
Su Zhang¹, Ziyuan Zhao¹,², Cuntai Guan¹,*
¹Nanyang Technological University
²Institute of Infocomm Research (I²R), A*STAR, Singapore
sorazcn@gmail.com, S210088@e.ntu.edu.sg, ctguan@ntu.edu.sg

Abstract

We used two multimodal models for continuous valence-arousal recognition using visual, audio, and linguistic information. The first model is the same as we used in ABAW2 and ABAW3, which employs the leader-follower attention. The second model has the same architecture for spatial and temporal encoding. As for the fusion block, it employs a compact and straightforward channel attention, borrowed from the End2You toolkit. Unlike our previous attempts that use Vggish feature directly as the audio feature, this time we feed the pre-trained VGG model using logmel-spectrogram and finetune it during the training. To make full use of the data and alleviate over-fitting, cross-validation is carried out. The code is available at https://github.com/sucv/ABAW3.

1. Introduction

Emotion recognition is the process of identifying human emotion. It plays a crucial role in behavioral modeling, human-computer interaction, and affective computing. By using the dimensional model [18], any emotional state can be taken as a point located in a continuous space, with the valence and arousal being the axes. Continuous emotion recognition seeks to map the $N$ sequential data points into $M$ sequential emotional state points, where $M$ usually equals $N$. The term “continuous” is endowed with two distinctive features. Spatially, it seeks to represent the emotional state as a point with continuous numerical values in the multi-dimensional space of the dimensional theory, as opposed to assigning categorical labels. Temporally, it incessantly forecasts the emotional state for a predetermined duration, forming a continuous record of the subject’s emotional trajectory over a specified period.

The task of Continuous Emotion Recognition (CER) is challenging, mainly due to the following reasons. Firstly, the complexity of emotions over an extended period can be seen as a composition of multiple facial muscular movements, which can be defined as one-actions [5] using the Facial Action Coding System (FACS) [19]. However, despite its atomic nature, human emotions, irrespective of the modality, typically exhibit large variations in their intensity, order, and duration, and take longer to unfold. Consequently, there is a need for models that can learn long-range temporal dependencies to address this issue. Secondly, emotions are highly subjective and vary significantly depending on individual experiences, making their perception susceptible to personal bias. For instance, children who have been subjected to physical abuse can detect signs of anger much faster than their counterparts [15]. Therefore, the data collected from subjects and the ground truth provided by annotators are prone to individual biases. One promising technique to mitigate this issue is multimodality, which involves the integration of visual, audio, and physiological data to develop reliable CER models.

In order to address the challenges posed by temporal dynamic learning and cross-subject generality, we adopt two strategies: large window resampling and multimodal fusion. Large window resampling involves selecting a window size that determines the amount of contextual information the model considers to make a prediction for each time step. On the other hand, multimodal fusion is a widely accepted technique that leverages the complementary nature of multiple data modalities to improve prediction accuracy. This is because multimodal data can disambiguate conflicting emotional cues, such as a crying face with joyful vocal expressions, or a neutral face with a harsh intonation. Additionally, multimodal data can enhance recognition robustness in real-world scenarios characterized by uncontrolled environments, subject diversity, various illumination conditions, spontaneous behaviors, background noise, and unclear speech. For instance, when an actor’s face is not visible, their voice or body gesture can serve as an emotional cue to support ongoing learning.

This report details our methodology for the valence-arousal estimation challenge from the fifth affective behav-

---

*This work is partially supported by the RIE2020 AME Programmatic Fund, Singapore (No. A20G860102).
Figure 1. The architecture of our proposed model. The model consists of four components, i.e., the visual, audio, linguistic, and co-attention blocks. The visual block has a cascaded 2DCNN-TCN structure, and the audio and linguistic blocks each contain a TCN. The three branches yield three independent spatial-temporal feature vectors. They are then fed to the attentive fusion block. Three independent attention encoders are used. For the $i$-th branch, its encoder consists of three independent linear layers, they adjust the dimension of the feature vector producing a query $Q_i$, a key $K_i$, and a value $V_i$. They are then regrouped and concatenated to form the cross-modal counterparts.

