TPFN: Applying Outer Product along Time to Multimodal Sentiment Analysis Fusion on Incomplete Data

Binghua Li[1,2,⋆,0000−0002−2595−4762], Chao Li[1,⋆], Feng Duan[2,⋆⋆], Ning Zheng[1,0000−0002−2282−1200], and Qibin Zhao[1,0000−0002−4442−3182]

[1] RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan
{chao.li, ning.zheng, qibin.zhao}@riken.jp
[2] College of Artificial Intelligence, Nankai University, Tianjin, China
nkvhua@outlook.com, duanf@nankai.edu.cn

Abstract. Multimodal sentiment analysis (MSA) has been widely investigated in both computer vision and natural language processing. However, studies on the imperfect data especially with missing values are still far from success and challenging, even though such an issue is ubiquitous in the real world. Although previous works show the promising performance by exploiting the low-rank structures of the fused features, only the first-order statistics of the temporal dynamics are concerned. To this end, we propose a novel network architecture termed Time Product Fusion Network (TPFN), which *takes the high-order statistics over both modalities and temporal dynamics into account*. We construct the fused features by the outer product along adjacent time-steps, such that richer modal and temporal interactions are utilized. In addition, we claim that the low-rank structures can be obtained by regularizing the Frobenius norm of latent factors instead of the fused features. Experiments on CMU-MOSI and CMU-MOSEI datasets show that TPFN can compete with state-of-the art approaches in multimodal sentiment analysis in cases of both random and structured missing values.

Keywords: Multimodal Sentiment Analysis, Multimodal Learning, Matrix/Tensor Decomposition, Incomplete Data

1 Introduction

Multimodal learning is recently one of the increasingly popular yet challenging tasks in both computer vision (CV) [1,12,14], and natural language processing (NLP) [5,52]. As its important application, multimodal sentiment analysis (MSA) is to predict the categories or intensity of the sentiment by jointly mining the data with various modalities, *e.g.*, visual, acoustic and language [44,50,51].

Although approaches on MSA have been well developed in an ideal situation, the imperfectness of the data is still a challenge we have to face in the real

⋆ Equal contribution.
⋆⋆ Feng Duan is the main corresponding author.

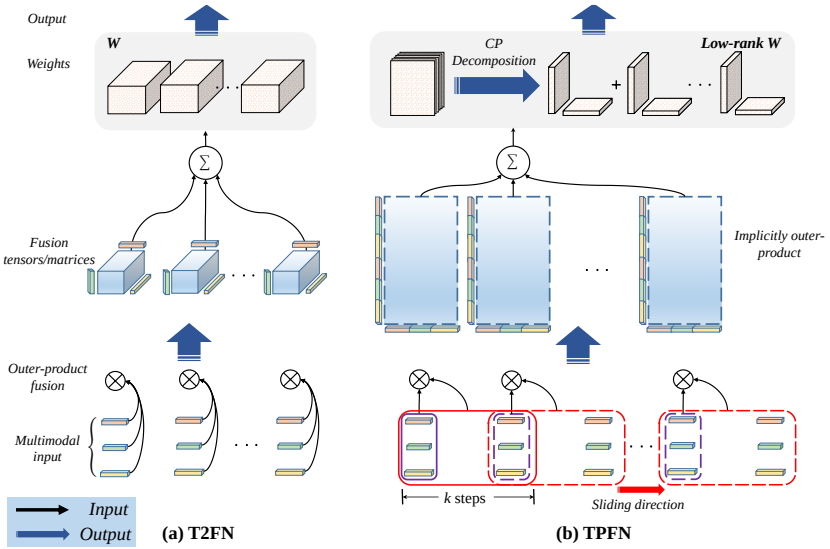

Fig. 1. Comparison between T2FN [23] and TPFN (our method). In the proposed model, we concatenate the features along adjacent k time-steps. In the fusion phase, the out product is implicitly applied to avoid the additional computational/storage consumption, and the corresponding weight tensor is also assumed with the low-rank CP format.

world, especially when there are missing values at unknown locations caused by mismatched modalities or sensor failures [35,40]. Recently, Liang *et al.* [23] demonstrates that the representation obtained from incomplete data has the low-rank structure in MSA, proposing a low-rank regularization based model termed temporal tensor fusion network (T2FN) against the bias caused by missing values. Despite the method achieves robust implementation against the imperfect data, there are still two aspects we can go further: 1) T2FN fully captures the interactions among all modalities of the data yet fails to exploit the data dynamics across the temporal domain and 2) the outer product is explicitly applied in the model, which would lead to heavy consumption of memory resources when increasing the time-steps or feature dimension.

To this end, we propose a novel model termed Time Product Fusion Network (TPFN) for the issue of incomplete MSA. Compared to previous works, TPFN 1) *captures additional interactions among different time-steps* to exploit higher-order statistics of temporal dynamics, and 2) *alleviates the issue of the unacceptable model size* by approximating the weight tensor using the well-known CANDECOMP/PARAFAC (CP) decomposition [21]. For the latter, we theoretically claim that the low-rank structure of the fused features can be controlled by the Frobenius norm of the latent factors. It allows us to avoid the explicit calculation of the outer product used in [23]. Fig. 1 gives the schematic diagram to demonstrate the difference between T2FN and the proposed TPFN. As shown

Table 1. Comparison of the outer-product-based methods. Each column corresponds one characteristic of the methods, where "*Decomposition*" indicates whether exploiting matrix/tensor decomposition for dimension reduction and "*Low-rankness*" means whether the low-rankness of features is considered in the model.

Methods	*Decomposition*	*Temporal dynamics*	*Imperfect data*	*Low-rankness*
TFN [49]	✗	✗	✗	✗
LMF [26]	✓	✗	✗	✗
HFFN [28]	✗	✗	✗	✗
HPFN [19]	✓	✓ (high-order)	✗	✗
T2FN [23]	✗	✓ (low-order)	✓	✓
TPFN(ours)	✓	✓ (high-order)	✓	✓

in Fig. 1, we construct the interaction among features by implicitly exploiting the outer-product along the adjacent time-steps. As for the learnable weights in the model, we apply the low-rank CP format to reducing the required model size.

