represented by speech segments, mainly due to the efficient latent information included in speech when describing depression. However, most of the existing works focus on stacking deep networks in audio-based depression detection, which may lead to insufficient knowledge for representing depression in speech. In this regard, we propose a deep learning model based on a parallel convolutional neural network and a transformer in order to mine effective information with an acceptable complexity. The proposed approach consists of a parallel convolutional neural network (parallel-CNN) module used to focus on local knowledge, while a transformer module is employed as the other parallel stream to perceive temporal sequential information using linear attention mechanisms with kernel functions. Then, we performed experiments on two datasets of Distress Analysis Interview Corpus-Wizard of OZ (DAIC-WOZ) and Multi-modal Open Dataset for Mental-disorder Analysis (MODMA). The experimental results indicate that the proposed approach achieves a better performance compared with the state-of-the-art strategies.

Keywords: depression detection; transformer; parallel convolutional neural networks; linear attention

check for updates

Citation: Yin, F.; Du, J.; Xu, X.; Zhao, L. Depression Detection in Speech Using Transformer and Parallel Convolutional Neural Networks. *Electronics* **2023**, *12*, 328. https://doi.org/10.3390/electronics12020328

Academic Editor: Stefanos Kollias

Received: 10 November 2022
Revised: 24 December 2022
Accepted: 3 January 2023
Published: 8 January 2023

1. Introduction

The mental disorder of depression affects a huge number of people in the world [1,2], making them suffer from psychological pain and pessimism, or even leading to an increasing tendency of self-abuse and suicide [3,4]. Depression itself is difficult to diagnose empirically, since it mainly depends on the patient's self-assessment [5] and other evaluation questionnaires [6,7], which may largely be affected by subjective factors. Hence, the design of effective depression detection systems with an objective standard of diagnosis is needed for the purpose of reducing these subjective effects [8].

Fortunately, the current research on automatic depression detection provides feasible methods for this scenario, including video-based [9–11], text-based [12,13], audio-based [5,14], electroencephalogram (EEG)-based [15,16], and multi-modal fusion [17–22] strategies. Within these strategies, speech in audio signals has been proved as effective in detecting depression due to the latent depression factors hidden in speech [23,24]. To design a speech-based depression detection system, existing research considers deep models employ stacking convolutional neural networks (CNNs) or recurrent neural network (RNN) structures [25–27]. This may lead to insufficient depression-related information obtained by the output layers due to the single-stream networks [28,29], and may also result in a large time complexity and memory footprint [30]. In addition, previous works on speech emotion recognition (SER) show the effectiveness in mining latent emotional information from multi-category features in speech [31–33], which implies the possibility of including parallel structures for the current depression detection task through learning from multiple representations [34].

In view of the limitations in the existing works, we propose a convolutional neural network based on a parallel structure and transformer model in this paper in order to learn effective depression representations through multi-source fusion with acceptable resource occupation. First, we set low-level mel-frequency cepstral coefficient (MFCC) features with a fixed length as the input to the network. Then, a parallel CNN structure was employed to process these inputs for the purpose of perceiving local knowledge. Afterwards, we added an improved transformer with a linear attention mechanism as a third stream of the proposed approach, instead of conventional RNNs, in order to capture temporal sequential information in speech. The linear attention mechanism utilizes the kernel function in constructing the inner product instead of the original scaled dot-product attention, in order to reduce the complexity of the attention mechanisms. Note that, compared with existing related works on depression detection in speech, the proposed approach utilizes CNN streams to learn local knowledge while setting a transformer stream for acquiring global sequential information.

The main contribution of this paper can be summarized as follows:

- We propose a transformer-CNN-CNN (TCC) model for depression detection in speech, with two modules of a parallel CNN and transformer.
- Within the proposed TCC, we employed linear attention mechanisms instead of the original attention in multi-head attention mechanisms to reduce time complexity.
- Within the proposed TCC, a fusion layer was then added to merge the outputs from the CNNs and transformer to obtain sufficient local and global information for fusion decisions.

The remainder of this paper is organized as follows. Section 2 introduces highly related works for the proposed approach, while Section 3 shows the proposed model consisting of a parallel CNN and a transformer based on a parallel structure, including the improvement scheme of the attention mechanisms. Then, Section 4 represents the databases, experimental setups, and analysis of the experimental results. The conclusion of this paper is provided in Section 5.

2. Related Works

2.1. Depression Detection in Speech

As a typical human psychological disorder, depression requires an accurate and fast diagnosis through speech communication [23,35], inspiring the works on automatic depression detection in speech [24,36,37]. The initial research on speech-based depression detection depends on shallow-structure algorithms of support vector machines (SVMs) [38] and Gaussian mixture models (GMMs) [39]. Then, Liu et al. [40] proposed a novel decision tree for depression detection. The emergence of deep learning makes it possible to utilize deep neural networks for performing more accurate depression detection in speech [14,34,41,42]. On exploring the theories for speech-based depression detection, Dubagunta et al. found that neurophysiological changes can occur during depression, which may affect laryngeal control (i.e., the behavior of the vocal folds) [43]. Along this direction, they used a CNN-based structure to demonstrate the effectiveness of depression detection only using the audio modality of speech [43]. Then, in this way, Zhao et al. also explored employing a structure using a CNN and long short-term memory (LSTM) with a multi-head attention mechanism [31], yet the inclusion of recurrent units may incur a high computational complexity [30]. In addition to the research on exploring learning algorithms, handcrafted features are also proved as effective when confronting speech-based depression detection [26,44].

2.2. Transformer

The transformer model proposed by Vaswani et al. shows a better performance and efficient parallelization in the field of natural language processing (NLP) compared with the conventional RNN and CNN-based networks [30]. Seeing the success of the transformer, following works also support its success in the fields of image, speech, and multi-modal processing [45–48]. Within the related works, a spatial transformer was proposed for image compositing [49]. Then, Chen et al. proposed a pre-trained image processing transformer that performs well in adapting to different image processing tasks [50]. In the field of speech-related works, Dong et al. expanded the transformer to speech recognition, which achieves a low word error rate and faster training [51]. Wang et al. presented an end-to-end model based on stacked transformer layers for SER tasks, which outperformed the prior arts by a large margin, still making use of the CNN-RNN-based architecture [52]. Related works on transformer-based SER also prove the feasibility of using transformer structures in computational paralinguistics [53,54].

2.3. Attention Mechanisms

Attention mechanisms initially achieved great success in the field of image processing and NLP [55–57]. In speech processing, Mirsamadi et al. introduced local attention based on a recurrent neural network for automatic speech recognition (ASR) [58], while Xie et al. and Jiang et al. utilized multiple attention mechanisms to achieve a better performance for SER tasks [59,60]. Regarding the transformer, multi-head attention is proposed for parallel computing, applied to the fields of image [61], language [62], speech [31,63], and multi-modal [64] processing, allowing the model to focus on the information of the representations from different locations. However, due to the computational complexity required for softmax calculation, the attention mechanism algorithm requires squared time-memory complexity with the input length [65,66]. Thus, in order to reduce the computational complexity of the attention mechanisms, Shen el al. proposed an efficient attention module through applying softmax functions to the query and key vectors, considering the association of matrix multiplication [67], while Tsai et al. proposed a kernel-based attention module through mathematical analysis on the attention-calculation formula [68].

