Dual Low-Rank Multimodal Fusion

Tao Jin[1*], **Siyu Huang**[2*], **Yingming Li**[1†], **and Zhongfei Zhang**[3]
[1]College of Information Science & Electronic Engineering, Zhejiang University, China
[2]Baidu Research, China
[3]Department of Computer Science, Binghamton University, USA
{jint_zju,yingming}@zju.edu.cn, huangsiyu@baidu.com, zzhang@binghamton.edu

Abstract

Tensor-based fusion methods have been proven effective in multimodal fusion tasks. However, existing tensor-based methods make a poor use of the fine-grained temporal dynamics of multimodal sequential features. Motivated by this observation, this paper proposes a novel multimodal fusion method called Fine-Grained Temporal Low-Rank Multimodal Fusion (FT-LMF). FT-LMF correlates the features of individual time steps between multiple modalities, while it involves multiplications of high-order tensors in its calculation. This paper further proposes Dual Low-Rank Multimodal Fusion (Dual-LMF) to reduce the computational complexity of FT-LMF through low-rank tensor approximation along dual dimensions of input features. Dual-LMF is conceptually simple and practically effective and efficient. Empirical studies on benchmark multimodal analysis tasks show that our proposed methods outperform the state-of-the-art tensor-based fusion methods with a similar computational complexity.

1 Introduction

Multimodal fusion aims to integrate information of multiple modalities as a compact but informative representation. Multimodal fusion is fundamentally significant for real-world multimodal applications like speech translation (Yuhas et al., 1989), emotion recognition (De Silva et al., 1997; Chen et al., 2018), and sentiment analysis (Morency et al., 2011). It is very challenging that it requires correlating the semantics of multiple modalities in an effective and efficient way. Recently, several methods have been proposed to learn joint embeddings of multiple modalities (Fukui et al., 2016; Nojavanasghari et al., 2016; Zadeh et al., 2017). There are two lines of fusion methods: early fusion and late fusion. In this paper, we mainly focus on the former, which aims to integrate information of different modalities before it is processed by the model.

Earlier work on early fusion employs a simple concatenation of input features (Pérez-Rosas et al., 2013; Park et al., 2014; Zadeh et al., 2016b). To construct a more compact representation, (Zadeh et al., 2017) introduces Tensor Fusion Network (TFN) which averages the features of each modality along the temporal dimension and transforms the multimodal features into a high-order tensor which is used for subsequent tasks. Although TFN achieves a better performance than the concatenation manner, its computational complexity increases exponentially with the number of modalities. (Liu et al., 2018) further proposes Low-Rank Multimodal Fusion (LMF) which employs low-rank approximation to reconstruct the high-order tensor. However, these tensor-based methods neglect the fine-grained temporal dynamics which include rich structured information for multimodal modeling. For example, if the facial expression of a man is happy at time step t, he will speak in a positive tone at time step $t + \Delta t$ more likely. The features of different time steps and different modalities are correlated.

Motivated by this observation, in this paper we introduce Fine-Grained Temporal Low-Rank Multimodal Fusion (FT-LMF). Instead of averaging the features along the temporal dimension, we associate the features of individual time steps between different modalities to form a high-order tensor. The tensor is then embedded to a low-dimensional matrix for subsequent tasks. Compared with LMF, FT-LMF is able to capture the cross-modal interactions at a finer granularity on the temporal dimension.

Since FT-LMF involves multiplications of high-

[*] means equal contribution
[†] corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 377–387
November 16 - 20, 2020.

order tensors in its calculation, its computational complexity increases exponentially with the number of modalities. To tackle this problem, we further introduce Dual Low-Rank Multimodal Fusion (Dual-LMF) which approximates the high-order tensor using low-rank tensor decomposition along both temporal and non-temporal dimensions. We show that Dual-LMF has a linear complexity w.r.t the number of modalities. In experiments, we have validated FT-LMF and Dual-LMF on four benchmark multimodal analysis datasets and they have shown promising results in comparison with the state-of-the-art methods.

The contributions of this paper can be summarized as follows:

(1) To address the ignorance of fine-grained temporal dynamics in the existing tensor-based fusion methods, we propose Fine-Grained Temporal Low-Rank Multimodal Fusion (FT-LMF) which correlates the features of different time steps between all the modalities.

(2) To reduce the computational complexity of FT-LMF, we propose Dual Low-Rank Multimodal Fusion (Dual-LMF) which employs low-rank decomposition to approximate the high-order tensor along its dual dimensions.

(3) Experimental results show that our methods outperform the most state-of-the-art methods on different multimodal analysis tasks.

2 Related Work

Multimodal analysis has attracted much attention recently. Thanks to the high-quality open-source datasets like CMU-MOSI, POM, YOUTUBE, and ICT-MMMO, many effective methods have been proposed and comprehensively evaluated. The key to multimodal analysis is the fusion of multimodal information. Generally, there are two lines of fusion methods, early fusion and late fusion. Early fusion methods integrate features of different modalities before feeding them to the model. For instance, concatenating different features (Zadeh et al., 2016b) is a simple way. However, the intra-modal dynamics cannot be effectively captured, and the temporal information of a single modality is ignored in early fusion. Late fusion methods (Nojavanasghari et al., 2016) utilize information of a single modality for inference, and then ensembling them by majority voting or weighted averaging (Wörtwein and Scherer, 2017). Unfortunately, the inter-modal interactions are not modeled in late fusion.

To address the drawbacks of the above methods, (Pham et al., 2019) investigates learning joint representations via cyclic translations from source to target modalities and only uses the source modality for prediction during testing. TFN (Zadeh et al., 2017) and its successive work(Liang et al., 2019) propose to embed multiple feature vectors into a high-order tensor to improve the modeling of inter-modal relationships. However, the computational complexity of TFN increases exponentially with the number of modalities. LMF (Liu et al., 2018) reduces the complexity of TFN by applying low-rank decomposition to the high-order tensor. While LMF simply averages the feature matrices along the temporal dimension or chooses a feature vector of one time step among all the time steps, ignoring the rich fine-grained temporal information.

