Department: Affective Computing and Sentiment Analysis
Editor: Erik Cambria, Nanyang Technological University

Adaptive Modality Distillation for Separable Multimodal Sentiment Analysis

Wei Peng
University of Oulu

Xiaopeng Hong
Xian Jiaotong University

Guoying Zhao
University of Oulu

***Abstract*—Multimodal sentiment analysis has increasingly attracted attention since with the arrival of complementary data streams, it has great potential to improve and go beyond unimodal sentiment analysis. In this paper, we present an efficient separable multimodal learning method to deal with the tasks with modality missing issue. In this method, the multimodal tensor is utilized to guide the evolution of each separated modality representation. To save the computational expense, Tucker decomposition is introduced, which leads to a general extension of the low-rank tensor fusion method with more modality interactions. The method, in turn, enhances our modality distillation processing. Comprehensive experiments on three popular multimodal sentiment analysis datasets, CMU-MOSI, POM, and IEMOCAP, show a superior performance especially when only partial modalities are available.**

SENTIMENT ANALYSIS is the study of opinions, emotions, appraisals, and attitudes towards different entities, e.g., objects or people. It is a potential but challenging research topic in affective computing [1]. Inspired by the humans cognition, which generally captures information from multiple perception ways, sentiment analysis with Multimodal Learning (MML) [2] has emerged to endow the computational agent the same ability.

Instead of only utilizing single input stream, MML leverages the complementary information from multiple modalities. Representing multimodal data, however, is not easy since both intra-modal and cross-modal dynamics should be well modeled to avoid errors in the final prediction [3], [4]. Benefited from deep neural networks, tensor-based approaches [5], [6], [7] for multimodal feature representations have already achieved superior performance.

Previous tensor-based methods [8], [9] compute the interactions of every dimension of each unimodal representation so that the interactions between modalities are fully involved. The output compact tensor is able to model the multimodality representation very well. However, these approaches are always computational expensive and will become invalid once one or more modalities are missing.

In this regard, we propose a computationally efficient method which is robust to the modality missing issues for sentiment analysis. Specifically, we provide a modality distillation method based on tensor fusion network [9], in which the learning from privileged information mechanism [10] is introduced. On one hand it builds a compact tensor to capture the intra-modality and cross-modality dynamics while the modalities are available.

On the other hand, complementary information is distilled from this tensor for each unimodal representation such that the representation for available modalities could also capture the interactions even one or more modalities are missing. Based on the message passing between individual modality representation and the multimodal tensor, we adaptively adjust the temperature of modality distillation. As this compact multimodal tensor will lead to a higher-order weight tensor for further prediction, we introduce Tucker decomposition for the weight tensor. Interestingly, this decomposition provides modality-specific weight for each modality.

Thus, modality-specific components could be captured from the multimodal tensor, which in return enhances the distillation progress. Finally, we conduct extensive experiments on three multimodal sentiment analysis datasets. Results show that our model has a comparable or even superior performance when all the modalities are available, and more importantly, it is still robust and gets overwhelming results when partial modalities are missing. Our contributions are as follows:

- We propose a novel strategy based to deal with sentiment analysis tasks with modality missing issue. This is the first separable tensor network with knowledge distillation.
- To improve the computational efficiency, we introduce Turker decomposition for the weight tensor, which also provides the modality-specific weight so that it enhances the distillation processing.
- Extensive experiments on three real-life multimodal sentiment analysis datasets show that our model gets comparable performance when all modalities available and superior results when one or more modalities are missing.

Related work

Sentiment analysis has been one of the most active research topics of affective computing in the last decade [11]. It is a valuable and potential task with various real-world applications, including financial market prediction, business review analysis, and even politics. According to [12], sentiment analysis tasks can be categorized into two parts, i.e., basic sentiment analysis tasks and sub-categories of the major tasks. From the method perspective, deep learning-based methods are becoming very popular for sentiment analysis due to their high performance in recent times. From the modality perspective, there have been multiple modalities which can be utilized for multimodal sentiment analysis. For instance, Liu et al. described a tensor networks for sentiment recognition from three modalities, including text, video and audio [13]. [14] uses deep model for emotion classification, with multiple input streams containing EEG and peripheral physiological signals. However, very few works explore the multimodal sentiment analysis task with modality missing issues, since it is much more challenging.