3. The Proposed Transformer-CNN-CNN

The overall structure of the proposed approach is shown in Figure 1. For the convenience of narration, we use transformer-CNN-CNN to refer to the proposed model, containing two modules of a parallel CNN and transformer. The parallel-CNN module consisted of three convolutional layers with pooling operations, and the transformer module included four transformer layers with multi-head attention, both utilizing the inputs from 40-dimensional low-level MFCC extraction [69], considering the usage of MFCC in existing speech-based depression detection works [70,71]. The outputs of the two modules were then processed through flattening (only for the parallel-CNN module) and concatenating, prior to feeding to the softmax layer to make a final decision.

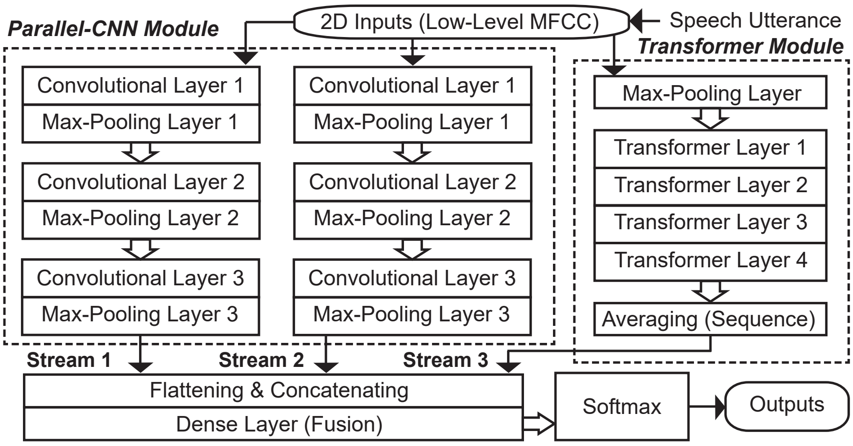


Figure 1. A systematic overview of the proposed TCC in this work for depression detection in speech, including the parallel-CNN and transformer modules.

3.1. The Parallel-CNN Module

Inspired by conventional CNN models for image and sequential signal processing (e.g., LeNet, AlexNet, and VGGNet) [72,73], the CNN architecture used in this paper set three stacking convolutional layers for each CNN stream, with 16, 32, and 64 channels, respectively. Using the parallel CNN's representations, the mel spectrogram X was processed as a grayscale image, with the two dimensions of time and frequency, noted as '$T \times d$', where the d mel-spectrogram point was set to 40. The size of the convolutional kernel for each convolutional layer was set to 3, whereas the stride step was 1. Each of the convolutional layers activated by a rectified linear unit (ReLU) function was followed by batch normalization (BN). Then, we performed maximum pooling and employed dropout operations afterwards for each convolutional layer with a rate of 0.3 in order to reduce the influence from overfitting.

3.2. The Transformer Module

Different from the original transformer, we improved the transformer structure in this paper and only a similarity-encoder structure was used in order to adapt to the current task of detecting depression in speech. In view of the existing research on transformers, a deep model does not always perform well when stacking a large number of transformer layers [74,75]. Hence, the transformer module of the proposed TCC contained four stacked transformer layers as a transformer encoder, including a multi-head attention sublayer and a feed-forward sublayer for each transformer layer. We further added residual skip connections for each sublayer, as shown in Figure 2.

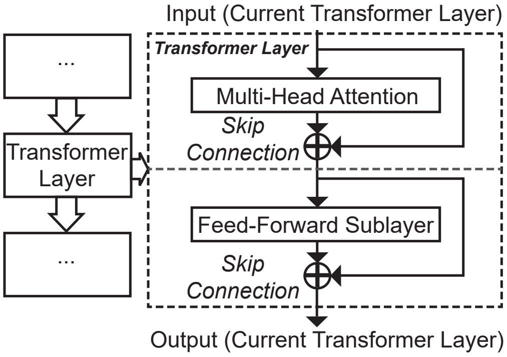


Figure 2. A systematic overview for each transformer layer from the transformer module within the proposed TCC.

Let the input of the lth transformer layer be $X_l^{(T)} \in \Re^{T_0 \times d}$, where T_0 is the length of the processed temporal sequence using 4×1 maximum pooling (see Figure 1) on the time dimension of the initial input X. Hence, the output of the corresponding multi-head attention sublayer is represented as

$$X_l^{(T,MA)} = X_l^{(T)} + MA(X_l^{(T)}), \tag{1}$$

where $MA(\cdot)$ is the mapping function for the multi-head attention. Thus, the output of the feed-forward sublayer can be calculated as

$$X_l^{(T,FF)} = X_l^{(T,MA)} + ReLU(X_l^{(T,MA)} \cdot W_1 + b_1) \cdot W_2 + b_2, \tag{2}$$

where $ReLU(\cdot)$ represents an ReLU activation, while the linear mappings are $W_1 \in \Re^{d \times d_0}$ and $W_2 \in \Re^{d_0 \times d}$ with there corresponding biases b_1 and b_2, using $d_0 = 512$ nodes for the hidden layer, with a dropout operation setting its dropout rate to 0.4.

3.2.1. Multi-Head Attention

Then, we focused on the multi-head attention unit as shown in Figure 2, which endows a model with sufficient information through parallel training. Within this unit, an input vector was divided into multiple spaces, and a self-attentional mechanism was employed to allow for a model's parallel training while capturing sufficient target knowledge. Compared with the current conventional single-head attention mechanism [76], multi-head attention can improve the effective resolution, thus enhancing the representation of speech using multiple spaces while reducing interference from noise through the average-pooling operator.

In a multi-head attention unit from the lth transformer layer, we first present the calculation process of query, key, and value vectors as $Q_{l,n^{(S)}} = X_l^{(T)} \cdot W_{Q_{l,n^{(S)}}}$, $K_{l,n^{(S)}} = X_l^{(T)} \cdot W_{K_{l,n^{(S)}}}$, and $V_{l,n^{(S)}} = X_l^{(T)} \cdot W_{V_{l,n^{(S)}}}$, respectively, with $n^{(S)} = 1, 2, \ldots, N^{(S)}$. Note that the $W_{Q_{l,n^{(S)}}} \in \Re^{d \times d_K}$, $W_{K_{l,n^{(S)}}} \in \Re^{d \times d_K}$, and $W_{V_{l,n^{(S)}}} \in \Re^{d \times d_V}$ represent the corresponding linear-mapping matrices for the three sorts of vectors, where we set $d_K = d_V = \lfloor \frac{d}{N^{(S)}} \rfloor$.

Afterwards, we performed concatenating (noted as '$Concat(\cdot)$') considering the outputs from all of the heads as

$$MA(X_l^{(T)}) = Concat(H_{l,1}, H_{l,2}, \ldots, H_{l,N^{(S)}}), \tag{3}$$

with the output of the $n^{(S)}$th head as

$$H_{l,n^{(S)}} = Atten(Q_{l,n^{(S)}}, K_{l,n^{(S)}}, V_{l,n^{(S)}}), \tag{4}$$

where $Atten(\cdot)$ represents the operator from an attention unit.

3.2.2. The Attention Unit

First, we employed scaled dot-product attention in the attention unit based on single-head attention mechanisms [30,76]. The main idea of the scaled dot-product attention is to enhance the representation of the current time step by introducing contextual information. For the $n^{(S)}$th head from the lth transformer layer, the query $Q_{l,n^{(S)}}$ represents the content of interest, while the key $K_{l,n^{(S)}}$ is equivalent to the tags of all of the words in the current dictionary. The influence degree of the context for the current time step is reflected through the dot-product result of $Q_{l,n^{(S)}}$ and $K_{l,n^{(S)}}$. Then, the relevance measure was obtained by softmax normalization, which was used to weight the value $V_{l,n^{(S)}}$ in order to obtain the attention score.