In this paper, we develop Fine-Grained Temporal Low-Rank Multimodal Fusion (FT-LMF) to correlate the features of different time steps between modalities. Furthermore, considering that the computational complexity of FT-LMF increases exponentially with the number of modalities, we propose Dual Low-Rank Multimodal Fusion (Dual-LMF) which performs low-rank decomposition to both dimensions of the input features. The performances of our methods on several tasks, i.e., multimodal sentiment analysis and speaker traits recognition, are improved with an acceptable complexity.

3 Multimodal Tensor Fusion

3.1 Tensor Fusion Network

We start by introducing TFN (Zadeh et al., 2017) which only adopts multimodal fusion on non-temporal dimension of input features. Suppose that the space of the m-th modality is $\mathbb{R}^{d_m \times t_m}$ and the number of modalities is M. We randomly choose one time step from features of each modality and denote it as $v_m \in \mathbb{R}^{d_m}$. As shown in Fig. 1, TFN transforms the input vectors v_1, v_2,...,v_M into a high-order tensor and then maps it back to a low-dimensional vector. The input tensor $\tilde{V}$ formed by the unimodal representation is calculated as:

$$\tilde{V} = \bigotimes_{m=1}^{M} v_m \tag{1}$$

where $\bigotimes$ denotes the tensor outer product operation over a set of vectors and $\tilde{V} \in \mathbb{R}^{\prod_{m=1}^{M} d_m}$ is the

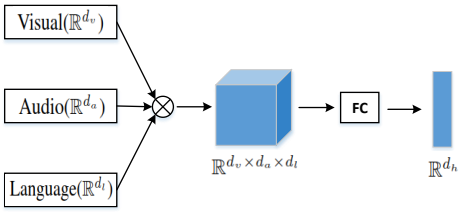

Figure 1: **Tensor Fusion Network (TFN).** The input vectors of three modalities are transformed into a 3-D tensor, and mapped back to a vector by a fully-connected layer.

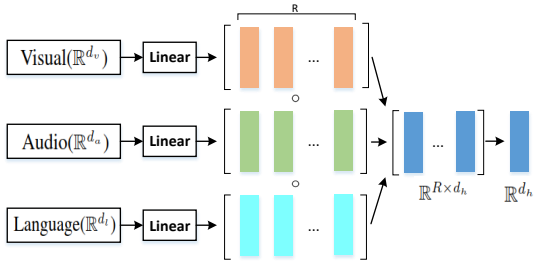

Figure 2: **Low-Rank Multimodal Fusion (LMF).** The input vectors are fed into linear layers (the number of layers is equal to rank value). The outputs are element-wise multiplied ($\circ$) with each other followed by a summation function along the rank dimension.

hybrid representation of the input vectors. Following the conventional setting of neural networks, $\tilde{V}$ is followed by a fully-connected layer for dimension reduction, as

$$h = W_h \cdot \tilde{V} + b_h \tag{2}$$

where $W_h \in \mathbb{R}^{d_h \times \prod_{m=1}^{M} d_m}$ and $b_h \in \mathbb{R}^{d_h}$ are learnable variables[1]. "$\cdot$" denotes linear operation. It is obvious that the computational complexity of TFN increases exponentially with the number of modalities.

3.2 Low-Rank Multimodal Fusion

To reduce the complexity of TFN, LMF (Liu et al., 2018) is proposed to utilize low-rank decomposition for approximating the high-order tensor W_h, as shown in Fig. 2. LMF first divides the $(M+1)$-order tensor W_h into a series of M-order tensors as

$$W_h = \left[W_h^1; W_h^2; ...; W_h^{d_h} \right] \tag{3}$$

For efficiently calculating the tensor multiplication $W_h^i \cdot \tilde{V}$, LMF applies low-rank decomposition to each W_h^i

$$W_h^i = \sum_{r=1}^{R} \bigotimes_{m=1}^{M} (W_h^i)_{m,r} \tag{4}$$

where $(W_h^i)_{m,r} \in \mathbb{R}^{d_m \times 1}$ and R is the value of rank. $W_h^i \cdot \tilde{V}$ is then computed based on Eqns. 1 and 4 [2]:

$$\begin{aligned} W_h^i \cdot \tilde{V} &= \left[\sum_{r=1}^{R} \bigotimes_{m=1}^{M} (W_h^i)_{m,r} \right] \cdot \left[\bigotimes_{m=1}^{M} v_m \right] \\ &= \sum \left[\sum_{r=1}^{R} \left[\bigotimes_{m=1}^{M} (W_h^i)_{m,r} \circ \bigotimes_{m=1}^{M} v_m \right] \right] \\ &= \sum_{r=1}^{R} \left[\sum \bigotimes_{m=1}^{M} \left[(W_h^i)_{m,r} \circ v_m \right] \right] \end{aligned} \tag{5}$$

where $\circ$ denotes element-wise multiplication and $\sum$ denotes the summation function for all the elements in the high-order tensor. To facilitate reading, we rewrite Eqn. 5 as

$$W_h^i \cdot \tilde{V} = \sum_{r=1}^{R} \bigwedge_{m=1}^{M} \left[(W_h^i)_{m,r}^{\mathsf{T}} v_m \right] \tag{6}$$

where $\Lambda_{m=1}^{M}$ denotes the element-wise multiplication $\circ$ over a sequence of tensors. For instance, $\Lambda_{m=1}^{3} x_m = x_1 \circ x_2 \circ x_3$. $W_h^i \cdot \tilde{V}$ is an element of $W_h \cdot \tilde{V}$, thus

$$W_h \cdot \tilde{V} = \sum_{r=1}^{R} \bigwedge_{m=1}^{M} \left[(W_h)_{m,r}^{\mathsf{T}} v_m \right] \tag{7}$$

where $(W_h)_{m,r} \in \mathbb{R}^{d_m \times d_h}$ consists of $(W_h^i)_{m,r} \in \mathbb{R}^{d_m \times 1}$. Through low-rank approximation to the high-order tensor, LMF scales linearly with the number of modalities.

