

Temporal Enhancement for Video Affective Content Analysis (Supplementary Material)

Anonymous Authors

In the supplementary materials, we first provide a detailed introduction to the datasets used in the experiments, and then list the hyperparameter settings of the experiments on each dataset. Finally, we further analyze the training process of the temporal enhancement method and the role of cross-modal temporal enhancement through visualization.

1 EXPERIMENTS

1.1 Datasets

The YF-E6 dataset is constructed by searching for six basic emotions as keywords on social video-sharing platforms like YouTube and Flickr. Initially, 3,000 videos are collected. After annotation and filtering processes conducted by annotators, a final set of 1,637 videos covering the six emotion categories is established. Specifically, the six basic emotion categories include anger, disgust, fear, joy, sadness, and surprise. The average duration of all videos is 112 seconds. In accordance with standard division protocols, we utilize 819 videos from the dataset for training, while the remaining 818 videos are assigned to the test set.

The VideoEmotion-8 dataset is constructed by searching 24 sub-category variant keywords representing 8 emotion categories across YouTube and Flickr websites, resulting in the retrieval of 7699 videos. Following meticulous filtering and annotation by annotators, 1101 videos are deemed suitable. These videos are then categorized into 8 distinct emotion categories. Each category contains a minimum of 100 videos, with an average duration of 107 seconds per video. Following the common experimental setup, the experiment is conducted in 10 runs in total. In each run, the dataset is randomly partitioned into training and test sets at a ratio of 2:1. The final result is determined by averaging the results obtained from the 10 experimental runs.

The LIRIS-ACCEDE dataset comprises 9,800 video clips extracted from 160 movies. Each clip lasts between 8 and 12 seconds. The MediaEval2016 task is based on this dataset and supplements with an additional 1,200 video clips, resulting in a total of 11,000 clips. These clips are annotated with continuous values using the 2D valence-arousal emotion model, ranging from 1 to 10. Following the standard division, we utilize 9,800 video clips as the training set, with the remaining 1,200 clips designated as the test set.

The VAD dataset consists of popular videos from the Chinese Bilibili website. Initially, a total of 7143 videos were downloaded. After subsequent rounds of filtering, segmenting, and evaluation by annotators, a total of 19,267 video clips from 3343 videos are retained. These video clips are labeled with five attributes: valence, arousal, primary emotion, valence comparison, and arousal comparison. Since our method focuses solely on emotion prediction, only the experiments of the first three labels are conducted. Following the experimental settings provided in the dataset, each label undergoes 5 rounds of testing. In each round, the dataset was randomly divided into training, validation, and test sets in a ratio of 7:1:2.

Table 1: Hyperparameter settings on YF-E6, VideoEmotion-8, and MediaEval2016 datasets. CMTE stands for the cross-modal temporal enhancement module. The following tables are the same.

hyperparameters	YF-E6	VideoEmotion-8	MediaEval2016
batch size	128	128	256
learning rate	0.0002	0.0002	0.0002
weight decay	0	0.0005	0.0005
optimizer	Adam	Adam	Adam
epochs	200	200	20
temporal length T	24	24	16
feature dimensions d	512	512	512
number of CMTE layer	1	2	2
dropout in classifier	0.5	0.5	0.5

Table 2: Hyperparameter settings for experiments on each label of the VAD dataset.

hyperparameters	VAD		
	Valence	Arousal	Primary Emotion
batch size	128	128	128
learning rate	0.0001	0.0001	0.0001
EMA weight decay	0.99	0.9	0.99
optimizer	Adam	Adam	Adam
epochs	20	20	20
danmu temporal length	23	23	23
visual temporal length	72	72	72
audio temporal length	48	48	48
feature dimensions d	1024	1024	1024
number of CMTE layer	1	2	2
dropout in classifier	0.25	0.25	0.25

The average result of five rounds is calculated to determine the performance of the model.

1.2 Implementation Details

For the YF-E6, VideoEmotion-8, and MediaEval2016 datasets, the feature sampling process employs a strategy of random sampling and sorting the features in temporal order during training. To ensure maximal retention of the original video’s modal information and reproducibility of experimental results, a sampling strategy is implemented at equal intervals during testing. The hyperparameter settings for the experiments on the aforementioned three datasets are presented in Table 1. Regarding the VAD dataset, its modal features are non-aligned. Following the previous method, the complete features of the video are utilized, thereby bypassing the feature sampling. Since the division of the VAD dataset is video-independent,