A general strategy is to infer the missing modality from the existing ones by modeling the probabilistic relationships among them [15]. Inspired by the insight that translation from a source to a target modality provides a method of learning joint representations using only the source modality as input, MCTN [16] learns robust joint representations by translating between modalities. Tsai et al. [17] proposed a factorized learning method by dividing representation into multimodal discriminative factors and modality-specific generative factors, which can deal with missing modality issues as well. Instead of utilizing a reconstruction method, we aim at a distillation method since reconstruction itself is a hard task.

Compared with traditional early- or late- level fusion, tensor-based methods [9], [13] are better approaches to model the intra- and inter-modal dynamics. They construct tensor by calculating the interaction of every dimension in each modality based on Cartesian-products. Meanwhile, one of the main drawback of tensor-based method is its high-dimensional property. Liu et al. [13] proposed a low-rank tensor-based method (LMF) for feature fusion. However, LMF does not deal with missing modality issues and it gives too much constraints for each tensor, e.g., requires that every modality representation shares the same length. Here, we introduce Tucker decomposition to reduce the computational expense and LMF tensor fusion network is actually a special case of our method.

Methodology

We form the problem as a separable multimodal learning issue with privileged information. Consider a machine learning processing with the training data formulated by a collection of triplets $\{(x_1, x_1^*, y_1), ..., (x_n, x_n^*, y_n)\}$, where the (x_i, y_i) is the data-label pair commonly used in supervised learning tasks and can be accessed during the whole learning procedure. x_i is the input data (feature) and y_i is its label. The novel element x_i^* represents the heterogeneous data of x_i with different modalities and it is missing during the inference procedure. Thus, x_i^* works as the privileged modalities to support the learning process. Then our goal is to find the best function f_s from the function space $\mathcal{F}_s$,

$$f_s = \underset{f_s \in \mathcal{F}_s}{\operatorname{argmin}} \, l(\sigma(f_s(x_i)), y_i, \mathcal{I}((x_i, x_i^*))), \quad (1)$$

where l is a classification loss function. σ works as the prediction layer, e.g., commonly used fully connected layer. $\mathcal{I}$ is a function to get side information. Like in [10], in analogy to a good Bayesian prior, the teacher $\mathcal{I}$ provides an opportunity to learn characteristics about the decision boundary which is not contained in the training samples. To simplify the formulation step, let x represent x_i and assume there is only one modality in it. This can be easily generalized to issue with more than one modality. Thus, its privileged information x^* contains $M-1$ modalities, here $M > 1$ is the total number of modalities.

Figure 1 represents the overall framework of our method. We build an independent model for each input modality and compute the interaction tensor $\mathcal{Z}$ (Multimodal Tensor in Figure 1) based on each modality representation. Instead of using the tensor directly for prediction, it is used to guide the learning process of single modality representation (z^m). This way, we expect that z^m could capture the complementary knowledge. So:

$$\begin{aligned} f_s = \underset{f_s \in \mathcal{F}_s}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} [l(y, \sigma(f_s(x))) \\ + \alpha l(s, \sigma(f_s(x))) + \beta l^*(z^*, f_s(x))], \end{aligned} \quad (2)$$

where n is the number of training samples, l^* is a loss function for tensor representation. α and β are the hyper-parameters used to balance the importance of each part. s is the soft label vector is the soft complementary tensor,

$$s = \hat{\sigma}(\mathcal{Z}/T_s), \quad z^* = f_p(\mathcal{Z}/T_z), \quad (3)$$

of which the $\hat{\sigma}$ is another prediction function and f_p is a project function that makes z^* sharing the same size with $f_s(x)$. Like [9], $\mathcal{Z}$ is a multimodal tensor computed by Cartesian-products. Temperature parameters T_s and T_z control how much to soften or smooth the class probability predictions and the complementary information, respectively. We alternatively update multimodal tensor and separated tensor. In this way, the separated modality representation can also capture the modality interaction information and thus could deal with missing modality cases.

