Efficient Low-rank Multimodal Fusion with Modality-Specific Factors

Zhun Liu*, **Ying Shen***, **Varun Bharadhwaj Lakshminarasimhan**,
Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency
School of Computer Science
Carnegie Mellon University
{zhunl,yshen2,vbl,pliang,abagherz,morency}@cs.cmu.edu

Abstract

Multimodal research is an emerging field of artificial intelligence, and one of the main research problems in this field is multimodal fusion. The fusion of multimodal data is the process of integrating multiple unimodal representations into one compact multimodal representation. Previous research in this field has exploited the expressiveness of tensors for multimodal representation. However, these methods often suffer from exponential increase in dimensions and in computational complexity introduced by transformation of input into tensor. In this paper, we propose the Low-rank Multimodal Fusion method, which performs multimodal fusion using low-rank tensors to improve efficiency. We evaluate our model on three different tasks: multimodal sentiment analysis, speaker trait analysis, and emotion recognition. Our model achieves competitive results on all these tasks while drastically reducing computational complexity. Additional experiments also show that our model can perform robustly for a wide range of low-rank settings, and is indeed much more efficient in both training and inference compared to other methods that utilize tensor representations.

1 Introduction

Multimodal research has shown great progress in a variety of tasks as an emerging research field of artificial intelligence. Tasks such as speech recognition (Yuhas et al., 1989), emotion recognition, (De Silva et al., 1997), (Chen et al., 1998), (Wöllmer et al., 2013), sentiment analysis, (Morency et al., 2011) as well as speaker trait analysis and media description (Park et al., 2014a) have seen a great boost in performance with developments in multimodal research.

However, a core research challenge yet to be solved in this domain is multimodal fusion. The goal of fusion is to combine multiple modalities to leverage the complementarity of heterogeneous data and provide more robust predictions. In this regard, an important challenge has been on scaling up fusion to multiple modalities while maintaining reasonable model complexity. Some of the recent attempts (Fukui et al., 2016), (Zadeh et al., 2017) at multimodal fusion investigate the use of tensors for multimodal representation and show significant improvement in performance. Unfortunately, they are often constrained by the exponential increase of cost in computation and memory introduced by using tensor representations. This heavily restricts the applicability of these models, especially when we have more than two views of modalities in the dataset.

In this paper, we propose the Low-rank Multimodal Fusion, a method leveraging low-rank weight tensors to make multimodal fusion efficient without compromising on performance. The overall architecture is shown in Figure 1. We evaluated our approach with experiments on three multimodal tasks using public datasets and compare its performance with state-of-the-art models. We also study how different low-rank settings impact the performance of our model and show that our model performs robustly within a wide range of rank settings. Finally, we perform an analysis of the impact of our method on the number of parameters and run-time with comparison to other fusion methods. Through theoretical analysis, we show that our model can scale linearly in the number of modalities, and our experiments also show a corresponding speedup in training when compared with

* equal contributions

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 2247–2256
Melbourne, Australia, July 15 - 20, 2018.