1.1 Related Works

Multimodal sentiment analysis. The studies on sentiment analysis (SA) are started from 2000s [34], and widely discussed in the NLP community [7,24,36,39]. Its extension, multimodal sentiment analysis (MSA), recently attracts more attention as the visual and audio features are capable of significantly improving the prediction performance compared to the conventional SA [31]. As a multimodal learning task, the core issue on the MSA study is how to efficiently fuse the features from multiple modalities [15,31,45,43,52]. In the existing methods, the outer-product-based fusion strategy shows impressive performance with very neat model [2,23,19,26,28,49]. The basic idea behind those methods is to exploit the outer-product to obtain the high-order statistics of features. The fused features reflect rich information about the interaction among multiple modalities. Table 1 compares several important characteristics of the state-of-the-art methods based on outer-product. As shown in Table 1, only T2FN [23] and our method can deal with the imperfect data, while other methods generally assume that both the training and test datasets are ideal. Compared to T2FN, our method takes the *high-order statistics of temporal dynamics* into account and further applies matrix/tensor decomposition on weights to reducing the number of parameters. Although there have been several studies on developing robust models for multimodal learning [6,25,29,40], they cannot trivially applied to the MSA task due to the difference of specific tasks and data.

Low-rankness in robust learning. Low-rank approximation is a collection of well-known methods to cope with the issue on the imperfect data like noise [13,32] and incompleteness [18,22,42,48]. The existing low-rank approximation methods can be roughly split into two categories: (a) matrix/tensor decomposition [8,17,41] and (b) nuclear norm minimization [11,16,27]. On the other side, in the studies on artificial neural networks (ANN), the low-rank assumption is generally im-

posed on weights for model compression [33,46] or for alleviating the overfitting issue [3,4]. Except for T2FN [23], there is seldom work in ANN to apply the low-rank regularization *directly* to feature maps for robust learning. Inspired by the prior arts, we also apply the low-rank regularization to coping with the imperfect data issue. However, we highlight the difference from T2FN that in this work the low-rank structure is obtained by regularizing the Frobenius norm of the latent factors instead of on the fused features as T2FN does. The new trick allows us to avoid the additional requirement on computation and storage of the model.

2 Preliminaries

To be self-contained of the paper, we concisely review the necessary multilinear algebra and operations, which play important roles to understand our model.

Notation. We use italic letters like a, b, A, B to denote scalars, boldface lowercase letters like $\mathbf{a}, \mathbf{b}, ...$ to denote vectors, boldface capital letters $\mathbf{A}$, $\mathbf{B}$, ... to denote matrices and calligraphic letters like $\mathcal{A}, \mathcal{B}, ...$ for tensors of arbitrary order. Time series are denoted by the underlined italic capital letters, *e.g.*, $\underline{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_T\}$. The operation "$\circ$" denotes the element-wise product, "$\bigotimes$" denotes outer product of vectors and "$\cdot$" denotes the matrix-tensor product. More details on multi-linear operations are given in [10] and the references therein.

CP format. In our model, we apply the well-known (CP) decomposition [8,17] to representing the weights in the last layer of the proposed network. Specifically, CP decomposition is to represent a tensor as a finite sum of rank-1 factors constructed by the outer products of vectors. Given an pth-order tensor $\mathcal{W}$ and the factor matrices $\mathbf{W}^{(p)}, p \in [P]$, the CP decomposition of $\mathcal{W}$ is given by:

$$\mathcal{W} = \sum_{r=1}^{R} \bigotimes_{p=1}^{P} \mathbf{w}_r^{(p)}, \tag{1}$$

where R denotes the CP-rank of $\mathcal{W}$[3], and $\mathbf{w}_r^{(p)}$ denotes the r-th column of $\mathbf{W}^{(p)}$. Note that if $P = 2$ then Eq. (1) is degenerated into the trivial product of two factor matrices, *i.e.*,

$$\mathcal{W} = \sum_{r=1}^{R} \bigotimes_{p=1}^{2} \mathbf{w}_r^{(p)} = \mathbf{W}^{(1)} \mathbf{W}^{(2),\top}, \tag{2}$$

where $\cdot^\top$ denotes the transpose operation of a matrix. In the rest of the paper, we also need to exploit the tensor nuclear norm $\|\mathcal{W}\|_*$, the dual norm of tensor spectral norm, to introduce a more efficient low-rank regularization into the model. As introduced in [37], the tensor nuclear norm is a good surrogate of

[3] Without ambiguity, we also use the notion of CP-rank to represent the number of rank-1 factors used in matrix/tensor approximation.

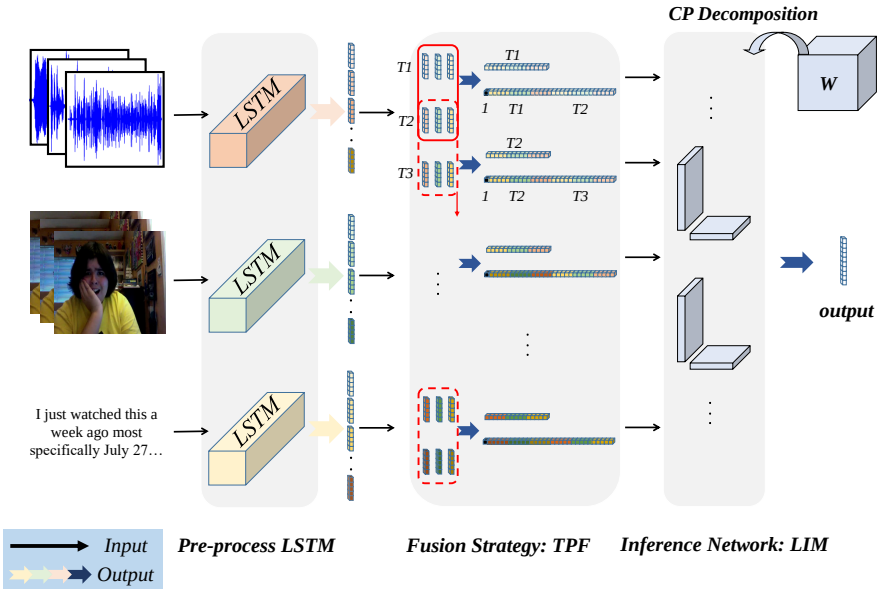

Fig. 2. Details of our TPFN with $k = 2$ and stride as 1. Three Modules are involved in our model: The Pre-process LSTM, the Time Product Fusion Module and the Low-rank Inference Module. Low-rank regularization is applied on the multimodal representation