Since different modalities may prefer different temperatures, instead of providing a soft target for the student network with the predefined learning directions, all the separated modality representations are dynamically learnt from the complementary tenor with adaptive temperatures based on the messages interaction between modalities. Suppose m_i is the message held by z^i and m_i is constructed based on z^i and its prediction,

$$m_i = f_{p_1}(\sigma(z^i)) || f_{p_2}(z^i). \quad (4)$$

Here the f_{p_1} and f_{p_2} are two project functions. They project the tensor prediction and representation into the same space. The $\cdot||\cdot$ indicates the concatenation operation.

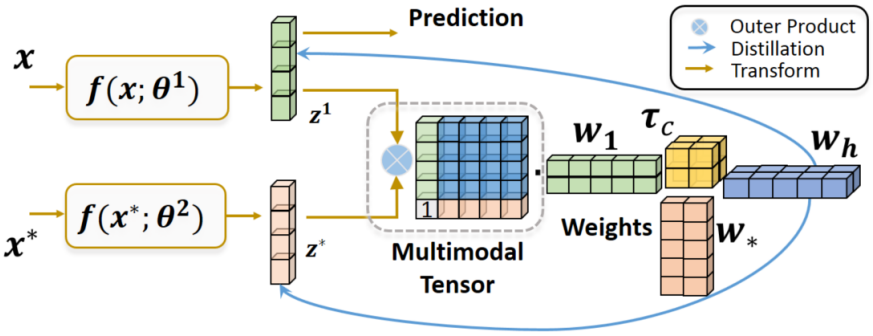

Figure 1. Illustration of the Framework. Since the 'Multimodal Tensor', including all modalities in training, is not available during inference, we build separable networks for different modalities and let the multimodal tensor dynamically guide each feature learning process. Then, the Tucker decomposition for the 'Weights' is introduced thus the heavy weight is replaced by $[\mathcal{T}_c; W_1, W_*, W_h]$ and the computational burden can be much lower.

Then the distillation temperature T is calculated by

$$T = f_{p_t}([m_i||m_{\mathcal{Z}}]), \tag{5}$$

where f_{p_t} is another projection function. $m_{\mathcal{Z}}$ is the message from the multimodal tensor. Both the temperatures T_s and T_z can be computed by this equation.

With the method formulated before, one could build a separable multimodal learning model with the capability of capturing modality interactions from the privileged information. However, the computational burden increases exponentially with the modality number and the dimension of each input representation. As the tensor $\mathcal{Z} \in \mathcal{R}^{d^1 \times d^2 \times \cdots \times d^M}$, one should apply a weight $\mathcal{W} \in \mathcal{R}^{d^1 \times d^2 \times \cdots \times d^M \times h}$ to map the tensor to a final representation vector $z_h \in \mathcal{R}^h$ and thus for the final prediction. So we get that

$$z_h = \mathcal{W} \cdot \bigotimes_{m=1}^{M} z^m. \tag{6}$$

The weight tensor is with order-(M+1). Thus, the computational cost for message $m_{\mathcal{Z}}$ is $\mathcal{O}(\prod_{m=1}^{M} d^m)$, which is expensive. Therefore, like [5], we conduct a Tucker decomposition for the weight matrix to improve the computational efficiency. Thus we get

$$z_h = ((\mathcal{T}_c \times_1 W_1) \times_2 \ldots \times_M W_M) \times_{M+1} W_h \cdot \bigotimes_{m=1}^{M} z^m \tag{7}$$

with $W_i \in \mathcal{R}^{d^i \times t^i}$, $W_h \in \mathcal{R}^{h \times t^{M+1}}$ and $\mathcal{T}_c \in \mathcal{R}^{\prod_{i=1}^{M+1} t^i}$ Here, operator $\times_i$ is the i-mode product.

This decomposition is very useful. Firstly, the computational expense now is $\mathcal{O}(\sum_{m=1}^{M} d^m + \prod_{m=1}^{M+1} t^m)$, which is much lower than previous one since $t^m << d^m$. Secondly, each weight component corresponds to a specific modality. Thus, modality-specific weight can be performed on each modality. That is

$$z_h = ((\mathcal{T}_c \times_1 (z_1^\top W_1) \ldots \times_M (z_M^\top W_M)) \times_{M+1} W_h. \tag{8}$$

Here, weight matrices project each modality to respective feature space and the weight tensor $\mathcal{T}_c$ is learned to capture the interaction between them. Once we get the the final representation z_h, we could perform the final prediction.

