DeepCU: Integrating both Common and Unique Latent Information for Multimodal Sentiment Analysis

Sunny Verma[1,2] , **Chen Wang**[2] , **Liming Zhu**[2] and **Wei Liu**[1,*]
[1]Advanced Analytics Institute, School of Computer Science, University of Technology Sydney, Australia
[2]Commonwealth Scientific and Industrial Research Organisation, CSIRO, Data61, Sydney, Australia
Sunny.Verma@student.uts.edu.au, Wei.Liu@uts.edu.au, {Chen.Wang, Liming.Zhu}@data61.csiro.au

Abstract

Multimodal sentiment analysis combines information available from visual, textual, and acoustic representations for sentiment prediction. The recent multimodal fusion schemes combine multiple modalities as a tensor and obtain either; the common information by utilizing neural networks, or the unique information by modeling low-rank representation of the tensor. However, both of these information are essential as they render inter-modal and intra-modal relationships of the data. In this research, we first propose a novel deep architecture to extract the common information from the multi-mode representations. Furthermore, we propose unique networks to obtain the modality-specific information that enhances the generalization performance of our multimodal system. Finally, we integrate these two aspects of information via a fusion layer and propose a novel multimodal data fusion architecture, which we call *DeepCU* (Deep network with both Common and Unique latent information). The proposed *DeepCU* consolidates the two networks for joint utilization and discovery of all-important latent information. Comprehensive experiments are conducted to demonstrate the effectiveness of utilizing both common and unique information discovered by *DeepCU* on multiple real-world datasets. The source code of proposed *DeepCU* is available at https://github.com/sverma88/DeepCU-IJCAI19.

1 Introduction

Recent developments in deep learning techniques has led tremendous success in Sentiment Analysis and emotion recognition. Despite of the recent multitude efforts utilizing language for sentiment analysis, a core research challenge for this domain is the efficient utilization of multimodal representations such as voice and visual gestures for sentiment prediction [Lahat *et al.*, 2015; Baltrušaitis *et al.*, 2018]. There is a growing trend of sharing opinion videos on social media platforms (Facebook, YouTube, etc.) which comprise of language, visual-gestures, and acoustic as multimodal representations. Combining the unimodal representation for sentiment analysis becomes crucial as the combined information from multiple modalities promises better generalization capabilities over traditional text-based schemes [Baltrušaitis *et al.*, 2018]. Figure 1 illustrates a typical multimodal sentiment analysis systems, where the utterance “That’s – that’s true” is ambiguous and can be perceived as positive or neutral sentiment. However, on combining speaker’s visual gesture and acoustic helps us in identifying the sentiment of the speaker.

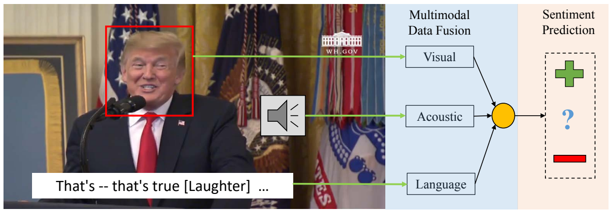

Figure 1: A typical Multimodal Sentiment Analysis System

Although the fusion of interacting modalities i.e. acoustic, visual, and language often improves the generalization performance, there are various scenarios with real-world datasets which must be handled properly while performing fusion, otherwise the joint representation might become futile. A common scenario in this regard is the occurrence of missing values in the unimodal representations [Lahat *et al.*, 2015] which leads to futile joint representations. For visual features missing values can occur due to several reasons for example poor lighting in the opinionated video, the speaker is wearing accessories (hat, glasses etc.) or covers his face while laughing. Similarly, for the auditory signal factors like voice-echo, ambient noise can cause missing values in the feature set. Figure 2, illustrates a motivating example presenting limitations with the current state of the art fusion techniques i.e. TFN [Zadeh *et al.*, 2017] shown as A., LMF [Liu *et al.*, 2018] shown as B.; and superiority of our proposed *DeepCU* shown as C. in Figure 2 when faced with missing values.

In Figure 2, to obtain the joint representation from acoustic and language modalities the TFN and the LMF utilizes an outer product on the augmented features. This results in both the bi-modal and the unimodal features in joint representation (as tensor). However, the joint representation in all cases is much sparse (contains more missing values) than the acoustic modality and the learning mechanisms of both the TFN and LMF fail to efficiently extract information in this scenario. Our proposed *DeepCU* can handle the missing value scenario