tensor rank. It therefore implies that bounding the nuclear norm generally results in a low-rank regularization.

Problem setting. In this paper, we consider the MSA task as a multimodal learning problem. Specifically, we assume a sample in the task to be a triple $(\underline{A}, \underline{V}, \underline{L})$, where $\underline{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_T\}$, $\underline{V} = \{\mathcal{V}_1, \ldots, \mathcal{V}_T\}$ and $\underline{L} = \{\mathcal{L}_1, \ldots, \mathcal{L}_T\}$ denote the time series of the length T *w.r.t.* the acoustic, visual and language data, respectively. The goal of our work is hence to learn a mapping from the multimodal data to an output associated with a specific task, such as classification or regression. Mathematically, we would like to learn a composite function

$$\widehat{\mathbf{y}} = f\left(\phi_a(\underline{A}), \phi_v(\underline{V}), \phi_l(\underline{L})\right), \tag{3}$$

where ϕ_a, ϕ_v and ϕ_l denotes the sub-mappings from the raw data to the features, and the function f includes the fusion phase and the mapping from the fused features to targets. As in the previous works of MSA, ϕ_a, ϕ_v and ϕ_l generally consist of deep neural networks like CNN [20] and RNN [30] to embed the raw data into the feature space. We also adopt the same architecture yet put the main focus on developing robust and efficient fusion method, *i.e.* the function $f(\cdot)$ in Eq. (3).

3 Time Product Fusion Network

The architecture of the proposed time product fusion network (TPFN) is shown in Fig. 2, where the whole network can be divided into three aspects: (a) pre-process LSTM, (b) time product fusion (TPF), and (c) low-rank inference module (LIM). Note that the aspect (a) corresponds the sub-mappings ϕ_a, ϕ_v and ϕ_l in Eq. (3), while the function f consists of the modules TPF and LIM. In addition, we also study a new low-rank regularization term to tackle the issue of imperfect data. Because we follow the basic structures as [23] in the pre-process LSTM module, below the focus of the paper is mainly on the rest and the new regularization strategy.

3.1 Main Idea: Outer Product through Time

To investigate the additional statistics of the feature dynamics, we construct the fusion operation by imposing the outer products among adjacent time-steps. In the MSA task, assume that we obtain the multimodal features per T time-steps by the sub-nets ϕ_a, ϕ_v and ϕ_l in Eq. (3), where the features are denotes by the matrices $\mathbf{A} \in \mathbb{R}^{d_a \times T}$ for acoustic, $\mathbf{V} \in \mathbb{R}^{d_v \times T}$ for visual, and $\mathbf{V} \in \mathbb{R}^{d_v \times T}$ for the language modal, respectively. Let $\mathbf{a}_t$, $\mathbf{v}_t$, and $\mathbf{l}_t$ be the t-th column of $\mathbf{A}$, $\mathbf{V}$ and $\mathbf{L}$, respectively, then we first concatenate the features from all modalities for the given time-step $t \in [T]$:

$$\mathbf{z}_t^\top = [\mathbf{a}_t^\top, \mathbf{v}_t^\top, \mathbf{l}_t^\top] \in \mathbb{R}^{1 \times L}, \tag{4}$$

where $L = d_a + d_v + d_l$. As shown in Eq. (4), all features in the t-th step are involved in the concatenated vector $\mathbf{z}_t$. Next, to model the interaction across time-steps, we construct enhanced vectors *w.r.t.* $\mathbf{z}_t$ by further concatenating the adjacent time-steps, *i.e.*,

$$\mathbf{z}_t^{e,\top} = [1, \mathbf{z}_t^\top, ..., \mathbf{z}_{t+k-1}^\top] \in \mathbb{R}^{1 \times (kL+1)}, k > 0, t \in [T], \tag{5}$$

where the element "1" is also padded to retain the intra-modal correlation for each modality. Using Eq. (4) and (5), the tth-step ingredient of the fused features is therefore calculated as

$$\mathbf{M}_t = \mathbf{z}_t \otimes \mathbf{z}_t^e = \mathbf{z}_t \otimes [1, \mathbf{z}_t, \ldots, \mathbf{z}_{t+k-1}] \in \mathbb{R}^{L \times (kL+1)}. \tag{6}$$

Note that $\mathbf{M}_t$ can be divided into three chunks: $\mathbf{z}_t \bigotimes 1$, $\mathbf{z}_t \bigotimes \mathbf{z}_t$, and $\mathbf{z}_t \bigotimes \mathbf{z}_{t+i}$. The first chunk keeps the *unimodal information* by the product with the identity. The second chunk is to model the *inter-modality interaction* in the local step, and the third attempts to explore the *dynamics across time*.

More interestingly, the calculation of $\mathbf{M}_t$ for all $t \in [T-k+1]$ is equivalent to sliding a window of the size k along the temporal domain. As shown in Fig. 2, we first collect the features from all modalities together by concatenation, and then use a window of the size k and stride 1 to slide the concatenated features from top to bottom. For each step, we extract the features contained in the window and

employ the outer-product to fuse the features. With the time window sliding, all intra-modality and inter-modality dynamics will be involved. As a special case, if k is set to 1, *i.e.* $\mathbf{z}_e^{\mathrm{e}\top} = [1, \mathbf{z}_t^\top]$, only the interaction within a local time-step would be captured.