Note that weight tensor $\mathcal{T}_c$ gives every dimension at every modality a weight to compute the intra- and inter- connections. As a trade-off between model complexity and efficiency, one could add sparsity constraints to $\mathcal{T}_c$. For any element $z_h[k]$ at dimension $k \in \mathcal{R}^h$, it can be treated as a correlation between elements from the different modalities weighted by the tensor slice $\mathcal{T}_c[:, : \ \ldots \ :, k]$, of which the structure definitely can be constrained by the rank constraint.

As a special case, this method is able to degrade to Low-rank fusion method when set the weight tensor $\mathcal{T}_c$ to an identity, which means different modalities share the same scale of the dimension in respective feature space and the interaction between each modality is only allowed at the same dimension.

According to Eq (8),

$$z_h[k] = \sum_{i=1}^{R} \bigotimes_{m=1}^{M} W_{m,k}^{(i)} \cdot \bigotimes_{m=1}^{M} z^m = \sum_{i=1}^{R} W_{1,k}^{(i)} z^1 \circ \sum_{i=1}^{R} W_{2,k}^{(i)} z^2 \cdots \circ \sum_{i=1}^{R} W_{M,k}^{(i)} z^M \tag{9}$$

Here, the $\mathcal{W}$ is decomposed with the rank of R. The $W_{m,k}^{(i)}$ is the $i-th$ decomposition of the $k-th$ slice of the weight tensor $\mathcal{W}$ for $m-$th modality. $\circ$ means element-wise product. So, it is clear that low-rank fusion network [13] is a special case of this Tucker decomposition method. Therefore, in this way, we find a modality specific weight-tensor for each modality representation. Then, we go one big step further. We apply the modality-specific weight to construct the individual modality representation and the final representation vector z_h is used as the complementary information to lead the distillation process.

Experiments

We perform the following experiments on three real-world datasets for multimodal sentiment analysis task. They are CMU-MOSI [9], POM [20], and IEMOCAP [21]. Each dataset contains three modalities, which are videos, audio and text. Following [13], we perform exactly the same pre-processing for fair comparison. First, we test with the case that all the modalities are available and compare with the state of the arts. Then we show the robustness of our method on the cases with one or more modalities missing. Finally, we also explore the effectiveness of each parts in the modality distillation.

Implementation

To compare with previous state-of-the-art methods, we build simple and straightforward neural networks for these three modalities. Like [13], for the visual and acoustic features, we build a two-layer fully connected network to get the embedded modality representation. For text feature, we employ LSTM to capture the sequences information in the feature. After that, according to Eq (8), we build a tensor-based module to capture the interactions between modalities. To model the message between a specific modality representation and the multimodal tensor, we employ two fully connected layers on the modality representation and its prediction. The outputs are unified to 32 after this projection. Then they are concatenated according to Eq (4). The same operation is performed on the multimodal tensor. Eventually, the output vectors are concatenated and thus the adaptive temperature is computed based on Eq (5).

During the training, we train the model for 50 epochs with a patience of 10, which stops the training once the loss has not decreased for 10 epochs. We also perform a grid search to find the best model on the validation dataset. For the grid search, like in [13], we search the best size for the hidden representation for the three modalities, the network training related hyper-parameters like batch size, learning rate, dropout rate and weight decay. We also search the rank R of the Eq (9) in the set of [1, 4, 8, 16].

All Modalities Available Issues

Since all modalities are available, the multimodal tensor is used for final prediction. Here, we compare with eight state-of-the-art methods. SVM-based method works as the baseline, which applies a SVM [22] classifier directly on the concatenated features. DF [6] builds one network for each modality and combines all the outputs with a joint network for final prediction. BC-LSTM [8] and MV-LSTM [18] are based on LSTM. MARN [19] and MFN [7] are two attention-based methods, while MARN models modality interactions by using a multi-attention block with a hybrid memory and the MFN introduces an attention mechanism along time and captures interactions with a multi-view gated memory. TFN [9] and LMF [13] are two tensor-based methods, which are most related to our work. TFN creates a multi-dimensional tensor to capture the view-specific and cross-view dynamics and LMF performs a high-efficient multimodal fusion method with low-rank decomposition.