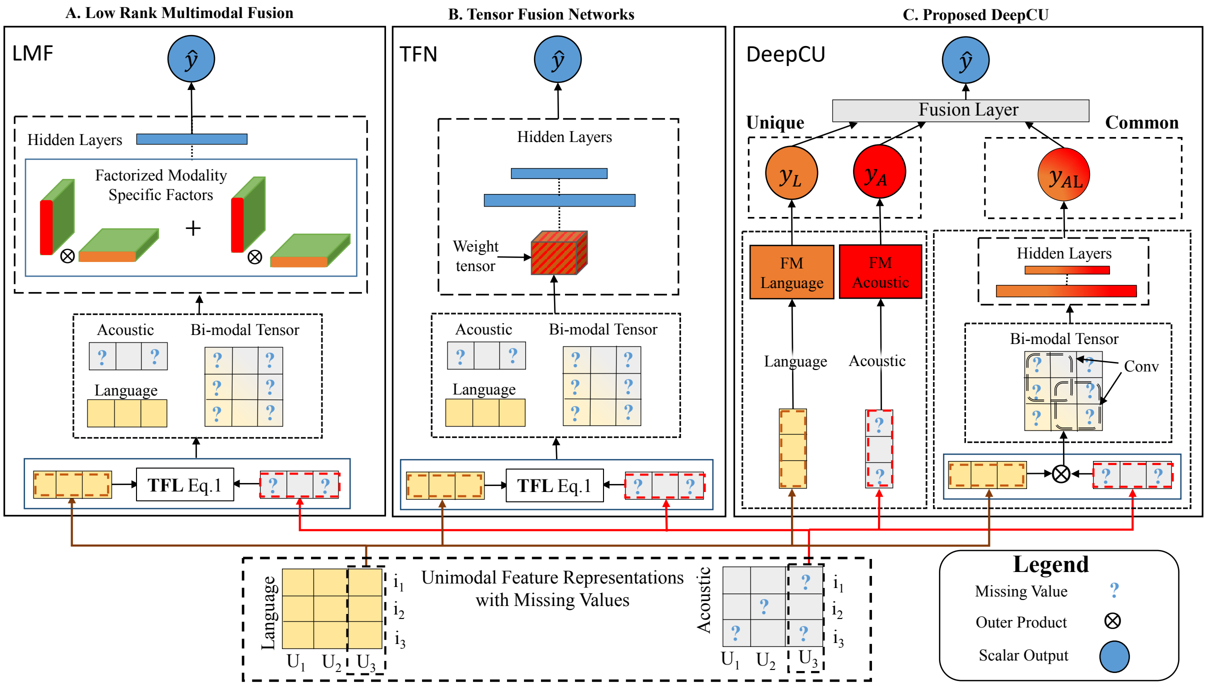

Figure 2: Comparison of missing values (interrogation mark '?') scenarios by State of the art A. Low-rank Multimodal Fusion (LMF), B. Tensor Fusion Networks (TFN), and C. our Proposed DeepCU.

due to the following:

1. The convolution kernels split the joint representation into overlapping segments while performing feature extraction which reduces the impact of missing values.
2. Factorization Machines (FMs) obtaining modality-specific unique information are robust with sparse feature vectors which subsides the impact on *DeepCU*'s performance and information discovery when the joint representation is futile.
3. Learning unshared latent representation for common and unique networks ensures that latent-embeddings of the superior representations remain unaffected by influences of inferior representations (i.e. gradient from futile representation). This restriction enforces latent-embeddings to attain complementary information and provides more expressiveness while performing fusion in the higher layers.

Motivated by the above points, we propose a novel deep common and unique feature extraction technique for multimodal data fusion, which we call as *DeepCU*. Our proposed *DeepCU* has two components 1) unique sub-network which obtains information specific to individual modalities and; 2) common sub-network which obtains combined information from joint (multi-mode) representations by using proposed deep-convolution tensor networks. Information from the common and the unique sub-networks is integrated by a fusion layer to obtain an integrated output.

The main contribution of this work are as follows:

I. We design a consolidated deep network for joint utilization and discovery of both the common (multi-mode) and unique (mode-specific) properties of the multimodal data for sentiment analysis.

II. Our proposed *DeepCU* is conceptually more expressive than existing state of the art (TFN and LMF) as it captures non-linear multi-mode interactions exhibiting in the tensorial representation within our common network sub-network. Moreover, our unique sub-network obtains both linear and factorized non-linear (quadratic) feature relations which mitigates the missing value scenarios and enhances the generalization capability of *DeepCU*.

III. We perform comprehensive experiments on multimodal CMU-MOSI and POM datasets and demonstrate the effectiveness of utilizing both common and unique latent information with comparisons to other techniques.

2 Related Work

We focus our review on recent neural based frameworks for multimodal data fusion proposed in the literature. In [Lin *et al.*, 2015] a bilinear-CNN is proposed to obtain bi-modal interactions among features obtained from two heterogeneous CNNs. However, the bilinear layer required parameter estimation of a quadratic number of neurons and hence prone to over-fitting. This limitation is alleviated in [Fukui *et al.*, 2016] which introduces an alternate formulation of the bilinear layer and obtains its compact representation by utilizing sophisticated neural based factorization schemes.

However, the above fusion schemes only express the bi-modal (or tri-modal) interactions from unimodal representations either as: a) inter-modal (outer product) or b) intra-model (simple concatenation) based representations. But, utilization of both the intra-modal and inter-modal representations are proven helpful in many machine learning tasks [Liu *et al.*, 2013; Verma *et al.*, 2017; Verma *et al.*, 2018]. In this regard, Tensor Fusion Layer (TFL) is proposed in [Zadeh *et al.*, 2017] which leverages the expressiveness of both the inter-model and the intra-model fusion schemes.

The TFL applies bilinear product by augmenting the unimodal representations with an additional feature of constant values equal to 1. The outer product on the augmented unimodal representations now yields two sets of information: 1) the bi-modal (or tri-modal) interactions in the form of 2D-tensor (3D-tensor) and 2) the raw unimodal representations of the modalities. Mathematically the TFL for bi-modal interactions can be expressed as in Equation (1), where $\boldsymbol{x}_1 \in \mathbb{R}^n$ and $\boldsymbol{x}_2 \in \mathbb{R}^m$ are feature vectors from two different modalities

$$TFL(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathcal{X} = \begin{bmatrix} \boldsymbol{x}_1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{x}_2 \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1 \otimes \boldsymbol{x}_2 & \boldsymbol{x}_1 \\ \boldsymbol{x}_2 & 1 \end{bmatrix} \quad (1)$$

'$\otimes$' represents the outer product and $\mathcal{X} \in \mathbb{R}^{(n+1)\times(m+1)}$.

