Fuchun Sun · Dewen Hu ·
Stefan Wermter · Lei Yang ·
Huaping Liu · Bin Fang (Eds.)

Communications in Computer and Information Science 1515

Cognitive Systems and Information Processing

6th International Conference, ICCSIP 2021
Suzhou, China, November 20–21, 2021
Revised Selected Papers

Springer

Communications in Computer and Information Science 1515

More information about this series at https://link.springer.com/bookseries/7899

Fuchun Sun · Dewen Hu · Stefan Wermter ·
Lei Yang · Huaping Liu · Bin Fang (Eds.)

Cognitive Systems and Information Processing

6th International Conference, ICCSIP 2021
Suzhou, China, November 20–21, 2021
Revised Selected Papers

Springer

Editors
Fuchun Sun
Tsinghua University
Beijing, China

Stefan Wermter
Universität Hamburg
Hamburg, Germany

Huaping Liu
Tsinghua University
Beijing, China

Dewen Hu
National University of Defense Technology
Changsha, China

Lei Yang
Tsingzhan Artificial Intelligence Research Institute
Nanjing, China

Bin Fang
Tsinghua University
Beijing, China

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-981-16-9246-8 ISBN 978-981-16-9247-5 (eBook)
https://doi.org/10.1007/978-981-16-9247-5

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

This volume contains the papers from the Sixth International Conference on Cognitive Systems and Information Processing (ICCSIP 2021), which was held in Suzhou, China, during November 20–21, 2021. ICCSIP is a prestigious biennial conference with past events held in Beijing (2012, 2014, 2016, 2018) and Zhuhai (2020). Over the past few years, ICCSIP has matured into a well-established series of international conferences on cognitive information processing and related fields. Similar to the previous event, ICCSIP 2021 provided an academic forum for the participants to share their new research findings and discuss emerging areas of research. It also established a stimulating environment for the participants to exchange ideas on the future trends and opportunities of cognitive information processing research.

Currently, cognitive systems and information processing are applied in an increasing number of research domains such as cognitive sciences and technology, visual cognition and computation, big data and intelligent information processing, bioinformatics, and applications. We believe that cognitive systems and information processing will certainly exhibit greater-than-ever advances in the near future. With the aim of promoting research and technical innovation in relevant fields, domestically and internationally, the fundamental objective of ICCSIP is defined as providing a premier forum for researchers and practitioners from academia, industry, and government to share their ideas, research results, and experiences.

ICCSIP 2021 received 105 submissions, all of which were written in English. After a thorough reviewing process, 41 papers were selected for presentation as full papers, resulting in an approximate acceptance rate of 39%. The accepted papers not only address challenging issues in various aspects of cognitive systems and information processing but also showcase contributions from related disciplines that illuminate the state of the art. In addition to the contributed papers, the ICCSIP 2021 technical program included four plenary speeches by Bo Zhang, Shiqing Chen, Yaonan Wang, and Lining Sun and six invited speeches by well-known scholars and entrepreneurs. We would also like to thank the members of the Advisory Committee for their guidance, the members of the International Program Committee and additional reviewers for reviewing the papers, and members of the Publications Committee for checking the accepted papers in a short period of time.

Last but not the least, we would like to thank all the speakers, authors, and reviewers as well as the participants for their great contributions that made ICCSIP 2021 successful

and all the hard work worthwhile. We also thank Springer for their trust and for publishing the proceedings of ICCSIP 2021.

November 2021

Fuchun Sun
Dewen Hu
Lei Yang
Stefan Wermter
Huaping Liu
Bin Fang

Organization

ICCSIP 2021 was hosted by the Chinese Association for Artificial Intelligence, the Chinese Association of Automation, and the IEEE Computational Intelligence Society. It was organized by the Cognitive Systems and Information Processing Society of Chinese Association for Artificial Intelligence, the Cognitive Computing and Systems Society of Chinese Association of Automation, Tsinghua University, and the Gusu Laboratory of Material Science with the help of the following co-organizers: the Nanjing Tsingzhan Institute of Artificial Intelligence, the China Center for Information Industry Development, and the Artificial Intelligence and Sensing Technology Institute (SIP) Co., Ltd.

Conference Committee

Honorary Chairs

Bo Zhang	Tsinghua University, China
Nanning Zheng	Xi'an Jiaotong University, China
Deyi Li	Chinese Association for Artificial Intelligence, China

Advisory Committee Chairs

Qionghai Dai	Tsinghua University, China
Fuji Ren	University of Tokyo, Japan
Shiming Hu	Tsinghua University, China
Hui Yang	Gusu Laboratory of Material Science, China

General Chairs

Fuchun Sun	Tsinghua University, China
Angelo Cangosi	University of Manchester, UK
Jianwei Zhang	University of Hamburg, Germany

Program Committee Chairs

Dewen Hu	National University of Defense Technology, China
Lei Yang	Tsingzhan Artificial Intelligence Research Institute, China
Stefan Wermter	University of Hamburg, Germany
Huaping Liu	Tsinghua University, China

