Modality-based Factorization for Multimodal Fusion

Elham J. Barezi, Pascale Fung
Center for Artificial Intelligence Research (CAiRE)
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
ejs@cse.ust.hk,pascale@ust.hk

Abstract

We propose a novel method, Modality-based Redundancy Reduction Fusion (MRRF), for understanding and modulating the relative contribution of each modality in multimodal inference tasks. This is achieved by obtaining an $(M+1)$-way tensor to consider the high-order relationships between M modalities and the output layer of a neural network model. Applying a modality-based tensor factorization method, which adopts different factors for different modalities, results in removing information present in a modality that can be compensated by other modalities, with respect to model outputs. This helps to understand the relative utility of information in each modality. In addition it leads to a less complicated model with less parameters and therefore could be applied as a regularizer avoiding overfitting. We have applied this method to three different multimodal datasets in sentiment analysis, personality trait recognition, and emotion recognition. We are able to recognize relationships and relative importance of different modalities in these tasks and achieves a 1% to 4% improvement on several evaluation measures compared to the state-of-the-art for all three tasks.

1 Introduction

Multimodal data fusion is a desirable method for many machine learning tasks where information is available from multiple source modalities, typically achieving better predictions through integration of information from different modalities. Multimodal integration can handle missing data from one or more modalities. Since some modalities can include noise, it can also lead to more robust prediction. Moreover, since some information may not be visible in some modalities or a single modality may not be powerful enough for a specific task, considering multiple modalities often improves performance (Potamianos et al., 2003; Soleymani et al., 2012; Kampman et al., 2018).

For example, humans assign personality traits to each other, as well as to virtual characters by inferring personality from diverse cues, both behavioral and verbal, suggesting that a model to predict personality should take into account multiple modalities such as language, speech, and visual cues.

Our method, Modality-based Redundancy Reduction multimodal Fusion (MRRF), builds on recent work in mutimodal fusion utilizing first an outer product tensor of input modalities to better capture inter-modality dependencies (Zadeh et al., 2017) and a recent approach to reduce the number of elements in the resulting tensor through low rank factorization (Liu et al., 2018). Whereas the factorization used in (Liu et al., 2018) utilizes a single compression rate across all modalities, we instead use Tuckers tensor decomposition (see the Methodology section), which allows different compression rates for each modality. This allows the model to adapt to variations in the amount of useful information between modalities. Modality-specific factors are chosen by maximizing performance on a validation set.

Applying a modality-based factorization method results in removing redundant information duplicated across modalities and leading to fewer parameters with minimal information loss. Through maximizing performance on a validation set, our method can work as a regularizer, leading to a less complicated model and reducing overfitting. In addition, our modality-based factorization approach helps to understand the differences in useful information between modalities for the task at hand.

We evaluate the performance of our approach using sentiment analysis, personality detection, and emotion recognition from audio, text and video frames. The method reduces the number of pa-

Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pages 260–269
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

rameters which requires fewer training samples, providing efficient training for the smaller datasets, and accelerating both training and prediction. Our experimental results demonstrate that the proposed approach can make notable improvements, in terms of accuracy, mean average error (MAE), correlation, and F_1 score, especially for the applications with more complicated inter-modality relations.

We further study the effect of different compression rates for different modalities. Our results on the importance of each modality for each task supports the previous results on the usefulness of each modality for personality recognition, emotion recognition and sentiment analysis.

In the sequel, we first describe related work. We elaborate on the details of our proposed method in Methodology section. In the following section we go on to describe our experimental setup. In the Results section, we compare the performance of MRRF and state-of-the-art baselines on three different datasets and discuss the effect of compression rate on each modality. Finally, we provide a brief conclusion of the approach and the results. Supplementary materials describe the methodology in greater detail.

Notation The operator $\otimes$ is the outer product operator where $z_1 \otimes \ldots \otimes z_M$ for $z_i \in \mathbb{R}^{d_i}$ leads to a M-way tensor in $\mathbb{R}^{d_1 \times \ldots \times d_M}$. The operator $\times_k$, for a given k, is k-mode product of a tensor $R \in \mathbb{R}^{r_1 \times r_2 \times \ldots \times r_M}$ and a matrix $W \in \mathbb{R}^{d_k \times r_k}$ as $W \times_k R$, which results in a tensor $\bar{R} \in \mathbb{R}^{r_1 \times \ldots \times r_{k-1} \times d_k \times r_{k+1} \times \ldots \times r_M}$.