Tensor Fusion Networks (TFN) proposed in [Zadeh *et al.*, 2017] learns a weight tensor $\mathcal{W} \in \mathbb{R}^{(n+1)\times(m+1)\times k}$ and a set of feed-forward layers to obtain the combined information from $\mathcal{X}$. The TFN outperformed all the previous fusion schemes for multimodal sentiment analysis on CMU-MOSI dataset as it leverages the expressiveness offered by both the bi-modal and unimodal information exhibiting in the joint representations obtained via TFL. However, the dimensionality of the weight tensor $\mathcal{W}$ increases exponentially by increasing the number of unimodal representations for fusion and hence the TFN is not scalable [Liu *et al.*, 2018].

Low-rank Multimodal Fusion (LMF) [Liu *et al.*, 2018] alleviates the scalability issues with TFN by approximating lower dimensional modality specific factors (commonly refereed as Rank-k tensors in CP decomposition [Liu *et al.*, 2013]). The LMF, the weight tensor $\mathcal{W}$ is equivalently expressed as $\mathcal{W} \equiv \big(\boldsymbol{W}_1 \otimes \boldsymbol{W}_2\big)$, where $\boldsymbol{W}_1 \in \mathbb{R}^{(n+1)\times k}$, $\boldsymbol{W}_2 \in \mathbb{R}^{(m+1)\times k}$. Extracting of information from $\mathcal{X}$ is now reformulated as: $\big(\hat{\boldsymbol{x}}_1 \times \boldsymbol{W}_1\big) \odot \big(\hat{\boldsymbol{x}}_2 \times \boldsymbol{W}_2\big)$, where $\hat{\boldsymbol{x}}_i = [\boldsymbol{x}_i, 1]^T$; and '$\odot$' is the element-wise product operator; and '$\times_i$' is the mode-i product between tensor and matrix. Hence, explicitly learning higher dimension weight tensor $\mathcal{W}$ with TFN is not required. The LMF is a current state of the art on CMU-MOSI dataset without any contextual information.

Approaches like [Zadeh *et al.*, 2018b; Poria *et al.*, 2017; Zadeh *et al.*, 2018a] incorporates contextual information from multimodal representations utilizes an attention mechanism to incorporate the information available from all utterances of the same speaker which enables them to model the complex dynamics of inter-modality relationships efficiently. Although these techniques are superior than the above schemes but they requires additional information like the identity of the speaker, the sequence of the utterance-sentiments while modelling their fusion schemes. This additional information might not be available in the general scenarios.

3 Proposed Methodology

Contrary to the existing fusion schemes we aim to utilize both the common and the unique information for multimodal data fusion. To this end, we first propose two sub-networks, i.e., 1) unique network for obtaining modality-specific features (described in Section 3.1) and; 2) common network which consists of proposed deep-convolution tensor networks (described in Section 3.2). The latent space for the unique information and the common information is unshared (i.e. influenced only by gradient of their respective sub-network) and allows *DeepCU* to obtain complementary information with both the sub-networks. Later, these two kinds of information is integrated via a fusion layer (described in Section 3.3 which allows joint optimization and information discovery in common and unique network's) to $\hat{y}$ as the final prediction from *DeepCU*. The differences and similarities between existing multimodal data fusion techniques and the proposed *DeepCU* are summarized in Section 3.

Fusion Schemes	Deep & Shallow	Inter Modality	Modality Specific	Convolution	Unshared Embeddings
DeepFM	✓	×	×	×	×
TFN	×	✓	✓	×	×
LMF	×	✓	✓	×	×
DeepCU (proposed)	✓	✓	✓	✓	✓

Table 1: Comparison of multimodal data fusion models

The raw feature vectors from a single utterance for acoustic and visual modalities are denoted as $z_a \in \mathbb{R}^{1\times k_a}$ and $z_v \in \mathbb{R}^{1\times k_v}$ respectively, where k_a and k_v represents the dimensionality of the feature vectors. For language modality the raw features are word-embeddings denoted as $z_l \in \mathbb{R}^{1\times s_l \times d_l}$, where s_l is the sequence length of the embeddings and k_l is the dimensionality of each sequence vector. The latent space (or embeddings) obtained from these features for the common and unique sub-networks are unshared and influenced only by their respective networks. This restriction allows both networks to learn complementary feature representations at lower layers which enhances their expressiveness in the fusion layer. Besides, optimizing unshared latent space is empirically shown beneficial in [He and Chua, 2017].

3.1 Unique Network

The modality-specific information is obtained by utilizing Factorization Machines (FMs). There are two main motivations behind utilizing FMs instead of any other shallow learning technique (Logistic Regression, SVM or, a single fully-connected layer etc.) for extracting the unique information from individual modalities as:

1. FMs has linear time complexity and it models both first and second-order factorized interactions from feature vector which enhances its expressive capabilities over other shallow techniques.
2. Real-world datasets often consist of missing values and FMs are capable of dealing with sparsity as they model feature interactions with factorized representations.

Prior to utilizing FMs, the feature vectors from unimodalities are processed via sub-embeddings vectors denoted as f_{FM_a}, f_{FM_v} and f_{FM_l} to extract latent features from the $\boldsymbol{z}_a$ (acoustic), $\boldsymbol{z}_v$ (visual), and $\boldsymbol{z}_l$ (language) respectively. The sub-embeddings network for acoustic and visual modalities is a single feed-forward linear layer. Whereas, for language modality the sub-embeddings network comprises of LSTM [Hochreiter and Schmidhuber, 1997] followed by a single feed-forward layer. FMs are then trained independently on f_{FM_v}, f_{FM_a}, f_{FM_l} to obtain y_V, y_A, and y_L as predicted sentiment from their respective modalities. We briefly discuss the details of FMs before presenting the procedure of unique information extraction.