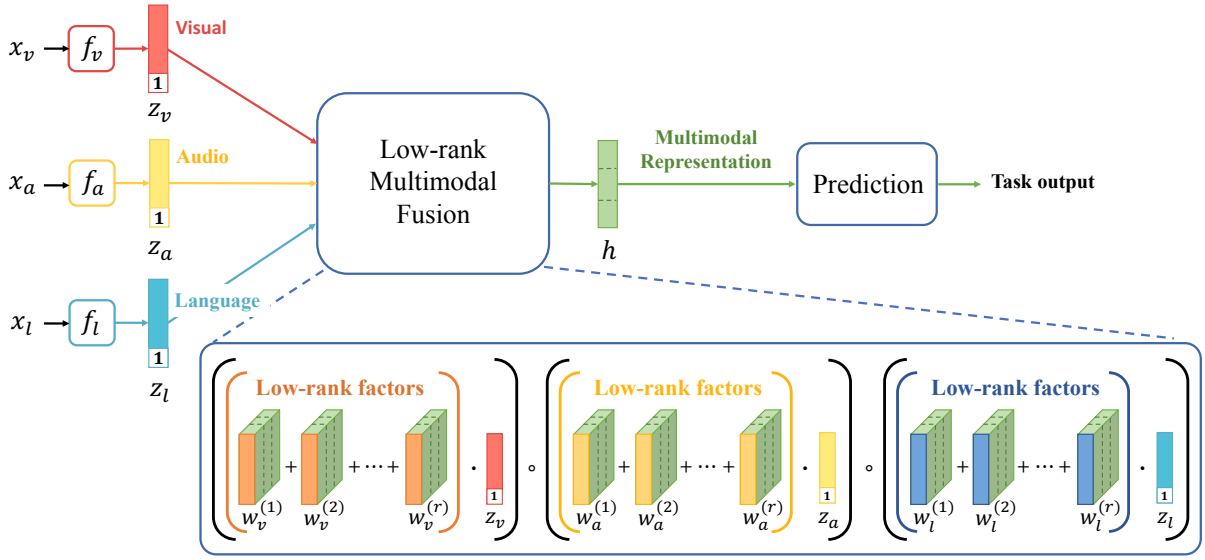

Figure 1: Overview of our Low-rank Multimodal Fusion model structure: LMF first obtains the unimodal representation z_a, z_v, z_l by passing the unimodal inputs x_a, x_v, x_l into three sub-embedding networks f_v, f_a, f_l respectively. LMF produces the multimodal output representation by performing low-rank multimodal fusion with modality-specific factors. The multimodal representation can be then used for generating prediction tasks.

other tensor-based models.

The main contributions of our paper are as follows:

- We propose the Low-rank Multimodal Fusion method for multimodal fusion that can scale linearly in the number of modalities.
- We show that our model compares to state-of-the-art models in performance on three multimodal tasks evaluated on public datasets.
- We show that our model is computationally efficient and has fewer parameters in comparison to previous tensor-based methods.

2 Related Work

Multimodal fusion enables us to leverage complementary information present in multimodal data, thus discovering the dependency of information on multiple modalities. Previous studies have shown that more effective fusion methods translate to better performance in models, and there's been a wide range of fusion methods.

Early fusion is a technique that uses feature concatenation as the method of fusion of different views. Several works that use this method of fusion (Poria et al., 2016) , (Wang et al., 2016) use input-level feature concatenation and use the concatenated features as input, sometimes even removing the temporal dependency present in the modalities (Morency et al., 2011). The drawback of this class of method is that although it achieves fusion at an early stage, intra-modal interactions are potentially suppressed, thus losing out on the context and temporal dependencies within each modality.

On the other hand, late fusion builds separate models for each modality and then integrates the outputs together using a method such as majority voting or weighted averaging (Wortwein and Scherer, 2017), (Nojavanasghari et al., 2016). Since separate models are built for each modality, inter-modal interactions are usually not modeled effectively.

Given these shortcomings, more recent work focuses on intermediate approaches that model both intra- and inter-modal dynamics. Fukui et al. (2016) proposes to use Compact Bilinear Pooling over the outer product of visual and linguistic representations to exploit the interactions between vision and language for visual question answering. Similar to the idea of exploiting interactions, Zadeh et al. (2017) proposes Tensor Fusion Network, which computes the outer product between unimodal representations from three different modalities to compute a tensor representation. These methods exploit tensor representations to model

inter-modality interactions and have shown a great success. However, such methods suffer from exponentially increasing computational complexity, as the outer product over multiple modalities results in extremely high dimensional tensor representations.

For unimodal data, the method of low-rank tensor approximation has been used in a variety of applications to implement more efficient tensor operations. Razenshteyn et al. (2016) proposes a modified weighted version of low-rank approximation, and Koch and Lubich (2010) applies the method towards temporally dependent data to obtain low-rank approximations. As for applications, Lei et al. (2014) proposes a low-rank tensor technique for dependency parsing while Wang and Ahuja (2008) uses the method of low-rank approximation applied directly on multidimensional image data (Datum-as-is representation) to enhance computer vision applications. Hu et al. (2017) proposes a low-rank tensor-based fusion framework to improve the face recognition performance using the fusion of facial attribute information. However, none of these previous work aims to apply low-rank tensor techniques for multimodal fusion.

Our Low-rank Multimodal Fusion method provides a much more efficient method to compute tensor-based multimodal representations with much fewer parameters and computational complexity. The efficiency and performance of our approach are evaluated on different downstream tasks, namely sentiment analysis, speaker-trait recognition and emotion recognition.

3 Low-rank Multimodal Fusion

In this section, we start by formulating the problem of multimodal fusion and introducing fusion methods based on tensor representations. Tensors are powerful in their expressiveness but do not scale well to a large number of modalities. Our proposed model decomposes the weights into low-rank factors, which reduces the number of parameters in the model. This decomposition can be performed efficiently by exploiting the parallel decomposition of low-rank weight tensor and input tensor to compute tensor-based fusion. Our method is able to scale linearly with the number of modalities.