2 Related Work

Multimodal Fusion: Multimodal fusion (Ngiam et al., 2011) has a very broad range of applications, including audio-visual speech recognition (Potamianos et al., 2003), classification of images and their captions (Srivastava and Salakhutdinov, 2012), multimodal emotion recognition (Soleymani et al., 2012), medical image analysis (James and Dasarathy, 2014), multimedia event detection (Lan et al., 2014), personality trait detection (Kampman et al., 2018), and sentiment analysis (Zadeh et al., 2017).

According to the recent work by (Baltrušaitis et al., 2018), the techniques for multimodal fusion can be divided into early, late and hybrid approaches. Early approaches combine the multimodal features immediately by simply concatenating them (D'mello and Kory, 2015). Late fusion combines the decision for each modality (either classification, or regression), by voting (Morvant et al., 2014), averaging (Shutova et al., 2016) or weighted sum of the outputs of the learned models (Glodek et al., 2011; Shutova et al., 2016). The hybrid approach combines the prediction by early fusion and unimodal predictions.

It has been observed that early fusion (feature level fusion) concentrates on the inter-modality information rather than intra-modality information (Zadeh et al., 2017) due to the fact that inter-modality information can be more complicated at the feature level and dominates the learning process. On the other hand, these fusion approaches are not powerful enough to extract the inter-modality integration model and they are limited to some simple combining methods (Zadeh et al., 2017).

Zadeh et al. (2017) proposed combining n modalities by computing an n-way tensor as a tensor product of the n different modality representations followed by a flattening operation, in order to include 1-st order to n-th order inter modality relations. This is then fed to a neural network model to make predictions. The authors show that their proposed method improves the accuracy by considering both inter-modality and intra-modality relations. However, the generated representation has a very large dimension which leads to a very large hidden layer and therefore a huge number of parameters.

The authors of (Poria et al., 2017a,b; Zadeh et al., 2018a,b) introduce attention mechanisms utilizing the contextual information available from the utterances for each speaker. They require additional information like the identity of the speaker, the sequence of the utterance-sentiments while integrating the multimodal data. Since these methods, despite our proposed method, need additional information might not be available in the general scenario, we do not include them in our experiments.

Low Rank Factorization: Recently (Liu et al., 2018) proposed a factorization approach in order to achieve a factorized version of the weight matrix which leads to fewer parameters while maintaining model accuracy. They use a CANDECOMP/PARAFAC decomposition (Carroll and Chang, 1970; Harshman, 1970) which follows Eq. 1 in order to decompose a tensor $W \in \mathbb{R}^{d_1 \times \ldots d_M}$

to several 1-dimensional vectors $w_m^i \in \mathbb{R}^{d_k}$:

$$\begin{aligned} W &= \sum_{i=1}^{r} \lambda_i w_1^i \otimes w_2^i \otimes \ldots \otimes w_M^i \\ &= \sum_{i=1}^{r} \lambda_i \otimes_{m=1}^{M} w_m^i \end{aligned} \tag{1}$$

where $\otimes$ is the outer product operator, λ_is are scalar weights to combine rank 1 decompositions. This approach used the same compression rate for all modalities, i.e. r is shared for all the modalities, and is not able to allow for varying compression rates between modalities. Previous studies have found that some modalities are more informative than others (De Silva et al., 1997; Kampman et al., 2018), suggesting that allowing different compression rates for different modalities should improve performance.

3 Methodology

3.1 Tucker Factorization for Multimodal Learning

Modality-based Redundancy Reduc- tion Fusion (MRRF): We have used Tucker's tensor decomposition method (Tucker, 1966; Hitchcock, 1927) which decomposes an M-way tensor $W \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_M}$ to a core tensor $R \in \mathbb{R}^{r_1 \times r_2 \times \ldots \times r_M}$ and M matrices $W_i \in \mathbb{R}^{r_i \times d_i}$, with $r_i \leq d_i$, as it can be seen in Eq. 2.

$$\begin{aligned} &W = R \times_1 W_1 \times_2 W_2 \times_3 \ldots \times_M W_M, \\ &\quad W \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_M} \\ &\quad R \in \mathbb{R}^{r_1 \times r_2 \times \ldots \times r_M}, \\ &\quad W_i \in \mathbb{R}^{d_i \times r_i} \end{aligned} \tag{2}$$

The operator $\times_k$ is a k-mode product of a tensor $R \in \mathbb{R}^{r_1 \times r_2 \times \ldots \times r_M}$ and a matrix $W \in \mathbb{R}^{d_k \times r_k}$ as $R \times_k W_k$, which results in a tensor $\bar{R} \in \mathbb{R}^{r_1 \times \ldots \times r_{k-1} \times d_k \times r_{k+1} \times \ldots \times r_M}$.