4 Fine-Grained Temporal LMF

The existing multimodal tensor-based fusion methods correlate multimodal features at a coarse granularity, while the rich temporal dynamic information underlying in different modalities is ignored. In this work, we propose Fine-Grained Temporal Low-Rank Multimodal Fusion (FT-LMF) for making a full use of the fine-grained information along the temporal dimension.

Based on the discussions in previous section, we can easily correlate the features of different

[1] In practice, the bias b_h is approximated by the concatenation of v_m and a scalar value of 1; thus, we omit b_h in the subsequent derivations of this paper.

[2] The detailed derivations are shown in Appendix.

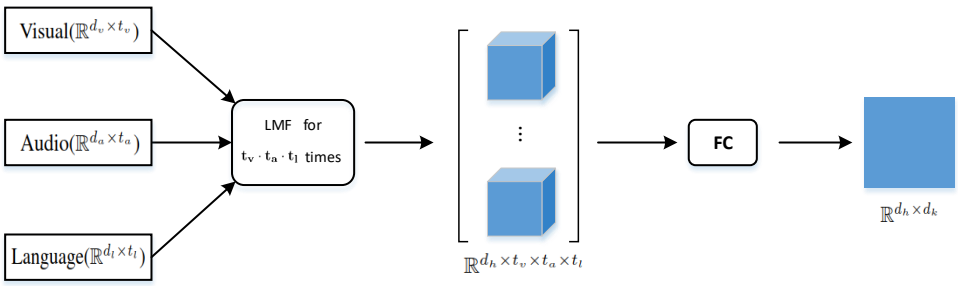

Figure 3: **Fine-Grained Temporal LMF (FT-LMF).** The inputs are feature matrices of multiple modalities. LMF is performed on all the time-step groups of different modalities. The number of groups is $t_v \times t_a \times t_l$. The space of an output of LMF is $\mathbb{R}^{d_h}$, thus the space of all the groups is $\mathbb{R}^{d_h \times t_v \times t_a \times t_l}$. A fully-connected layer follows the high-order tensor to reduce the tensor space from $\mathbb{R}^{d_h \times t_v \times t_a \times t_l}$ to $\mathbb{R}^{d_h \times d_k}$. In practice, the parameters W_k on FC layer are generated by attention mechanism for better effectiveness.

time steps between multiple modalities. We use $H[l_1, l_2, ..., l_M] \in \mathbb{R}^{d_h}$ with index $[l_1, l_2, ..., l_M]$ to denote the correlation result of selected time-steps of M modalities. We can obtain a high-order tensor $H \in \mathbb{R}^{\prod_{m=1}^{M} t_m \times d_h}$ which carries the interactive information of different time steps between modalities. Following Eqn. 2, we calculate the values of tensor as:

$$H[l_1, l_2, ..., l_M] = W_h \cdot \left[\bigotimes_{m=1}^{M} (V_m)_{l_m} \right] \tag{8}$$

where $V_m \in \mathbb{R}^{d_m \times t_m}$ denotes the feature matrix of m-th modality and $(V_m)_{l_m}$ denotes the l_m-th time step of V_m.

We map H to a 2-D matrix

$$K = W_k \cdot H + b_k \tag{9}$$

where the spaces of W_k and b_k are $\mathbb{R}^{d_k \times \prod_{m=1}^{M} t_m}$ and $\mathbb{R}^{d_k}$, respectively; thus the space of K is $\mathbb{R}^{d_k \times d_h}$. For the convenience of subsequent derivations, we rewrite Eqn. 9 as

$$K_i = W_k \cdot H_i + b_k \tag{10}$$

where $H_i \in \mathbb{R}^{\prod_{m=1}^{M} t_m}$ is just one channel of H, and $K_i \in \mathbb{R}^{d_k}$ is one channel of K. In practice, we employ attention mechanism to generate W_k for better capturing the importance of each time-step group:

$$W_k[l_1, ..., l_M] = \frac{e^{\left\{W_2 \tanh(W_1 H[l_1,...,l_M] + b_1)\right\}}}{\sum_{o_1,...,o_M=1}^{t_1,...,t_M} e^{\left\{W_2 \tanh(W_1 H[o_1,...,o_M] + b_1)\right\}}} \tag{11}$$

where $W_1 \in \mathbb{R}^{d_h \times d_h}$, $b_1 \in \mathbb{R}^{d_h}$, $W_2 \in \mathbb{R}^{d_k \times d_h}$ are trainable variables, $W_k \in \mathbb{R}^{d_k \times \prod_{m=1}^{M} t_m}$ consists of $W_k[l_1, ..., l_M] \in \mathbb{R}^{d_k}$. The numerator is element-wise divided by the denominator.

FT-LMF shown in Fig. 3 is able to capture the fine-grained temporal interactions between different modalities, while the computational complexity of its high-order tensor H increases exponentially with the number of modalities. To tackle this problem, we further propose Dual-LMF as discussed in the next section.

5 Dual-LMF

Based on FT-LMF, Dual-LMF further performs low-rank decomposition on both temporal dimension and non-temporal dimensions. First, we follow LMF to divide the $(M+1)$-order tensor W_k into a series of M-order tensors. The number of the tensors is d_k:

$$W_k = \left[W_k^1; W_k^2; ...; W_k^{d_k} \right] \tag{12}$$

We apply low-rank decomposition to each W_k^j,

$$W_k^j = \sum_{r_2=1}^{R_2} \bigotimes_{m=1}^{M} (W_k^j)_{m,r_2} \tag{13}$$

where $(W_k^j)_{m,r_2} \in \mathbb{R}^{t_m \times 1}$. We then rewrite $W_k^j H_i$ as

$$W_k^j \cdot H_i = \left[\sum_{r_2=1}^{R_2} \bigotimes_{m=1}^{M} (W_k^j)_{m,r_2} \right] \cdot H_i \tag{14}$$