Publication Chair

Bin Fang	Tsinghua University, China

Program Committee

Chenguang Yang	University of the West of England, UK
Guang-Bin Huang	Nanyang Technological University, Singapore
Katharina Rohlfing	University of Paderborn, Germany
Antonio Chella	Università degli Studi di Palermo, Italy
Yufei Hao	EPFL, Switzerland
Zhen Deng	Fuzhou University, China
Jun Ren	Hubei University of Technology, China
Chunfang Liu	Beijing University of Technology, China
Changsheng Li	Beijing Institute of Technology, China
Mingjie Dong	Beijing University of Technology, China
Peng Su	Beijing Information Science and Technology University, China
Rui Huang	University of Electronic Science and Technology of China, China
Tian Liu	Beijing Information Science and Technology University, China
Haiyuan Li	Beijing University of Posts and Telecommunications, China
Yong Cao	Northwestern Polytechnical University, China
Taogang Hou	Beijing Jiaotong University, China

Technical Sponsor

NVIDIA-IM

Contents

Algorithm

Vision

Robotics & Application

Algorithm

WeaveNet: End-to-End Audiovisual Sentiment Analysis

Yinfeng Yu[1](✉), Zhenhong Jia[1], Fei Shi[1], Meiling Zhu[2], Wenjun Wang[3], and Xiuhong Li[1]

[1] College of Information Science and Engineering, Xinjiang University, Urumqi, China
yuyinfeng@xju.edu.cn
[2] No. 59 Middle School of Urumqi, Urumqi, China
[3] Shanxi Datong University, Datong, China

Abstract. The way of analyzing sentiment by the proposed model in this paper is strikingly similar to the mechanism by which one person perceives another's sentiment. In this paper, We proposed a novel neural architecture named WeaveNet to "listen" and "watch" a person's sentiment. The main strength of our model comes from capturing both intra-interactions of one modal and inter-interactions of different modals stage by stage. Intra-interactions were modeled by convolution operations in the first few stages for each modality respectively and by bidirectional LSTM in the final stage for both audio clips and video clips. Inter-interactions were recognized at each stage applying various fusion effectively. At the same time, our model concentrated on the delicate design of the neural network rather than handcrafted features. The inputs of the network in our model were raw audios and natural images. In addition, audio clips and frames of a video were aligned by keyframe rather than by time in time order. We performed extensive comparisons on three publicly available datasets for both sentiment analysis and emotion recognition. WeaveNet outperformed state-of-the-art results in three publicly available datasets.

Keywords: Audiovisual · End-to-end · Sentiment analysis

1 Introduction

Sentiment analysis is one of the most active research areas in natural language processing and video processing. It is the computational study of individuals' emotions, sentiments, and so on [33]. With the rapid development of social networks, individuals can widely express their opinions. These opinions provide precious resources for sentiment analysis, which assists the development of automatic sentiment analysis [23]. The early methods were based on just only one modal extracting features by external tools [2,12,27]. Then the models transforming one modal to another were developed. For example, text-based sentiment analysis continued advanced by utilizing automatic speech recognition technology to convert speech into texts. With the development of automatic speech recognition, bimodal-based methods were introduced to sentiment analysis tasks [5,21,28].

Supported by Xinjiang Natural Science Foundation under Grant 2020D01C026 and Grant 2015211C288.

F. Sun et al. (Eds.): ICCSIP 2021, CCIS 1515, pp. 3–16, 2022.
https://doi.org/10.1007/978-981-16-9247-5_1

Recently, with the rapid development of communication technology, large amounts of data were uploaded by web users in the form of audios or videos, rather than just only texts. A large number of videos were readily available, which considerably promoted the research in the field of multimodal sentiment analysis and emotion recognition. Multimodal sentiment analysis has achieved advancement in performance and become an emerging research field of artificial intelligence [3,8,11,14,15,18,19,29,31].

At the same time, there are still some challenges that need to be overcome.

The first challenge is how to capture the inter-interactions of different modals effectively. Sometimes, there exists a contradiction among three modals in multimodal sentiment analysis. For example, "Cry with joy". In the above scene, the sentiment analysis result of textual is neutral, which was contradicted with that of both visual(negative) and acoustic(positive). Some researchers have observed that bimodal sub-tensors are more informative when used without other sub-tensors during their second set of ablation experiments in Tensor Fusion Network [30]. Whether it is the best choice to join audio, visual and textual together or not in multimodal emotion recognition and sentiment analysis is still a question.

The second challenge is how to capture the intra-interactions of one modal effectively. The intra-interactions of one modal vary from person to person. The way of expressiveness of affection varies widely from person to person. The amount of sentiment information in a specific modal varies widely from scene to scene. For example, some people express their affections more vocally, some more visually and others rely heavily on logic express little emotion [20].

Much of research work in multimodal sentiment analysis was based on handcrafted features extracted from raw videos, raw audios, and texts. The affection representation of these models was learned from the handcrafted features. Whether it is reasonable or not is still a question. To overcome the above challenges, lots of research has been done. These research achievements are mainly divided into the following three categories. The first category of models has relied densely on handcrafted features. The input of the neural network in these models [4,9,13,17,22] is handcrafted features extracted by external tools. The second category of models has relied lightly on handcrafted features, which is called end-to-end. The input of the neural network in these models is raw data. Meanwhile, there is a layer for extricating handcrafted features at the following layer of the neural network in these models [24,26,34]. The third category of models named end-to-end audiovisual model has never relied on any handcrafted features. They rely heavily on the design of models with very deep neural networks, using raw data as their input. They are different from the first category models since they accept raw audios and videos as the input of the model. They are also different from the second category models since it never uses any handcrafted features from the input layer to the output layer through the whole network. The representation completely got through neural networks without any handcrafted features. The proposed model in this paper was endeavored to overcome some of the above challenges via well designing a third category model. The model was performed to capture intra-interactions by a highlight in several multistage fusion. The model aimed to capture inter-interactions by fusing and summarizing in several multistage fusion. The inputs of the proposed model were

raw audios and videos. The model never used any handcrafted features from the input layer to the output layer through the whole network.