For M modalities with representations D_1, D_2, $\ldots$ and D_M of size d_1, d_2, $\ldots$ and d_M, an M-modal tensor fusion approach as proposed by the authors of (Zadeh et al., 2017) leads to a tensor $D = D_1 \otimes D_2 \otimes \ldots \otimes D_m \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_M}$. The authors proposed flattening the tensor layer in the deep network which results in loss of the information included in the tensor structure. In this paper, we propose to avoid the flattening and follow Eq. 3 with weight tensor $W \in \mathbb{R}^{h \times d_1 \times d_2 \times \ldots \times d_M}$, where leads to an output layer H of size h.

$$H = WD \tag{3}$$

The above equation suffers from a large number of parameters ($O(\prod_{i=1} d_i h)$) which requires a large number of the training samples, huge time and space, and easily overfits. In order to reduce the number of parameters, we propose to use Tucker's tensor decomposition (Tucker, 1966; Hitchcock, 1927) as shown in Eq. 4, which works as a low-rank regularizer (Fazel, 2002).

$$\begin{aligned} &W = R \times_1 W_1 \times_2 W_2 \times_3 \ldots \times_{M+1} W_{M+1}, \\ &\quad W \in \mathbb{R}^{h \times d_1 \times d_2 \times \ldots \times d_M}, \\ &\quad R \in \mathbb{R}^{r_1 \times r_2 \times r_3 \times \ldots \times r_M}, \\ &\quad W_i \in \mathbb{R}^{r_i \times d_i},\ i = \{1, \ldots, M\}, \\ &\quad W_{M+1} \in \mathbb{R}^{r_{M+1} \times h} \end{aligned} \tag{4}$$

The non-diagonal core tensor R maintain inter-modality information after compression, despite the factorization proposed by (Liu et al., 2018) which loses part of inter-modality information.

3.2 Proposed MRRF framework

We propose Modality-based Redundancy Reduction Fusion (MRRF), a tensor fusion and factorization method allowing for modality specific compression rates, combining the power of tensor fusion methods with a reduced parameter complexity. Without loss of generality, we will consider the number of modalities to be 3 in this discussion.

Our method first forms an outer product tensor from input modalities D, then projects this via a tensor W to a feature vector H passed as input to a neural network which performs the desired inference task.

$$H = WD \tag{5}$$

The trainable projection tensor W represents a large number of parameters, and in order to reduce this number, we propose to use Tucker's tensor decomposition (Tucker, 1966; Hitchcock, 1927), which works as a low-rank regularizer (Fazel, 2002). This results in a decomposition of W into a core tensor R of reduced dimensionality and three modality specific matrices W_i.

$$W = R \times_1 W_1 \times_2 W_2 \times_3 W_3 \tag{6}$$

where $\times_k$ is a k-mode product of a tensor and a matrix. Equation 5 can then be re-written

$$\begin{aligned} &Z = W_1 \times_1 W_2 \times_2 W_3 \times_3 D \\ &H = ZR \end{aligned} \tag{7}$$

See Figure 1 for an overview of this process for the case of three separate channels for audio, text, and video. In practice we flatten tensors Z and R to reduce this last operation to a matrix multiplication. Further details of the decomposition strategy can be found in the supplementary materials.

Note that a simple outer product of the input features leads only to the high-order trimodal dependencies. In order to also obtain the unimodal and bimodal dependencies, the input feature vectors for each modality are padded by 1. This also provides a constant element whose corresponding factors in W act as a bias vector.

Algorithm 1 shows the whole MRRF process.