Hence, the calculation process of the scaled dot-product attention can be represented as

$$Atten(Q_{l,n^{(S)}}, K_{l,n^{(S)}}, V_{l,n^{(S)}}) = Softmax\left(\frac{Q_{l,n^{(S)}} \cdot K_{l,n^{(S)}}^T}{\sqrt{d_K}}\right) \cdot V_{l,n^{(S)}}, \tag{5}$$

in which $Softmax(\cdot)$ represents the softmax function.

Further, by applying the softmax function to

$$\begin{cases} Q_{l,n^{(S)}} = [q_{l,n^{(S)},1}, q_{l,n^{(S)},2}, \ldots, q_{l,n^{(S)},T_0}]^T \in \Re^{T_0 \times d_K}, \\ K_{l,n^{(S)}} = [k_{l,n^{(S)},1}, k_{l,n^{(S)},2}, \ldots, k_{l,n^{(S)},T_0}]^T \in \Re^{T_0 \times d_K}, \\ V_{l,n^{(S)}} = [v_{l,n^{(S)},1}, v_{l,n^{(S)},2}, \ldots, v_{l,n^{(S)},T_0}]^T \in \Re^{T_0 \times d_V}, \end{cases} \tag{6}$$

we obtained the ith row ($i = 1, 2, \ldots, T_0$) of the attention unit's output as

$$\begin{aligned} &Atten_i(Q_{l,n^{(S)}}, K_{l,n^{(S)}}, V_{l,n^{(S)}}) \\ &= \frac{[e^{\frac{q^T_{l,n^{(S)},i} k_{l,n^{(S)},1}}{\sqrt{d_K}}}, e^{\frac{q^T_{l,n^{(S)},i} k_{l,n^{(S)},2}}{\sqrt{d_K}}}, \ldots, e^{\frac{q^T_{l,n^{(S)},i} k_{l,n^{(S)},T_0}}{\sqrt{d_K}}}]}{\sum_{j=1}^{T_0} e^{\frac{q^T_{l,n^{(S)},i} k_{l,n^{(S)},j}}{\sqrt{d_K}}}} \cdot V_{l,n^{(S)}}, \end{aligned} \tag{7}$$

resulting in the form of

$$Atten_i(Q_{l,n^{(S)}}, K_{l,n^{(S)}}, V_{l,n^{(S)}}) = \frac{\sum_{m=1}^{T_0} \left(e^{\frac{q^T_{l,n^{(S)},i} k_{l,n^{(S)},m}}{\sqrt{d_K}}} v^T_{l,n^{(S)},m} \right)}{\sum_{j=1}^{T_0} e^{\frac{q^T_{l,n^{(S)},i} k_{l,n^{(S)},j}}{\sqrt{d_K}}}}. \tag{8}$$

Then, considering the similarity measurement of the scaled dot-product attention [77], we further rewrote Equation (8) as

$$Atten_i(Q_{l,n^{(S)}}, K_{l,n^{(S)}}, V_{l,n^{(S)}}) = \frac{\sum_{m=1}^{T_0} \left(Sim(q_{l,n^{(S)},i}, k_{l,n^{(S)},m}) v^T_{l,n^{(S)},m} \right)}{\sum_{j=1}^{T_0} Sim(q_{l,n^{(S)},i}, k_{l,n^{(S)},j})}, \tag{9}$$

where $Sim(\cdot)$ represents the similarity measurement between two arguments of vectors. In view of the possible $O(T_0^2)$ computational complexity for the softmax calculation in processing long sequences, we further set non-negative mappings and inner products in order to replace the softmax for representing the similarity measurement [68]. Hence, we induced two non-negative mappings $\phi(\cdot)$ and $\varphi(\cdot)$ to perform linear inner products for $q_{l,n^{(S)},i}$ and $k_{l,n^{(S)},j}$ (or $k_{l,n^{(S)},m}$), respectively, leading to $Sim(q_{l,n^{(S)},i}, k_{l,n^{(S)},j}) = \phi^T(q_{l,n^{(S)},i})\varphi(k_{l,n^{(S)},j})$. Employing this similarity form, we obtained

$$Atten_i(Q_{l,n^{(S)}}, K_{l,n^{(S)}}, V_{l,n^{(S)}}) = \frac{\sum_{m=1}^{T_0} \left(\phi^T(q_{l,n^{(S)},i})\varphi(k_{l,n^{(S)},m}) v^T_{l,n^{(S)},m} \right)}{\sum_{j=1}^{T_0} \phi^T(q_{l,n^{(S)},i})\varphi(k_{l,n^{(S)},j})}, \tag{10}$$

which was further transformed into

$$Atten_i(Q_{l,n^{(S)}}, K_{l,n^{(S)}}, V_{l,n^{(S)}}) = \frac{\phi^T(q_{l,n^{(S)},i}) \sum_{m=1}^{T_0} \left(\varphi(k_{l,n^{(S)},m}) v^T_{l,n^{(S)},m} \right)}{\phi^T(q_{l,n^{(S)},i}) \sum_{j=1}^{T_0} \varphi(k_{l,n^{(S)},j})}. \tag{11}$$

When assuming that, for the matrix forms,

$$\varphi(K_{l,n^{(S)}}) = [\varphi(k_{l,n^{(S)},1}), \varphi(k_{l,n^{(S)},2}), \ldots, \varphi(k_{l,n^{(S)},T_0})]^T \in \Re^{T_0 \times d_K}, \tag{12}$$

we rewrote the ith row of the attention unit's output as

$$Atten_i(Q_{l,n^{(S)}}, K_{l,n^{(S)}}, V_{l,n^{(S)}}) = \frac{\phi^T(q_{l,n^{(S)},i}) \left(\varphi^T(K_{l,n^{(S)}}) \cdot V_{l,n^{(S)}} \right)}{\phi^T(q_{l,n^{(S)},i}) \left(\varphi^T(K_{l,n^{(S)}}) \mathbf{e} \right)}, \tag{13}$$

where all of the elements $\mathbf{e} \in \Re^{T_0 \times 1}$ are equal to one. Hence, the computational complexity changes into $O(T_0(d_K)^2)$. For both of the non-negative mappings $\phi(\cdot)$ and $\varphi(\cdot)$, we employed the exponential linear unit (ELU) activation as in [68].

4. Experiments

4.1. Experimental Setups

We employed the datasets of Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) [78] and Multi-modal Open Dataset for Mental-disorder Analysis (MODMA) [43,79] in this paper. The setup of one-time training-test splitting was included in the experiments in accordance with existing related research [31,80,81].