The main contributions of this paper are as follows:

- We proposed a novel model named WeaveNet for audiovisual sentiment analysis. Our model was designed to capture both intra-interactions and inter-interactions stage by stage through the whole audio clips and video clips. Intra-interactions were modeled by convolution operations in the first few steps and by bidirectional LSTM in the final stage. Inter-interactions were identified at each stage using multistage fusion.
- Our model was concentrated on the delicate design of the neural network rather than handcrafted features. The inputs of the network in our model were raw audios and natural images. Furthermore, audio clips and frames of a video were aligned by keyframe rather than by time in time order.
- WeaveNet achieved state-of-the-art results for audiovisual sentiment analysis in three publicly available datasets.

The rest of the paper is organized as follows. In the following section, we will review related work. In Sect. 3, we will exhibit more details of our methodology. In Sect. 4, experiments and results are presented, and the conclusion follows in Sect. 5.

2 Related Work

2.1 Multistage Fusion

Multistage fusion is a divide-and-conquer approach which distributes the fusion burden at several stages, letting any stage to perform in a more specialized and effective way [10].

2.2 Fusion Strategies

With the development of multimodal sentiment analysis, the fusion strategies have increased in quantity. There are concatenation fusion, add fusion, dot multiply fusion, and so on. Here z^a denotes the representation tensor of audios. z^v denotes the representation tensor of images. z^f denotes the fusion tensor.

Concatenation Fusion. The paper [7] provided a fusion formula for the representation of an utterance generated by concatenating all three multimodal features. The formula used in our model is as follows: $z^f = \tanh((W^f[z^a; z^v]) + b^f)$.

Add Fusion. Add fusion should make sure the dimension is identical between z^a and z^v, $z^f = \tanh((W^f(z^a + z^v)) + b^f)$.

Dot Multiply Fusion. Dot Multiply fusion should make sure the dimension is the same between z^a and z^v, $z^f = \tanh((W^f(z^a \odot z^v)) + b^f)$, Where $\odot$ indicates dot multiple by element-wise. W^f denotes the weight parameters. b^f denotes the parameters of bias.

3 Proposed Approach

In this section, we first describe the overall architecture of our proposed model in Sect. 3.1. Section 3.2 provides formulation and alignment. Then we give the detailed formulas of every module in Sect. 3.3.

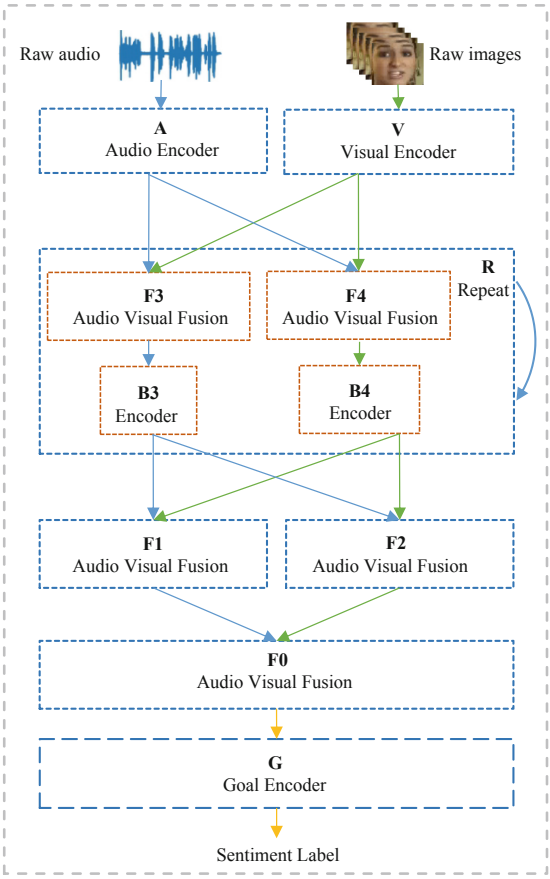

Fig. 1. The architecture of WeaveNet.

3.1 Overview of Network Architecture

The overall architecture is detailed in Fig. 1. **A** in bold is short for **A**udio module; **V** in bold is short for **V**isual module; **R** in bold is short for **R**epeatable module; **F0, F1, F2, F3, F4** in bold is short for submodule of **F**usion between audio and visual; **B3, B4** in bold is short for submodule of **B**roadcast of fusion information between audio and visual; **G** in bold is short for module of **G**oal output. It is an end-to-end network with the inputs of raw audios and raw images. The raw audios processed by the module **A**and the