ior analysis in-the-wild (ABAW5) workshop [6–9, 11–14, 25]. ABAW5 aims for the affective behavior analysis in-the-wild. The valence-arousal estimation sub-challenge bases on the extended version of the Aff-Wild2 database, in which the valence and arousal are annotated continuously for each video frame. Our work is an extension of the last year’s attempt [26] on ABAW5 [14]. The extension are twofold. First, a pre-trained audio VGGnet is employed to serve as the audio backbone for the vggish [4] extraction on-the-go. Thus, the expressiveness of deep feature on visual and audio modalities can be further improved through finetuning. Second, a compact and straightforward fusion block is borrowed from the End2You toolkit, in which the channel attention is employed to fuse the multimodal information. The results of using leader-follower attention and channel attention are reported.

The remainder of the paper is arranged as follows. Section 2 details the model architecture including the visual, aural, linguistic, and attentive fusion blocks. Section 3 elaborates the implementation details including the data preprocessing, training settings, and post-processing. Section 4 provides the continuous emotion recognition results. Section 5 concludes the work.

2. Model Architecture

There are two architectures for our model, the leader-follower attention network (LFAN) and channel attention network (CAN). Both of them consists of independent branches to extract the spatiotemporal encoding for each modality, followed by the fusion block using leader-follower attention and channel attention, respectively. For each branch, a CNN backbone is employed to extract the spatial deep feature on-the-go for video frames and logmel-spectrogram, followed by a temporal convolutional network (TCN) [1] to further learn the spatiotemporal encoding. For the linguistic branch, the pre-extracted bert features [17] are taken as the input to feed the TCN. For branches fed by low-level features, no backbone is employed as well. The illustration of LFAN using video frames, logmel-spectrograms, and speech tokens are shown in Fig 1. As for the details of the channel attention-based fusion block, please refer to the End2You paper [22] and the code at https://github.com/end2you/end2you.

3. Implementation Details

3.1. Database

The ABAW5 competition uses the Aff-Wild2 database. The corpora of the valence-arousal estimation sub-challenge includes 564 trials. The database is split into the training, validation and test sets. The partitioning is done in a subject independent manner so that any subject’s data are included in only one partition. The partitioning produces 341, 71, and 152 trials for the training, validation, and test sets. Four experts annotate the videos using the method proposed in [2]. In addition to the annotations (for the training and validation sets only) and the raw videos, the bounding boxes and landmarks for each raw video are also available.
The audio preprocessing firstly converts all the videos to mono with a 16K sampling rate in wav format. The logmel-spectrogram are then extracted using the preprocessing code from the Vggish repository. The only change is that we specified the hop length to be 1/fps of the raw video, in order to synchronize with other modalities and annotations.

The linguistic preprocessing is carried out as follows. The mono wav file obtained from the audio preprocessing is fed to a pretrained speech recognition model from the Vosk toolkit, from which the recognized words and the word-level timestamp are obtained. The recognized words are then fed to a pretrained punctuation and capitalization model from the deepmultilingualpunctuation toolkit. After which a pretrained BERT model from the Pytorch library is employed to extract the word-level linguistic features. The linguistic features are obtained by summing together the last four layers of the BERT model [21]. To synchronize, the word-level linguistic features are populated according to the timestamp of each word and each frame. Specifically, a word usually has a larger time span than that for a frame. Therefore, for one word, its feature is repetitively assigned to the time steps of all the frames within the time span.

For the valence-arousal labels, all the rows containing −5 are excluded. To ensure that the features and annotations have the same length, the feature matrices are either repeatedly padded (using the last feature points) or trimmed, depending on whether the feature length is shorter and longer than the trial length, respectively.