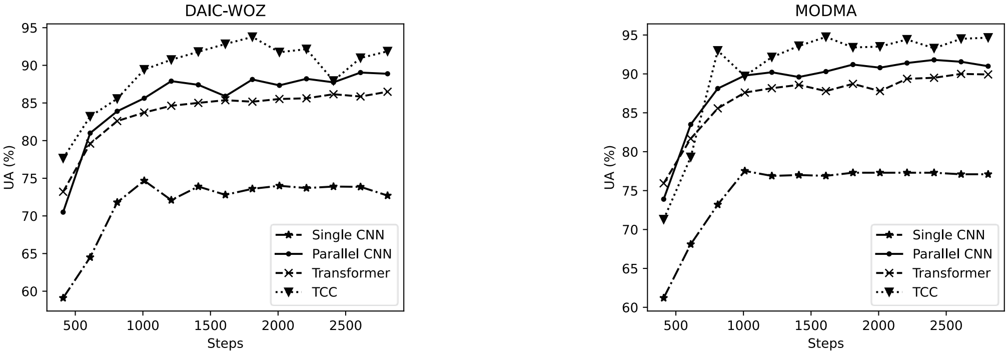
The DAIC-WOZ dataset was composed of interview recordings for the clinical diagnosis of psychological stress, conducted by Ellie, an animated virtual interviewer, with a sampling rate of 16 kHz. The dataset included 42 randomly selected patients with depression and 47 healthy speakers. In order to ensure the validity of the sample, we eliminated the abnormal speech segments containing a small amount of information—that is, less than 3 s and more than 12 s—leading to a total of 4401 samples (2156 depression and 2245 normal). Then, 1000 samples were randomly selected as the test set, whereas the rest of the samples formed the training part, considering the balance of data distribution between the two classes as in Ref. [31]. The MODMA dataset contained 52 subjects, including 23 subjects (7 females; 16–55 years) with depression and 29 healthy-control subjects (9 females; 18–55 years). The speakers were recorded by interviewing, reading, and picture description. After segmenting the original speech, the average duration for the segments exceeded 10 s. Hence, a total of 1321 samples were obtained through eliminating the abnormal samples from the data (590 depression and 731 normal), in which, we randomly selected 350 samples as the test set. Note that, despite the partially speaker-dependent cases in splitting the sets, we employed the random splitting, since the unsupervised side information (e.g., speaker, text, duration, and noise) provides relatively low impact on the supervised depression-related information.

Within the experiment, we employed the Hanning window, with a window length of 25 ms in speech framing, and a window shift of 10 ms. The speech segments with a length of 9 s were intercepted by the Python library of Librosa to extract the MFCC features [82]. The network employed a batch size of 32 using the adaptive moment estimation (Adam) optimizer, with the initial learning rate equal to 0.0005.

4.2. Experimental Results

4.2.1. Experiments for TCC

First, for the purpose of verifying the validity of the multi-stream structure within the proposed model, we make a comparison on the depression-detection performance between a single CNN, parallel CNN, transformer, and the proposed TCC model (using the scaled dot-product attention). For the TCC model, its parameter settings keep consistent with the parallel CNN, and the multi-head numbers used in the transformer and TCC are both set to 4. In this regard, we present Figure 3 in order to show the unweighted accuracy (UA) or equivalently unweighted average recall (UAR) convergence curves of the four models on the two datasets, respectively.



(**a**) (**b**)

Figure 3. The unweighted accuracies (%) of single CNN, parallel CNN, transformer, and the proposed TCC model on (**a**) DAIC-WOZ and (**b**) MODMA datasets.

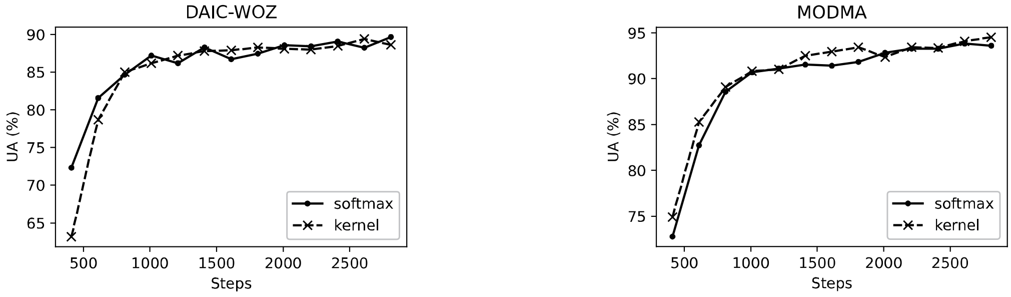
It is learnt from the figure that the proposed TCC model achieves the best UA performance, followed by the parallel CNN, transformer, and finally the single-CNN model. Among them, the TCC performs better compared with the parallel CNN, while there is a slight gap between the transformer and the two models. According to the performance of the parallel CNN, we still believe that the CNN has a strong feature extraction ability and can learn effective feature representation from almost any input to complete the classification task. The transformer is inferior in feature extraction compared to the parallel CNN, yet it can be implemented using paralleled computing so as to improve the training efficiency of the model. As for the proposed TCC model, the multi-head attention mechanism in the transformer can strengthen a simple CNN model's ability to capture temporal features by dividing the input into multiple spaces for processing. Considering the analysis and experimental results, the proposed TCC model may be optimal among these models in comparison.

To further investigate the optimal performance of the proposed model, we changed the number of the attention heads $N^{(S)}$ used in the experiments for the two datasets. The models can be represented as 'TCC-Single' and 'TCC-$N^{(S)}$', respectively, where TCC-Single represents the single-head attention mechanism used by the model and TCC-$N^{(S)}$ indicates that the model utilizes a multi-head attention mechanism. The attention unit in TCC-Single and TCC-$N^{(S)}$ was set to scaled dot-product attention. We directly trained the models on the two datasets, and used UAs to evaluate the effectiveness of the models, as shown in Table 1. It can be seen from the table that the TCC-$N^{(S)}$ models achieve better UA results compared with the TCC-Single model on the two datasets. Specifically, TCC-4 achieves the best UA performance (93.8%) on the DAIC-WOZ dataset, whereas TCC-2 performs the best (95.8%) on the MODMA dataset.

Table 1. The unweighted accuracies (%) of the proposed TCC approach with different numbers of heads in the multi-head attention mechanism (TCC-Single, TCC-2, TCC-4, TCC-8, and TCC-16).

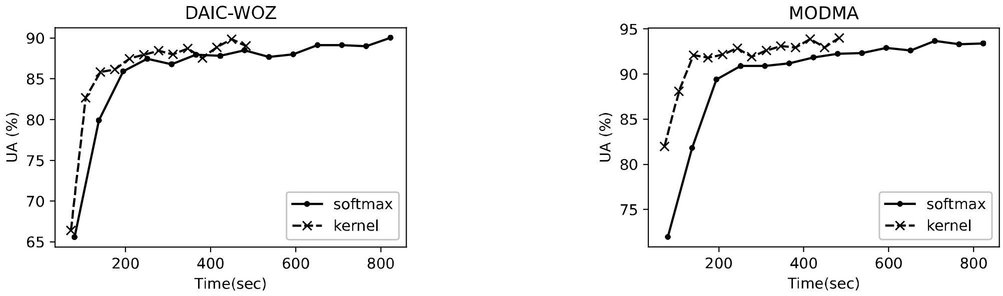
Models\Datasets	**DAIC-WOZ**	**MODMA**
TCC-Single	84.5	89.0
TCC-2	90.4	**95.8**
TCC-4	**93.8**	94.6
TCC-8	91.5	93.8
TCC-16	91.0	91.5

In view of the experimental results in Table 1, we chose the number of heads in the multi-head attention as $N^{(S)} = 4$ in the following experiments. Then, in order to compare both the accuracy and time-complexity performance of different attention computing units more precisely, we only used the transformer stream without the parallel-CNN streams, with the UA convergence curves shown in Figures 4 and 5, with respect to iteration steps and the training time (within 3000 iteration steps) on the two datasets. We employed the two setups of scaled dot-product attention (noted as 'softmax') and linear inner-product attention (noted as 'kernel' due to the similar mapping assumption as in kernel methods [83]). In Figure 4, we show the UAs starting from 400 iteration steps, and then present the results every 200 steps. Similarly, in Figure 5, we set 80 s as the start point for presenting the UAs, also using 200 steps as the interval. It can be seen that the UA performance of the proposed model when using the linear attention is similar to that when using the scaled dot-product attention, and even slightly better compared with the scaled dot-product attention. In addition, the 'kernel' setup leads to a lower training time compared with the 'softmax' setup due to the reduced complexity in the matrix calculation.