raw images handled by the module **V** were woven with each other. Both module **A** and **V** were designed to highlight intra-interactions in audios and images respectively. The module **R** included four sub modules named **F3, F4, B3, B4**. Audios handled by module **A** and images processed by module **V** performed an fusion (e.g., add) by sub module **F3** and did a fusion(e.g., dot multiply) by sub module **F4**. Both sub module **F3** and sub module **F4** aimed to capture inter-interactions between audios and images. Then the output of sub module **F3** was processed by the sub module **B3**, and the output of sub module **F4** was processed by sub module **B4**. Both sub module **B3** and sub module **B4** were also designed to highlight intra-interactions in audios and images respectively. The procedure in module **R** was called weave. The process of weaving could iterate n times according to a specific situation. The module **R** performed to capture both intra-interactions and inter-interactions between audios and images. In our experiments, we only repeated one time. When finished the process of weaving, the results processed by sub module **B3** and the results processed by sub module **B4** would make a weave fusion again. Both module **F1** and module **F2** also aimed to capture inter-interactions between audios and images. The output of module **F1** and the product of module **F2** would make a fusion (e.g., concatenate). The module **F0** performed to make a summarize between audios and images. The results of concatenating fusion would process by module **G**. The output of module **G** was a sentiment score, which was transformed to sentiment label.

3.2 Formulation and Alignment

Problem Formulation. Given a dataset with data, $X = (X^a, X^v)$, where X^a, X^v stand for auditory and visual modality inputs, respectively. Usually, a data set is indexed by videos, which means that if we have N videos, then $X = (X_0, X_1, ..., X_{N-1})$, where $X_i = (X_i^a, X_i^v), 0 \leq i < N$. The corresponding labels for these N videos are $Y = (Y_0, Y_1, ..., Y_{N-1}), Y_i \in \mathbb{R}$ [17]. C denotes the number of all the classes (e.g. sentiment or emotion type) in a data set. In this work, we were tackling to learn a prediction function h, such that $h : X \longrightarrow Y$.

$y_{i,c}$ denotes an binary indicator, if the class label Y is the correct classification for observation X_i it is "1", else it is "0". Where $0 \leq c < C$.

$p_{i,c}$ denotes the predicted probability that observation X_i is of class c by prediction function h. The $p_{i,c}$ is as follows:

$$p_{i,c} = \frac{exp^{h(X_i^a, X_i^v; y_{i,c})}}{\sum_{k=0}^{C-1} exp^{h(X_k^a, X_k^v; y_{k,c})}} \tag{1}$$

The cross-entropy was our optimization goal.

Alignment of Audio Clips and Frames. How to align audio clips and frames of a video? f^v denotes the frame rate per second of a video clip. f^a denotes the sampling frequency of an audio clip. t^a denotes the length of an audio clip in the time domain. N_s^a denotes the number of slices for an audio clip. s_t^a denotes the stride of a slice in an audio clip in the continuous domain. w_t^a denotes the length of a slice in an audio clip in a continuous domain. s^a denotes the stride of a slice in an audio clip in the discrete domain. w^a denotes the length of a slice in an audio clip in the discrete domain.

$$N_s^a = \lceil \frac{t^a - w_t^a}{s_t^a} \rceil + 1 \tag{2}$$

When w_t^a and s_t^a have the following relation:

$$w_t^a = 2 \times s_t^a \tag{3}$$

Then the N_s^a is calculated by the following:

$$N_s^a = \lceil \frac{t^a}{s_t^a} \rceil - 1 \tag{4}$$

When $t^a = 1$ and $f^v = 30.0$, it satisfy the following equation:

$$\lceil \frac{1}{s_t^a} \rceil - 1 = f^v \tag{5}$$

By solving Eq. (5), $s_t^a = \frac{1}{31}$ in second can be derived. w^a and s^a are calculated by the following:

$$w^a = \lceil f^a \times w_t^a \rceil \tag{6}$$

$$s^a = \lceil f^a \times s_t^a \rceil \tag{7}$$

The above is a solution for aligning audio clips and frames of a video in time.

3.3 Modules in Details

Module A. $X_i^a(t)$ denotes the i-th raw audio, where $0 \leq i < N, t \in \mathbb{R}$. $X_i^a(n)$ denotes the i-th sampled audio. N_{i1}^a denotes the length of $X_i^a(n)$.

$$X_i^a(n) = X_i^a(t) \times \delta(t - \frac{n}{f^a}) \tag{8}$$

where $\delta(t)$ is unit impulse function, $0 \leq n < N_{i1}^a, n \in \mathbb{N}$. Discrete audio $X_i^a(n)$ was obtained. $X_i^a(:, w^a)$ denotes the reshaped one of X_i^a. N_{i2}^a denotes the first dimension size of $X_i^a(:, w^a)$.

$$N_{i2}^a = \lceil \frac{N_{i1}^a - w^a}{s^a} \rceil + 1 \tag{9}$$

$$X_i^a(m, w^a) = X_i^a[m \times s^a : m \times s^a + w^a] \tag{10}$$

where $0 \leq m < N_{i2}^a, m \in \mathbb{N}$. We segmented the discrete audio by the width of w^a and the stride s^a to make sure each segment have overlapped with its neighbors. Then we flattened all the segments in order and reshaped them to the length of f^a. So, every audio has several segments at the length of f^a. Then we padded $X_i^a(:, w^a)$ with zeros to make sure it's length is integer times of f^a. N^b denotes the number of segments of sampled audio in a second.

$$N_i^{ap} = N_{i2}^a - N^b \times \lfloor \frac{N_{i2}^a}{N^b} \rfloor \tag{11}$$