The CCC loss is used as the loss function. The Adam optimizer with a weight decay of 0.001 is employed. The maximal epoch number and early stopping counter are set to 100 and 20, respectively. The learning rate (LR) and min-
Table 1. The CCC results from the 6-fold cross-validation on the validation set. Fold 0 is exactly the original data partitioning provided by ABAW5.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Method</th>
<th>Fold 0</th>
<th>Fold 1</th>
<th>Fold 2</th>
<th>Fold 3</th>
<th>Fold 4</th>
<th>Fold 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valence</td>
<td>LFAN</td>
<td>0.441</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>CAN</td>
<td>0.423</td>
<td>0.612</td>
<td>0.484</td>
<td>0.529</td>
<td>0.543</td>
<td>0.488</td>
</tr>
<tr>
<td>Arousal</td>
<td>LFAN</td>
<td>0.645</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>CAN</td>
<td>0.670</td>
<td>0.680</td>
<td>0.592</td>
<td>0.618</td>
<td>0.655</td>
<td>0.610</td>
</tr>
</tbody>
</table>

The warmup scheme with ReduceLROnPlateau scheduler with a patience of 5 and factor of 0.1 is employed based on the validation CCC. Three groups of layers for the visual ResNet50 backbone and audio VGG backbone are manually selected for further fine-tuning. When epoch = 0, the first group is unfrozen. The learning rate is then linearly warmed up to $1e - 5$ within an epoch. The repetitive warm-up is carried out until epoch = 5. After which the ReduceLROnPlateau takes over to update the learning rate. It gradually drops the learning rate in a factor of 0.1 should no higher validation CCC appears for a consecutive 5 epochs. When the learning rate is already $1e - 8$ and no higher validation CCC appears for the latest 5 epochs, the second group is unfrozen, the learning rate is also reset to $1e - 5$ before undergoing the warm-up scheme and ReduceLROnPlateau again. The procedure is repeated until no groups are available to unfreeze. Note that at the end of each epoch, the current best model state dictionary (i.e., the one with the greatest validation CCC) is loaded. By doing so, when there is no improvement on validation CCC, the training CCC cannot keep increasing and ends up with over-fitting.

Note that unlike our previous attempts on ABAW2 and ABAW3, no post-processing is employed. Given the results from 6 folds, we simply choose the ones with the highest validation CCC to submit. Also note that the results for any fold are selected from 15 instances using different seeds.

4. Result and Discussion

The validation results of our methods against the baseline are reported in Table 1. Note that for one emotion type, the reported result for each fold is the best instance selected from 15 instances using different seeds. Finally, the best results of Fold 0 from CAN and LFAN, together with other 3 best folds (measured in the average CCC) from CAN, are submitted to ABAW5.

In this study, two multimodal models are compared, namely CAN and LFAN, with the former utilizing only video frames and logmel-spectrograms while the latter incorporates an additional Bert feature. Validation results indicate that the CAN model with two modalities outperforms the LFAN model with three modalities, achieving a higher CCC score in Fold 0. It is worth noting that attempts to include the Bert feature as a third modality in the CAN model did not result in improvement. This suggests that the current linguistic information encoded by the pre-trained Bert model does not provide useful information to the model, or at least not in its current form. The issue lies in the fact that the length of tokens is not equivalent to the length of the sampling window, rendering fine-tuning impossible during training. Future research is needed to explore how linguistic information can be effectively incorporated into the fusion process. Furthermore, the study highlights the efficacy of fine-tuning. Both ResNet50 and VGG16 models, trained using large public datasets and fed by video frames and logmel-spectrograms, show improvements in performance when the last few layers are fine-tuned during training. This improvement can be as high as 10% or more, compared to models that use pre-extracted CNN and Vggish features, albeit at a cost of 300% — 500% longer training time.

The best results from all the teams are reported in Table 2.

Table 2. The overall test results in CCC. The bold fonts indicate the best results. Citations for several teams are not available from Google Scholar by the time when this report was drafted.
5. Conclusion

In our study, we aimed to develop effective models for continuous valence-arousal recognition, utilizing multimodal information from visual, audio, and linguistic sources. To overcome challenges posed by temporal dynamic learning and cross-subject generality, we employed two strategies: large window resampling and multimodal fusion. Specifically, we designed the CAN and LFAN models, with a resampling window length and hop length of 300 and 200, respectively, and multimodal fusion block to effectively integrate information from different modalities. Moving forward, our research will explore the incorporation of transformer-based networks for improved temporal dynamics learning.

References