Factorization Machine FMs were originally proposed for recommendation systems [Rendle, 2010s]. They are widely utilized for information extraction especially when dealing with extremely sparse feature sets. Given a sparse real valued feature $\boldsymbol{x} \in \mathbb{R}^n$, FMs estimates the target i.e. $\hat{y}_{FM(\boldsymbol{x})} \in \mathbb{R}$ by modelling all interactions between each pair of features via factorized interaction parameters as below:

$$\hat{y}_{FM(\boldsymbol{x})} = w_0 + \sum_{i=1}^{n} \boldsymbol{w}_i \boldsymbol{x}_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \boldsymbol{v}_i^T \boldsymbol{v}_j \ . \ \boldsymbol{x}_i \boldsymbol{x}_j \tag{2}$$

where w_0 is the global bias, $\boldsymbol{w} \in \mathbb{R}^n$ models the interaction of the *i*-th feature to the target. The $\boldsymbol{v}_i^T \boldsymbol{v}_j$ term denotes the factorized interaction, where $\boldsymbol{v}_i \in \mathbb{R}^k$ denotes the latent vector of size k for feature *i*, and $\hat{y}_{FM(\boldsymbol{x})}$ is the predicted value.

Extracting Acoustic-Specific Unique Information

The latent embeddings denoted as $\boldsymbol{f}_{FM_a} \in \mathbb{R}^{1 \times k_a}$ are obtained from the acoustic features z_a as below:

$$\boldsymbol{f}_{FM_a} = \sigma\Big(\boldsymbol{z}_a \times W_{FM_a} + \boldsymbol{b}_{0_{FM_a}}\Big) \tag{3}$$

where W_{FM_a} and $\boldsymbol{b}_{0_{FM_a}}$ are the sub-embedding network hyper-parameters and σ is the activation function. The unique acoustic information is the obtained by utilizing FM in Equation (2) on the latent embedding obtained as $y_A = \hat{y}_{FM(\boldsymbol{f_{FM_a}})}$.

Extracting Visual-Specific Unique Information

The latent embeddings, $\boldsymbol{f}_{FM_v}$ from z_v are obtained analogous to the acoustic sub-embedding network. The unique visual information is then obtained as $y_V = \hat{y}_{FM(\boldsymbol{f_{FM_v}})}$.

$$\boldsymbol{f}_{FM_v} = \sigma\Big(\big(\boldsymbol{z}_v \times W_{FM_v}\big) + \boldsymbol{b}_{0_{FM_v}}\Big) \tag{4}$$

Extracting Language-Specific Unique Information

The latent embeddings denoted as $\boldsymbol{f}_{FM_l} \in \mathbb{R}^{1 \times k_a}$ are obtained from the language features z_l as below:

$$\boldsymbol{f}_{FM_l} = \sigma\Big(LSTM(\boldsymbol{z}_l) \times W_{FM_l} + \boldsymbol{b}_{0_{FM_l}}\Big) \tag{5}$$

where W_{FM_l} and $\boldsymbol{b}_{0_{FM_l}}$ are the sub-embeddings networks hyper-parameters. The unique language specific information is then obtained as $y_L = \hat{y}_{FM(\boldsymbol{f_{FM_l}})}$.

3.2 Common Network

To obtain the common information from multi-mode representations we propose a deep convolution-tensor network. In this regard, we first obtain joint representation as tensors from modalities by performing outer product on their latent embeddings. These tensors are naturally multi-dimensional where each element of the tensor represents the interaction strength between the elements of the fusion-modalities. Therefore we applied convolution-kernels on these tensors as they are non-linear feature extractors and generalize better than feed-forward layers [Kim *et al.*, 2016]. Utilizing convolutions on the joint representations alleviates the need of factorization in *DeepCU* and also makes it highly scalable.

Analogous to the unique network the unimodal representations are processed via sub-embeddings networks to obtain latent embeddings. Then the outer product is utilized to capture joint representations as tensors from these embeddings. Convolution kernels of appropriate dimensions are then applied to the tensors for feature extraction. To reduce the impact of missing values in our common network we obtain multiple sets of combined representation as below:

- T_{AV} bi-modal representation from acoustic & visual.
- T_{AL} bi-modal representation from acoustic & language.
- T_{VL} bi-modal representation from visual & language.
- T_{AVL} tri-modal representation from acoustic, visual, & language.

The motivation to obtain multiple sets of tensor representation is that if assuming any one of the modalities (for example acoustic) has missing values. Then, the tensorial representations obtained with the latent embeddings of this modality (i.e. T_{AV}, T_{AL} and T_{AVL}) are affected but not the other tensor representations (i.e. T_{VL}). Moreover this information loss is further subsided by the information obtained by the corresponding unique network. Again, the latent embeddings for each tensor pair in the common network are unshared which enables *DeepCU* to obtain complementary information within each tensorial representation.

Extracting Combined Information from the Bi-Modal Interactions of Acoustic and Visual Modalities

The latent embeddings for the acoustic ($f_{AV} \in \mathbb{R}^{1 \times k_a}$) and visual features ($f_{VA} \in \mathbb{R}^{1 \times k_v}$) are obtained as below:

$$\begin{aligned} f_{AV} &= \sigma\Big(\boldsymbol{z}_a \times W_{av} + b_{av}\Big) \\ f_{VA} &= \sigma\Big(\boldsymbol{z}_v \times W_{va} + \boldsymbol{b}_{va}\Big) \\ T_{AV} &= f_{AV} \otimes f_{VA} \end{aligned} \tag{6}$$

where $[W_{av}, b_{av}]$ and $[W_{va}, b_{va}]$ represents the sub-embeddings networks hyper-parameters. $T_{AV} \in \mathbb{R}^{k_v \times k_l}$ represents the bi-modal representation obtained by taking outer product of the latent embeddings. Convolution filters are then applied to capture the non-linear interactions in T_{AV} as:

$$\mathcal{G}_{AV} = \sigma\Big(Conv(T_{AV})\Big) \tag{7}$$

where $\mathcal{G}_{AV}$ represents the output from convolution layer which is then processed through fully-connected layer as:

$$\boldsymbol{h}_{AV} = \sigma\Big(\hat{\boldsymbol{g}}_{AV} \times W_{AV} + \boldsymbol{b}_{AV}\Big) \tag{8}$$