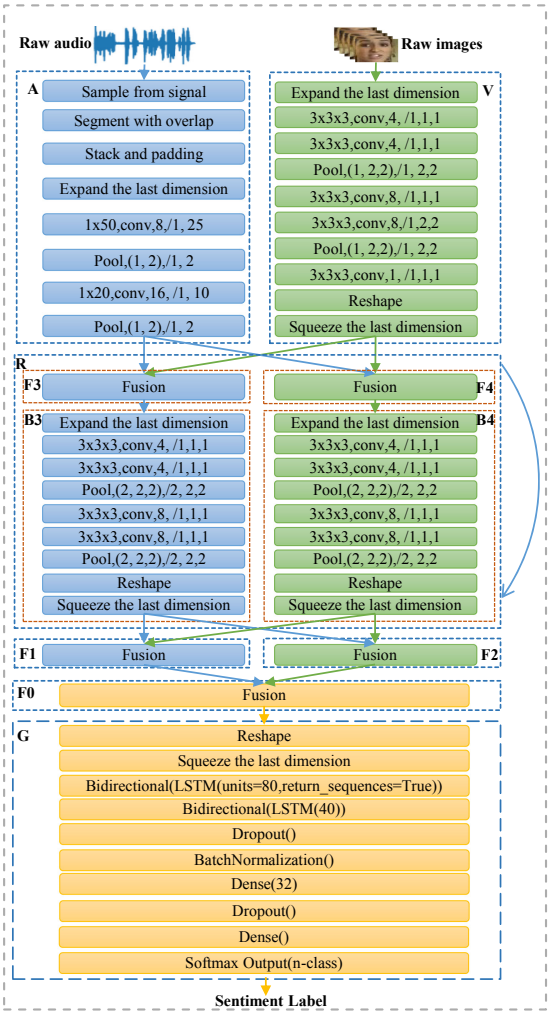

Fig. 2. The details of WeaveNet.

Where N_i^{ap} denotes the padding length of X_i^a.

$$X_i^{ap} = zeros((N_i^{ap}, w^a)) \tag{12}$$

Where X_i^{ap} denotes the padding part of X_i^a.

$$X_i^a = stack[X_i^a; X_i^{ap}] \tag{13}$$

Then we reshaped X_i^a to the shape (N_{i3}^a, f^a). Where N_{i3}^a is compute as follows:

$$N_{i3}^a = \lfloor \frac{N_{i2}^a + N_i^{ap}}{N^b} \rfloor \tag{14}$$

Then we sampled T times at an identical time interval from 0 to N_{i3}^a. k^a denotes the index of audio segments in an audio clip. Here $0 \leq k^a < N_{i3}^a$, $k^a \in \mathbb{N}$.

$$k^a = \lfloor \frac{N_{i3}^a}{T} \times j \rfloor \tag{15}$$

where $0 \leq j < T$, $j \in \mathbb{N}$. Then we sorted all the k^a in ascendant order and put them in a list named $\mathbf{K}^a$.

$$Z_i^a(j,:) = X_i^a(\mathbf{K}^a[j],:) \tag{16}$$

Z_i^a denotes the embedding of raw audio with the shape of (T, f^a), which can conserve the time information of intra-interactions. We made re-sample techniques to make sure every audio have T segments at the length of f^a. The above procedure for raw audios named "slice" in this paper. After the slicing procedure, we expanded the last dimension to make every audio has the shape(T, f^a, 1). Then 1-d convolution and 1-d max-pooling were used twice to make a summarize for raw audios.

Module V. $X_i^v(:,:,:)$ denotes the i-th video clip, where $0 \leq i < N$. N_i^v denotes the number of frames extracted from the i-th video clip. h^v indicates the height of every face got from a video clip. w^v means the width of every face got from a video clip. To make a balance between computation complexity and the intra-interactions of one modal, we obtained faces from all the frames in the i-th video applying python package face recognition[1]. Succeeding, We re-sized all the images with the shape of (h^v, w^v). Then we sampled T times at an identical time interval from 0 to N_i^v with the same solution as a process of audios. k^v denotes the index of frames in a video clip. Here $0 \leq k^v < N_i^v$, $k^v \in \mathbb{N}$.

$$k^v = \lfloor \frac{N_i^v}{T} \times j \rfloor \tag{17}$$

where $0 \leq j < T$, $j \in \mathbb{N}$. Then we sorted all the k^v in ascendant order and put them in a list named $\mathbf{K}^v$, which conserve the time information of intra-interactions.

$$Z_i^v(j,:,:) = X_i^v(\mathbf{K}^v[j],:,:) \tag{18}$$

Z_i^v denotes the embedding of raw faces with the shape of (T, h^v, w^v). We made re-sample techniques to make sure every video have T segments at the shape of (h^v, w^v) maintaining the time information of intra-interactions. Following the slicing process, We expanded the last dimension to make every video with the shape(T, h^v, w^v, 1). Later 2-d convolution, 2-d convolution, and 2-d max-pooling were used twice to make a summarize for the raw images. Following made a reshape and squeezing transformation to the output of 2-d convolution.