Algorithm 1 Tensor Factorization Layer.

```
Input: $n$ input modalities $D_1, D_2, \dots, D_n$ of size $d_1, d_2, \dots, d_n$, correspondingly.
Initialization: factorization size for each modality $r_1, r_2, \dots, r_n$.

1: Compute tensor $D = D_1 \otimes D_2 \otimes \dots \otimes D_n$
2: Generate the layers for $out = WD$ which $W = \hat{R} \times_1 W_1 \times_2 \dots \times_M W_M$ in order to
   transform the high-dimensional tensor $D$ to the output $h$.
3: Use Adam optimizer for the differentiable tensor factorization layer to find the unknown
   parameters $W_1, W_2, \dots, W_n, \hat{R}$.
Output: Factors for Weight Matrix $W$: $W_1, W_2, \dots, W_n, R$.
```

The original tensor fusion approach as proposed in (Zadeh et al., 2017) flattened the tensor D which results in loss of the information included in the tensor structure, which is avoided in our approach. Liu et al. (2018) developed a similar approach to ours using a diagonal core tensor R, losing much inter-modality information. Our non-diagonal core tensor maintains key inter-modality information after compression.

Note that the factorization step is task dependent, included in the deep network structure and learned during network training. Thus, for follow-up learning tasks, we would learn a new factorization specific to the task at hand, typically also estimating optimal compression ratios as described in the discussion section. In this process, any shared, helpful information is retained, as demonstrated by our results.

Analysis of parameter complexity: Following our proposed approach, we have decomposed the trainable W tensor to four substantially smaller trainable matrices (W_1, W_2, W_3, R) leading to $O(\sum_{i=1}^{M}(d_i * r_i) + \prod_{i=1}^{M} r_i * h)$ parameters. Concat fusion (CF) leads to a layer size of $O(\sum_{i=1}^{M} d_i)$ and $O(\sum_{i=1}^{M} d_i * h)$ parameters.

The tensor fusion approach (TF), leads to a layer size of $O(\prod_{i=1}^{M} d_i)$, and $O(\prod_{i=1}^{M} d_i * h)$ parameters. The LMF approach (Liu et al., 2018) requires training $O(\sum_{i=1}^{M} r * h * d_i)$ parameters, where r is the rank used for all the modalities.

It can be seen that the number of parameters in the proposed approach is substantially fewer than the simple tensor fusion (TF) approach and comparable to the LMF approach.

4 Experimental Setup

4.1 Datasets

We perform our experiments on the following multimodal datasets: CMU-MOSI (Zadeh et al., 2016), POM (Park et al., 2014), and IEMOCAP (Busso et al., 2008) for sentiment analysis, speaker traits recognition, and emotion recognition, respectively. These tasks can be done by integrating both verbal and nonverbal behaviors of the persons.

The CMU-MOSI dataset is annotated on a seven-step scale as highly negative, negative, weakly negative, neutral, weakly positive, positive, highly positive which can be considered as a 7 class classification problem with 7 labels in the range $[-3, +3]$. The dataset is an annotated dataset of 2199 opinion utterances from 93 distinct YouTube movie reviews, each containing several opinion segments. Segments average of 4.2 seconds in length.

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.

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, disgust and neutral).

Each dataset consists of three modalities, namely language, visual, and acoustic. The visual and acoustic features are calculated by taking the average of their feature values over the word time