3.1 Multimodal Fusion using Tensor Representations

In this paper, we formulate multimodal fusion as a multilinear function $f : V_1 \times V_2 \times ... \times V_M \rightarrow H$ where $V_1, V_2, ..., V_M$ are the vector spaces of input modalities and H is the output vector space. Given a set of vector representations, $\{z_m\}_{m=1}^{M}$ which are encoding unimodal information of the M different modalities, the goal of multimodal fusion is to integrate the unimodal representations into one compact multimodal representation for downstream tasks.

Tensor representation is one successful approach for multimodal fusion. It first requires a transformation of the input representations into a high-dimensional tensor and then mapping it back to a lower-dimensional output vector space. Previous works have shown that this method is more effective than simple concatenation or pooling in terms of capturing multimodal interactions (Zadeh et al., 2017), (Fukui et al., 2016). Tensors are usually created by taking the outer product over the input modalities. In addition, in order to be able to model the interactions between any subset of modalities using one tensor, Zadeh et al. (2017) proposed a simple extension to append 1s to the unimodal representations before taking the outer product. The input tensor $\mathcal{Z}$ formed by the unimodal representation is computed by:

$$\mathcal{Z} = \bigotimes_{m=1}^{M} z_m, z_m \in \mathbb{R}^{d_m} \quad (1)$$

where $\bigotimes_{m=1}^{M}$ denotes the tensor outer product over a set of vectors indexed by m, and z_m is the input representation with appended 1s.

The input tensor $\mathcal{Z} \in \mathbb{R}^{d_1 \times d_2 \times ... d_M}$ is then passed through a linear layer $g(\cdot)$ to to produce a vector representation:

$$h = g(\mathcal{Z}; \mathcal{W}, b) = \mathcal{W} \cdot \mathcal{Z} + b, \; h, b \in \mathbb{R}^{d_y} \quad (2)$$

where $\mathcal{W}$ is the weight of this layer and b is the bias. With $\mathcal{Z}$ being an order-M tensor (where M is the number of input modalities), the weight $\mathcal{W}$ will naturally be a tensor of order-$(M+1)$ in $\mathbb{R}^{d_1 \times d_2 \times ... \times d_M \times d_h}$. The extra $(M+1)$-th dimension corresponds to the size of the output representation d_h. In the tensor dot product $\mathcal{W} \cdot \mathcal{Z}$, the weight tensor $\mathcal{W}$ can be then viewed as d_h order-M tensors. In other words, the weight $\mathcal{W}$ can be partitioned into $\widetilde{\mathcal{W}}_k \in \mathbb{R}^{d_1 \times ... \times d_M}, \; k = 1, ..., d_h$. Each $\widetilde{\mathcal{W}}_k$ contributes to one dimension in the output vector h, i.e. $h_k = \widetilde{\mathcal{W}}_k \cdot \mathcal{Z}$. This interpretation of tensor fusion is illustrated in Figure 2 for the bi-modal case.

One of the main drawbacks of tensor fusion is that we have to explicitly create the high-dimensional tensor $\mathcal{Z}$. The dimensionality of $\mathcal{Z}$

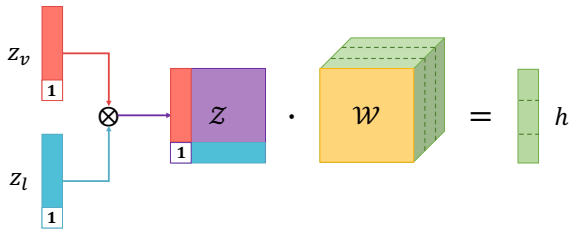

Figure 2: Tensor fusion via tensor outer product

will increase exponentially with the number of modalities as $\prod_{m=1}^{M} d_m$. The number of parameters to learn in the weight tensor $\mathcal{W}$ will also increase exponentially. This not only introduces a lot of computation but also exposes the model to risks of overfitting.

3.2 Low-rank Multimodal Fusion with Modality-Specific Factors

As a solution to the problems of tensor-based fusion, we propose Low-rank Multimodal Fusion (LMF). LMF parameterizes $g(\cdot)$ from Equation 2 with a set of modality-specific low-rank factors that can be used to recover a low-rank weight tensor, in contrast to the full tensor $\mathcal{W}$. Moreover, we show that by decomposing the weight into a set of low-rank factors, we can exploit the fact that the tensor $\mathcal{Z}$ actually decomposes into $\{z_m\}_{m=1}^{M}$, which allows us to directly compute the output h without explicitly tensorizing the unimodal representations. LMF reduces the number of parameters as well as the computation complexity involved in tensorization from being exponential in M to linear.