At last, we conduct the pooling by summation on all $\mathbf{M}_t$'s to obtain the final fused features. If we define the factor matrices $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2 \ldots, \mathbf{z}_{T-k+1}]$ and $\mathbf{Z}^{\mathrm{e}} = [\mathbf{z}_1^{\mathrm{e}}, \mathbf{z}_2^{\mathrm{e}}, ..., \mathbf{z}_{T-k+1}^{\mathrm{e}}]$, then the fused feature matrix can be written as

$$\mathbf{M} = \sum_{t=1}^{N} \mathbf{M}_t = \mathbf{Z}\mathbf{Z}^{e,\top} \in \mathbb{R}^{L\times(kL+1)}, \tag{7}$$

where $N = T - k + 1$. In summary, the feature matrix $\mathbf{M}$ in our model reflects the interaction across not only multiple modalities but also time steps, which supplies more information to tackle the issue of imperfect data compared to T2FN [23].

However, it leads to troubles if we directly use $\mathbf{M}$ in Eq. (7) as the inputs of the sequential layers. It is because the size of $\mathbf{M}$ quadratically grows when increasing L, which equals the sum of dimension of features through all modalities. To alleviate such an issue, we will show in the following section that the acceptable feature size can be obtained by leveraging the inherent low-rank structure of $\mathbf{M}$.

3.2 Low-rank Inference Module

Below, we introduce the low-rank inference module (LIM), which not only maps the "fused features" to the final output but also avoids using numerous parameters in the model.

Consider a collection of affine functionals $g_i(\cdot)$, $i \in [d_o]$, each of which maps a feature matrix to a scalar. Given i, the functional can be thus written as

$$g_i\left(\mathbf{M}; \mathbf{W}_i, b_i\right) = \langle \mathbf{W}_i, \mathbf{M} \rangle + b_i, \tag{8}$$

where $\langle \cdot, \cdot \rangle$ denotes the trivial inner product, and $\mathbf{W}_i \in \mathbb{R}^{L\times(kL+1)}$ and b_i denote the weight and bias in the context of neural networks, respectively. Note that a large size of the feature matrix $\mathbf{M}$ also leads to the weight $\mathbf{W}_i$ with the same size, which would be unaffordable in practice. To tackle the issue, we decompose $\mathbf{W}_i$, $\forall i$ as the aforementioned CP format of the rank equaling R. As given in Eq. (1) and (2), the weights can be decomposed as

$$\mathbf{W}_i = \mathbf{W}_i^{(1)}\mathbf{W}_i^{(2),\top} = \sum_{r=1}^{R} \mathbf{w}_{i,r}^{(1)} \otimes \mathbf{w}_{i,r}^{(2)}, \tag{9}$$

where $\mathbf{w}_{i,r}^{(j)}$, $j = 1, 2$ denotes the r-th column of $\mathbf{W}_i^{(j)}$. Note from Eq. (7) and (9) that both $\mathbf{W}_i$ and $\mathbf{M}$ can be expressed by summation on outer products of

two vectors shaped $\mathbb{R}^L$ and $\mathbb{R}^{kL+1}$. Hence we modify Eq. (8) as

$$\begin{aligned}\widehat{y}_i = g_i\left(\mathbf{M}; \mathbf{W}_i, b_i\right) &= \left\langle \sum_{r=1}^{R} \mathbf{w}_{i,r}^{(1)} \otimes \mathbf{w}_{i,r}^{(2)}, \sum_{t=1}^{N} \mathbf{z}_t \otimes \mathbf{z}_t^e \right\rangle + b_i \\ &= \sum_{r=1}^{R}\sum_{t=1}^{N} tr\left((\mathbf{w}_{i,r}^{(1)} \otimes \mathbf{w}_{i,r}^{(2)})^\top (\mathbf{z}_t \otimes \mathbf{z}_t^e) \right) + b_i \\ &= \sum_{r=1}^{R}\sum_{t=1}^{N} (\mathbf{w}_{i,r}^{(1)\top} \cdot \mathbf{z}_t)(\mathbf{w}_{i,r}^{(2)\top} \cdot \mathbf{z}_t^e) + b_i \\ &= \left\langle \underbrace{\mathbf{W}_i^{(1),\top}\mathbf{Z}}_{R\times N}, \underbrace{\mathbf{W}_i^{(2),\top}\mathbf{Z}^e}_{R\times N} \right\rangle + b_i \end{aligned}, \tag{10}$$

where $tr(\cdot)$ denotes the trace function, and the equality in the third line is obtained due to the cyclic property. At last, $\widehat{\mathbf{y}}^\top = [\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_{d_o}]^\top$ gives the final output of the proposed network. As shown in Eq. (10), $\widehat{y}_i$ is obtained by calculating the inner product between $\mathbf{W}_i^{(1),\top}\mathbf{Z}$ and $\mathbf{W}_i^{(2),\top}\mathbf{Z}^{\mathrm{e}}$, of which the size $R \times N$ would generally far smaller than the size $L \times (kL+1)$ *w.r.t.* $\mathbf{M}$, especially when L is large (it usually happens when dealing with multimodal data including visual and linguistic features). As for the computational complexity, Eq. (10) results in $\mathcal{O}(kRNL)$, while totally $\mathcal{O}(kNL^2)$ is needed if we directly calculate the feature matrix $\mathbf{M}$ and use Eq. (8) to obtain the output. In the experimental section, we will also empirically prove that the proposed TPFN incorporates more interactions across the temporal domain yet with fewer parameters due to our low-rank inference module.

3.3 Low-rank Regularization

As mentioned in T2FN [23], the existence of missing values would increase the rank of the fused feature matrix $\mathbf{M}$. Inspired by the claim, we also impose the low-rank regularization into the model to handle the issue of imperfect data. However, unlike regularizing the Frobenius norm of $\mathbf{M}$ directly, we argue that the rank can be bounded by the norm of its latent factor matrix $\mathbf{Z}$.