Module R. The module **R** included four sub modules named **F3, F4, B3, B4**. The sub module **F3** and sub module **F4** were designed to capture inter-interactions between audios and images. However, there was a subtle difference between them. The sub module **F3** focused on audio representation with the complementary of images. The sub module **F4** centered on visual representation with the attention of audio. The fusion strategy in sub module **F3** and sub module **F4** were choosable. The fusion strategy

[1] https://pypi.org/project/face_recognition/.

could be tensor fusion, add fusion, dot multiply fusion and so on. The sub module **B3** and sub module **B4** aimed to capture intra-interactions. The architecture of them could be identical or distinctive. In our model, weighing the convenience of the repeated of module **R**, the architecture of sub module **B3** and sub module **B4** were sketched to the same. In the sub module **B3** or sub module **B4**, we first expanded the last dimension of the input. Then repeated twice of the procedure of 3-d max-pooling after two 3-d convolutions. Then the output of the last 3-d max-pooling was reshaped. When completed the repeated of module R, the output of both sub module **B3** and sub module **B4** were squeezed the last dimension respectively after reshaped.

Module F1 and F2. The module **F1** and module **F2** were designed to capture inter-interactions between audios and images. There was a subtle difference between them. The module **F1** concentrated on audio representation with the complementary of images. The module **F2** focused on visual representation with the attention of audio. The fusion strategy in module **F1** and module **F2** were choosable. The fusion strategy could be tensor fusion, add fusion, dot multiply fusion and so on. In our model, the fusion strategy in module **F1,F2,F3,F4** are as follows:

$$z^f = \lambda_k \times \tanh((W^f(z^a + z^v)) + b^f) + (1 - \lambda_k) \times \tanh((W^f(z^a \odot z^v)) + b^f) \quad (19)$$

where $1 \leq k \leq 4$, $\lambda_k \in [0, 1]$ is a parameter corresponding with module $\mathbf{Fk}$ respectively.

Module F0. The module **F0** was designed to capture inter-interactions between audios and images. The fusion strategy in this model could be tensor fusion, add fusion, dot multiply fusion, concatenate fusion and so on.

Module G. The module **G** aimed to capture the temporal information from the summary of audio and images. To make an alignment the representation of both audios and images, we reshaped the input of this module. Then the last dimension of the output of reshaping transformation was squeezed. After that, two layers of bidirectional Long short time memory Layer were designed to capture the temporal information. The result of the procedure by the two bidirectional LSTM was transferred to a Batch Normalization and two dense layers. The last layer was a dense layer with a "softmax" activation function, which acted to transform a sentiment score of more than two class to sentiment label.

4 Experiments

4.1 Experimental Setup

Datasets. To examine the effectiveness of the proposed model, we designed various experiments to evaluate the performance of WeaveNet. We had chosen two domains: sentiment analysis and emotion recognition. The first two datasets were sentiment analysis. The final one was emotion recognition. All benchmarks involved two modals with raw audios and raw images.

MOSI. Multi-modal Opinion level Sentiment Intensity [32]. It is an opinion-level annotated corpus of sentiment and subjectivity analysis in online videos, which including 2199 segments. Sentiment intensity is assigned from strongly negative to strongly positive with a linear range from minus 3 to plus 3. For every video clip, the annotators possed seven choices: strongly positive, positive, weakly positive, neutral, weakly negative, negative, strongly negative. We transformed them into 3, 5, 7 labels with identical distribution, and conducted 3-class, 5-class, and 7-class sentiment classification using MOSI video clips.

MOUD. Multimodal Opinion Utterances Dataset [16]. The MOUD dataset contains 498 video clips. Every video clip was labeled to be either positive, negative or neutral. However, there are 20 video clips in MOUD, which has an extraordinary name. We failed to extract frames from these video clips. We dropped these video clips in our experiments. In MOUD experiment, we conducted binary sentiment classification using 478 video clips.

IEMOCAP. Interactive Emotional dyadic Motion Capture database [1]. It was collected by the Speech Analysis and Interpretation Laboratory at the University of Southern California. The database recorded from ten actors in dyadic sessions with markers on the face, head, and hands. The actors performed selected emotional scripts and also improvised hypothetical scenarios designed to elicit specific types of emotions. The dataset had 7532 video clips. Every video clip was annotated for the presence of 10 emotions (anger, disgust, excited, fear, frustration, happiness, neutral state, sadness, surprise, and others). We dropped the clips annotated as "others". We conducted binary (anger, happiness, neutral state) sentiment classification using IEMOCAP video clips. We got frames from video clips of MOUD and IEMOCAP and then extracted faces from frames using face recognition python package.

Evaluation Criteria. Different datasets in our experiments have different labels. For binary classification and multiclass classification, we reported accuracy A^C, where C is the number of all the classes in a dataset. Higher values denote better performance.

Implementation Details. T denotes video time-steps. The time-steps T in our experiments was set to 32. Every raw audio was sampled at the sample rate of 16K Hz. $f^a = 16000$. $s^a = \frac{1}{31}$. $N^b = 5$. Every face was re-sized with h^v(height) = 128, w^v(width) = 128. We randomly select 8:1:1 for training, validation, and test set for all three datasets. One NVIDIA Tesla K80 GPU was used for training and testing. Our model was trained using Adam with an initial learning rate at 1e-3 and with an original epoch at 44. We combined drop out with early stopping to get our model rid of overfitting. The parameters of module **F1,F2,F3,F4** are $\lambda_1 = 0.9$, $\lambda_2 = 0.1$, $\lambda_3 = 0.9$, $\lambda_4 = 0.1$ respectively. The fusion tactics of module **F0** in our experiments was concatenation fusion. The details were in Fig. 2.