Finally, the hidden representation $\boldsymbol{h}_{AV}$ is processed through feed-forward layer to obtain the final predicted value y_{AV} as:

$$y_{AV} = \big(\boldsymbol{h}_{AV} \times \boldsymbol{w}_{AV}\big) + b_{0_{AV}} \tag{9}$$

Extracting Combined Information from the Bi-Modal Interactions of Visual and Language Modalities

The latent embeddings for the visual ($f_{VL} \in \mathbb{R}^{1\times k_v}$) and language features ($f_{LV} \in \mathbb{R}^{1\times k_l}$) are obtained as shown below:

$$\begin{aligned} f_{VL} &= \sigma\Big(\boldsymbol{z}_v \times W_{vl} + b_{vl}\Big) \\ f_{LV} &= \sigma\Big(LSTM(\boldsymbol{z}_l) \times W_{lv} + \boldsymbol{b}_{lv}\Big) \end{aligned} \tag{10}$$

$T_{VL} \in \mathbb{R}^{k_v \times k_l}c$ is then obtained by taking outer product of the latent embeddings representing their bi-modal interactions. Analogous to Equations (7) to (9) the bi-modal interactions are processed to obtain y_{VL} as the predicted output.

Extracting Combined Information from the Bi-Modal Interactions of Acoustic and Language Modalities

Analogous to the above y_{AL} is obtained as the predicted output from bi-modal acoustics and visual interactions.

Extracting Combined Information from the Tri-Modal Interactions of Acoustic, Visual and Language Modalities

The tri-modal interactions are obtained by taking outer product between latent embeddings of acoustic, visual and language; i.e. $T_{AVL} = \big(f_{AVL} \otimes f_{VLA} \otimes f_{LAV}\big) \in \mathbb{R}^{k_a \times k_v \times k_l}$. Convolution filters and fully connected layers are then applied on T_{AVL} to obtain the predicted values y_{AVL} as below.

$$\begin{aligned} \mathcal{G}_{AVL} &= \sigma\Big(Conv\big(T_{AVL}\big)\Big) \\ \boldsymbol{h}_{AVL} &= \sigma\Big(\hat{\boldsymbol{g}}_{AVL} \times W_{AVL} + \boldsymbol{b}_{AVL}\Big) \\ y_{AVL} &= \big(\boldsymbol{h}_{AVL} \times \boldsymbol{w}_{AVL}\big) + b_{0_{avl}} \end{aligned} \tag{11}$$

3.3 Fusion Layer

The scalar outputs from the common and the unique sub-networks are integrated by applying $\hat{y} = h^T Z$, where the vector Z is obtained by concatenating the predicted scalar outputs from the unique and common sub-networks as $Z = [y_A, y_V, y_L, y_{AL}, y_{VL}, y_{AV}, y_{AVL}]$, and $h = \big[\hat{h}_A, \hat{h}_V, \hat{h}_L, \hat{h}_{AL}, \hat{h}_{VL}, \hat{h}_{AV}, \hat{h}_{AVL}\big]$ is a vector of appropriate dimension consisting of fusion weights. For simplicity, all the weights in h can be set to one i.e. $h = J_{1,7}$ and are not optimized while training. We refer this model as static fusion denoted as *DeepCU$_{SF}$*. Otherwise, the weights in h can be randomly initialized (simply a fully connected layer with number of neurons equal to seven) and optimized via the loss on the target function and the model is referred as dynamic fusion denoted as *DeepCU$_{DF}$*.

Our proposed *DeepCU* can be applied to a variety of tasks such as for classification, ranking etc. However, for this work we estimate the parameters of *DeepCU* via minimizing the mean square error (MSE) loss in Equation (12).

$$L = \frac{1}{n} \sum_{\forall x \in \chi} \left(\hat{y}(x) - y(x)\right)^2 \tag{12}$$

where χ denotes the set of multimodal training data instances, $y(x)$ denotes the target of instance x, and $\hat{y}(x)$ denotes the prediction obtained from *DeepCU*.

3.4 Complexity Analysis

Theoretically, the paramount computational cost in *DeepCU* is feature extraction from the multimodal tensor which is $\mathcal{O}(N \times K \times S^2 \times M^2)$ as described in [He and Sun, 2015]) where N and K are the number of input and output feature maps respectively and; S represents the spatial size of the filter and M represents the spatial size of the output feature map. If we fix the dimensionality of the latent space for each modality as 32 (as in LMF), then the number of parameters in *DeepCU* are 1.06e6 whereas the number of parameters in LMF and TFN is equal to 1.1e6 and 12.5e6 respectively.

4 Experimental Settings

Dataset. We perform experiments on the CMU-MOSI [Zadeh *et al.*, 2016] and POM [Park *et al.*, 2014] datasets consisting of YouTube videos for movie reviews. The CMU-MOSI dataset consists of movie reviews videos from 93 distinct speakers. Each video consists of multiple opinion segments with a total of 2199 segments in the whole dataset, annotated with the sentiment in the range $[-3,\ 3]$. The POM dataset consists of 903 movie review videos where each video is annotated 16 sentiments of the speaker. To evaluate the generalization capability of models, the training, validation, and testing splits of the dataset are speaker independent.