All the results are listed in the Table 1. There are 12 results from each method for these three datasets. Though our model is not designed directly for the complete information task, one can still find the comparable capability of our model. For instance, we get two best and two second best results on the IEMOCAP dataset. In all of these 12 results, 67% of our results can get the top-

Table 1. Performance comparison(all modalities are existing) on CMU-MOSI, POM, and IEMOCAP with eight current best results. Top-3 results are marked in bold. Note that our model is not directly designed for this multimodel fusion task. Nonetheless, our model gets a comparable or even superior performance when compared with them.

	CMU-MOSI					POM			IEMOCAP			
Method	MAE	Corr	Acc-2	F1	Acc-7	MAE	Corr	Acc	F1-Hp	F1-Sd	F1-Ag	F1-Nu
SVM-Based	1.864	0.057	50.2	50.1	17.5	0.887	0.104	33.9	81.5	78.8	82.4	64.9
DF [6]	1.143	0.518	72.3	72.1	26.8	0.869	0.144	34.1	81.0	81.2	65.4	44.0
BC-LSTM [8]	1.079	0.581	73.9	73.9	28.7	0.840	0.278	34.8	81.7	81.7	84.2	64.1
MV-LSTM [18]	1.019	0.601	73.9	74.0	33.2	0.891	0.270	34.6	81.3	74.0	**84.3**	66.7
MARN [19]	**0.968**	0.625	**77.1**	**77.0**	**34.7**	-	-	39.4	83.6	81.2	84.2	65.9
MFN [7]	**0.965**	**0.632**	**77.4**	**77.3**	**34.1**	**0.805**	**0.349**	**41.7**	**84.0**	82.1	83.7	**69.2**
TFN [9]	0.970	**0.633**	73.9	73.4	32.1	0.886	0.093	31.6	83.6	**82.8**	84.2	65.4
LMF [13]	**0.912**	**0.668**	**76.4**	**75.7**	32.8	**0.796**	**0.396**	**42.8**	**85.8**	**85.9**	**89.0**	**71.7**
Ours	0.996	0.606	74.0	74.0	**34.3**	**0.836**	**0.313**	**36.2**	**86.0**	**85.9**	**88.7**	**72.3**

3 performance. All of our results can achieve a top-5 results. Overall, our model is comparable with the state-of-the-arts when all modalities are available.

Modality Missing Issues

Here, we evaluate our model on the cases with one or more modalities missing. As there is little works in this case, we implement four approaches here, including two traditional methods (SVM and Sparse Tikhonov-Regularized Hashing (STRH) [23]) and two deep models (Deep-S and LMF-S). SVM gives a prediction directly on the available modalities. Likewise, Deep-S also directly builds the network for the available modalit(ies). For fair comparison, we keep all other neural modules consistent with our model excepts for the distillation related ones. STRH enforces both the 0-norm induced sparsity constraints and the Tikhonov regularization on the binary solution vectors to maximize cross-modal correlation. Since LMF is a special case of our model, we extend LMF to a separated one, LMF-S. We keep same depth for each sub-network for fair comparison. Likewise, we also conduct score-level fusion for the final output for cases with more than one modality available.

It can be seen from Tables 2 and 3 that, for the case with only one modality, our method gets a overwhelming performance. For instance, On the dataset of CMU-MOSI, when there is only text information, with the modality distillation, our model achieves the best results. This performance is even comparable to the cases where all modalities are available, which shows the robustness of our model. For the cases with two modalities, our method can also achieve superior performance. For example, on the IEMOCAP datatset, when both audio and text data available, our model gets F1-Ag with 87.1%, which is even better than most methods in the Table 1 with all the modality available. However, the SVM and LMF-S methods only achieve 72.6% in the same case. Meanwhile, there are a few failed cases(5/72), like F1-Nu value with visual and audio in the IEMOCAP dataset. It may be caused by the fusion catastrophe, which is a common issue in MML. Nonetheless, for most of the cases, we could get a superior performance.