(**a**) (**b**)

Figure 4. The unweighted accuracies (%) of the transformer module with softmax and kernel setups for different iteration steps in training on (**a**) DAIC-WOZ and (**b**) MODMA datasets.



(**a**) (**b**)

Figure 5. The unweighted accuracies (%) of the transformer module with softmax and kernel setups for different iteration steps in training on the time scale on (**a**) DAIC-WOZ and (**b**) MODMA datasets.

Finally, we present the confusion matrices for the proposed TCC (including TCC-softmax and TCC-kernel) on DAIC-WOZ and MODMA datasets, considering the experimental results of the metrics, as shown in Figure 6. The results indicate that the proposed TCC performs well in classifying both normal and depression speech. In addition, it is observed that the TCC-kernel outperforms TCC-softmax on the MODMA dataset and is insignificant for the UAs of the two setups on both of the datasets (at the significance level of 0.05 using a one-tailed *z*-test); hence, it is feasible to use the TCC-kernel directly due to its lower computational complexity.

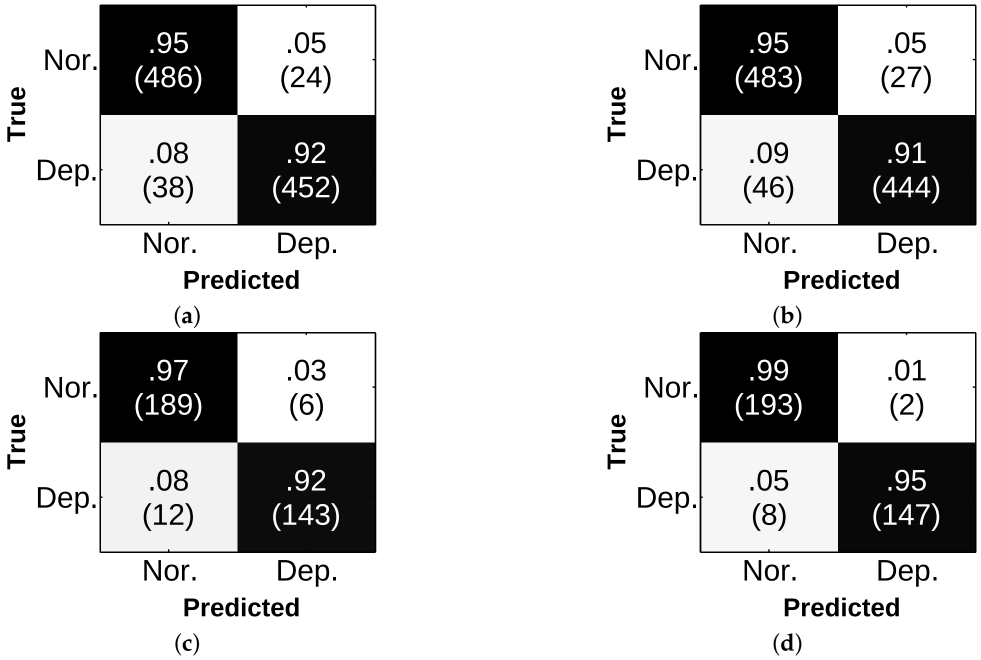


Figure 6. The confusion matrices of the proposed (**a**) TCC-softmax on DAIC-WOZ, (**b**) TCC-kernel on DAIC-WOZ, (**c**) TCC-softmax on MODMA, and (**d**) TCC-kernel on MODMA for the normal (noted as 'Nor.') and depression (noted as 'Dep.') classes.

4.2.2. Experimental Comparisons

Afterwards, in order to prove the validity of the proposed depression detection approach, we present an experimental comparison between the proposed TCC approach (with 'softmax' and 'kernel' setups) and the state-of-the-art models of 2D-CNN-LSTM [84], CNN-bi-directional LSTM (BLSTM) using end-to-end (E2E) multi-task learning (MTL) with self-attention (noted as 'CNN-BLSTM+E2EMTL+SA') [27], the speech-based depression detection approach decision tree (DT) [40], and DepAudioNet [85], with the same parametric setups as in the corresponding works for the compared existing approaches. Note that all of these approaches in the comparison share the same experimental settings and data inputs. The evaluation metrics of precision, recall, and F1-score for the positive class (depression) were considered in the result comparison, with the corresponding experimental results shown in Table 2. Note that the precision and recall metrics can be acquired based on true and predicted positive labels, thus obtaining the F1-score through the harmonic average for the precision and recall. As can be learned from Table 2, the proposed approach exceeds the model of LSTM architectures (see 2D-CNN-LSTM, CNN-BLSTM+E2EMTL+SA, and DepAudioNet) in all of the metrics. For the DAIC-WOZ dataset, the best F1-score of the proposed TCC model is 93.6%, whereas, for the MODMA dataset, the TCC-kernel achieves the best F1-score of 96.7%, which is 2.6% higher than TCC-softmax.

Table 2. The precisions, recalls, and F1-scores (%) of the proposed TCC approach (including 'softmax' and 'kernel' setups) and the state-of-the-art approaches (2D-CNN-LSTM, CNN-BLSTM+E2EMTL+SA, DT, and DepAudioNet) on the DAIC-WOZ and MODMA datasets.

Approaches\Metrics	**Precision**	**Recall**	**F1-Score**
DAIC-WOZ Dataset:			
2D-CNN-LSTM [84]	91.2	92.0	91.6
CNN-BLSTM+E2EMTL+SA [27]	82.2	89.1	85.5
DT [40]	78.7	80.0	79.2
DepAudioNet [85]	82.4	84.0	83.2
TCC-softmax (Proposed)	**95.0**	**92.2**	**93.6**
TCC-kernel (Proposed)	94.3	90.6	92.4
MODMA Dataset:			
2D-CNN-LSTM [84]	92.9	93.5	93.2
CNN-BLSTM+E2EMTL+SA [27]	93.5	90.1	91.8
DT [40]	79.6	81.3	80.4
DepAudioNet [85]	83.9	87.6	85.7
TCC-softmax (Proposed)	96.0	92.3	94.1
TCC-kernel (Proposed)	**98.7**	**94.8**	**96.7**

To make further comparisons, we present the average results of these metrics as shown in Table 3. Note that the average results in Table 3 refer to the average values of the precision, recall (equivalent to UA), and F1-score across the two classes in depression detection. It is proved that the proposed TCC approach is effective for the prediction of depression and that the proposed TCC approach significantly outperforms the existing CNN-BLSTM+E2EMTL+SA, DT, and DepAudioNet approaches (at the significance level of 0.01 using the one-tailed z-test) when using the UA (or UAR) metric. Specifically, the TCC-softmax setup significantly outperforms these approaches at the significance level of 0.001. Therefore, through the experimental results and analysis, it is proved that the proposed TCC approach performs well for depression detection in speech.