Baselines. We extensively evaluate the performance of both neural based and non-neural based fusion schemes for multimodal sentiment analysis. Thus we trained our *DeepCU* as well as other baselines with mse loss (i.e., Equation (12)) but C-MKL which is trained for binary classification (due to the objective function utilized in [Poria *et al.*, 2015]). For calculating the binary and multi-class accuracies, we followed the protocol in [Liu *et al.*, 2018] and map the predicted output (and the target values) to integer values.

4.1 Parameter Settings in DeepCU

We train our model by minimizing the MSE loss with RMS optimizer with learning rate equal to 6×10^{-3} and batch-size of 64. To avoid over-fitting we applied dropout [Srivastava *et al.*, 2014] in our model and tune the dropout probability from [0.1, 0.9] with a step size of 0.05. The optimal dimensions of latent spaced within each sub-embeddings network was searched in [5,10,15,20,25,30], while the number of convolution filters was searched in [1, 2, 3, 4, 5]. We also varied the size of convolution filter between 3 and 5. Moreover, to reduce covariance shift and improve performance we applied batch normalization [Ioffe and Szegedy, 2015] to all hidden layers of *DeepCU*. For acoustic and visual modalities the sub-embeddings network is a single feed-forward layer, while for language we used LSTM [Hochreiter and Schmidhuber, 1997] (basic uni-directional LSTM cell) with 128 units.

4.2 Evaluation Metrics

We evaluate the performance of the baselines and *DeepCU* for regression, binary and multi-class classification problems. For regression, we report Mean Absolute Error (MAE) and Pearson's Correlation (Correlation). For binary classification, we report accuracy and F1 score, where as for multi-class classification we only report accuracy. For all metrics higher

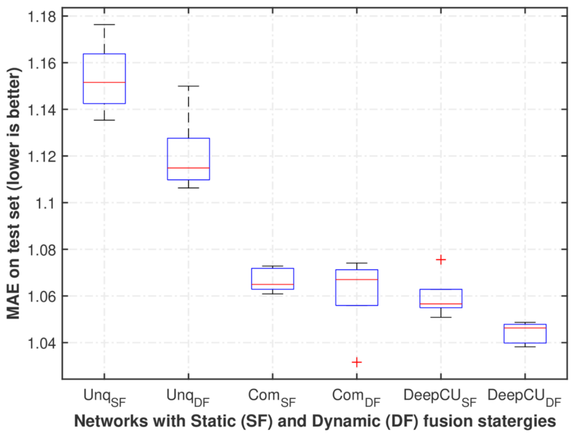

Figure 3: Performance comparison of DeepCU vs common (Com) and unique (Unq) networks on the CMU-MOSI dataset.

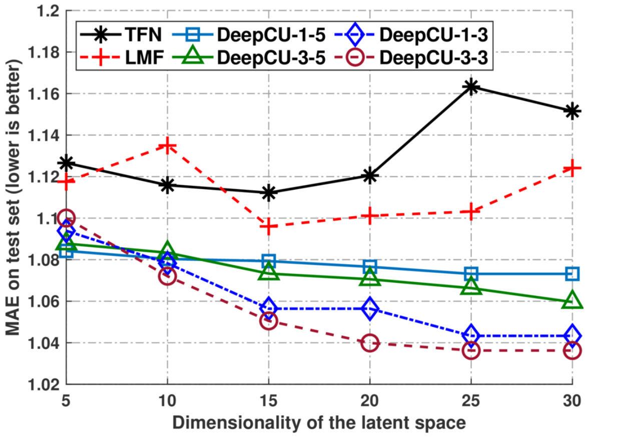

Figure 4: Performance of DeepCU, TFN, and LMF by varying hyperparameters on the CMU-MOSI dataset. The legend DeepCU-x-y represents, x = number of convolution filters and y = filter size.

value is better except for MAE. Similar to [Zadeh *et al.*, 2017; Liu *et al.*, 2018] we employed early stopping strategy, where we terminated training *DeepCU* and all baselines if the MAE on validation-set did not improved in 5 consecutive epochs.

4.3 Results and Explainability Analysis

The key contribution of this work is utilization of both unique and common information for multimodal data fusion. We performed experiments to study the significance of our proposed fusion scheme under the following research questions:

Q1: Does the integration of common and unique latent information actually beneficial or their integration deteriorates the performance of *DeepCU* over individual sub-networks?

To evaluate this, we studied whether fusing the common and unique information is actually beneficial or their integration deteriorates the performance of *DeepCU* over individual sub-networks. To achieve so, we evaluate the performance of common network on all the hyper-parameter settings as explained in Section 4.1. While the unique network were evaluated by varying the size of latent dimensions and dropout probabilities and optimizers. We also applied both the dynamic and static fusion schemes to the common and unique networks. We present the MAE of the optimized networks with box-plot in Figure 3.

It is clearly visible that integrating both the common and the unique information improves the performance of proposed *DeepCU*. The common network exploits the information from both bi-modal and tri-modal interactions by applying deep-convolution operations which drastically reduces its MAE compared to unique networks. Besides, the plot suggests that for all the networks the dynamic fusion performs slightly better than static fusion. However the network with dynamic fusion layer required more epochs for convergence.

Besides, the integration of common and unique information further achieves reduction in MAE and is visible in the box plots for both the fusion schemes in *DeepCU*. Moreover, *DeepCU* with dynamic fusion scheme achieves the lowest MAE and confirms that integration of common and unique information is actually beneficial for multimodal data fusion.

Q2: Are convolutions able to efficiently capture the information from non-linear interactions exhibiting in the multi-mode representation? Moreover, how does the hyper-parameters affect the performance of *DeepCU*?