3.2.1 Low-rank Weight Decomposition

The idea of LMF is to decompose the weight tensor $\mathcal{W}$ into M sets of modality-specific factors. However, since $\mathcal{W}$ itself is an order-$(M+1)$ tensor, commonly used methods for decomposition will result in $M+1$ parts. Hence, we still adopt the view introduced in Section 3.1 that $\mathcal{W}$ is formed by d_h order-M tensors $\widetilde{\mathcal{W}}_k \in \mathbb{R}^{d_1 \times \ldots \times d_M}, k = 1, \ldots, d_h$ stacked together. We can then decompose each $\widetilde{\mathcal{W}}_k$ separately.

For an order-M tensor $\widetilde{\mathcal{W}}_k \in \mathbb{R}^{d_1 \times \ldots \times d_M}$, there always exists an exact decomposition into vectors in the form of:

$$\widetilde{\mathcal{W}}_k = \sum_{i=1}^{R} \bigotimes_{m=1}^{M} w_{m,k}^{(i)}, \; w_{m,k}^{(i)} \in \mathbb{R}^{d}_{m} \tag{3}$$

The minimal R that makes the decomposition valid is called the **rank** of the tensor. The vector sets $\{\{w_{m,k}^{(i)}\}_{m=1}^{M}\}_{i=1}^{R}$ are called the rank R decomposition factors of the original tensor.

In LMF, we start with a fixed rank r, and parameterize the model with r decomposition factors $\{\{w_{m,k}^{(i)}\}_{m=1}^{M}\}_{i=1}^{r}, k = 1, \ldots, d_h$ that can be used to reconstruct a low-rank version of these $\widetilde{\mathcal{W}}_k$.

We can regroup and concatenate these vectors into M modality-specific low-rank factors. Let $\mathbf{w}_m^{(i)} = [w_{m,1}^{(i)}, w_{m,2}^{(i)}, \ldots, w_{m,d_h}^{(i)}]$, then for modality m, $\{\mathbf{w}_m^{(i)}\}_{i=1}^{r}$ is its corresponding low-rank factors. And we can recover a low-rank weight tensor by:

$$\mathcal{W} = \sum_{i=1}^{r} \bigotimes_{m=1}^{M} \mathbf{w}_m^{(i)} \tag{4}$$

Hence equation 2 can be computed by

$$h = \left(\sum_{i=1}^{r} \bigotimes_{m=1}^{M} \mathbf{w}_m^{(i)} \right) \cdot \mathcal{Z} \tag{5}$$

Note that for all m, $\mathbf{w}_m^{(i)} \in \mathbb{R}^{d_m \times d_h}$ shares the same size for the second dimension. We define their outer product to be over only the dimensions that are not shared: $\mathbf{w}_m^{(i)} \otimes \mathbf{w}_n^{(i)} \in \mathbb{R}^{d_m \times d_n \times d_h}$. A bimodal example of this procedure is illustrated in Figure 3.

Nevertheless, by introducing the low-rank factors, we now have to compute the reconstruction of $\mathcal{W} = \sum_{i=1}^{r} \bigotimes_{m=1}^{M} \mathbf{w}_m^{(i)}$ for the forward computation. Yet this introduces even more computation.

3.2.2 Efficient Low-rank Fusion Exploiting Parallel Decomposition

In this section, we will introduce an efficient procedure for computing h, exploiting the fact that tensor $\mathcal{Z}$ naturally decomposes into the original input $\{z_m\}_{m=1}^{M}$, which is parallel to the modality-specific low-rank factors. In fact, that is the main reason why we want to decompose the weight tensor into M modality-specific factors.

Using the fact that $\mathcal{Z} = \bigotimes_{m=1}^{M} z_m$, we can simplify equation 5:

$$\begin{aligned} h &= \left(\sum_{i=1}^{r} \bigotimes_{m=1}^{M} \mathbf{w}_m^{(i)} \right) \cdot \mathcal{Z} \\ &= \sum_{i=1}^{r} \left(\bigotimes_{m=1}^{M} \mathbf{w}_m^{(i)} \cdot \mathcal{Z} \right) \\ &= \sum_{i=1}^{r} \left(\bigotimes_{m=1}^{M} \mathbf{w}_m^{(i)} \cdot \bigotimes_{m=1}^{M} z_m \right) \\ &= \bigwedge_{m=1}^{M} \left[\sum_{i=1}^{r} \mathbf{w}_m^{(i)} \cdot z_m \right] \end{aligned} \tag{6}$$