$H_i[l_1, l_2, ..., l_M] \in \mathbb{R}^1$ is an element of $H_i \in$

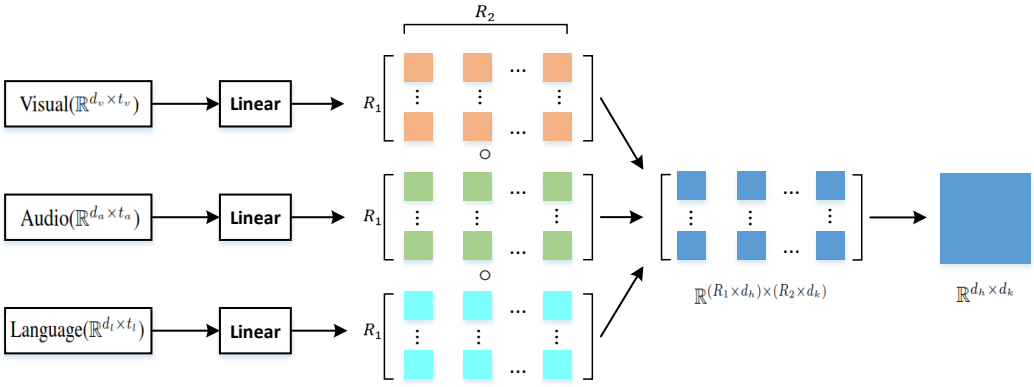

Figure 4: **Dual-LMF.** Dual-LMF performs dimension reduction on both the temporal dimension and the non-temporal dimension of input features, while FT-LMF only performs dimension reduction on the non-temporal dimension. After linear mapping, features of all the modalities have the same space $\mathbb{R}^{R_1 \times d_h \times R_2 \times d_k}$, and we perform element-wise multiplication to combine them. Finally, we sum over both rank dimensions (R_1 and R_2) of the high-order tensor to obtain the multimodal fusion matrix.

$\mathbb{R}^{\prod_{m=1}^{M} t_m}$ and it can be calculated by Eqn. 6:

$$H_i[l_1, ..., l_M] = \sum_{r_1=1}^{R_1} \bigwedge_{m=1}^{M} \left[(W_h^i)_{m,r_1}^{\mathsf{T}} (V_m)_{l_m} \right] \quad (15)$$

where $(W_h^i)_{m,r_1}$ is the decomposed component of W_h. Then H_i which consists of $H_i[l_1, l_2, ..., l_M]$ is obtained as,

$$H_i = \sum_{r_1=1}^{R_1} \bigotimes_{m=1}^{M} \left[(W_h^i)_{m,r_1}^{\mathsf{T}} V_m \right] \quad (16)$$

$(W_h^i)_{m,r_1}^{\mathsf{T}} V_m$ is a t_m-dimensional vector. We substitute Eqn. 16 into Eqn. 14 [2]:

$$\begin{aligned}
& W_k^j \cdot H_i \\
&= \left[\sum_{r_2=1}^{R_2} \bigotimes_{m=1}^{M} (W_k^j)_{m,r_2} \right] \cdot \left[\sum_{r_1=1}^{R_1} \bigotimes_{m=1}^{M} \left[(W_h^i)_{m,r_1}^{\mathsf{T}} V_m \right] \right] \\
&= \sum \left[\sum_{r_2=1}^{R_2} \sum_{r_1=1}^{R_1} \bigotimes_{m=1}^{M} \left[(W_k^j)_{m,r_2} \circ [(W_h^i)_{m,r_1}^{\mathsf{T}} V_m] \right] \right] \\
&= \sum_{r_2=1}^{R_2} \sum_{r_1=1}^{R_1} \left[\sum \left[\bigotimes_{m=1}^{M} (W_k^j)_{m,r_2} \circ [(W_h^i)_{m,r_1}^{\mathsf{T}} V_m] \right] \right] \\
&= \sum_{r_2=1}^{R_2} \sum_{r_1=1}^{R_1} \bigwedge_{m=1}^{M} \left[(W_h^i)_{m,r_1}^{\mathsf{T}} V_m (W_k^j)_{m,r_2} \right]
\end{aligned} \quad (17)$$

where we treat both $(W_k^j)_{m,r_2}$ and $(W_h^i)_{m,r_1}^{\mathsf{T}} V_m$ as t_m-dimensional vectors in the second and third rows of Eqn. 17. While in the forth row, we utilize the original sizes, i.e., $(W_k^j)_{m,r_2} \in \mathbb{R}^{t_m \times 1}$, $(W_h^i)_{m,r_1}^{\mathsf{T}} V_m \in \mathbb{R}^{1 \times t_m}$.

$W_k^j \cdot H_i$ is an element of $W_k \cdot H$ and $[i, j]$ is the corresponding index. We refer to the derivations of LMF and employ a simple transformation to Eqn. 17 to obtain the output fusion matrix $W_k H$:

$$\begin{aligned}
& W_k \cdot H = \\
& \sum_{r_2=1}^{R_2} \sum_{r_1=1}^{R_1} \bigwedge_{m=1}^{M} \left[(W_h)_{m,r_1}^{\mathsf{T}} V_m (W_k)_{m,r_2} \right]
\end{aligned} \quad (18)$$

Similar to Eqn. 11, $(W_k)_m \in \mathbb{R}^{t_m \times (R_2 \times d_k)}$ is computed with element-wise attention mechanism,

$$\begin{aligned}
& [(W_k)_m]_{l_m} = \\
& \frac{e^{\left\{ W_4 \tanh\left(W_3 \left[(W_h)_m^{\mathsf{T}} (V_m)_{l_m} \right] + b_3 \right) \right\}}}{\sum_{o_m=1}^{t_m} e^{\left\{ W_4 \tanh\left(W_3 \left[(W_h)_m^{\mathsf{T}} (V_m)_{o_m} \right] + b_3 \right) \right\}}}
\end{aligned} \quad (19)$$

where the space of $(W_h)_m$ is $\mathbb{R}^{d_m \times (R_1 \times d_h)}$, the space of W_3 is $\mathbb{R}^{(R_1 \times d_h) \times (R_1 \times d_h)}$, the space of W_4 is $\mathbb{R}^{(R_2 \times d_k) \times (R_1 \times d_h)}$, the space of b_3 is $\mathbb{R}^{R_1 \times d_h}$, $(W_k)_m \in \mathbb{R}^{t_m \times (R_2 \times d_k)}$ consists of $[(W_k)_m]_{l_m} \in \mathbb{R}^{(R_2 \times d_k)}$, the numerator is element-wise divided by the denominator.