Table 3. The (unweighted) average precisions, recalls (UA), and F1-scores (macro) (%) of the proposed TCC approach (including 'softmax' and 'kernel' setups) and the state-of-the-art approaches (2D-CNN-LSTM, CNN-BLSTM+E2EMTL+SA, DT, and DepAudioNet) on the DAIC-WOZ dataset.

Approaches\Metrics (Average)	**Precision**	**Recall (UA)**	**F1-Score (Macro)**
DAIC-WOZ Dataset:			
2D-CNN-LSTM [84]	90.8	92.6	91.7
CNN-BLSTM+E2EMTL+SA [27]	85.3	89.5	87.3
DT [40]	78.9	80.2	79.6
DepAudioNet [85]	82.8	84.9	83.8
TCC-softmax (Proposed)	**93.9**	**93.8**	**93.8**
TCC-kernel (Proposed)	92.8	92.7	92.7

5. Conclusions

This paper proposed a transformer-CNN-CNN (TCC) approach for depression detection in speech. The proposed approach employed a parallel convolutional neural network (CNN) and a transformer using a parallel structure, resulting in three parallel streams to obtain local and temporal information, considering linear attention to reduce the complexity of the model. Then, we evaluated the proposed approach on the DAIC-WOZ and MODMA datasets, proving the effectiveness for the proposed TCC approach. The future works will focus on two aspects. First, it will be possible to explore effective acoustic low-level features in presenting depression in speech. Then, further parallel structures based on the proposed TCC can be investigated in order to improve the detection performance through parallel computing.

Author Contributions: Conceptualization, F.Y. and J.D.; methodology, F.Y. and J.D.; validation, J.D. and X.X.; formal analysis, X.X.; writing—original draft preparation, F.Y., J.D., and X.X.; writing—review and editing, X.X. and L.Z.; visualization, F.Y., J.D., and X.X.; supervision, X.X. and L.Z.; project administration, F.Y. and X.X.; funding acquisition, F.Y., X.X., and L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Qinglan Project for Colleges and Universities in Jiangsu Province, China, the China Postdoctoral Science Foundation under Grant 2022M711693, the Natural Science Foundation of China under Grants No. 61801241, No. 62071242, and No. 62172235, and the Natural Science Foundation of Jiangsu under Grant BK20191381.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable help.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

EEG	ElectroEncephaloGram
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
SER	Speech Emotion Recognition
MFCC	Mel-Frequency Cepstral Coefficients
TCC	Transformer-CNN-CNN
SVM	Support Vector Machine
GMM	Gaussian Mixture Model
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
ASR	Automatic Speech Recognition
ReLU	Rectified Linear Unit
BN	Batch Normalization
ELU	Exponential Linear Unit
DAIC-WOZ	Distress Analysis Interview Corpus-Wizard of OZ
MODMA	Multi-modal Open Dataset for Mental-disorder Analysis
Adam	Adaptive Moment Estimation
UA	Unweighted Accuracy
UAR	Unweighted Average Recall
BLSTM	Bi-directional LSTM
E2E	End-to-End
MTL	Multi-Task Learning
DT	Decision Tree

References

1. Tiller, J.W. Depression and anxiety. *Med. J. Aust.* **2013**, *199*, S28–S31. [CrossRef] [PubMed]
2. Liu, C.H.; Zhang, E.; Wong, G.T.F.; Hyun, S. Factors associated with depression, anxiety, and PTSD symptomatology during the COVID-19 pandemic: Clinical implications for US young adult mental health. *Psychiatry Res.* **2020**, *290*, 113172. [CrossRef] [PubMed]
3. Buason, A.; Norton, E.C.; McNamee, P.; Thordardottir, E.B.; Asgeirsdóttir, T.L. *The Causal Effect of Depression and Anxiety on Life Satisfaction: An Instrumental Variable Approach*; Technical Report; National Bureau of Economic Research: Cambridge, MA, USA, 2021.
4. Hawton, K.; i Comabella, C.C.; Haw, C.; Saunders, K. Risk factors for suicide in individuals with depression: A systematic review. *J. Affect. Disord.* **2013**, *147*, 17–28. [CrossRef] [PubMed]

5. Scherer, S.; Lucas, G.M.; Gratch, J.; Rizzo, A.S.; Morency, L.P. Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews. *IEEE Trans. Affect. Comput.* **2015**, *7*, 59–73.
6. Sharp, R. The Hamilton rating scale for depression. *Occup. Med.* **2015**, *65*, 340. [CrossRef]
7. Kroenke, K.; Spitzer, R.L. The PHQ-9: A new depression diagnostic and severity measure. *Psychiatr. Ann.* **2002**, *32*, 509–515. [CrossRef]
8. Valstar, M.; Gratch, J.; Schuller, B.; Ringeval, F.; Lalanne, D.; Torres Torres, M.; Scherer, S.; Stratou, G.; Cowie, R.; Pantic, M. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In Proceedings of the International Workshop on Audio/Visual Emotion Challenge, Amsterdam, The Netherlands, 16 October 2016; pp. 3–10.
9. Chao, L.; Tao, J.; Yang, M.; Li, Y. Multi task sequence learning for depression scale prediction from video. In Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 526–531.
10. De Melo, W.C.; Granger, E.; Hadid, A. Depression detection based on deep distribution learning. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 4544–4548.
11. Pampouchidou, A.; Pediaditis, M.; Maridaki, A.; Awais, M.; Vazakopoulou, C.M.; Sfakianakis, S.; Tsiknakis, M.; Simos, P.; Marias, K.; Yang, F.; et al. Quantitative comparison of motion history image variants for video-based depression assessment. *EURASIP J. Image Video Process.* **2017**, *2017*, 64. [CrossRef]
12. Sun, B.; Zhang, Y.; He, J.; Yu, L.; Xu, Q.; Li, D.; Wang, Z. A random forest regression method with selected-text feature for depression assessment. In Proceedings of the Annual Workshop on Audio/Visual Emotion Challenge (AVEC), Mountain View, CA, USA, 23–27 October 2017; pp. 61–68.
13. Wolohan, J.; Hiraga, M.; Mukherjee, A.; Sayyed, Z.A.; Millard, M. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP. In Proceedings of the International Workshop on Language Cognition and Computational Models, Santa Fe, NM, USA, 20 August 2018; pp. 11–21.
14. He, L.; Cao, C. Automated depression analysis using convolutional neural networks from speech. *J. Biomed. Inform.* **2018**, *83*, 103–111. [CrossRef]
15. Li, X.; Hu, B.; Sun, S.; Cai, H. EEG-based mild depressive detection using feature selection methods and classifiers. *Comput. Methods Programs Biomed.* **2016**, *136*, 151–161. [CrossRef]
16. Cai, H.; Han, J.; Chen, Y.; Sha, X.; Wang, Z.; Hu, B.; Yang, J.; Feng, L.; Ding, Z.; Chen, Y.; et al. A pervasive approach to EEG-based depression detection. *Complexity* **2018**, *2018*, 5238028. [CrossRef]
17. Pampouchidou, A.; Simantiraki, O.; Vazakopoulou, C.M.; Chatzaki, C.; Pediaditis, M.; Maridaki, A.; Marias, K.; Simos, P.; Yang, F.; Meriaudeau, F.; et al. Facial geometry and speech analysis for depression detection. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju Island, Republic of Korea, 11–15 July 2017; pp. 1433–1436.
18. Yang, L.; Jiang, D.; Xia, X.; Pei, E.; Oveneke, M.C.; Sahli, H. Multimodal measurement of depression using deep learning models. In Proceedings of the Annual Workshop on Audio/Visual Emotion Challenge (AVEC), Mountain View, CA, USA, 23–27 October 2017; pp. 53–59.
19. Rodrigues Makiuchi, M.; Warnita, T.; Uto, K.; Shinoda, K. Multimodal fusion of Bert-CNN and gated CNN representations for depression detection. In Proceedings of the International on Audio/Visual Emotion Challenge and Workshop (AVEC), Nice, France, 21 October 2019; pp. 55–63.
20. Yin, S.; Liang, C.; Ding, H.; Wang, S. A multi-modal hierarchical recurrent neural network for depression detection. In Proceedings of the International on Audio/Visual Emotion Challenge and Workshop, Nice, France, 21 October 2019; pp. 65–71.
21. Williamson, J.R.; Young, D.; Nierenberg, A.A.; Niemi, J.; Helfer, B.S.; Quatieri, T.F. Tracking depression severity from audio and video based on speech articulatory coordination. *Comput. Speech Lang.* **2019**, *55*, 40–56. [CrossRef]
22. Jan, A.; Meng, H.; Gaus, Y.F.B.A.; Zhang, F. Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Trans. Cogn. Dev. Syst.* **2017**, *10*, 668–680. [CrossRef]
23. Cummins, N.; Scherer, S.; Krajewski, J.; Schnieder, S.; Epps, J.; Quatieri, T.F. A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* **2015**, *71*, 10–49. [CrossRef]
24. Zhao, Z.; Bao, Z.; Zhang, Z.; Deng, J.; Cummins, N.; Wang, H.; Tao, J.; Schuller, B. Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders. *IEEE J. Sel. Top. Signal Process.* **2019**, *14*, 423–434. [CrossRef]
25. Zhao, Z.; Li, Q.; Cummins, N.; Liu, B.; Wang, H.; Tao, J.; Schuller, B.W. Hybrid Network Feature Extraction for Depression Assessment from Speech. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Shanghai, China, 25–29 October 2020; pp. 4956–4960.
26. Mao, K.; Zhang, W.; Wang, D.B.; Li, A.; Jiao, R.; Zhu, Y.; Wu, B.; Zheng, T.; Qian, L.; Lyu, W.; et al. Prediction of Depression Severity Based on the Prosodic and Semantic Features with Bidirectional LSTM and Time Distributed CNN. *IEEE Trans. Affect. Comput.* **2022**. [CrossRef]
27. Li, Y.; Zhao, T.; Kawahara, T. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, Austria, 15–19 September 2019; pp. 2803–2807.