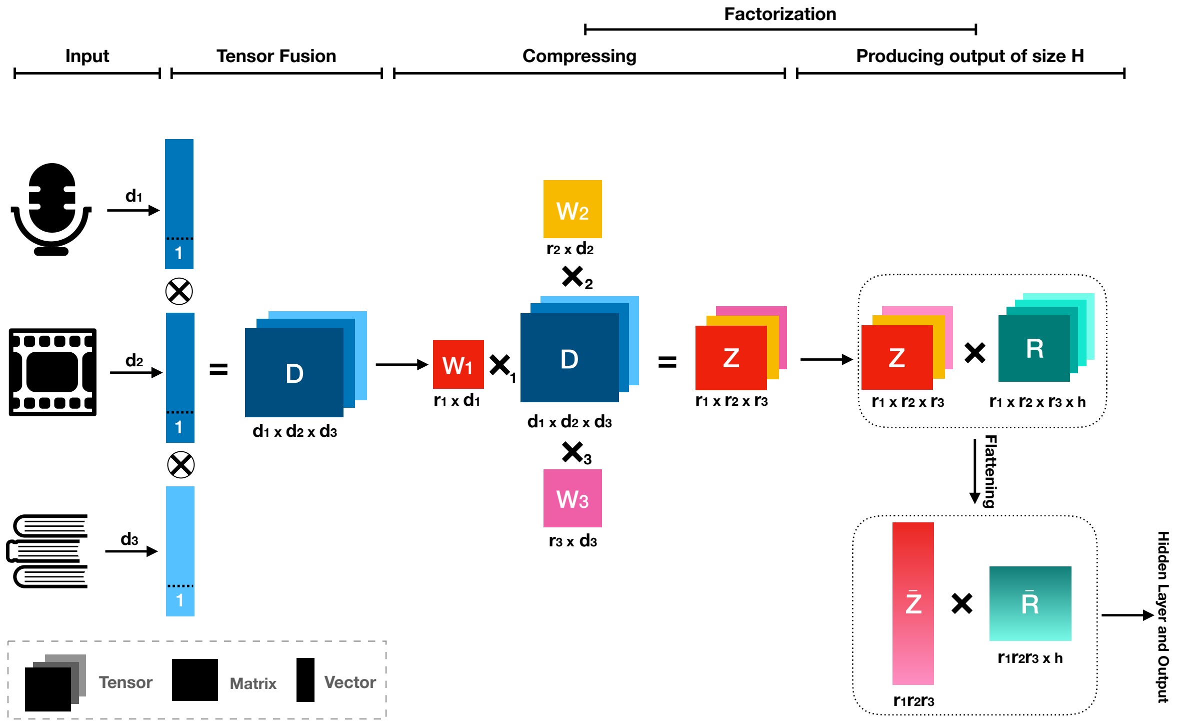

Figure 1: Diagram of Modality-based Redundancy Reduction Multimodal Fusion (MRRF).

interval (Chen et al., 2017). In order to perform time alignment across modalities, the three modalities are aligned using P2FA (Yuan and Liberman, 2008) at the word level.

Pre-trained 300-dimensional Glove word embeddings (Chen et al., 2017) were used to extract the language feature representations, which encodes a sequence of the transcribed words into a sequence of vectors.

Visual features for each frame (sampled at 30Hz) are extracted using the library Facet[1] which includes 20 facial action units, 68 facial landmarks, head pose, gaze tracking and HOG features (Zhu et al., 2006).

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

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

Dataset Level	CMU-MOSI Segment	IEMOCAP Segment	POM 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

4.2 Model Architecture

Similarly to (Liu et al., 2018), we use a simple model architecture for extracting the representations for each modality. We used three unimodal sub-embedding networks to extract representations z_a, z_v and z_l for each modality, respectively. For acoustic and visual modalities, the sub-embedding network is a simple 2-layer feed-forward neural network, and for language, we used a long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997).

We tuned the layer sizes, the learning rates and the compression rates, by checking the mean average error for the validation set by grid search. We trained our model using the Adam optimizer (Kingma and Ba, 2014). All models were implemented with Pytorch (Paszke et al., 2017).

[1] goo.gl/1rh1JN

Dataset	CMU-MOSI					POM			IEMOCAP			
Metric	MAE	Corr	Acc-2	F1	Acc-7	MAE	Corr	Acc	F1-Happy	F1-Sad	F1-Angry	F1-Neutral
CF	1.140	0.52	72.3	72.1	26.5	0.865	0.142	34.1	81.1	81.2	65.1	44.1
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
MRRF	0.912	0.772	77.46	76.73	33.02	0.69	0.44	43.02	87.71	85.9	90.02	73.7

Table 2: Results for Sentiment Analysis on CMU-MOSI, emotion recognition on IEMOCAP and personality trait recognition on POM. (CF, TF, and LMF stand for concat, tensor and low-rank fusion respectively).

5 Experimental Results and Comparing with State-of-the-art

We compared our proposed method with three baseline methods. Concat fusion (CF) (Baltrušaitis et al., 2018) proposes a simple concatenation of the different modalities followed by a linear combination. The tensor fusion approach (TF) (Zadeh et al., 2017) computes a tensor including uni-modal, bi-modal, and tri-modal combination information. LMF (Liu et al., 2018) is a tensor fusion method that performs tensor factorization using the same rank for all the modalities in order to reduce the number of parameters. Our proposed method aims to use different factors for each modality.