Thanks to the low-rank decomposition on both temporal and non-temporal dimensions of input features, Dual-LMF shown in Fig. 4 is much more efficient than FT-LMF and has a good scalability to the increasing number of modalities.

6 Experiments

6.1 Datasets

We evaluate FT-LMF and Dual-LMF on several benchmark datasets of multimodal analysis tasks,

including CMU-MOSI (Zadeh et al., 2016a), POM (Park et al., 2014), YOUTUBE (Morency et al., 2011), and ICT-MMMO (Wöllmer et al., 2013).

CMU-MOSI (Zadeh et al., 2016a) is created for sentiment analysis, which contains 63 long videos with a sentiment label in range [-3,3]. During the training and testing, we divide the 63 videos into 2199 chunks for label alignment. Following the existing work, we divide the whole dataset into three parts, for training, validation, and testing. Note that the same speaker does not appear in multiple sets.

POM (Park et al., 2014) is created for speaker traits recognition. It contains 903 movie review videos and each video is annotated with 16 speaker traits, including confident, passionate, voice pleasant, dominant, credible, vivid, expertise, entertaining, reserved, trusting, relaxed, outgoing, thorough, nervous, persuasive and humorous.

YouTube (Morency et al., 2011) is created for sentiment analysis. It contains 47 videos from the social media website YouTube and each video is annotated at the segment level for sentiment.

ICT-MMMO (Wöllmer et al., 2013) is created for sentiment analysis. It contains 370 movie review videos and each video is annotated at the video level for sentiment.

6.2 Features

In this paper, we follow the existing methods to do empirical studies on three different modalities, including audio, visual, and text. In addition, P2FA (Yuan and Liberman, 2008) is utilized to align the three modalities at the word granularity. The visual and audio features are aligned by computing their average value over the utterance interval of each word.

To extract audio, visual, and text features, we follow the methods of LMF. Specifically, for audio modality, we use COVAREP (Degottex et al., 2014) to extract a set of low-level audio features. For visual modality, we use Facet (iMotions, 2017) to extract a set of visual features for each frame. For text modality, we use pre-trained 300-dimension glove word vectors (Pennington et al., 2014) to extract word representations.

For audio and visual features, we use a 2-layer feed-forward neural network to handle the features of all time steps. For text features, we use an LSTM (Hochreiter and Schmidhuber, 1997) to capture the semantic information. After encoding the features, we send them to fusion models.

6.3 Metrics

For different datasets, we compare methods under different metrics. For CMU-MOSI, we report Mean Absolute Error (MAE), Pearson correlation (Corr), binary accuracy, F1-Score, 7-class accuracy. For POM, we report average MAE, average Corr, average binary-accuracy for speaker traits. For YouTube, we report 3-class accuracy and F1-Score. For ICT-MMMO, we report binary accuracy and F1-Score.

6.4 Model and Optimization

For a fair comparison, we implement FT-LMF and Dual-LMF similarly to LMF, while we keep all the time steps of the three modalities. The output of our FT-LMF and Dual-LMF is $K \in R^{d_k \times d_h}$. In the experiments, we set d_k to 1 and d_h to the number of attributes. We employ MAE loss function to optimize the learnable variables.

6.5 Experimental Setting

For CMU-MOSI, the output dimension is 1; we train the model for at most 500 epochs. If MAE does not increase for 20 epochs, we stop the training. The other hyper-parameters (i.e., hidden size, learning rate, batch size) are determined by the grid search method. The best hyper-parameters are different for TFN, LMF, FT-LMF, Dual-LMF.

For POM, the output dimension is 16, since we treat the predictions for 16 speakers as a multi-label task. We also train the model for at most 500 epochs and the patience is 20. The other hyper-parameters are determined by the grid search method.

For YOUTUBE and ICT-MMMO, the output dimension is 1. We also train the model for at most 500 epochs and the patience is 20. The other hyper-parameters are determined by the grid search method.

6.6 Comparison Baselines

We use TFN (Zadeh et al., 2017) and LMF (Liu et al., 2018) as our baselines. In addition, we compare our methods with several other state-of-the-art methods which employ simple feature-encoding ways, like LSTM and fully-connected layer, since we also use these simple ways and focus on the fusion method before final prediction.

SVM is trained on simply concatenated multimodal features for prediction (Zadeh et al., 2016b; Park et al., 2014; Pérez-Rosas et al., 2013).

Table 1: Experimental results on CMU-MOSI and POM

Dataset	CMU-MOSI					POM		
Model	Acc(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc-7(↑)	MAE(↓)	Acc(↑)	Corr(↑)
SVM	50.2	50.1	1.864	0.057	17.5	0.887	33.9	0.104
DF	74.2	74.2	1.143	0.518	26.8	0.869	34.1	0.144
MV-LSTM	73.9	74.0	1.019	0.601	33.2	0.891	34.6	0.270
BC-LSTM	73.9	73.9	1.079	0.614	28.7	0.840	34.8	0.278
MCTN	**79.3**	**79.1**	0.909	0.676	-	-	-	-
MARN	77.1	77.0	0.968	0.625	34.7	-	39.4	-
TFN	73.9	73.4	0.970	0.633	32.1	0.886	31.6	0.093
LMF	76.4	75.7	0.912	0.668	32.8	0.796	**42.8**	0.396
FT-LMF	78.7	78.8	**0.901**	0.693	35.1	0.788	42.1	0.395
Dual-LMF	78.4	78.3	**0.901**	**0.700**	**35.8**	**0.777**	**42.8**	**0.398**