28. Wang, Y.; Zhao, X.; Li, Y.; Hu, X.; Huang, K.; CRIPAC, N. Densely Cascaded Shadow Detection Network via Deeply Supervised Parallel Fusion. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, 13–19 July 2018; pp. 1007–1013.
29. Du, X.; El-Khamy, M.; Lee, J.; Davis, L. Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 953–961.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Long Beach, CA, USA, 2017; pp. 5998–6008.
31. Zhao, Y.; Liang, Z.; Du, J.; Zhang, L.; Liu, C.; Zhao, L. Multi-Head Attention-Based Long Short-Term Memory for Depression Detection From Speech. *Front. Neurorobotics* **2021**, *15*, 684037. [CrossRef] [PubMed]
32. Xu, X.; Deng, J.; Zhang, Z.; Fan, X.; Zhao, L.; Devillers, L.; Schuller, B.W. Rethinking auditory affective descriptors through zero-shot emotion recognition in speech. *IEEE Trans. Comput. Soc. Syst.* **2022**, *9*, 1530–1541. [CrossRef]
33. Xu, X.; Deng, J.; Cummins, N.; Zhang, Z.; Zhao, L.; Schuller, B.W. Exploring zero-shot emotion recognition in speech using semantic-embedding prototypes. *IEEE Trans. Multimed.* **2022**, *24*, 2752–2765. [CrossRef]
34. Zhao, Z.; Li, Q.; Zhang, Z.; Cummins, N.; Wang, H.; Tao, J.; Schuller, B.W. Combining a parallel 2D CNN with a self-attention dilated residual network for CTC-based discrete speech emotion recognition. *Neural Netw.* **2021**, *141*, 52–60. [CrossRef]
35. Goldman, L.S.; Nielsen, N.H.; Champion, H.C. Awareness, diagnosis, and treatment of depression. *J. Gen. Intern. Med.* **1999**, *14*, 569–580. [CrossRef]
36. Niu, M.; Tao, J.; Liu, B.; Huang, J.; Lian, Z. Multimodal spatiotemporal representation for automatic depression level detection. *IEEE Trans. Affect. Comput.* **2020**. [CrossRef]
37. Huang, Z.; Epps, J.; Joachim, D.; Sethu, V. Natural language processing methods for acoustic and landmark event-based features in speech-based depression detection. *IEEE J. Sel. Top. Signal Process.* **2019**, *14*, 435–448. [CrossRef]
38. Long, H.; Guo, Z.; Wu, X.; Hu, B.; Liu, Z.; Cai, H. Detecting depression in speech: Comparison and combination between different speech types. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 13–16 November 2017; pp. 1052–1058.
39. Jiang, H.; Hu, B.; Liu, Z.; Wang, G.; Zhang, L.; Li, X.; Kang, H. Detecting depression using an ensemble logistic regression model based on multiple speech features. *Comput. Math. Methods Med.* **2018**, *2018*, 6508319. [CrossRef] [PubMed]
40. Liu, Z.; Wang, D.; Zhang, L.; Hu, B. A Novel Decision Tree for Depression Recognition in Speech. *arXiv* **2020**, arXiv:2002.12759.
41. Dong, Y.; Yang, X. A hierarchical depression detection model based on vocal and emotional cues. *Neurocomputing* **2021**, *441*, 279–290. [CrossRef]
42. Cummins, N.; Baird, A.; Schuller, B.W. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods* **2018**, *151*, 41–54. [CrossRef]
43. Dubagunta, S.P.; Vlasenko, B.; Doss, M.M. Learning voice source related information for depression detection. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6525–6529.
44. Stasak, B.; Joachim, D.; Epps, J. Breaking Age Barriers With Automatic Voice-Based Depression Detection. *IEEE Pervasive Comput.* **2022**, *21*, 10–19. [CrossRef]
45. Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; Tran, D. Image Transformer. In Proceedings of the International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; pp. 4055–4064.
46. Karita, S.; Chen, N.; Hayashi, T.; Hori, T.; Inaguma, H.; Jiang, Z.; Someki, M.; Soplin, N.E.Y.; Yamamoto, R.; Wang, X.; et al. A comparative study on Transformer vs RNN in speech applications. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Guadeloupe, France, 15–18 December 2019; pp. 449–456.
47. Gabeur, V.; Sun, C.; Alahari, K.; Schmid, C. Multi-modal Transformer for video retrieval. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 214–229.
48. Zhou, L.; Fan, X.; Tjahjadi, T.; Das Choudhury, S. Discriminative attention-augmented feature learning for facial expression recognition in the wild. *Neural Comput. Appl.* **2022**, *34*, 925–936. [CrossRef]
49. Lin, C.H.; Yumer, E.; Wang, O.; Shechtman, E.; Lucey, S. ST-GAN: Spatial Transformer generative adversarial networks for image compositing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 9455–9464.
50. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-trained image processing Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 12299–12310.
51. Dong, L.; Xu, S.; Xu, B. Speech-Transformer: A no-recurrence sequence-to-sequence model for speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5884–5888.
52. Wang, X.; Wang, M.; Qi, W.; Su, W.; Wang, X.; Zhou, H. A Novel End-to-End Speech Emotion Recognition Network with Stacked Transformer Layers. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 6–11 June 2021; pp. 6289–6293.