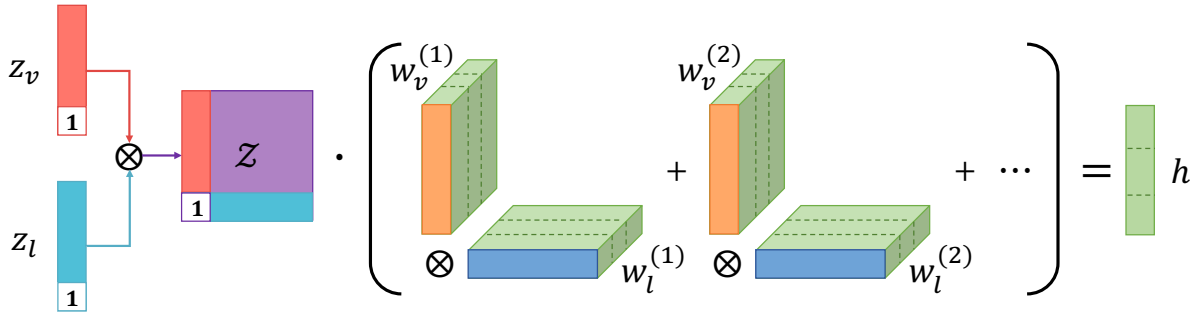

Figure 3: Decomposing weight tensor into low-rank factors (See Section 3.2.1 for details.)

where $\bigwedge_{m=1}^{M}$ denotes the element-wise product over a sequence of tensors: $\bigwedge_{t=1}^{3} x_t = x_1 \circ x_2 \circ x_3$.

An illustration of the trimodal case of equation 6 is shown in Figure 1. We can also derive equation 6 for a bimodal case to clarify what it does:

$$
\begin{aligned}
h &= \left(\sum_{i=1}^{r} \mathbf{w}_a^{(i)} \otimes \mathbf{w}_v^{(i)} \right) \cdot \mathcal{Z} \\
&= \left(\sum_{i=1}^{r} \mathbf{w}_a^{(i)} \cdot z_a \right) \circ \left(\sum_{i=1}^{r} \mathbf{w}_v^{(i)} \cdot z_v \right)
\end{aligned} \tag{7}
$$

An important aspect of this simplification is that it exploits the parallel decomposition of both $\mathcal{Z}$ and $\mathcal{W}$, so that we can compute h without actually creating the tensor $\mathcal{Z}$ from the input representations z_m. In addition, different modalities are decoupled in the simplified computation of h, which allows for easy generalization of our approach to an arbitrary number of modalities. Adding a new modality can be simply done by adding another set of modality-specific factors and extend Equation 7. Last but not least, Equation 6 consists of fully differentiable operations, which enables the parameters $\{\mathbf{w}_m^{(i)}\}_{i=1}^{r}\, m = 1, ..., M$ to be learned end-to-end via back-propagation.

Using Equation 6, we can compute h directly from input unimodal representations and their modal-specific decomposition factors, avoiding the weight-lifting of computing the large input tensor $\mathcal{Z}$ and $\mathcal{W}$, as well as the r linear transformation. Instead, the input tensor and subsequent linear projection are computed implicitly together in Equation 6, and this is far more efficient than the original method described in Section 3.1. Indeed, LMF reduces the computation complexity of tensorization and fusion from $O(d_y \prod_{m=1}^{M} d_m)$ to $O(d_y \times r \times \sum_{m=1}^{M} d_m)$.

In practice, we use a slightly different form of Equation 6, where we concatenate the low-rank factors into M order-3 tensors and swap the order in which we do the element-wise product and summation:

$$
h = \sum_{i=1}^{r} \left[\bigwedge_{m=1}^{M} \left[\mathbf{w}_m^{(1)}, \mathbf{w}_m^{(2)}, ..., \mathbf{w}_m^{(r)} \right] \cdot \hat{z}_m \right]_{i,:} \tag{8}
$$

and now the summation is done along the first dimension of the bracketed matrix. $[\cdot]_{i,:}$ indicates the i-th slice of a matrix. In this way, we can parameterize the model with M order-3 tensors, instead of parameterizing with sets of vectors.

4 Experimental Methodology

We compare LMF with previous state-of-the-art baselines, and we use the Tensor Fusion Networks (TFN) (Zadeh et al., 2017) as a baseline for tensor-based approaches, which has the most similar structure with us except that it explicitly forms the large multi-dimensional tensor for fusion across different modalities.

We design our experiments to better understand the characteristics of LMF. Our goal is to answer the following four research questions:

(1) Impact of Multimodal Low-rank Fusion: Direct comparison between our proposed LMF model and the previous TFN model.

(2) Comparison with the State-of-the-art: We evaluate the performance of LMF and state-of-the-art baselines on three different tasks and datasets.

(3) Complexity Analysis: We study the modal complexity of LMF and compare it with the TFN model.

(4) Rank Settings: We explore performance of LMF with different rank settings.

The results of these experiments are presented in Section 5.