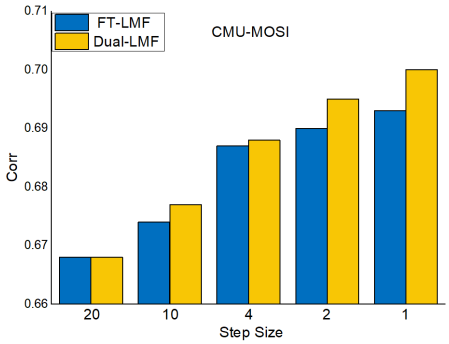

Figure 5: The performances (Corr) of different time step sizes on CMU-MOSI. The step size of 1 is standard FT-LMF/Dual-LMF proposed in this paper. The step size of 20 is LMF (Liu et al., 2018).

DF (Nojavanasghari et al., 2016) uses multiple fully-connected layers to predict the results for each modality, respectively, and ensembles the results.

BC-LSTM (Poria et al., 2017) correlates multiple modalities with a proposed context-dependent fusion method.

MV-LSTM (Rajagopalan et al., 2016) is an extension to LSTM, designed to model both view-specific and cross-view dynamic by partitioning internal representations to mirror the multiple input modalities.

MCTN (Pham et al., 2019) investigates learning joint representations via cyclic translations from source to target modalities and only uses the source modality for prediction during testing.

MARN (Zadeh et al., 2018) discovers the interaction between modalities through time with a neural module called Multi-attention Block and stores them in a hybrid memory component called Long-short Term Hybrid Memory. Although MARN considers temporal information, it is not tensor-based.

7 Results

7.1 Compared with State-of-the-Art

Table 1 shows the performances of the methods on CMU-MOSI and POM datasets. On CMU-MOSI, FT-LMF and Dual-LMF outperform the state-of-the-art methods on MAE, Corr, and Acc-7; and Dual-LMF has a better overall performance than FT-LMF. On POM, we report the average performances on 16 speakers and find that Dual-LMF outperforms the state-of-the-art methods on all the metrics. Table 2 shows the performances of the methods on ICT-MMMO and YOUTUBE datasets. The observed results are similar to those of POM. The promising empirical results demonstrate the effectiveness of our methods.

Table 2: Experimental results on ICT-MMMO and YOUTUBE.

Dataset	ICT-MMMO		YOUTUBE	
Model	Acc(↑)	F1(↑)	Acc-3(↑)	F1(↑)
SVM	68.8	68.7	42.4	37.9
DF	65.0	58.7	45.8	32.0
MV-LSTM	72.5	72.3	45.8	43.3
BC-LSTM	70.0	70.1	47.5	47. 3
MCTN	81.3	80.8	51.7	52.4
MARN	86.3	85.9	54.2	52.9
TFN	72.5	72.6	47.5	41.0
LMF	83.7	84.0	49.2	47.8
FT-LMF	85.0	85.0	52.5	50.3
Dual-LMF	**87.5**	**87.7**	**55.9**	**54.3**

7.2 Effect of Fine-Grained Temporal Information

To further validate the effect of fine-grained temporal information, we show the performances of FT-LMF and Dual-LMF with different time-step sizes[3]. In our experiments, t_v=t_a=t_l=20. The time-

[3]In practice, we average the features along the temporal dimension every time-step size.

Table 3: Computational complexity of tensor-based fusion methods. The FLOPs do not include the preprocessing modules (i.e., 2-layer feed-forward neural network and LSTM). The space complexity is discussed w.r.t. the number of modalities.

Method	FLOPs	Complexity
TFN	1.3×10^5	Exp
LMF	2.6×10^3	Linear
FT-LMF	1.4×10^5	Exp
Dual-LMF	1.8×10^4	Linear

step sizes of standard FT-LMF and Dual-LMF are 1. Here we select a series of time-step sizes including 2, 4, 10, and 20 for comparison. Note that FT-LMF, Dual-LMF, and LMF are equivalent when the time-step size is 20. As shown in Fig. 5, we find that the performances of the models are improved as the step size decreases. The results demonstrate the effectiveness of incorporating fine-grained temporal dynamics into the multimodal fusion scheme.

7.3 Complexity Analysis

Space Complexity

We analyze the space complexity of different methods theoretically. Following the supposition in the approach section, we focus on the sizes of learnable variables and the output of each layer. Note that we omit the variables with relatively small size, i.e., bias.

TFN The size of a vector of m-th modality is d_m. Therefore, the size of the high-order tensor is $\prod_{m=1}^{M} d_m$. The size of variables in the fully-connected layer is $d_h \times \prod_{m=1}^{M} d_m$. The space complexity is $O(d_h \times \prod_{m=1}^{M} d_m)$.

LMF We map all the vectors to a dimension of d_h. The rank is set to R; the size of variables in linear layers is $d_h \times R \times \sum_{m=1}^{M} d_m$; the size of the output is $d_h \times R \times M$. The space complexity is $O(d_h \times R \times \sum_{m=1}^{M} d_m)$.

FT-LMF We use LMF for $\prod_{m=1}^{M} t_m$ groups of time steps in total. The size of variables in a fully-connected layer of LMF is $d_h \times R \times \sum_{m=1}^{M} d_m$; thus, the size of the generated high-order tensor is $\prod_{m=1}^{M} t_m \times d_h$. The size of variables in the subsequent attention-based fully-connected layer is $d_h \times (d_h + d_k)$. The space complexity is $O(d_h \times (\prod_{m=1}^{M} t_m + R \times \sum_{m=1}^{M} d_m + d_h + d_k))$.

Dual-LMF The size of variables in the linear layer is $d_h \times R_1 \times \sum_{m=1}^{M} d_m + d_h \times (d_h + d_k) \times R_1 \times R_2$. The size of the output is $M \times R_2 \times d_k \times R_1 \times d_h$. The space complexity is $O(d_h \times R_1 \times \sum_{m=1}^{M} d_m + R_2 \times R_1 \times d_h \times ((M+1) \times d_k + d_h))$.