Ablation study

We also perform corresponding ablation study to investigate each parts of adaptive modality distillation. Firstly, we investigate the effectiveness of the distillation method in our model. In fact, the framework LMF-S is equal to our model without the distillation mechanism. Thus it can be used to evaluate the effectiveness of our modality distillation method. Then, we evaluate the influence of the adaptive temperature mechanism. To this end, we remove the message-interaction based distillation by giving a fixed temperature (2) for all the distillation processing. This method is referred as Our-fixed. All the comparison results are listed in the Table 3.

When compared to the LMF-S, our performance is very competitive. For example, on the IEMCAP dataset, we get F1 score of 82.6 % for the class sad (F1-Sd) when there is only video data, while LMF-S only get 70.6%. One could find more examples in the Table 3. This proves the effectiveness of our distillation module. Then

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MIS.2021.3057757, IEEE Intelligent Systems

Table 2. Performance comparison(One modality is missing). The best results for different inputs are marked with: Visual+Audio, Audio+Text, and Visual+Text, respectively.

		CMU-MOSI					POM			IEMOCAP			
Methods	Available Modality	MAE	Corr	Acc-2	F1	Acc-7	MAE	Corr	Acc	F1-Hp	F1-Sd	F1-Ag	F1-Nu
SVM	Visual+Audio	1.553	0.006	51.7	12.6	16.3	0.951	0.054	29.7	73.4	73.8	67.8	55.2
	Audio+Text	1.982	0.116	53.5	11.6	14.7	1.038	0.059	30.9	78.9	76.8	72.6	51.3
	Visual+Text	1.977	0.091	54.1	19.7	15.7	1.047	0.053	30.1	75.4	72.9	66.5	56.1
STRH	Visual+Audio	1.479	0.009	53.5	17.4	16.7	0.929	0.032	34.2	71.0	78.8	71.9	52.3
	Audio+Text	1.447	0.089	51.3	18.9	19.5	0.912	0.013	32.7	79.3	79.6	78.6	54.2
	Visual+Text	1.420	0.005	51.9	31.4	18.7	0.923	0.012	32.5	73.6	71.3	73.5	56.2
Deep-S	Visual+Audio	1.429	0.113	44.7	27.6	15.7	1.100	0.141	29.7	79.4	73.8	67.8	**59.2**
	Audio+Text	1.311	0.313	56.5	47.7	21.6	1.155	0.159	31.3	78.9	76.8	72.6	58.5
	Visual+Text	1.323	0.331	56.5	47.7	21.3	1.124	0.166	31.1	79.4	72.9	66.5	64.1
LMF-S	Visual+Audio	1.467	0.074	44.7	47.4	18.5	0.988	0.120	32.6	81.0	78.8	81.9	54.5
	Audio+Text	**1.103**	0.541	68.9	69.0	29.0	0.903	0.171	32.9	79.3	79.6	83.6	64.4
	Visual+Text	1.165	0.527	68.2	68.2	28.5	0.904	0.183	33.1	82.6	81.0	82.4	63.9
Ours-fixed	Visual+Audio	1.457	0.070	51.6	50.5	17.6	0.878	0.209	34.3	79.7	82.2	84.8	49.5
	Audio+Text	1.091	0.547	69.2	70.1	30.9	0.881	0.247	33.1	80.7	82.3	85.7	65.6
	Visual+Text	1.076	0.531	67.9	67.9	27.8	0.887	0.230	32.8	81.6	82.3	86.4	63.5
Ours	Visual+Audio	**1.374**	**0.126**	**55.7**	**54.7**	**22.1**	**0.865**	**0.243**	**34.8**	**82.2**	**81.5**	**85.2**	54.2
	Audio+Text	1.117	**0.569**	**70.7**	**70.8**	**29.8**	**0.863**	**0.248**	**33.6**	**83.6**	**83.2**	**87.1**	**68.5**
	Visual+Text	**1.082**	**0.569**	**69.4**	**69.8**	**33.4**	**0.881**	**0.256**	**33.7**	**83.5**	**81.7**	**84.0**	**65.2**

Table 3. Performance comparison(Two modalities are missing). The best results are markedThe best results for different inputs are marked with: Visual, Audio, and Text, respectively.