4.2 Performance Comparison with State-of-the-art

Baseline Models. We compared the performance of WeaveNet with current state-of-the-art models for audiovisual sentiment analysis. Due to space constraints, each baseline name denoted by a symbol(in parenthesis) which used in Table 1 to refer to specific baseline results.

EndToEnd1(▷). This was an end-to-end model that applied a convolutional neural network to get an affectionate representation of the acoustic modality, which employed a deep residual network to get an emotional description of the visual modality. This model did a regression for both arousal and valence of emotion using the RECOLA database of the AVEC 2016 research challenge on emotion recognition [25].

EndToEnd2(◁). This was an end-to-end audiovisual deep residual network for multimodal apparent personality trait recognition. This model was evaluated on the dataset that was released as part of the ChaLearn First Impressions Challenge. The network was trained end-to-end for predicting the five personality traits of people from their videos. It made five continuous prediction values corresponding to each trait for the video clip [6]. The type of the model used in above two papers(◁, ▷) was a regression rather than classification. We modified the activation function of the output layer of the model in those papers to "softmax" and remained all the same as the original models. To make a comparison with the proposed model, we built the model of those papers and tested them on MOSI, MOUD, and IEMOCAP.

Quantitative Evaluation. We performed comparisons of the proposed model and two state-of-the-art methods in audiovisual sentiment analysis on three datasets. The comparison results in Table 1 illustrated that our model consistently outperformed others, which demonstrate the effectiveness of our proposed model.

Table 1. Results were for sentiment analysis on both MOSI and MOUD, emotion recognition on IEMOCAP. SOTA1 refer to the previous best state of the art. Best results were highlighted in bold. $\triangle_{SOTA}$ showed the change in performance over SOTA1. Improvements were highlighted in green. Those to be improved were highlighted in red. The WeaveNet slightly outperformed some of SOTA.

Model	MOSI			MOUD	IEMOCAP		
	Accuracy(%)			Accuracy(%)	Accuracy(%)		
	A^7	A^5	A^3	A^2	A^2		
					ang.[a]	hap.[b]	neu.[c]
[25]	17.73	30.91	55.91	59.09	85.53	92.77	77.87
[25]	22.73	34.55	51.36	63.64	84.82	90.21	76.60
WeaveNet	**25.00**	**36.36**	**55.92**	**68.18**	**87.66**	**92.78**	**77.88**
$\triangle_{SOTA}$	↑ 2.27	↑ 1.81	↑ 0.01	↑ 4.54	↑ 2.13	↑ 0.01	↑ 0.01

[a]Denotes anger. [b]Denotes happiness. [c]Denotes neutral state.

4.3 Analysis of the Proposed Approach

The most significant factor of the proposed model outperformance than baseline models is that the proposed model fusion between audios and images stage by stage. Apprehend all the detailed characterization of the sentiment in a video is an unrealistic idea since the computation cost is so high that the result is unachievable. Coarser much detailed

information gravely makes the fine classification impossible. We dropped some information subtly after the fusion step by step, which makes the ability to capture the sentiment improved very much. The second critical factors are the time alignment between audios and images. We aligned the keyframe of both audio and images from the entire video clip rather than by time, which has a summarize function. The high accuracy of keyframe alignment between audio clips and video clips performs a sparse representation with a low computational cost, which made the proposed model possess a high ability to capture the inter-interactions. The convolution layer located before the bidirectional LSTM at the head of the whole network in our model is a powerful way to reduce computation cost. Convolution operation holds a very high-level share of parameters. It does well at taking the characteristic of sentiment in both audios and images. At the same time, the parameters of convolution are relatively less than other networks such as RNN with the same ability in representation. However, both the audios and images are a sequence. So in our model, we captured the intra-interactions stage by stage with convolution operation. After summarizing, the data volume of the representation for sentiment decreased. We sent the representation to the bidirectional LSTM to capture the intra-interactions. The order of the convolution layer and bidirectional LSTM is reasonable.

5 Conclusion

In this paper, we proposed an effective model named WeaveNet for audiovisual sentiment analysis. Our model was designed to capture both intra-interactions and inter-interactions stage by stage through the whole audio clips and video clips. Intra-interactions modeled by convolution operations in the first few steps and by bidirectional LSTM in the final stage. Inter-interactions were identified at each stage using multistage fusion. Our model concentrated on the delicate design of the neural network rather than handcrafted features. The inputs of the network in our model were raw audios and natural images. Besides, audio clips and frames of a video were aligned by keyframe rather than by time in time order. We performed extensive comparisons on three publicly available datasets for both sentiment analysis and emotion recognition. WeaveNet achieved state-of-the-art results in three publicly available datasets.