To do so, we first introduce the key lemma, which appears as Lemma 1 in [38] and is popularly used in collaborative filtering [47]. Assuming a matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, we have

Lemma 1 (from [38]) *Assume matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, which can be represented by arbitrary decomposition $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$, then we have*

$$\|\mathbf{X}\|_* = \min_{\mathbf{U},\mathbf{V}} \|\mathbf{U}\|_F \|\mathbf{V}\|_F, \quad s.t.\, \mathbf{X} = \mathbf{U}\mathbf{V}^\top. \tag{11}$$

Lemma 1 implies that the nuclear norm of the matrix $\mathbf{X}$ is upper bounded by the product of Frobenius norm of its factor matrices. Using Lemma 1, we propose

the claim that the nuclear norm of $\mathbf{M}$ defined in Eq. (7) is upper bounded by the Frobenius norm of $\mathbf{Z}$. Specifically,

Claim 1 *Define the matrices* $\mathbf{M}$ *and* $\mathbf{Z}$ *as above, then the following inequality holds:*

$$\|\mathbf{M}\|_* \leq \sqrt{N + k\|\mathbf{Z}\|_F}\|\mathbf{Z}\|_F. \tag{12}$$

The proof is trivial by exploiting the relation $\|\mathbf{Z}^{\mathrm{e}}\|_F^2 \leq N + k\|\mathbf{Z}\|_F^2$. Since the matrix nuclear norm is the convex envelope of matrix rank [37], Claim 1 implies that the rank of the feature matrix $\mathbf{M}$ would be "controlled" by the Frobenius norm of its factor matrix $\mathbf{Z}$. More importantly, Claim 1 allow us to avoid explicitly calculating the fused matrix $\mathbf{M}$ for the low-rank regularization yet the model still results in the robustness against the imperfect data. Therefore, in our model we multiply $\|\mathbf{Z}\|_F$ with a norm factor λ (tuning parameter) and add it to the loss function as regularization to train the network.

4 Experiments

Dataset. We evaluate our method on two datasets: CMU-MOSI [53] and CMU-MOSEI [54]. *CMU-MOSI* is a multimodal sentiment analysis dataset containing 93 videos, which are then split into 2,199 short video clips by sentence. *CMU-MOSEI* is a dataset that can be applied to both emotion recognition and sentiment analysis, and it is the largest dataset on multimodal sentiment analysis at present. It contains 23453 labeled movies collected from 1000 different speakers in YouTube, covering 250 hot topics. CMU-MOSI and CMU-MOSEI datasets we use are pre-trained by the methods in [9] and [54], respectively [4]. Dataset statistics are shown in Table 2.

Table 2. Dataset statistics of **CMU-MOSI** and **CMU-MOSEI**. Number of samples and size of features for each modality are listed.

	Number of Samples			Size of Features		
	Train	**Val**	**Test**	**Acoustic**	**Visual**	**Language**
MOSEI	15,290	2,291	4,832	74	35	300
MOSI	1,284	229	686	5	20	300

Incompleteness modelling. Like [23], we exploit two strategies for dropping the data to simulate the incompleteness, *i.e.* random drop and structured drop. Given the missing percentage $p \in \{0.0, 0.1, ..., 0.9\}$, we *i.i.d.* drop the entries of the data at random for the former, and randomly remove the whole time step for the latter.

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

Table 3. Comparison of classification accuracy (ACC-2) and the total number of parameters used in the model for the *CMU-MOSI* task. **RD** and **SD** represent the random and structural drop task, respectively. The "**Low**", "**Medium**" and "**High**" columns correspond the missing percentage equaling 0.1, 0.5, 0.9, respectively. The "**Params**" column shows the number of parameters used in the methods.

Task	Method	Low	Medium	High	Params
RD	TFN	0.7361	0.7172	0.4475	759,424
	LMF	0.7346	0.7317	0.5218	**2,288**
	HPFN	0.7565	0.6982	0.5568	4,622,039
	T2FN	0.7769	0.7113	0.5962	19,737
	TPFN/reg(ours)	0.7638	0.7594	0.5845	8,652
	TPFN(ours)	**0.7915**	**0.7609**	**0.6559**	19,488
SD	TFN	0.7317	0.6880	0.5758	390,784
	LMF	0.7346	0.7128	0.5976	**792**
	HPFN	0.7463	0.7186	0.6151	1,168,247
	T2FN	0.7478	0.7142	0.6137	19,737
	TPFN/reg(ours)	**0.7682**	0.7288	0.6151	11,360
	TPFN(ours)	0.7594	**0.7434**	**0.6516**	7,344

General setting. We select the window size k from $\{1, 2, 3, 4, 5\}$, and keep the stride equalling 1. The CP-rank r is tuned from $\{4, 8, 12, 16, 24, 32\}$ and regularization parameter λ (if has) is tuned from $\{0.0, 0.0001, 0.001, 0.003, 0.01\}$. The hidden size of the pre-process LSTM is selected from $\{8, 16\}$, $\{4, 8, 16\}$, $\{64, 128\}$ for acoustic, visual and language, respectively. We train our method for 200 epochs in all experiments and employ early stop when the model does achieve the minimum loss on valid set for over 20 times. Adam Optimizer is used in our paper and the learning rate is tuned from $\{0.0003, 0.001, 0.003\}$.
Goal. The aim of our experiments include two aspects: First, we demonstrate the effectiveness of our methods on the incomplete multimodal sentiment analysis task, comparing with the results by the current state-of-the-art (SOTA) approaches; Second, we discuss the impact of tuning parameters such as the window size k, regularization parameter λ and the rank r on performance.