With respect to the number of modalities, we can easily find that TFN and FT-LMF have an exponential space complexity, while LMF and Dual-LMF have a linear space complexity.

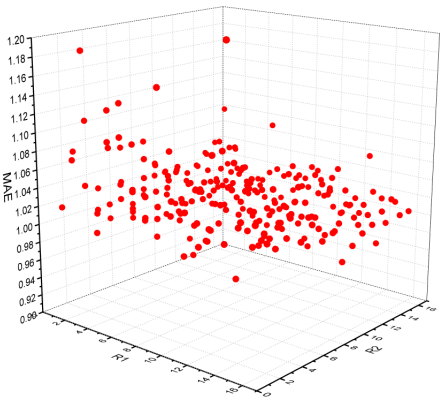

Figure 6: The performances (MAE) of combinations of rank values R_1 and R_2 on CMU-MOSI

Practical FLOPs

Table 4 shows the float point operation (FLOPs) of different methods on CMU-MOSI. Specifically, we use a set of hyper-parameters as t_v=t_a=t_l=20, d_v=d_a=d_l=32, R_1=4, R_2=4, d_h=1, d_k=1. The FLOPs of TFN and FT-LMF are much more than those of LMF and Dual-LMF.

7.4 Empirical Study on Rank Value

The selection of rank is important in multimodal fusion. We utilize the hyper-parameters mentioned above and evaluate Dual-LMF with combinations of different rank values R_1 and R_2. We start by setting both R_1 and R_2 to 1, and gradually increase them. The results on CMU-MOSI are shown in Fig. 6. We find that only a single R_1 or R_2 cannot well determine the final performance. Thus, a careful selection of R_1 and R_2 is necessary. In addition, Dual-LMF with low rank values can achieves similar results to that with high rank values and the computational complexity is reduced.

8 Conclusion

In this paper, we have proposed novel multimodal fusion methods, including FT-LMF and Dual-LMF, for multimodal analysis tasks. FT-LMF is a fine-grained version of Low-Rank Multimodal Fusion which particularly associates the features of individual time steps between multiple modalities. Based on FT-LMF, Dual-LMF performs low-rank tensor approximation along dual dimensions of input features to reduce the exponential computational com-

plexity of FT-LMF to a linear complexity w.r.t. the number of modalities. The experimental results show that our methods achieve superior performance compared with the state-of-the-art methods with similar computational cost.

Acknowledgments

This work is supported in part by Science and Technology Innovation 2030 –"New Generation Artificial Intelligence" Major Project No.(2018AAA0100904), NSFC (No. 61672456, 61702448, U19B2043), Artificial Intelligence Research Foundation of Baidu Inc., the funding from HIKVision and Horizon Robotics, and ZJU Converging Media Computing Lab.

References

Lawrence S Chen, Thomas S Huang, Tsutomu Miyasato, and Ryohei Nakatsu. 2018. Multimodal human emotion/expression recognition. In *IEEE FG*.

Liyanage C De Silva, Tsutomu Miyasato, and Ryohei Nakatsu. 1997. Facial emotion recognition using multi-modal information. In *ICICS*.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *ICASSP*.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*.

iMotions. 2017. Facial expression analysis.

Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019. Learning representations from imperfect time series data via tensor rank regularization. In *ACL*.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *ICMI*.

Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *ICMI*.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *ICMI*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *ACL*.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *ACL*.

Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke. 2016. Extending long short-term memory for multi-view structured learning. In *ECCV*.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*.

Torsten Wörtwein and Stefan Scherer. 2017. What really matters—an information gain analysis of questions and reactions in automated ptsd screenings. In *ACII*.

Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*.

Ben P Yuhas, Moise H Goldstein, and Terrence J Sejnowski. 1989. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. In *AAAI*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016a. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016b. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*.

.1 Reproducibility of the paper

We implement experiments on GTX 1080Ti. The main hyperparameters include audio hidden dimension, video hidden dimension, text hidden dimension, audio dropout rate, video dropout rate, text dropout rate, learning rate, weight decay, rank1 and rank2. Grid search is employed to find the appropriate combination of parameters. For each method, we randomly try 2000 combinations, since the model is small and the running time is short as shown in Table 4. The feature extraction method and the division of training and test sets follow (Zadeh et al., 2018). If the paper is accepted, we promise to open the source code and the best-performing hyperparameters.

Table 4: The model size and execution time of our methods.

Method	Size(MB)	Time(s)
TFN	1163	19.3
LMF	627	11.7
FT-LMF	1097	17.9
Dual-LMF	721	13.0

.2 Derivations for Eqns. 5 and 6 in the paper

$W_h^i \cdot \tilde{V}$ can be rewritten as:

$$W_h^i \cdot \tilde{V} = \left[\sum_{r=1}^{R}\bigotimes_{m=1}^{M}(W_h^i)_{m,r}\right] \cdot \left[\bigotimes_{m=1}^{M} v_m\right] \tag{20}$$

where "$\cdot$" denotes linear operation for $\bigotimes_{m=1}^{M} v_m$. Since $\sum_{r=1}^{R}\bigotimes_{m=1}^{M}(W_h^i)_{m,r}$ and $\bigotimes_{m=1}^{M} v_m$ have the same size $R^{\prod_{m=1}^{M} d_m}$, we can rewrite the linear operation as the combination of element-wise multiplication and summation. The two formations are equivalent.

$$\begin{aligned} W_h^i \cdot \tilde{V} &= \sum\left[\sum_{r=1}^{R}\left[\bigotimes_{m=1}^{M}(W_h^i)_{m,r} \circ \bigotimes_{m=1}^{M} v_m\right]\right] \\ &= \sum_{r=1}^{R}\left[\sum\left[\bigotimes_{m=1}^{M}(W_h^i)_{m,r} \circ \bigotimes_{m=1}^{M} v_m\right]\right] \end{aligned} \tag{21}$$

where $\bigotimes_{m=1}^{M}(W_h^i)_{m,r} \circ \bigotimes_{m=1}^{M} v_m$ can be rewritten as another formation, $\bigotimes_{m=1}^{M}\left[(W_h^i)_{m,r} \circ v_m\right]$. The equivalence can be proven by element-wise comparison:

Proposition 1.

$$\bigotimes_{m=1}^{M}(W_h^i)_{m,r} \circ \bigotimes_{m=1}^{M} v_m = \bigotimes_{m=1}^{M}\left[(W_h^i)_{m,r} \circ v_m\right] \tag{22}$$

Proof.

$$\begin{aligned} &\left[\bigotimes_{m=1}^{M}(W_h^i)_{m,r} \circ \bigotimes_{m=1}^{M} v_m\right]_{c_1,c_2,...,c_M} \\ &= \left[\bigotimes_{m=1}^{M}(W_h^i)_{m,r}\right]_{c_1,c_2,...,c_M} \circ \left[\bigotimes_{m=1}^{M} v_m\right]_{c_1,c_2,...,c_M} \\ &= \left[(W_h^i)_{1,r}\right]_{c_1} \circ ... \circ \left[(W_h^i)_{M,r}\right]_{c_M} \circ (v_1)_{c_1} \circ ... \circ (v_M)_{c_M} \\ &= \left\{\left[(W_h^i)_{1,r}\right]_{c_1} \circ (v_1)_{c_1}\right\} \circ ... \circ \left\{\left[(W_h^i)_{M,r}\right]_{c_M} \circ (v_M)_{c_M}\right\} \\ &= \left[(W_h^i)_{1,r} \circ v_1\right]_{c_1} \circ ... \circ \left[(W_h^i)_{M,r} \circ v_M\right]_{c_M} \\ &= \left[\bigotimes_{m=1}^{M}\left[(W_h^i)_{m,r} \circ v_m\right]\right]_{c_1,c_2,...,c_M} \end{aligned} \tag{23}$$

where $c_1, c_2, ..., c_M (c_m \in [1, 2, ..., d_m])$ denotes the index of the elements in high-order tensor. □

$W_h^i \cdot \tilde{V}$ can be rewritten as follows:

$$W_h^i \cdot \tilde{V} = \sum_{r=1}^{R}\left[\sum\bigotimes_{m=1}^{M}\left[(W_h^i)_{m,r} \circ v_m\right]\right] \tag{24}$$

where $\sum\bigotimes_{m=1}^{M}\left[(W_h^i)_{m,r} \circ v_m\right]$ can be rewritten as another formation, $\bigwedge_{m=1}^{M}\left[(W_h^i)_{m,r}^{\mathsf{T}} v_m\right]$. The equivalence can be proven as follows:

Proposition 2.

$$\sum\bigotimes_{m=1}^{M}\left[(W_h^i)_{m,r} \circ v_m\right] = \bigwedge_{m=1}^{M}\left[(W_h^i)_{m,r}^{\mathsf{T}} v_m\right] \tag{25}$$

Proof.

$$\bigwedge_{m=1}^{M}\left[(W_h^i)_{m,r}^{\mathsf{T}} v_m\right] = \bigwedge_{m=1}^{M}\sum\left[(W_h^i)_{m,r} \circ v_m\right] \tag{26}$$

Following the simple transformation like $(a + b)(c + d) = ac + ad + bc + bd$, we can easily transform $\bigwedge_{m=1}^{M} \sum \left[(W_h^i)_{m,r} \circ v_m \right]$ to $\sum \bigotimes_{m=1}^{M} \left[(W_h^i)_{m,r} \circ v_m \right]$. These two formations are equal, just with different operation orders. The former utilizes summation($\sum$) first, while the later uses multiplication($\bigotimes$) between different elements first. □

Therefore, we obtain the final formation of $W_h^i \tilde{V}$:

$$W_h^i \cdot \tilde{V} = \sum_{r=1}^{R} \bigwedge_{m=1}^{M} \left[(W_h^i)_{m,r}^{\mathsf{T}} v_m \right] \tag{27}$$

.3 Derivations for Eqn. 17 in the paper

$W_k^j \cdot H_i$ can be rewritten as:

$$W_k^j \cdot H_i = \left[\sum_{r_2=1}^{R_2} \bigotimes_{m=1}^{M} (W_k^j)_{m,r_2} \right] \cdot \left[\sum_{r_1=1}^{R_1} \bigotimes_{m=1}^{M} \left[(W_h^i)_{m,r_1}^{\mathbf{T}} V_m \right] \right] \tag{28}$$

similar to Eqns. 21, 24, and 27, we obtain the final formation of $W_k^j \cdot H_i$,

$$\begin{aligned}
&W_k^j \cdot H_i \\
&= \sum \left[\sum_{r_2=1}^{R_2} \sum_{r_1=1}^{R_1} \left[\bigotimes_{m=1}^{M} (W_k^j)_{m,r_2} \circ \bigotimes_{m=1}^{M} [(W_h^i)_{m,r_1}^{\mathbf{T}} V_m] \right] \right] \\
&= \sum \left[\sum_{r_2=1}^{R_2} \sum_{r_1=1}^{R_1} \bigotimes_{m=1}^{M} \left[(W_k^j)_{m,r_2} \circ [(W_h^i)_{m,r_1}^{\mathbf{T}} V_m] \right] \right] \\
&= \sum_{r_2=1}^{R_2} \sum_{r_1=1}^{R_1} \left[\sum \bigotimes_{m=1}^{M} \left[(W_k^j)_{m,r_2} \circ [(W_h^i)_{m,r_1}^{\mathbf{T}} V_m] \right] \right] \\
&= \sum_{r_2=1}^{R_2} \sum_{r_1=1}^{R_1} \bigwedge_{m=1}^{M} \left[(W_h^i)_{m,r_1}^{\mathbf{T}} V_m (W_k^j)_{m,r_2} \right]
\end{aligned} \tag{29}$$