Pattern Recognition Letters
journal homepage: www.elsevier.com

Supplementary Material for Cross-modal Context-gated Convolution for Multi-modal Sentiment Analysis

Huanglu Wen[a], Shaodi You[b], Ying Fu[a,**]

[a] *Beijing Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China*
[b] *Computer Vision Research Group in the Institute of Informatics, University of Amsterdam, Amsterdam, Netherlands*

Article history:

2020 MSC: 68T07, 68T10, 97C20, 82C32

Keywords: Artificial Neural Networks, Pattern recognition, Affective behavior, Multi-modal temporal sequences

ABSTRACT

When inferring sentiments, using verbal clues only is problematic because of the ambiguity. Adding related vocal and visual contexts as complements for verbal clues can be helpful. To infer sentiments from multi-modal temporal sequences, we need to identify both sentiment-related clues and their cross-modal interactions. However, sentiment-related behaviors of different modalities may not happen at the same time. These behaviors and their interactions are also sparse in time, making it hard to infer the correct sentiments. Besides, unaligned sequences from sensors also have varying sampling rates, amplifying the misalignment and sparsity mentioned above. While most previous multi-modal sentiment analysis works only focus on word-aligned sequences, we propose cross-modal context-gated convolution for unaligned sequences. Cross-modal context-gated convolution models the cross-modal interaction locally, dealing with the misalignment while reducing the effect of unrelated information. Cross-modal context-gated convolution introduces the concept of cross-modal context gate, enabling it to catch useful cross-modal interactions more effectively. Cross-modal context-gated convolution also brings more possibilities to the layer design for multi-modal sequential modeling. Experiments on multi-modal sentiment analysis datasets under both word-aligned and unaligned conditions show the validity of our approach.

1. Dataset Introduction and Feature Pre-processing

The CMU-MOSI dataset consists of sentiment utterances extracted from opinion videos. Each utterance has real-valued sentiment intensity annotations from [-3,+3] where positive values indicate positive sentiments and vice versa. The CMU-MOSEI dataset consists of sentiment utterances taken from monologue videos on YouTube. Each utterance is labelled with sentiment intensity like CMU-MOSI as well as 6 emotions (happiness, sadness, anger, fear, disgust and surprise). Here we only use the sentiment intensity labels.

We use the processed data from the authors of MulT[1] (1). Here we briefly describe how they process the raw data.

The unaligned acoustic and visual features in CMU-MOSI are extracted at the sampling rate of 12.5 and 15 Hz. For CMU-MOSEI, the sampling rates of unaligned acoustic and visual features are 20 and 15 Hz. Besides, the text features are having a varying sampling rate in both datasets.

After the unaligned raw feature extraction, GloVe (2) are used to extract the word embeddings out of the texts, resulting in 300 dimensional embedding vectors.

COVAREP (3) are used to extract low-level acoustic features from voices, including 12 Mel-frequency cepstral coefficients (MFCCs), pitch tracking and voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters and maxima dispersion quotients. The total dimension of the acoustic feature is 74.

[**] Corresponding author: Ying Fu (fuying@bit.edu.cn)

[1] http://immortal.multicomp.cs.cmu.edu/raw_datasets/processed_data/

Table 1. The training settings for our model. CCC, CIL, SA, SAL and TC are the abbreviation of cross-modal context-gated convolution, cross-modal interaction layer, self-attention, self-attention layers and Temporal Convolution separately. - means not available.

Dataset	CMU-MOSI		CMU-MOSEI	
Setting	Word-aligned	Unaligned	Word-aligned	Unaligned
Optimizer	Adam	Adam	Adam	Adam
Batch Size	128	128	16	16
Learning Schedule	Cosine Annealing	Cosine Annealing	Constant	Constant
Initial Learning Rate	6e-3	6e-3	1e-3	1e-3
End Learning Rate	0	0	-	-
# Epochs for Early Stop	10	10	10	10
CCC Kernel Size	3	3	15	5
CCC&SA Width	24	40	40	40
# of CCC&SA Heads	2	10	10	10
# of CILs&SALs	4	4	4	4
TC Kernel Size (L/V/A)	1/3/3	1/3/3	1/3/3	1/3/3
Textual Embedding Dropout	0.19	0.29	0.2	0.2
CCC Dropout (Language)	0.16	0.08	0.2	0.2
CCC Dropout (Acoustic)	0.06	0.30	0.2	0.2
CCC Dropout (Visual)	0.12	0.28	0.2	0.2
SA Dropout	0.29	0.23	0.2	0.2
Output Dropout	0.22	0.13	0.1	0.1
Gradient Clip	0.8	0.8	1	1
# of Epochs	50	50	20	20