We now present a detailed study on how hyper-parameters affects the performance of *DeepCU*. In this regard, we plot the mean MAE obtained by varying hyper-parameters in Figure 4(b). The x-axis in plot represents the dimensionality of latent-embeddings and the curves represents combinations on a) the number of convolution filters, b) filter-size, and c) fusion scheme. We also plot the performances of TFN and LMF obtained on the same latent dimensions.

A clear trend can be seen in all the curves reflecting performance of *DeepCU*, where the MAE tends to decrease with increase in the latent dimensions. This is because the lower dimensions tensor is equal to the size of convolution kernel and hence the performance of *DeepCU* is not significantly better than TFN and LMF. However, the marginal improvement can be attributed to the unshared latent space and the unique information. Besides, the performance gradually improves with the increase in the latent dimensions which supports the learning requirement of convolution kernels.

Another trend can be noticed in the performance curves of *DeepCU* where convolutions of filter-size 3 performs slightly better that filter-size 5. This may be due to the increase in overlapping regions between segments which might be better for applying convolution on multi-mode representations.

Q3: Does *DeepCU* provide a better mulit-modal fusion technique compared to state of the art such as TFN and LMF?

We compare the performance of multiple SOTA (and other baselines) and *DeepCU* on the CMU-MOSI and POM datasets for this requirement and the results are reported in Sections 4.3 and 4.3.

On CMU-MOSI dataset we improve the state of the art by **4.68**% for regression and **2.25**% for correlation and on multiclass the accuracy improvement is **9.63**%. On POM dataset we improve the correlation by **23.10**% and for regression the improvement is **2**% compared to state of the art.

The above results confirms our hypothesis on the advantages of *DeepCU*: a) utilizing both the common and unique

MOSI Dataset	Regression		Binary		7-class
	MAE (lower is better)	Correlation	Accuracy	F1	Accuracy
RF	$1.4095 \pm 1.09\times10^{-4}$	$0.2041 \pm 3.29\times10^{-4}$	$53.98 \pm 6.57\times10^{-1}$	52.75 ± 1.48	18.27 ± 1.41
SVM	$1.4259 \pm 1.43\times10^{-5}$	$0.1288 \pm 3.36\times10^{-4}$	47.74 ± 5.78	36.59 ± 4.37	$13.98 \pm 4.12\times10^{-1}$
DNN$_{JR}$ [Pérez-Rosas *et al.*, 2013]	$1.1801 \pm 2.31\times10^{-4}$	$0.4973 \pm 2.41\times10^{-4}$	$68.59 \pm 2.27\times10^{-1}$	$68.67 \pm 2.27\times10^{-1}$	25.48 ± 3.75
RF-MD	$1.1993 \pm 1.63\times10^{-4}$	$0.4636 \pm 2.41\times10^{-4}$	$66.11 \pm 5.81\times10^{-1}$	$66.16 \pm 6.02\times10^{-1}$	$26.03 \pm 3.60\times10^{-1}$
SVM-MD [Zadeh *et al.*, 2016]	$1.2749 \pm 2.97\times10^{-4}$	$0.4950 \pm 1.71\times10^{-4}$	$67.60 \pm 2.59\times10^{-1}$	$67.68 \pm 2.64\times10^{-1}$	$17.49 \pm 1.00\times10^{-1}$
C-MKL [Poria *et al.*, 2015]	—	—	$66.85 \pm 4.65\times10^{-1}$	$68.30 \pm 6.43\times10^{-1}$	—
ELM [Poria *et al.*, 2016]	$1.1786 \pm 2.28\times10^{-4}$	$0.4935 \pm 1.22\times10^{-4}$	69.70 ± 1.08	71.61 ± 1.66	24.42 ± 1.68
DeepFM [Guo *et al.*, 2017]	$1.1038 \pm 1.81\times10^{-5}$	$0.5227 \pm 1.73\times10^{-4}$	$69.14 \pm 7.64\times10^{-1}$	$69.10 \pm 7.3\times10^{-1}$	$28.90 \pm 4.54\times10^{-1}$
TFN (SOTA 1) [Zadeh *et al.*, 2017]	$1.1111 \pm 3.03\times10^{-4}$	$0.5341 \pm 1.02\times10^{-4}$	$69.59 \pm 7.06 \times 10^{-1}$	$68.48 \pm 7.93 \times 10^{-1}$	31.98 ± 1.13
LMF (SOTA 2) [Liu *et al.*, 2018]	$1.0960 \pm 2.11\times10^{-4}$	$0.5555 \pm 3.28\times10^{-5}$	$70.25 \pm 2.05 \times 10^{-1}$	$70.31 \pm 1.98 \times 10^{-1}$	$30.76 \pm 3.39 \times 10^{-1}$
DeepCU$_{SF}$ (static fusion)	$1.0595 \pm 7.08\times10^{-5}$	$0.5536 \pm 7.66\times10^{-5}$	$71.49 \pm 2.00 \times 10^{-1}$	$71.42 \pm 1.98 \times 10^{-1}$	$33.54 \pm 6.39 \times 10^{-1}$
DeepCU$_{DF}$ (dynamic fusion)	$\mathbf{1.0442} \pm \mathbf{1.71}\times10^{-5}$	$\mathbf{0.5609} \pm \mathbf{1.05}\times10^{-5}$	$\mathbf{73.54} \pm \mathbf{1.10} \times 10^{-1}$	$\mathbf{73.52} \pm \mathbf{1.14} \times 10^{-1}$	$\mathbf{34.04} \pm \mathbf{3.61} \times 10^{-1}$

Table 2: Performance comparison of DeepCU vs other fusion techniques on CMU-MOSI dataset. The mean and variance for each baseline and DeepCU are obtained by executing them for five times. This superiority of DeepCU is specifically visible in the case of 7-class classification.