		CMU-MOSI					POM			IEMOCAP			
Methods	Available Modality	MAE	Corr	Acc-2	F1	Acc-7	MAE	Corr	Acc	F1-Hp	F1-Sd	F1-Ag	F1-Nu
SVM	Visual	1.481	0.010	48.1	17.4	14.7	0.995	0.015	32.5	68.3	70.1	66.8	54.2
	Audio	1.595	0.007	46.9	13.7	16.0	0.893	0.047	33.7	67.9	72.1	71.9	51.3
	Text	1.983	0.091	54.2	17.9	15.1	0.999	0.051	30.1	72.9	75.6	65.3	41.4
STRH	Visual	1.458	0.004	52.5	11.3	16.3	0.939	0.023	34.0	70.2	61.3	57.3	51.4
	Audio	1.522	0.003	51.9	10.7	16.2	0.885	0.037	32.4	68.5	69.4	72.4	41.9
	Text	1.433	0.002	53.4	25.4	17.1	0.921	0.018	33.5	71.2	70.2	71.9	53.7
Deep-S	Visual	1.441	0.130	44.7	27.6	15.4	1.069	0.148	29.5	79.9	70.3	67.8	64.8
	Audio	1.417	0.095	44.8	27.6	16.0	1.131	0.134	29.9	78.9	78.1	79.9	53.7
	Text	1.206	0.532	68.3	67.9	27.2	1.180	0.184	32.7	78.9	75.6	65.3	63.4
LMF-S	Visual	1.398	0.066	55.2	**52.3**	**21.1**	1.050	0.104	29.8	80.4	70.6	66.8	54.3
	Audio	1.458	0.040	49.1	48.2	16.0	0.930	**0.151**	32.4	78.9	79.7	82.3	48.4
	Text	1.108	0.535	69.3	68.4	27.9	0.912	0.228	33.2	81.4	80.2	81.2	63.5
Ours-fixed	Visual	1.466	0.180	45.5	30.9	15.5	0.869	0.144	34.3	81.8	81.7	85.6	65.3
	Audio	1.475	0.075	51.6	49.9	18.9	0.867	0.204	33.8	82.4	80.9	86.3	64.8
	Text	1.084	0.541	70.3	70.4	30.4	0.912	0.218	33.8	83.6	83.2	85.9	64.9
Ours	Visual	**1.390**	**0.137**	**55.2**	50.7	18.2	**0.875**	**0.136**	**34.1**	**82.5**	**82.6**	**86.1**	**66.7**
	Audio	**1.377**	**0.160**	**56.7**	**56.2**	**22.7**	**0.879**	0.147	**33.2**	**83.2**	**81.4**	**85.7**	**66.1**
	Text	**1.074**	**0.570**	**71.8**	**70.9**	**30.5**	**0.884**	**0.273**	**35.4**	**84.1**	**82.0**	**86.2**	**66.9**

we compare with Our-fixed. From Table 3 we know that, more than 80% (70/84) results benefits from the adaptive temperature mechanism and have been improved, which definitely proves the effectiveness of our adaptive mechanism.

Conclusion

In this paper, a novel message-interaction adaptive modality distillation is proposed to deal with the multimodal sentiment analysis, which is an important research topic in affective computing. The proposed method is successfully applied to multimodal learning problems with modality missing issues, which are common cases in the real-life but rarely considered in previous works. The method constructs a separable tensor fusion network with learning from the privileged information, and provides an adaptive distillation strategy based on the modality messages. In this way, our method could successfully capture from even missing modalities the intra- and inter- modality interaction dynamics, thus improves the performance for modality missing issues. To further

enhance the computational efficiency, we perform a Turker decomposition for the weight tensor. Interestingly, we found that the famous low-rank tensor fusion method is a specific case of our model. Comprehensive experiments on three real-life multimodal sentiment analysis datasets prove that the proposed method is superior to the existing methods with modality missing issues and it is also comparable to the state-of-the-art approaches when all modalities are available.