Table 2. The search ranges of hyper-parameters on different settings. - means not searched

Dataset	CMU-MOSI		CMU-MOSEI	
Setting	Word-aligned	Unaligned	Word-aligned	Unaligned
Learning Rate	[1e-3,1e-2]	[1e-3,1e-2]	-	-
(6x) Dropouts	[0,0.3]	[0,0.3]	-	-
CCC Kernel SIze	[1,49]	[1,49]	[1,49]	[1,49]
CCC&SA Width	[1,100]	-	-	-
# of CCC&SA Heads	[1,10]	-	-	-

Facet[2] are used to extract 35 facial action unit activities from videos. The unaligned multi-modal sequences compose of these extracted features.

To get the word-aligned sequences, P2FA (4) are used to extract the start and stop timestamps w.r.t. words and calculate the average of acoustic and visual features based on the timestamps.

In addition, for both unaligned and word-aligned sequences, to ensure each sample has the same sequence length, we truncate or left-pad these sequences, since each utterance may have different length.

Note that for features on CMU-MOSI, the authors of MulT may perform some kind of feature selection but there is no explanation of how it is being done.

We check the processed data and find that there is (only) one unnormalized acoustic features with very large mean (about 100) and variance, which different from any other features in CMU-MOSEI dataset, and we perform min-max normalization for that feature so that the value lies in [0,1], which is noted as

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{1}$$

where x is the original feature and z is the normalized feature.

2. Training Settings

Here we describe how we train the models in detail.

First, the data we used is described above.

Second, for reproducing the results in Multi-modal Transformer (MulT) (1), we use the code provided by the authors[3] and the hyper-parameters in the original paper. We slightly modify the code for it to run properly on the latest PyTorch (v 1.5.0).

Third, the training settings for our model is listed in Table 1. Note that the CCC Kernel Size here is the actual kernel size on the word-aligned condition but this is not the same for the unaligned condition. On the unaligned condition, because we

[2] https://imotions.com/

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

want to keep that every convolution layer can access to every element in the input sequence, we add a base kernel size offset to this kernel size value to get the final kernel size. For example, if there is a cross-modal context-gated convolution with sequences from source modality $X_S \in \mathbb{R}^{t_S \times d_S}$ and target modality $X_T \in \mathbb{R}^{t_T \times d_T}$, the kernel size offset is calculated as

$$offset = \left\lfloor \frac{t_S}{t_T} \right\rfloor / 2 * 2, \tag{2}$$

where / means integer division.

Fourth, we perform Bayesian Optimization on the evaluation set to search the best hyper-parameters using Ray tune (5) and Ax[4]. The search space is listed in Table 2. Due to the restriction of the computational power, for every search task, we search about 400 hyper-parameter combinations. On the CMU-MOSI dataset, we additionally train the model with the best 16 combinations and report the best result on the test set. We do not perform the above operation on the CMU-MOSEI dataset since 400 is larger than the size of the whole search space on CMU-MOSEI dataset.

References

[1] Y. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the Conference of the Association for Computational Linguistics, ACL 2019, pp. 6558–6569. doi:10.18653/v1/p19-1656.
URL https://doi.org/10.18653/v1/p19-1656

[2] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, pp. 1532–1543. doi:10.3115/v1/d14-1162.
URL https://doi.org/10.3115/v1/d14-1162

[3] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, COVAREP - A collaborative voice analysis repository for speech technologies, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, pp. 960–964. doi:10.1109/ICASSP.2014.6853739.
URL https://doi.org/10.1109/ICASSP.2014.6853739

[4] J. Yuan, M. Liberman, Speaker identification on the scotus corpus, The Journal of the Acoustical Society of America 123 (5) (2008) 3878–3878.

[5] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, I. Stoica, Tune: A research platform for distributed model selection and training, arXiv preprint arXiv:1807.05118 (2018).

[4] https://github.com/facebook/Ax