4.1 Datasets

We perform our experiments on the following multimodal datasets, CMU-MOSI (Zadeh et al., 2016a),

Dataset	CMU-MOSI	IEMOCAP	POM
Level	Segment	Segment	Video
# Train	1284	6373	600
# Valid	229	1775	100
# Test	686	1807	203

Table 1: The speaker independent data splits for training, validation, and test sets.

POM (Park et al., 2014b), and IEMOCAP (Busso et al., 2008) for sentiment analysis, speaker traits recognition, and emotion recognition task, where the goal is to identify speakers emotions based on the speakers' verbal and nonverbal behaviors.

CMU-MOSI The CMU-MOSI dataset is a collection of 93 opinion videos from YouTube movie reviews. Each video consists of multiple opinion segments and each segment is annotated with the sentiment in the range [-3,3], where -3 indicates highly negative and 3 indicates highly positive.

POM The POM dataset is composed of 903 movie review videos. Each video is annotated with the following speaker traits: confident, passionate, voice pleasant, dominant, credible, vivid, expertise, entertaining, reserved, trusting, relaxed, outgoing, thorough, nervous, persuasive and humorous.

IEMOCAP The IEMOCAP dataset is a collection of 151 videos of recorded dialogues, with 2 speakers per session for a total of 302 videos across the dataset. Each segment is annotated for the presence of 9 emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed and neutral).

To evaluate model generalization, all datasets are split into training, validation, and test sets such that the splits are speaker independent, i.e., no identical speakers from the training set are present in the test sets. Table 1 illustrates the data splits for all datasets in detail.

4.2 Features

Each dataset consists of three modalities, namely language, visual, and acoustic modalities. To reach the same time alignment across modalities, we perform word alignment using P2FA (Yuan and Liberman, 2008) which allows us to align the three modalities at the word granularity. We calculate the visual and acoustic features by taking the average of their feature values over the word time interval (Chen et al., 2017).

Language We use pre-trained 300-dimensional Glove word embeddings (Pennington et al., 2014) to encode a sequence of transcribed words into a sequence of word vectors.

Visual The library Facet[1] is used to extract a set of visual features for each frame (sampled at 30Hz) including 20 facial action units, 68 facial landmarks, head pose, gaze tracking and HOG features (Zhu et al., 2006).

Acoustic We use COVAREP acoustic analysis framework (Degottex et al., 2014) to extract a set of low-level acoustic features, including 12 Mel frequency cepstral coefficients (MFCCs), pitch, voiced/unvoiced segmentation, glottal source, peak slope, and maxima dispersion quotient features.

4.3 Model Architecture

In order to compare our fusion method with previous work, we adopt a simple and straightforward model architecture [2] for extracting unimodal representations. Since we have three modalities for each dataset, we simply designed three unimodal sub-embedding networks, denoted as f_a, f_v, f_l, to extract unimodal representations z_a, z_v, z_l from unimodal input features x_a, x_v, x_l. For acoustic and visual modality, the sub-embedding network is a simple 2-layer feed-forward neural network, and for language modality, we used an LSTM (Hochreiter and Schmidhuber, 1997) to extract representations. The model architecture is illustrated in Figure 1.

4.4 Baseline Models

We compare the performance of LMF to the following baselines and state-of-the-art models in multimodal sentiment analysis, speaker trait recognition, and emotion recognition.

Support Vector Machines Support Vector Machines (SVM) (Cortes and Vapnik, 1995) is a widely used non-neural classifier. This baseline is trained on the concatenated multimodal features for classification or regression task (Pérez-Rosas et al., 2013), (Park et al., 2014a), (Zadeh et al., 2016b).

Deep Fusion The Deep Fusion model (DF) (Nojavanasghari et al., 2016) trains one deep neural model for each modality and then combine the output of each modality network with a joint neural network.

Tensor Fusion Network The Tensor Fusion Network (TFN) (Zadeh et al., 2017) explicitly models view-specific and cross-view dynamics by creating a multi-dimensional tensor that captures uni-

[1] goo.gl/1rh1JN

[2] The source code of our model is available on Github at https://github.com/Justin1904/Low-rank-Multimodal-Fusion

modal, bimodal and trimodal interactions across three modalities.

Memory Fusion Network The Memory Fusion Network (MFN) (Zadeh et al., 2018a) accounts for view-specific and cross-view interactions and continuously models them through time with a special attention mechanism and summarized through time with a Multi-view Gated Memory.

Bidirectional Contextual LSTM The Bidirectional Contextual LSTM (BC-LSTM) (Zadeh et al., 2017), (Fukui et al., 2016) performs context-dependent fusion of multimodal data.