ACKNOWLEDGMENT

This work is supported by the Academy of Finland for ICT 2023 project (grant 328115) and project MiGA (grant 316765) and Infotech Oulu. As well, the authors wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

REFERENCES

1. E. Cambria et al., "Affective Computing and Sentiment Analysis," A Practical Guide to Sentiment Analysis, Socio-Affective Computing, chap. 1, Springer, 2017, pp. 1–10.
2. S. Poria et al., "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intelligent Systems*, vol. 33, 2018, pp. 17–25.
3. I. Chaturvedi et al., "Fuzzy Commonsense Reasoning for Multimodal Sentiment Analysis," *Pattern Recognition Letters*, vol. 125, no. 264-270, 2019.
4. L. Stappen et al., "Sentiment Analysis and Topic Recognition in Video Transcriptions," *IEEE Intelligent Systems*, vol. 36, no. 2, 2021.
5. H. Ben-younes et al., "MUTAN: Multimodal Tucker Fusion for Visual Question Answering," *ICCV*, 2017, pp. 2631–2639.
6. B. Nojavanasghari et al., "Deep multimodal fusion for persuasiveness prediction," *ICMI*, 2016, pp. 284–288.
7. A. Zadeh et al., "Memory fusion network for multi-view sequential learning," *AAAI, vol. 32*, 2018, pp. 5634–5641.
8. A. Fukui et al., "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding," *EMNLP*, 2016, pp. 457–468.
9. A. Zadeh et al., "Tensor Fusion Network for Multimodal Sentiment Analysis," *EMNLP*, 2017, pp. 1103–1114.
10. D. Lopez-Paz et al., "Unifying distillation and privileged information," *ICLR*, 2015.
11. Y. Susanto et al., "Ten Years of Sentic Computing," *Cognitive Computation*, vol. 13, 2021.
12. A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," *Artificial Intelligence Review*, vol. 53, no. 6, 2020, pp. 4335–4385.
13. Z. Liu et al., "Efficient low-rank multimodal fusion with modality-specific factors," *ACL*, 2018, pp. 2247–2256.
14. S. Tripathi et al., "Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset," *AAAI*, 2017, pp. 4746–4752.
15. J. Hoffman, S. Gupta, and T. Darrell, "Learning with Side Information through Modality Hallucination," *In CVPR*, 2016, pp. 826–834.
16. H. Pham et al., "Found in translation: Learning robust joint representations by cyclic translations between modalities," *AAAI, vol. 33*, 2019, pp. 6892–6899.
17. Y. Tsai et al., "Learning factorized multimodal representations," *ICLR*, 2019.
18. S. Shyam et al., "Extending long short-term memory for multi-view structured learning," *ECCV, vol. 9911*, 2016, pp. 338–353.
19. A. Zadeh et al., "Multi-attention recurrent network for human communication comprehension," *AAAI, vol. 32*, 2018, pp. 5642–5649.
20. S. Park et al., "Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach," *ICMI*, 2014, pp. 50–57.
21. C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, 2008, pp. 335–359.
22. C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, 1995, pp. 273–297.
23. L. Tian et al., "Sparse Tikhonov-Regularized Hashing for Multi-Modal Learning," *IEEE ICIP*, 2018, pp. 3793–3797.

Wei Peng is currently a Ph.D. candidate with the Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland. He received the M.S. degree in computer science from the Xiamen University, Xiamen, China, in 2016. His current research interests include machine learning, affective computing, medical imaging, and human action analysis.

Xiaopeng Hong received the Ph.D. degree in computer application and technology from the Harbin Institute of Technology, Harbin, China, in 2010. He is currently an Associate Professor with Xian Jiaotong University, Xian, China. His current research interests include multi-modal learning, affective computing, intelligent medical examination, and humancomputer interaction, etc.

Guoying Zhao is the corresponding author and is currently a Professor with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. She received the Ph.D. degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005. Her current research interests include affective computing, facial-expression and micro-expression recognition, emotional gesture analysis, and human computer interaction.