In Table 2, we present mean average error (MAE), the correlation between prediction and true scores, binary accuracy (Acc-2), multi-class accuracy (Acc-7) and F1 measure. The proposed approach outperforms baseline approaches in nearly all metrics, with marked improvements in Happy and Neutral recognition. The reason is that the inter-modality information for these emotions is more complicated than the other emotions and requires a non-diagonal core tensor to extract the complicated information. It is worth to note that for the equivalent setting and equal ranks for all the modalities, the result of the proposed method is always marginally better than LMF method.

5.1 Investigating the Effect of Compression Rate on Each Modality

In this section, we aim to investigate the amount of redundant information in each modality. To do this, after obtaining a tensor which includes the combinations of all modalities with the equivalent size, we factorize a single dimension of the tensor while keeping the size for the other modalities fixed. By observing how the performance changes by compression rate, one can find how much redundant information is contained in the corresponding modality relative to the other modalities.

The results can be seen in Fig. 2, 3 and 4. The horizontal axis is the compressed size and the vertical axis shows the accuracy for each modality. Note that due to the padding of each D_i with 1, we have used $r_i + 1$ as the new embedding size.

The first point that could be perceived clearly from the different modality diagrams is that each of the modalities changes in a different way when getting compressed, which means they each have a different amount of information that can not be compensated by the non-compressed modalities. In other words, a high accuracy when a modality is highly compressed means that there is a lot of redundant information in this modality — the information loss resulting from factorization could be compensated by the other modalities so performance was not reduced.

Fig. 2 shows results for the CMU-MOSI sentiment analysis dataset. For this dataset, a notable decrease in accuracy can be seen by compressing the video modality, while the audio and text modalities are not notably sensitive to compression. This shows that for sentiment analysis based on CMU-MOSI dataset, the information in Video modality cannot be compensated by other modalities, however most information in the audio and language modalities is covered in video modality. In other words, the video contains essential information for this task whereas information from audio and language can be recovered from video.

Fig. 3 shows the average accuracy over 16 personality types for the POM personality trait recognition dataset. For this dataset also, each of the modalities has a different behavior for different compression rates. We can see that the audio modality includes more non-redundant information for personality recognition as accuracy is highly affected by audio compression. In addition, there is a notable accuracy reduction when the language modality is highly compressed, which shows a small amount of non-redundant information for this task. Note that the POM data does not contain sufficient information for an effective analysis of the 16 personality sub types individually.

Fig. 4 shows the results for the IEMOCAP emo-

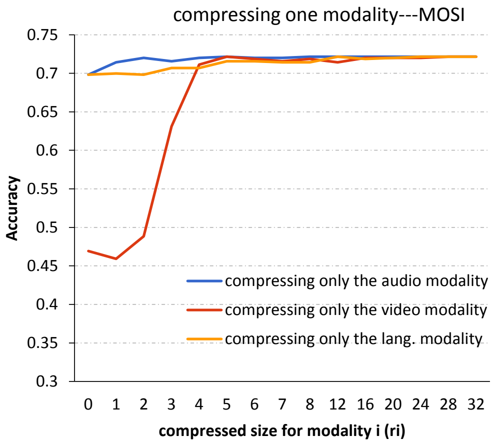

Figure 2: CMU-MOSI sentiment analysis dataset: Effect of different compression rates on accuracy for single modalities.

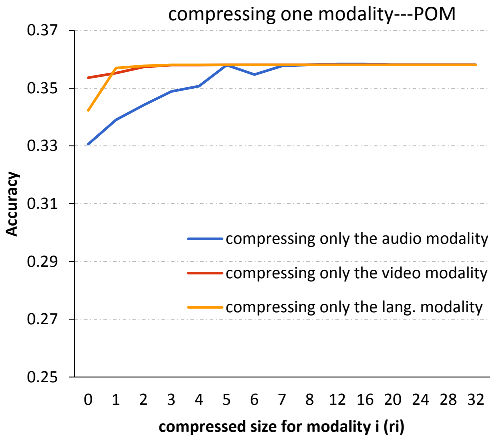

Figure 3: POM personality recognition dataset: Effect of different compression rates on accuracy for single modalities.

tion recognition dataset for each of the four emotional categories: happy, angry, sad, and neutral. Looking at the sad category, we see notable accuracy reduction for small sizes (high compression) for all the modalities, showing that each contains at least some non-redundant information. However, high compression of audio and especially language modalities results in strong accuracy reduction whereas video compression results in relatively minor reduction. It can be concluded that for this emotion, the language modality has the most non-redundant information and the video modality very little — it's information can be compensated by the other two modalities. Moving on to the angry emotion, small sizes (high compression) result in accuracy reduction for audio and language modalities, showing that they contain some non-redundant information, with the audio modality containing more. Again the information in video can be almost completely compensated by the other two modalities.