4.1 Performance on MOSI and MOSEI

For comparison, we implement TFN [49], LMF [26], HPFN [19] and T2FN [23] as baselines to evaluate the performance of our TPFN method. We also show the performance of TPFN without regularization (TPFN/reg), demonstrating that the low-rank regularization does improve the performance on the incomplete data. Table 3 and 4 show the classification accuracy (ACC-2) of the methods on CMU-MOSI and CMU-MOSEI, respectively. We select the missing percentage $p = 0.1, 0.5, 0.9$ to represent the low, medium and high incompleteness strength, respectively. Also, we show in Fig. 3 the performance change of our method under a full range of missing percentage p on CMU-MOSI dataset.
Results on accuray. Overall, our methods obtain the superior performance among all methods. Similar results goes in CMU-MOSEI. Fig. 3 also shows that

Table 4. Comparison of classification accuracy (ACC-2) and the total number of parameters used in the model for the *CMU-MOSEI* task. **RD** and **SD** represent the random and structural drop task, respectively. The "**Low**", "**Medium**" and "**High**" columns correspond the missing percentage equaling 0.1, 0.5, 0.9, respectively. The "**Params**" column shows the number of parameters used in the methods.

Task	Method	Low	Medium	High	Params
RD	TFN	0.7195	0.7193	0.6705	1,353,856
	LMF	0.7307	0.7233	0.6684	**1,208**
	HPFN	0.7371	0.7189	0.7119	1,295,895
	T2FN	0.7394	**0.7382**	0.7104	18,785
	TPFN/reg(ours)	0.7375	0.7297	0.7156	14,240
	TPFN(ours)	**0.7411**	0.7367	**0.7334**	16,842
SD	TFN	0.7295	0.7121	0.6968	759,424
	LMF	0.7313	0.7067	0.7057	**1,304**
	HPFN	0.7311	0.7245	0.7003	1,296,423
	T2FN	0.7350	0.7295	0.7173	9,945
	TPFN/reg(ours)	**0.7437**	0.7301	0.7007	7,056
	TPFN(ours)	0.7386	**0.7382**	**0.7301**	5,796

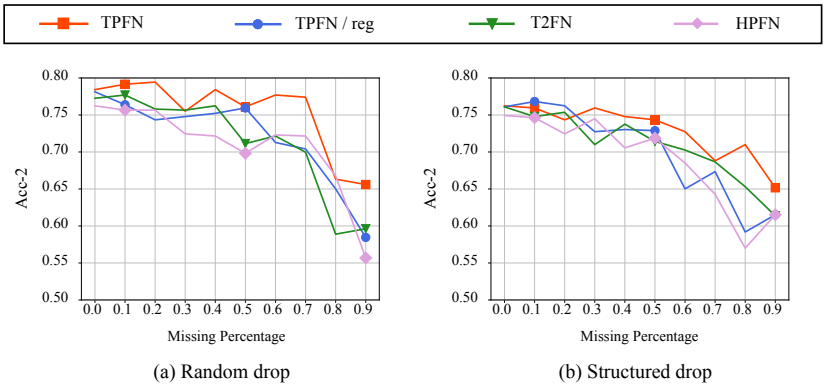

Fig. 3. Classification accuracy on CMU-MOSI in the full range of the missing percentage. The marker points corresponds to the results shown in Table 3.

our method can maintain a relatively stable performance as missing percentage increases. Meanwhile, it is shown that the low-rank regularization is helpful for the task on incomplete multimodal data since TPFN without the reg. term performs worse than the one equipped with the regularization. We can also see from the "**Params**" that TFPN uses less number of parameters than T2FN in the experiment. Although LMF used the least number of parameters, our methods significantly outperforms LMF specifically when the missing percentage is high.

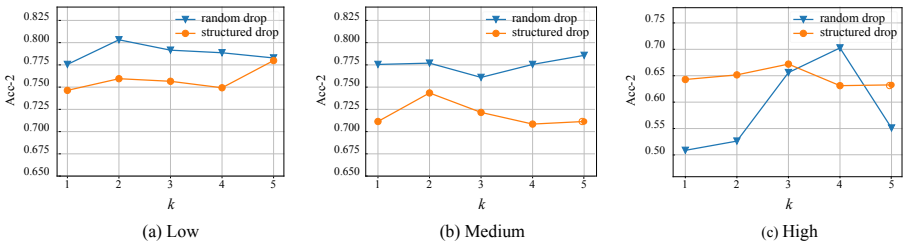

Fig. 4. Comparison on CMU-MOSI as k varies in the random drop task and the structured drop task.

4.2 Effect of Time Window Size k

To investigate how the performance changes as the time window size k varies, we keep all parameters in Sec. **4.1** except for k, and conduct the random and structured drop task on CMU-MOSI. The results are given in Fig. 4.

It can be observed when the missing percentage p is small, that the accuracy floats within a limited range as k varies in $[1, 2, 3, 4, 5]$, which indicates that our method is relatively robust with respect to k values in those cases. While in the case with high missing percentage, a suitable value of k can remarkably improve the performance in both tasks. We infer that choosing suitable k can help the method to"see" more information when the modalities get sparse. Assume that $\mathbf{a}_t$ is missing yet $\mathbf{v}_t$ and $\mathbf{l}_t$ are not, setting $k = 1$ implies that the visual and language modality can hardly interact with the acoustic one in the t-th step, while a bigger k incorporates additional information of modalities from the adjacent time-steps (such as $\mathbf{a}_{t+1}$), allowing exploring the inter-modality dynamics. Generally, $k = 1$ means the dynamics across time is neglected, while more interactions would be taken into account for a larger k. Note that the performance tends to degradation when the given k is too large as shown in Fig. 4. We conjecture that the domination by dynamics across time weakens the local dynamics. As Eq. (6) shows, only $\mathbf{z}_\mathrm{t} \bigotimes 1$ and $\mathbf{z}_\mathrm{t} \bigotimes \mathbf{z}_\mathrm{t}$ contain the interaction within time-steps. When k increasing, more correlations between series are involved, leading to the weakness on local dynamics.