References

1. Busso, C., et al.: Iemocap: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**, 335–359 (2008)
2. Cambria, E.: Affective computing and sentiment analysis. IEEE Intell. Syst. **31**, 102–107 (2016)
3. Chen, M., Wang, S., Liang, P.P., Baltrusaitis, T., Zadeh, A., Morency, L.P.: Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: ICMI (2017)
4. Etienne, C., Fidanza, G., Petrovskii, A., Devillers, L., Schmauch, B.: Speech emotion recognition with data augmentation and layer-wise learning rate adjustment. CoRR abs/1802.05630 (2018)

5. Gievska, S., Koroveshovski, K., Tagasovska, N.: Bimodal feature-based fusion for real-time emotion recognition in a mobile context. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 401–407 (2015)
6. Güçlütürk, Y., Güçlü, U., van Gerven, M., van Lier, R.: Deep impression: audiovisual deep residual networks for multimodal apparent personality trait recognition. In: ECCV Workshops (2016)
7. Hazarika, D., Poria, S., Mihalcea, R., Cambria, E., Zimmermann, R.: Icon: Interactive conversational memory network for multimodal emotion detection. In: EMNLP (2018)
8. Kim, D.H., Lee, M.K., Choi, D.Y., Song, B.C.: Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild. In: ICMI (2017)
9. Kim, J., Englebienne, G., Truong, K.P., Evers, V.: Deep temporal models using identity skip-connections for speech emotion recognition. In: ACM Multimedia (2017)
10. Liang, P.P., Liu, Z., Zadeh, A., Morency, L.P.: Multimodal language analysis with recurrent multistage fusion. CoRR abs/1808.03920 (2018)
11. Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., Morency, L.P.: Efficient low-rank multimodal fusion with modality-specific factors. In: ACL (2018)
12. Ma, X., Yang, H., Chen, Q., Huang, D., Wang, Y.: Depaudionet: an efficient deep model for audio based depression classification. In: AVEC@ACM Multimedia (2016)
13. Mistry, K., Zhang, L., Neoh, S.C., Lim, C.P., Fielding, B.: A micro-GA embedded PSO feature selection approach to intelligent facial emotion recognition. IEEE Trans. Cybern. **47**, 1496–1509 (2017)
14. Nasir, M., Jati, A., Shivakumar, P.G., Chakravarthula, S.N., Georgiou, P.G.: Multimodal and multiresolution depression detection from speech and facial landmark features. In: AVEC@ACM Multimedia (2016)
15. Nguyen, D.L., Nguyen, K., Sridharan, S., Ghasemi, A., Dean, D., Fookes, C.: Deep spatio-temporal features for multimodal emotion recognition. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1215–1223 (2017)
16. Pérez-Rosas, V., Mihalcea, R., Morency, L.P.: Utterance-level multimodal sentiment analysis. In: ACL (2013)
17. Pham, H., Manzini, T., Liang, P.P., Póczos, B.: Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. CoRR abs/1807.03915 (2018)
18. Poria, S., Cambria, E., Hazarika, D., Mazumder, N., Zadeh, A., Morency, L.P.: Multi-level multiple attentions for contextual multimodal sentiment analysis. In: 2017 IEEE International Conference on Data Mining (ICDM), pp. 1033–1038 (November 2017). https://doi.org/10.1109/ICDM.2017.134
19. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: from unimodal analysis to multimodal fusion. Inf. Fusion **37**, 98–125 (2017)
20. Poria, S., Chaturvedi, I., Cambria, E., Hussain, A.: Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 439–448 (2016)
21. Seng, K.P., Ang, L.M., Ooi, C.S.: A combined rule-based & machine learning audio-visual emotion recognition approach. IEEE Trans. Affect. Comput. **9**, 3–13 (2018)
22. Sivaprasad, S., Joshi, T., Agrawal, R., Pedanekar, N.: Multimodal continuous prediction of emotions in movies using long short-term memory networks. In: ICMR (2018)
23. Soleymani, M., García, D., Jou, B., Schuller, B.W., Chang, S.F., Pantic, M.: A survey of multimodal sentiment analysis. Image Vis. Comput. **65**, 3–14 (2017)
24. Trigeorgis, G., et al.: Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5200–5204 (2016)

25. Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B.W., Zafeiriou, S.: End-to-end multimodal emotion recognition using deep neural networks. IEEE J. Sel. Top. Sign. Proces. **11**, 1301–1309 (2017)
26. Tzirakis, P., Zhang, J., Schuller, B.W.: End-to-end speech emotion recognition using deep neural networks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5089–5093 (2018)
27. Wu, A., Huang, Y., Zhang, G.: Feature fusion methods for robust speech emotion recognition based on deep belief networks. In: ICNCC 2016 (2016)
28. Yan, J., Zheng, W., Xu, Q., Lu, G., Li, H., Wang, B.: Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech. IEEE Trans. Multimedia **18**, 1319–1329 (2016)
29. Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M.C., Sahli, H.: Multimodal measurement of depression using deep learning models. In: AVEC@ACM Multimedia (2017)
30. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. In: Empirical Methods in Natural Language Processing, EMNLP (2017)
31. Zadeh, A., Liang, P.P., Poria, S., Vij, P., Cambria, E., Morency, L.P.: Multi-attention recurrent network for human communication comprehension. CoRR abs/1802.00923 (2018)
32. Zadeh, A., Zellers, R., Pincus, E., Morency, L.P.: Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. CoRR abs/1606.06259 (2016)
33. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis : a survey. Wiley Interdisc. Rew. Data Min. Knowl. Discov. **8**, e1253 (2018)
34. Zhu, B., Zhou, W., Wang, Y., Wang, H., Cai, J.J.: End-to-end speech emotion recognition based on neural network. In: 2017 IEEE 17th International Conference on Communication Technology (ICCT), pp. 1634–1638 (2017)