By comparing the highest accuracy values for various emotion categories, it is observed that neutral is hard to predict in comparison to the other categories. Again, the audio and Language modalities both include non-redundant information leading to a severe accuracy reduction with high compression of these modalities, with video containing almost no information not compensated by audio and language.

The happy category is the easiest to predict emotion, and it slightly suffers for very small sizes of audio and video and language modalities, indicating a small amount of non-redundant information in all modalities.

6 Conclusion

We proposed a tensor fusion method for multimodal media analysis by obtaining an $M + 1$-way tensor to consider the high-order relationships between M input modalities and the output layer. Our modality-based factorization method removes the redundant information in this high-order dependency structure and leads to fewer parameters with minimal loss of information. In addition, a modality-based factorization approach helps to understand the relative quantities of non-redundant information in each modality through investigation sensitivity to modality-specific compression rates. As the proposed compression method leads to a less complicated model, it can be applied as a regularizer which avoiding overfitting.

We have provided experimental results for combining acoustic, text, and visual modalities for three different tasks: sentiment analysis, personality trait recognition, and emotion recognition. We have seen that the modality-based tensor compression approach improves the results in comparison to the simple concatenation method, the tensor fusion method and tensor fusion using the same factorization rank for all modalities, as proposed in the LMF method. In other words, the proposed method enjoys the same benefits as the tensor fusion method and avoids suffering from having a large number of parameters, which leads to a more complex model, needs many training samples and is more prone to overfitting. We have investigated the effect of the compression rate on single modalities while