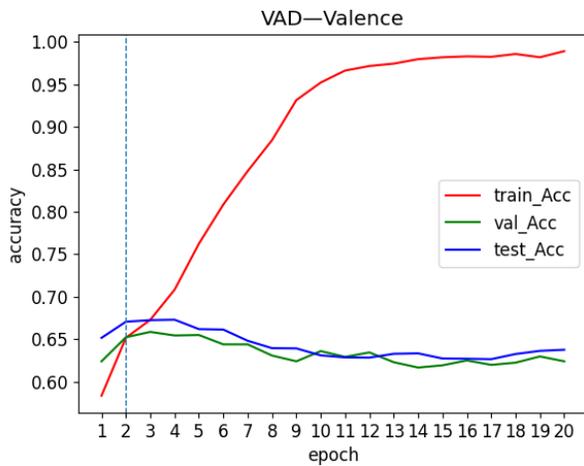
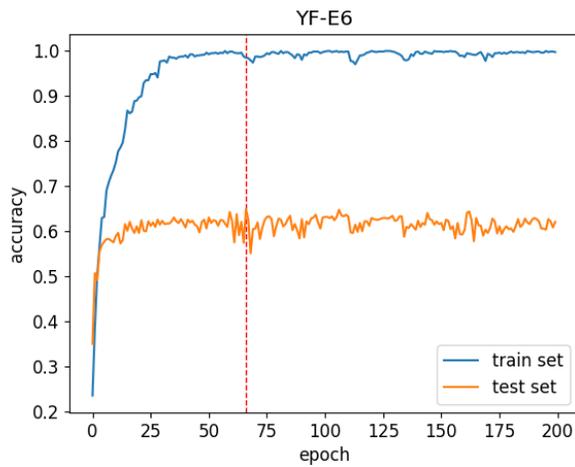


Figure 1: Training process on YF-E6 dataset and the Valence label of VAD dataset.

it means video clips from the same video do not appear simultaneously in the training, validation, or test sets. The Exponential Moving Average (EMA) strategy is applied during testing to integrate the model simply, enhancing its generalization performance and preventing overfitting to the training set. The hyperparameter settings for experiments on each label of the VAD dataset are provided in Table 2.

1.3 Further Analysis

The main text provides a quantitative analysis of the method based on temporal enhancement, while the supplementary material examines the method from a visualization perspective. Fig 1 illustrates the training process of the model on the YF-E6 dataset and the Valence label of the VAD dataset. First, for the graph above, the accuracy of the model on both the training and test sets gradually increases with the number of training epochs. Around the 25th epoch, the accuracy of the training set essentially reaches 100%. Afterward, the accuracy of both training and test sets fluctuates

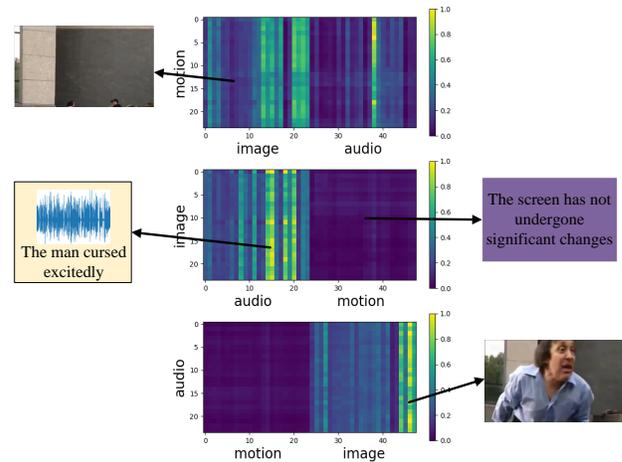


Figure 2: Visualization of attention weights for cross-modal temporal enhancement module

within a small range. This represents an ideal training curve, indicating that the model can train stably and effectively learn the input data's potential patterns during continuous training. Around 60 to 70 epochs, as indicated by the red dotted line, the test set accuracy reaches its maximum value. At this point, the training set accuracy exhibits a slight downward trend with brief fluctuations. This is attributed to temporal feature sampling, where sampled features may vary in validity across epochs, contributing to model robustness. It is evident that the temporal length T is a key parameter. Second, for the graph below, because the dataset division is independent of videos, it leads to diverse video styles across training, validation, and test sets, and poses a challenge. The model continues to fit the training data, with accuracy improving and stabilizing. However, the accuracy of the test and validation sets initially increases but later decreases. It indicates that the model struggles to generalize beyond the training set, potentially due to the absence of temporal enhancement operations and the lack of utilization of powerful semantic encoders like CLIP. A simple motion encoder cannot bridge the semantic gap between the training and test sets.

In order to confirm that the cross-modal temporal enhancement module can indeed integrate all modal information in interactions, and emphasize the most relevant temporal fragments while suppressing other irrelevant temporal fragments, we visualize the attention weights in three different cross-modal temporal enhancement modules within the same layer. As shown in Fig 2, the modality identified by the column of each graph acts as the Query in the cross-modal attention, and the modality identified by the row acts as the Key. The weights are normalized within the range $[0,1]$. For the weight graph of the first row, a specific temporal fragment in the audio modality is emphasized among all temporal fragments across all modalities. However, the image modality includes numerous important fragments that are not the most prominent. By examining the corresponding image from the original video at this timestamp, we intercept and place it in the upper left corner. It becomes apparent that the image primarily consists of background information, akin to noise. Hence, its weight normalization close to

233 0 is expected. For the weight graphs of the second and third rows,
234 all temporal fragments within the motion modality exhibit lower
235 weights. Likewise, by identifying the original signals correspond-
236 ing to important temporal fragments and suppressed fragments in
237 the video, the arrows in the figure indicate the text descriptions
238

239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290

or images corresponding to these original signals. As observed by
humans, temporal fragments within the suppressed motion fea-
ture do not offer clues that facilitate emotion recognition, whereas
highlighted fragments do have rich emotional information.

291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348