POM Dataset	MAE	Correlation	Multi-Class Accuracy
TFN (SOTA 1)	1.0481 ± 0.0030	0.0866 ± 0.023	28.62 ± 0.127
LMF (SOTA 2)	0.8739 ± 0.0051	0.2311 ± 0.024	33.61 ± 0.314
DeepCU$_{DF}$	$\mathbf{0.8568 \pm 0.0045}$	$\mathbf{0.2845 \pm 0.009}$	$\mathbf{34.77 \pm 0.493}$

Table 3: Performance comparison on the POM dataset.

latent information obtained using unshared-embeddings; b) the use of convolutions to capture utmost expressiveness offered by multi-mode representation; and c) the use of factorized representations in unique networks to reduced the impact of missing values present in the individual modalities.

As expected the dynamic fusion schemes performs better than the static fusion scheme in *DeepCU*. Conceptually, this is because the weights in the static fusion layer were not optimal and optimizing these weights via back-propagation allows the proposed *DeepCU$_{DF}$* network to obtain better mixing weights for integrating common and unique information.

4.4 Case Study with Missing Values from the Acoustic Modality in the CMU-MOSI Dataset

As a qualitative analysis on the performance of the fusion schemes, we perform an investigative study of TFN, LMF, and *DeepCU* when facing missing values in the feature sets. In this regard, we selected two examples with highest percentage of missing values from the actual dataset in the acoustic modality and reported their predicted sentiment obtained from each of the fusion schemes in Section 4.4.

In the first example, the absolute error with the prediction from TFN is 0.5118, from LMF is 0.3387; and from *DeepCU* is 0.0154. The predicted sentiment value from *DeepCU* achieves the lowest error when the corresponding feature set contains a large fraction of missing values. In the second example, the absolute error with the prediction from TFN is 0.3475, from LMF is 0.4417; and from *DeepCU* is 0.1209. Again the predicted sentiment value from *DeepCU* achieves the lowest error when the corresponding feature set contains moderate fraction of missing values.

These examples confirms the effectiveness of utilizing both common and unique information for multimodal data fusion.

Missing values in acoustic modality	Ground-Truth of Sentiment	TFN	LMF	**DeepCU$_{DF}$**
63.51 %	0.0	0.5118	-0.3387	**-0.0154**
21.62 %	-1.0	-1.3475	-1.4417	**-1.1209**

Table 4: Affect of missing values on DeepCU$_{DF}$, TFN, and LMF. These feature vectors are taken from the actual CMU-MOSI dataset.

Moreover, they also exhibit the importance of handling missing values with real-world datasets, as their proper consideration might boost the performance of multimodal systems.

5 Conclusions and Future Work

In this paper, we have introduced *DeepCU* which utilizes both common and unique latent information for sentiment analysis on multimodal data. The *DeepCU* consolidates two subnetworks a) deep convolution-tensor networks for obtaining common information from multi-model data; and b) unique subnetwork to obtain information offered by the individual modalities. Both the sub-networks are integrated via a fusion layer, and the parameters are optimized by back-propagation on the target loss function. The *DeepCU* outperformed state of the art approaches as it leverages the expressiveness of all-types of information by enforcing the two sub-networks to learn complimentary information in the embeddings layer. Comprehensive experiments demonstrate the effectiveness of our proposed *DeepCU* for multimodal data fusion. In future, we plan to introduce attention networks to integrate the information obtained by the two sub-networks.

References

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

[Fukui *et al.*, 2016] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings* *of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, 2016.

[Guo *et al.*, 2017] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. Deepfm: a factorization-machine based neural network for ctr prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, 2017.

[He and Chua, 2017] Xiangnan He and Tat-Seng Chua. Neural factorization machines for sparse predictive analytics. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 355–364. ACM, 2017.

[He and Sun, 2015] Kaiming He and Jian Sun. Convolutional neural networks at constrained time cost. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5353–5360, 2015.

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

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*. JMLR. org, 2015.

[Kim *et al.*, 2016] Donghyun Kim, Chanyoung Park, Jinoh Oh, Sungyoung Lee, and Hwanjo Yu. Convolutional matrix factorization for document context-aware recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 233–240. ACM, 2016.

[Lahat *et al.*, 2015] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 2015.

[Lin *et al.*, 2015] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015.

[Liu *et al.*, 2013] Wei Liu, Jeffrey Chan, James Bailey, Christopher Leckie, and Kotagiri Ramamohanarao. Mining labelled tensors by discovering both their common and discriminative subspaces. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, 2013.

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

[Park *et al.*, 2014] Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 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, 2014.

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

[Poria *et al.*, 2015] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544, 2015.

[Poria *et al.*, 2016] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59, 2016.

[Poria *et al.*, 2017] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 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)*, 2017.

[Rendle, 2010s] Steffen Rendle. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 995–1000. IEEE, 2010s.

[Srivastava *et al.*, 2014] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[Verma *et al.*, 2017] Sunny Verma, Wei Liu, Chen Wang, and Liming Zhu. Extracting highly effective features for supervised learning via simultaneous tensor factorization. In *AAAI*, pages 4995–4996, 2017.

[Verma *et al.*, 2018] Sunny Verma, Wei Liu, Chen Wang, and Liming Zhu. Hybrid networks: Improving deep learning networks via integrating two views of images. In *Neural Information Processing - 25th International Conference, ICONIP , Proceedings, Part I*, pages 46–58, 2018.

[Zadeh *et al.*, 2016] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

[Zadeh *et al.*, 2017] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017.

[Zadeh *et al.*, 2018a] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.

[Zadeh *et al.*, 2018b] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.