53. Lian, Z.; Liu, B.; Tao, J. CTNet: Conversational Transformer network for emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 985–1000. [CrossRef]
54. Wang, Y.; Shen, G.; Xu, Y.; Li, J.; Zhao, Z. Learning Mutual Correlation in Multimodal Transformer for Speech Emotion Recognition. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Brno, Czech Republic, 30 August–3 September 2021; pp. 4518–4522.
55. Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; Zhang, Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 842–850.
56. Huang, P.Y.; Liu, F.; Shiang, S.R.; Oh, J.; Dyer, C. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, Berlin, Germany, 16–23 August 2016*; ACL: Berlin, Germany, 2016; pp. 639–645.
57. Zhao, B.; Wu, X.; Feng, J.; Peng, Q.; Yan, S. Diversified visual attention networks for fine-grained object classification. *IEEE Trans. Multimed.* **2017**, *19*, 1245–1256. [CrossRef]
58. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2227–2231.
59. Xie, Y.; Liang, R.; Liang, Z.; Huang, C.; Zou, C.; Schuller, B. Speech emotion classification using attention-based LSTM. *IEEE/ACM Trans. Audio, Speech Lang. Process.* **2019**, *27*, 1675–1685. [CrossRef]
60. Jiang, P.; Xu, X.; Tao, H.; Zhao, L.; Zou, C. Convolutional-Recurrent Neural Networks with Multiple Attention Mechanisms for Speech Emotion Recognition. *IEEE Trans. Cogn. Dev. Syst.* **2022**, *14*, 1564–1573. [CrossRef]
61. Wang, J.; Peng, X.; Qiao, Y. Cascade multi-head attention networks for action recognition. *Comput. Vis. Image Underst.* **2020**, *192*, 102898. [CrossRef]
62. Tao, C.; Gao, S.; Shang, M.; Wu, W.; Zhao, D.; Yan, R. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, 13–19 July 2018; pp. 4418–4424.
63. Nediyanchath, A.; Paramasivam, P.; Yenigalla, P. Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–9 May 2020; pp. 7179–7183.
64. Chen, H.; Jiang, D.; Sahli, H. Transformer Encoder with Multi-modal Multi-head Attention for Continuous Affect Recognition. *IEEE Trans. Multimed.* **2020**, *23*, 4171–4183. [CrossRef]
65. Katharopoulos, A.; Vyas, A.; Pappas, N.; Fleuret, F. Transformers are RNNs: Fast autoregressive Transformers with linear attention. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 13–18 July 2020; pp. 5156–5165.
66. Luo, S.; Li, S.; Cai, T.; He, D.; Peng, D.; Zheng, S.; Ke, G.; Wang, L.; Liu, T.Y. Stable, fast and accurate: Kernelized attention with relative positional encoding. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 22795–22807.
67. Shen, Z.; Zhang, M.; Zhao, H.; Yi, S.; Li, H. Efficient attention: Attention with linear complexities. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3-8 January 2021; pp. 3531–3539.
68. Tsai, Y.H.H.; Bai, S.; Yamada, M.; Morency, L.P.; Salakhutdinov, R. Transformer Dissection: An Unified Understanding for Transformer's Attention via the Lens of Kernel. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Hong Kong, China, 3–7 November 2019; ACL: Hong Kong, China, 2019.
69. Jati, A.; Georgiou, P. Neural predictive coding using convolutional neural networks toward unsupervised learning of speaker characteristics. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **2019**, *27*, 1577–1589. [CrossRef]
70. Rejaibi, E.; Komaty, A.; Meriaudeau, F.; Agrebi, S.; Othmani, A. MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomed. Signal Process. Control* **2022**, *71*, 103107. [CrossRef]
71. Zhang, P.; Wu, M.; Dinkel, H.; Yu, K. DEPA: Self-supervised audio embedding for depression detection. In Proceedings of the ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 135–143.
72. Khan, A.; Sohail, A.; Zahoora, U.; Qureshi, A.S. A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* **2020**, *53*, 5455–5516. [CrossRef]
73. Alippi, C.; Disabato, S.; Roveri, M. Moving convolutional neural networks to embedded systems: The AlexNet and VGG-16 case. In Proceedings of the ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), Porto, Portugal, 11–13 April 2018; pp. 212–223.
74. Zhou, D.; Kang, B.; Jin, X.; Yang, L.; Lian, X.; Jiang, Z.; Hou, Q.; Feng, J. DeepViT: Towards deeper vision Transformer. *arXiv* **2021**, arXiv:2103.11886.
75. Liu, Y.; Zhang, J.; Fang, L.; Jiang, Q.; Zhou, B. Multimodal motion prediction with stacked Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7577–7586.
76. Liu, L.; Liu, J.; Han, J. Multi-head or single-head? An empirical comparison for Transformer training. *arXiv* **2021**, arXiv:2106.09650.
77. Song, K.; Jung, Y.; Kim, D.; Moon, I.C. Implicit kernel attention. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 9713–9721.

78. Gratch, J.; Artstein, R.; Lucas, G.; Stratou, G.; Scherer, S.; Nazarian, A.; Wood, R.; Boberg, J.; DeVault, D.; Marsella, S.; et al. The distress analysis interview corpus of human and computer interviews. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland, 26–31 May 2014; ELRA: Reykjavik, Iceland, 2014; pp. 3123–3128.
79. Cai, H.; Gao, Y.; Sun, S.; Li, N.; Tian, F.; Xiao, H.; Li, J.; Yang, Z.; Li, X.; Zhao, Q.; et al. MODMA dataset: A Multi-modal Open Dataset for Mental-disorder Analysis. *arXiv* **2020**, arXiv:2002.09283.
80. Liu, Z.; Li, C.; Gao, X.; Wang, G.; Yang, J. Ensemble-based depression detection in speech. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 13–16 November 2017; pp. 975–980.
81. Lopez-Otero, P.; Docio-Fernandez, L. Analysis of gender and identity issues in depression detection on de-identified speech. *Comput. Speech Lang.* **2021**, *65*, 101118. [CrossRef]
82. Fahad, M.S.; Deepak, A.; Pradhan, G.; Yadav, J. DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features. *Circuits Syst. Signal Process.* **2021**, *40*, 466–489. [CrossRef]
83. Xu, X.; Deng, J.; Coutinho, E.; Wu, C.; Zhao, L.; Schuller, B. Connecting Subspace Learning and Extreme Learning Machine in Speech Emotion Recognition. *IEEE Trans. Multimed.* **2019**, *21*, 795–808. [CrossRef]
84. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323.
85. Ma, X.; Yang, H.; Chen, Q.; Huang, D.; Wang, Y. DepAudioNet: An efficient deep model for audio based depression classification. In Proceedings of the International Workshop on Audio/Visual Emotion Challenge (AVEC), Amsterdam, The Netherlands, 16 October 2016; ACM: Amsterdam, The Netherlands, 2016; pp. 35–42.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.