4.3 Effect of Regularization Parameter λ

We keep all parameters used in Sec. **4.1** unchanged except for λ, and conduct the random and structured drop task on CMU-MOSI. Results are shown in Fig. 5.

As shown in Fig. 5, TPFN is stable with the change of λ, yet a suitable value of λ is indeed able to improve the performance compared to the one without the low-rank regularization ($\lambda = 0$). As our aforementioned discussion, the regularization on $\|\mathbf{Z}\|_F$ is able to bound the low-rank structure of the fused features, and therefore results in the "restored" features, which is closer to the clean features than their incomplete counterparts. However, when λ gets too large, the

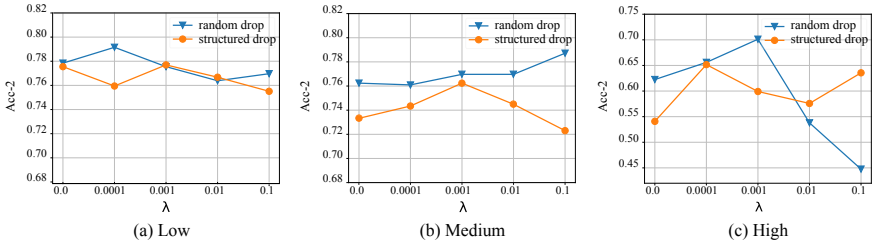

Fig. 5. Comparison on CMU-MOSI as λ varies in the random drop task and the structured drop task.

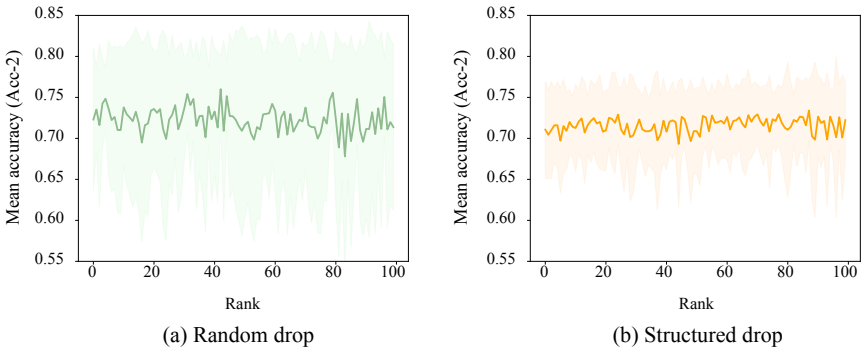

Fig. 6. Performance on CMU-MOSI as the rank r varies. The results are obtained by averaging the all missing percentage and varying the CP-rank r from 1 to 100.

loss function would be dominated by the regularization terms, and therefore degrades the performance as the result. Note that the significant improvement by the regularization term appears when the missing percentage is high. It is because the low-rank prior could guide the method to revise the bias by the severe incompleteness of the features.

4.4 Discussion on CP-rank

To discuss how the CP-rank effects the performance, we also keep all parameters in Sec. **4.1** except for r, and conduct the random and structured drop tasks on CMU-MOSI. We use the mean value of accuracy obtained from 10 different missing percentages to illustrate the overall performance, and the experimental results are shown in Fig. 6, where the CP-rank r varies from 1 to 100.

It is shown that TPFN is robust with respect to the CP-rank r for both the random and structural drop tasks. In other words, the performance seems not to be remarkably influenced as the rank increases. The similar phenomenon has

been discussed in the previous study on LMF [26]. More supplemental materials and codes are available in the webpage https://qibinzhao.github.io.

5 Conclusions

The main focus of this paper is on a new fusion strategy for multimodal sentiment analysis with the imperfect data. Compared to the existing outer-product-based fusion methods, the proposed TPFN can capture the high-order dynamics along the temporal domain by applying the outer product within time windows. Additionally, in the low-rank inference module, our method achieves the competitive performance with less parameters than T2FN [23]. Also, we have introduced a new low-rank regularization for the model. In contrast to T2FN, we have claimed that the Frobenius norm regularization on the factor matrix can obtain a low-rank fused feature matrix. In the experiments, We have not only shown that the proposed method outperforms the state-of-the-arts, but also further discussed how the window size k, the regularization parameter λ and the CP-rank r affect the performance, showing that a moderate determination of tuning parameters are helpful for the task with the incomplete data.

Acknowledgment. Binghua and Chao contributed equally. We thank our colleagues Dr. Ming Hou and Zihao Huang for discussions that greatly improved the manuscript. This work was partially supported by the National Key R&D Program of China (No. 2017YFE0129700), the National Natural Science Foundation of China (No. 61673224) and the Tianjin Natural Science Foundation for Distinguished Young Scholars (No. 18JCJQJC46100). This work is also supported by JSPS KAKENHI (Grant No. 20H04249, 20H04208, 20K19875).

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proc. ICCV (2015)
2. Barezi, E.J., Fung, P.: Modality-based factorization for multimodal fusion. arXiv preprint arXiv:1811.12624 (2018)
3. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: Proc, ICCV (2017)
4. Ben-Younes, H., Cadene, R., Thome, N., Cord, M.: Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In: Proc. AAAI (2019)
5. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**(4), 335 (2008)
6. Cai, L., Wang, Z., Gao, H., Shen, D., Ji, S.: Deep adversarial learning for multi-modality missing data completion. In: Proc. SIGKDD (2018)
7. Cambria, E., Poria, S., Bajpai, R., Schuller, B.: Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In: Proc. COLING (2016)

8. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. Psychometrika **35**(3), 283–319 (1970)
9. Chen, M., Wang, S., Liang, P.P., Baltrušaitis, T., Zadeh, A., Morency, L.P.: Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: Proc. ICMI (2017)
10. Cichocki, A., Lee, N., Oseledets, I., Phan, A.H., Zhao, Q., Mandic, D.P., et al.: Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions. Foundations and Trends® in Machine Learning **9**(4-5), 249–429 (2016)
11. Dong, J., Zheng, H., Lian, L.: Low-rank laplacian-uniform mixed model for robust face recognition. In: Proc. CVPR (2019)
12. Duong, C.T., Lebret, R., Aberer, K.: Multimodal classification for analysing social media. arXiv preprint arXiv:1708.02099 (2017)
13. Fan, H., Chen, Y., Guo, Y., Zhang, H., Kuang, G.: Hyperspectral image restoration using low-rank tensor recovery. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **10**(10), 4589–4604 (2017)
14. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proc. CVPR (2017)
15. Gu, Y., Li, X., Chen, S., Zhang, J., Marsic, I.: Speech intention classification with multimodal deep learning. In: Canadian Conference on Artificial Intelligence. pp. 260–271 (2017)
16. Guo, J., Zhou, Z., Wang, L.: Single image highlight removal with a sparse and low-rank reflection model. In: Proc. ECCV (September 2018)
17. Harshman, R.A., et al.: Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. UCLA Working Phonetics Paper (1970)
18. He, W., Yao, Q., Li, C., Yokoya, N., Zhao, Q.: Non-local meets global: An integrated paradigm for hyperspectral denoising. In: Proc. CVPR (2019)
19. Hou, M., Tang, J., Zhang, J., Kong, W., Zhao, Q.: Deep multimodal multilinear fusion with high-order polynomial pooling. In: Proc. NeurIPS (2019)
20. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proc. MM (2014)
21. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM review **51**(3), 455–500 (2009)
22. Li, C., He, W., Yuan, L., Sun, Z., Zhao, Q.: Guaranteed matrix completion under multiple linear transformations. In: Proc. CVPR (2019)
23. Liang, P.P., Liu, Z., Tsai, Y.H.H., Zhao, Q., Salakhutdinov, R., Morency, L.P.: Learning representations from imperfect time series data via tensor rank regularization. arXiv preprint arXiv:1907.01011 (2019)
24. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Mining text data, pp. 415–463. Springer (2012)
25. Liu, H., Lin, M., Zhang, S., Wu, Y., Huang, F., Ji, R.: Dense auto-encoder hashing for robust cross-modality retrieval. In: Proc. MM (2018)
26. Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., Morency, L.P.: Efficient low-rank multimodal fusion with modality-specific factors. arXiv preprint arXiv:1806.00064 (2018)
27. Lu, C., Peng, X., Wei, Y.: Low-rank tensor completion with a new tensor nuclear norm induced by invertible linear transforms. In: Proc. CVPR (2019)

28. Mai, S., Hu, H., Xing, S.: Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In: Proc. ACL (2019)
29. Miech, A., Laptev, I., Sivic, J.: Learning a text-video embedding from incomplete and heterogeneous data. arXiv preprint arXiv:1804.02516 (2018)
30. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
31. Morency, L.P., Mihalcea, R., Doshi, P.: Towards multimodal sentiment analysis: Harvesting opinions from the web. In: Proc. ICMI (2011)
32. Nimishakavi, M., Jawanpuria, P.K., Mishra, B.: A dual framework for low-rank tensor completion. In: Proc. NeurIPS (2018)
33. Pan, Y., Xu, J., Wang, M., Ye, J., Wang, F., Bai, K., Xu, Z.: Compressing recurrent neural networks with tensor ring for action recognition. In: Proc. AAAI (2019)
34. Pang, B., Lee, L., et al.: Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval **2**(1–2), 1–135 (2008)
35. Pham, H., Liang, P.P., Manzini, T., Morency, L.P., Póczos, B.: Found in translation: Learning robust joint representations by cyclic translations between modalities. In: Proc. AAAI (2019)
36. Poria, S., Cambria, E., Winterstein, G., Huang, G.B.: Sentic patterns: Dependency-based rules for concept-level sentiment analysis. Know.-Based Syst. **69**, 45–63 (2014)
37. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM review **52**(3), 471–501 (2010)
38. Srebro, N., Shraibman, A.: Rank, trace-norm and max-norm. In: Proc. COLT (2005)
39. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. CL **37**(2), 267–307 (2011)
40. Tran, L., Liu, X., Zhou, J., Jin, R.: Missing modalities imputation via cascaded residual autoencoder. In: Proc. CVPR (2017)
41. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika **31**(3), 279–311 (1966)
42. Wang, A., Li, C., Jin, Z., Zhao, Q.: Robust tensor decomposition via orientation invariant tubal nuclear norms. In: Proc. AAAI (2020)
43. Wang, H., Meghawat, A., Morency, L.P., Xing, E.P.: Select-additive learning: Improving generalization in multimodal sentiment analysis. In: Proc. ICME (2017)
44. Wang, Y., Shen, Y., Liu, Z., Liang, P.P., Zadeh, A., Morency, L.P.: Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In: Proc. AAAI (2019)
45. Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., Morency, L.P.: Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intell. Syst. **28**(3), 46–53 (2013)
46. Yang, Y., Krompass, D., Tresp, V.: Tensor-train recurrent neural networks for video classification. In: Proc. ICML (2017)
47. Yu, K., Zhu, S., Lafferty, J., Gong, Y.: Fast nonparametric matrix factorization for large-scale collaborative filtering. In: Proc. SIGIR (2009)
48. Yuan, L., Li, C., Mandic, D., Cao, J., Zhao, Q.: Tensor ring decomposition with rank minimization on latent space: An efficient approach for tensor completion. In: Proc. AAAI (2019)
49. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250 (2017)

50. Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E., Morency, L.P.: Memory fusion network for multi-view sequential learning. In: Proc. AAAI (2018)
51. Zadeh, A., Liang, P.P., Poria, S., Vij, P., Cambria, E., Morency, L.P.: Multi-attention recurrent network for human communication comprehension. In: Proc. AAAI (2018)
52. Zadeh, A., Zellers, R., Pincus, E., Morency, L.P.: Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259 (2016)
53. Zadeh, A., Zellers, R., Pincus, E., Morency, L.P.: Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intell. Syst. **31**(6), 82–88 (2016)
54. Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proc. ACL (2018)