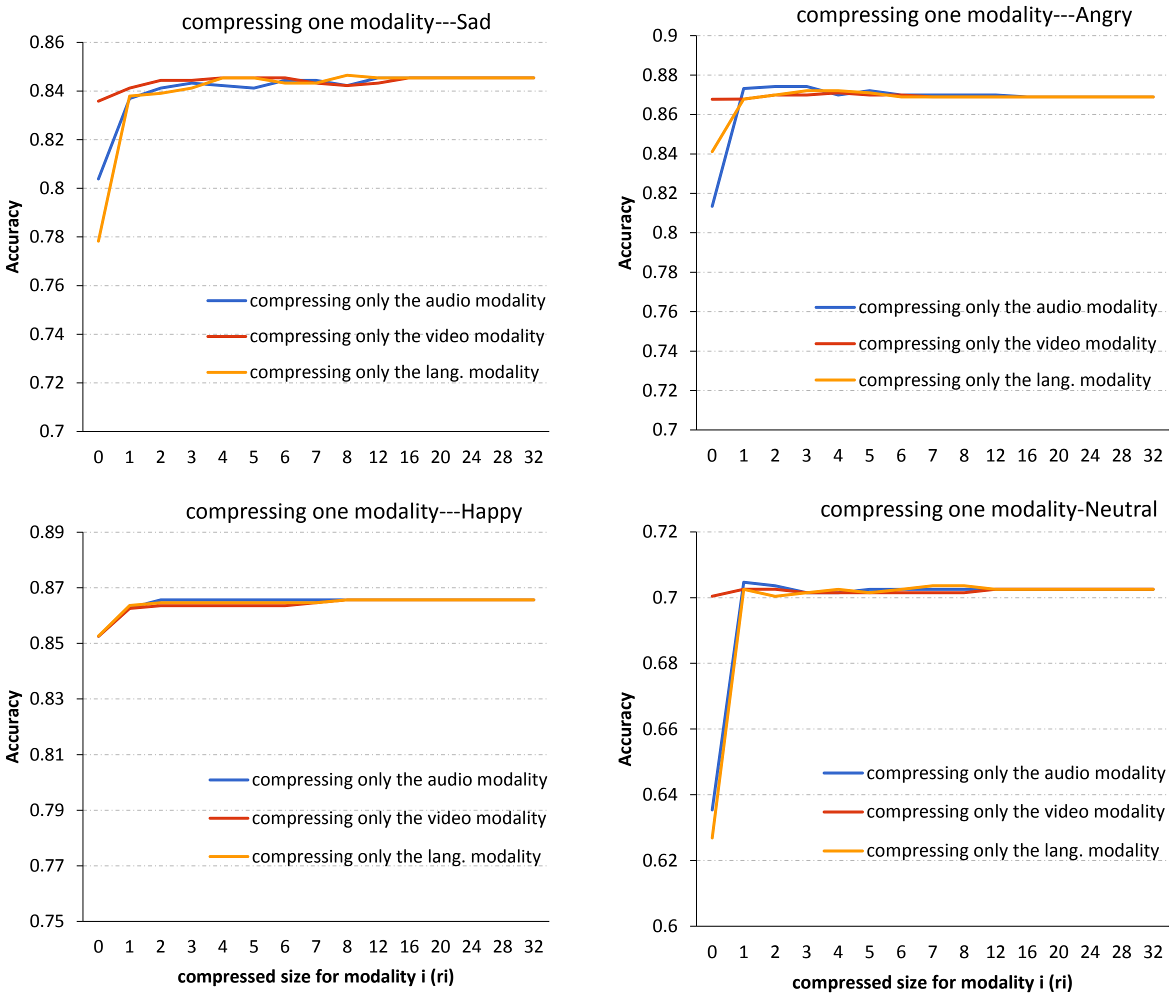

Figure 4: IEMOCAP Emotion Recognition Dataset: Effect of different compression rates on accuracy for single modalities.

fixing the other modalities helping to understand the amount of useful non-redundant information in each modality. Moreover, we have evaluated our method by comparing the results with state-of-the-art methods, achieving a 1% to 4% improvement across multiple measures for the different tasks.

In future work, we will investigate the relation between dataset size and compression rate by applying our method to larger datasets. This will help to understand the trade-off between the model size and available training data, allowing more efficient training and avoiding under- and overfitting.

As the availability of data with more and more modalities increases, both finding a trade-off between cost and performance and effective and efficient utilization of available modalities will be vital. Exploring compression methods promises to help identify and remove highly redundant modalities.

Acknowledgments

This work was partially funded by grants #16214415 and #16248016 of the Hong Kong Research Grants Council, ITS/319/16FP of Innovation Technology Commission, and RDC 1718050-0 of EMOS.AI. Thanks to Dr. Ian D. Wood and Peyman Momeni for their helpful commnets.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

J Douglas Carroll and Jih-Jie Chang. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. *Psychometrika*, 35(3):283–319.

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*, pages 163–171. ACM.

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*, volume 1, pages 397–401. IEEE.

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*, pages 960–964. IEEE.

Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, 47(3):43.

Maryam Fazel. 2002. *Matrix rank minimization with applications*. Ph.D. thesis, PhD thesis, Stanford University.

Michael Glodek, Stephan Tschechne, Georg Layher, Martin Schels, Tobias Brosch, Stefan Scherer, Markus Kächele, Miriam Schmidt, Heiko Neumann, Günther Palm, et al. 2011. Multiple classifier systems for the classification of audio-visual emotional states. In *Affective Computing and Intelligent Interaction*, pages 359–368. Springer.

RA Harshman. 1970. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84.

Frank L Hitchcock. 1927. The expression of a tensor or a polyadic as a sum of products. *Studies in Applied Mathematics*, 6(1-4):164–189.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Alex Pappachen James and Belur V Dasarathy. 2014. Medical image fusion: A survey of the state of the art. *Information Fusion*, 19:4–19.

Onno Kampman, Elham J Barezi, Dario Bertero, and Pascale Fung. 2018. Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 606–611.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Zhen-Zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G Hauptmann. 2014. Multimedia classification and event detection using double fusion. *Multimedia tools and applications*, 71(1):333–347.

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. *arXiv preprint arXiv:1806.00064*.

Emilie Morvant, Amaury Habrard, and Stéphane Ayache. 2014. Majority vote of diverse classifiers for late fusion. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 153–162. Springer.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 689–696, New York, NY, USA. ACM.

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 *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 50–57. ACM.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017a. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017b. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1033–1038. IEEE.

Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior. 2003. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the*

2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 160–170.

Mohammad Soleymani, Maja Pantic, and Thierry Pun. 2012. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing*, 3(2):211–223.

Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems 25*, pages 2222–2230. Curran Associates, Inc.

Ledyard R Tucker. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

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

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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

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

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*, volume 2, pages 1491–1498. IEEE.