Multi-View LSTM The Multi-View LSTM (MV-LSTM) (Rajagopalan et al., 2016) aims to capture both modality-specific and cross-modality interactions from multiple modalities by partitioning the memory cell and the gates corresponding to multiple modalities.

Multi-attention Recurrent Network The Multi-attention Recurrent Network (MARN) (Zadeh et al., 2018b) explicitly models interactions between modalities through time using a neural component called the Multi-attention Block (MAB) and storing them in the hybrid memory called the Long-short Term Hybrid Memory (LSTHM).

4.5 Evaluation Metrics

Multiple evaluation tasks are performed during our evaluation: multi-class classification and regression. The multi-class classification task is applied to all three multimodal datasets, and the regression task is applied to the CMU-MOSI and the POM dataset. For binary classification and multi-class classification, we report F1 score and accuracy Acc$-k$ where k denotes the number of classes. Specifically, Acc-2 stands for the binary classification. For regression, we report Mean Absolute Error (MAE) and Pearson correlation (Corr). Higher values denote better performance for all metrics except for MAE.

5 Results and Discussion

In this section, we present and discuss the results from the experiments designed to study the research questions introduced in section 4.

5.1 Impact of Low-rank Multimodal Fusion

In this experiment, we compare our model directly with the TFN model since it has the most similar structure to our model, except that TFN explicitly forms the multimodal tensor fusion. The comparison reported in the last two rows of Table 2 demonstrates that our model significantly outperforms TFN across all datasets and metrics. This competitive performance of LMF compared to TFN emphasizes the advantage of Low-rank Multimodal Fusion.

5.2 Comparison with the State-of-the-art

We compare our model with the baselines and state-of-the-art models for sentiment analysis, speaker traits recognition and emotion recognition. Results are shown in Table 2. LMF is able to achieve competitive and consistent results across all datasets.

On the multimodal sentiment regression task, LMF outperforms the previous state-of-the-art model on MAE and Corr. Note the multiclass accuracy is calculated by mapping the range of continuous sentiment values into a set of intervals that are used as discrete classes.

On the multimodal speaker traits Recognition task, we report the average evaluation score over 16 speaker traits and shows that our model achieves the state-of-the-art performance over all three evaluation metrics on the POM dataset.

On the multimodal emotion recognition task, our model achieves better results compared to the state-of-the-art models across all emotions on the F1 score. F1-emotion in the evaluation metrics indicates the F1 score for a certain emotion class.

5.3 Complexity Analysis

Theoretically, the model complexity of our fusion method is $O(d_y \times r \times \sum_{m=1}^{M} d_m)$ compared to $O(d_y \prod_{m=1}^{M} d_m)$ of TFN from Section 3.1. In practice, we calculate the total number of parameters used in each model, where we choose $M = 3$, $d_1 = 32$, $d_2 = 32$, $d_3 = 64$, $r = 4$, $d_y = 1$. Under this hyper-parameter setting, our model contains about 1.1e6 parameters while TFN contains about 12.5e6 parameters, which is nearly 11 times more. Note that, the number of parameters above counts not only the parameters in the multimodal fusion stage but also the parameters in the subnetworks.

Furthermore, we evaluate the computational complexity of LMF by measuring the training and testing speeds between LMF and TFN. Table 3 illustrates the impact of Low-rank Multimodal Fusion on the training and testing speeds compared with TFN model. Here we set rank to be 4 since it can generally achieve fairly competent performance.

Dataset	CMU-MOSI					POM			IEMOCAP			
Metric	MAE	Corr	Acc-2	F1	Acc-7	MAE	Corr	Acc	F1-Happy	F1-Sad	F1-Angry	F1-Neutral
SVM	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	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	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	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	0.968	0.625	77.1	77.0	34.7	-	-	39.4	83.6	81.2	84.2	65.9
MFN	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	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	**0.912**	**0.668**	76.4	75.7	32.8	**0.796**	**0.396**	**42.8**	**85.8**	**85.9**	**89.0**	**71.7**

Table 2: Results for sentiment analysis on CMU-MOSI, emotion recognition on IEMOCAP and personality trait recognition on POM. Best results are highlighted in bold.

Model	Training Speed (IPS)	Testing Speed (IPS)
TFN	340.74	1177.17
LMF	1134.82	2249.90

Table 3: Comparison of the training and testing speeds between TFN and LMF. The second and the third columns indicate the number of data point inferences per second (IPS) during training and testing time respectively. Both models are implemented in the same framework with equivalent running environment.

Based on these results, performing a low-rank multimodal fusion with modality-specific low-rank factors significantly reduces the amount of time needed for training and testing the model. On an NVIDIA Quadro K4200 GPU, LMF trains with an average frequency of 1134.82 IPS (data point inferences per second) while the TFN model trains at an average of 340.74 IPS.

5.4 Rank Settings

To evaluate the impact of different rank settings for our LMF model, we measure the change in performance on the CMU-MOSI dataset while varying the number of rank. The results are presented in Figure 4. We observed that as the rank increases, the training results become more and more unstable and that using a very low rank is enough to achieve fairly competent performance.

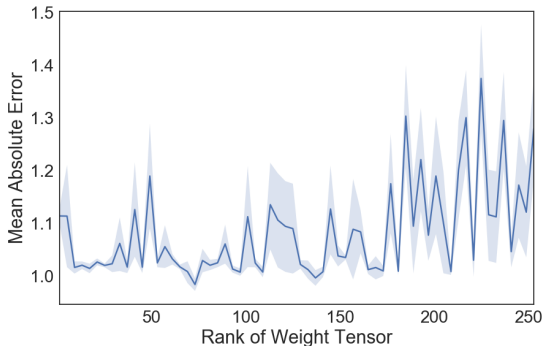

Figure 4: The Impact of different rank settings on Model Performance: As the rank increases, the results become unstable and low rank is enough in terms of the mean absolute error.

6 Conclusion

In this paper, we introduce a Low-rank Multimodal Fusion method that performs multimodal fusion with modality-specific low-rank factors. LMF scales linearly in the number of modalities. LMF achieves competitive results across different multimodal tasks. Furthermore, LMF demonstrates a significant decrease in computational complexity from exponential to linear time. In practice, LMF effectively improves the training and testing efficiency compared to TFN which performs multimodal fusion with tensor representations.

Future work on similar topics could explore the applications of using low-rank tensors for attention models over tensor representations, as they can be even more memory and computationally intensive.

Acknowledgements

This material is based upon work partially supported by the National Science Foundation (Award # 1833355) and Oculus VR. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation or Oculus VR, and no official endorsement should be inferred.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Lan-*

guage Resources and Evaluation 42(4):335–359. https://doi.org/10.1007/s10579-008-9076-6.

Lawrence S Chen, Thomas S Huang, Tsutomu Miyasato, and Ryohei Nakatsu. 1998. Multimodal human emotion/expression recognition. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, pages 366–371.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, New York, NY, USA, ICMI 2017, pages 163–171. https://doi.org/10.1145/3136755.3136801.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20(3):273–297.

Liyanage C De Silva, Tsutomu Miyasato, and Ryohei Nakatsu. 1997. Facial emotion recognition using multi-modal information. In *Information, Communications and Signal Processing, 1997. ICICS., Proceedings of 1997 International Conference on*. IEEE, volume 1, pages 397–401.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarepa collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, pages 960–964.

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. *arXiv preprint arXiv:1606.01847* .

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Guosheng Hu, Yang Hua, Yang Yuan, Zhihong Zhang, Zheng Lu, Sankha S Mukherjee, Timothy M Hospedales, Neil M Robertson, and Yongxin Yang. 2017. Attribute-enhanced face recognition with neural tensor fusion networks.

Othmar Koch and Christian Lubich. 2010. Dynamical tensor approximation. *SIAM Journal on Matrix Analysis and Applications* 31(5):2360–2375.

Tao Lei, Yu Xin, Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. 2014. Low-rank tensors for scoring dependency structures. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 1381–1391.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interactions*. ACM, pages 169–176.

Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, pages 284–288.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014a. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, pages 50–57.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014b. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, New York, NY, USA, ICMI '14, pages 50–57. https://doi.org/10.1145/2663204.2663260.

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

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 973–982.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, pages 439–448.

Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrušaitis, and Roland Goecke. 2016. Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision*.

Ilya Razenshteyn, Zhao Song, and David P Woodruff. 2016. Weighted low rank approximations with provable guarantees. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM, pages 250–263.

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. 2016. Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis. *arXiv preprint arXiv:1609.05244* .

Hongcheng Wang and Narendra Ahuja. 2008. A tensor approximation approach to dimensionality reduction. *International Journal of Computer Vision* 76(3):217–229.

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* 28(3):46–53.

Torsten Wortwein and Stefan Scherer. 2017. What really mattersan information gain analysis of questions and reactions in automated ptsd screenings. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, pages 15–20.

Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America* 123(5):3878.

Ben P Yuhas, Moise H Goldstein, and Terrence J Sejnowski. 1989. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine* 27(11):65–71.

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

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. *arXiv preprint arXiv:1802.00927* .

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. *arXiv preprint arXiv:1802.00923* .

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* 31(6):82–88.

Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, and Shai Avidan. 2006. Fast human detection using a cascade of histograms of oriented gradients. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, volume 2, pages 1491–1498.