

“©2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”

Deep-HOSeq: Deep Higher Order Sequence Fusion for Multimodal Sentiment Analysis

Sunny Verma^{†,§}, Jiwei Wang[‡], Zhefeng Ge[‡], Rujia Shen[‡], Fan Jin[‡], Yang Wang[†], Fang Chen[†], and Wei Liu[†]

[†]University of Technology Sydney, Sydney, Australia

[‡]Hangzhou Dianzi University, Hangzhou, China

{sunny.verma, yang.wang, fang.chen, wei.liu}@uts.edu.au

{wangjiwei, gezhefeng, rujiashen, fanjin}@hdu.edu.cn

Abstract—Multimodal sentiment analysis utilizes multiple heterogeneous modalities for sentiment classification. The recent multimodal fusion schemes customize LSTMs to discover intra-modal dynamics and design sophisticated attention mechanisms to discover the inter-modal dynamics from multimodal sequences. Although powerful, these schemes completely rely on attention mechanisms which is problematic due to two major drawbacks 1) deceptive attention masks, and 2) training dynamics. Nevertheless, strenuous efforts are required to optimize hyperparameters of these consolidate architectures, in particular their custom-designed LSTMs constrained by attention schemes. In this research, we first propose a common network to discover both intra-modal and inter-modal dynamics by utilizing basic LSTMs and tensor based convolution networks. We then propose unique networks to encapsulate temporal-granularity among the modalities which is essential while extracting information within asynchronous sequences. We then integrate these two kinds of information via a fusion layer and call our novel multimodal fusion scheme as *Deep-HOSeq* (Deep network with higher order Common and Unique Sequence information). The proposed *Deep-HOSeq* efficiently discovers all-important information from multimodal sequences and the effectiveness of utilizing both types of information is empirically demonstrated on CMU-MOSEI and CMU-MOSI benchmark datasets. The source code of our proposed *Deep-HOSeq* is and available at <https://github.com/sverma88/Deep-HOSeq--ICDM-2020>.

Index Terms—multimodal data fusion, sentiment analysis, tensor analysis, convolution neural networks.

I. INTRODUCTION

There is increasing popularity with sharing opinionated videos on social media platforms such as YouTube, Facebook, etc. where the speaker’s sentiments are available via multiple heterogeneous forms of information such as language (spoken words), visual-gestures, and acoustic (voice). While there has been significant development in 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 [1], [2]. Since utilizing cues from these interacting modalities often presents a more complete view of the underlying phenomenon and thus enhances the generalization performance for sentiment prediction [1], [2]. Although performing multimodal

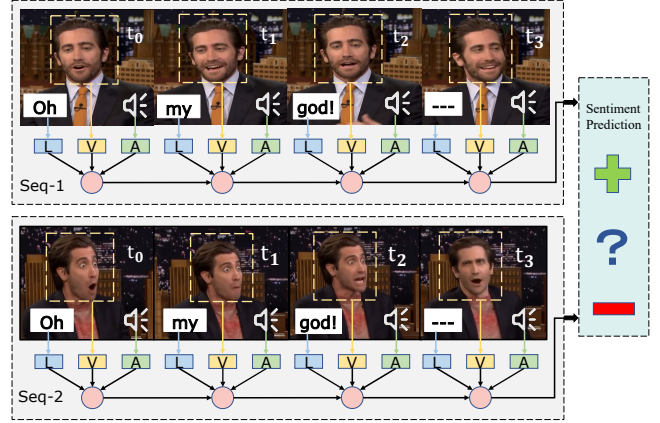


Fig. 1. A typical scenario illustrating different sentiments expressed with same spoken-utterance but visual gestures and vocal intonations. The asynchronous visual-gesture (occurring after the end of spoken words) at time t_3 paramountly aids in identification of the speaker’s sentiment in the two sequences. Efficient processing of such asynchronous (and synchronous) temporal-interactions are a necessity for sentiment analysis through multimodal fusion.

fusion for sentiment prediction is itself a challenging task due to multiple recurrent issues such as missing-values in the visual and acoustic modalities, misalignment, and etc. [3], [4]. The challenge is exacerbated when the fusion is required in the temporal domain as the multimodal temporal-interaction possesses the dual nature of promising the data-granularity and concealing its ambiguity as peril. A motivating example for this scenario is presented in Fig. 1 where the speakers in both the sequences utilize the same words to express their sentiments differently.

The speakers in both the sequences of Fig. 1 utilize the same utterances (spoken words) to express their sentiments. Although both the sequences contain the same spoken words, the interactions between facial expressions and vocal intonations asynchronously occurring with each spoken word unveil critical information that necessitates their disparate labeling of the sequences. In particular, the facial expression at time t_3 drives the identification of the speaker’s sentiment in the sequences. Therefore, discarding such temporal-granularity will result in the loss of critical information that substan-

Jiwei Wang[‡], Zhefeng Ge[‡], Rujia Shen[‡] contributed equally to this work and share the second authorship of this work.

Sunny Verma^{†,§} is the corresponding author of this work.

tially helps in identifying the speaker’s true sentiment. While these interactions can occur in the form of synchronous and asynchronous¹ multimodal interactions and hence, combining these temporal-cues will enhance the robustness of sentiment prediction with multimodal temporal sequences.

In this regard, to enhance the predictive power by utilizing such temporal-cues recent multimodal approaches such as MARN (Multi-Attention Recurrent Network) [5] and MFN (Memory Fusion Network) [6] combined both the inter-modal and intra-modal interaction while performing multimodal fusion. These schemes utilize series of LSTMs to obtain intra-modal dynamics and constrain them with sophisticated attention schemes (multi-attention head in MARN and delta-attention memory in MFN) to exploit the inter-modal temporal-interactions. Although both of these techniques unanimously conclude that the utilization of both these types of information positively impacts multimodal sentiment analysis however these schemes entirely rely on the attention mechanism to discover inter-modal information and amalgamate it with intra-modal information. Their complete reliance on attention scheme is problematic due to two reasons: 1) they have deceptive attention masks (in MFN), and hence it is obscure whether the gain in prediction is attributable to inter-modal interactions [7] and, 2) the role of training dynamics (in MARN) instead of multiple-heads [8]. Nevertheless, both these schemes require substantial efforts to optimize the hyperparameters of their consolidated architectures to perform multimodal sequence fusion efficiently. To alleviate these drawbacks, we propose *Deep-HOSeq* to perform multimodal fusion, in particular when the modalities are available as temporal sequences.

The *Deep-HOSeq* performs multimodal fusion by extracting two kinds of contrasting information from multimodal temporal sequences. The first kind of information is the amalgamation of both inter-modal and intra-modal information and can be perceived as the common² information extracted from the modality interaction. The second type of information exploits the temporal-granularity (synchronous and asynchronous interactions within modalities, as shown in Fig.1) among the multimodal sequences and is derived as unique information while performing multimodal fusion. To aid the understating of proposed *Deep-HOSeq* we illustrate its workflow in Fig. 2.

To extract these two kinds of information, we design a common network that first utilizes basic LSTM to obtain the intra-modal information from each unimodality. Then the obtained intra-modal information from each modality is amalgamated as multi-mode tensors by taking their outer-

TABLE I
COMPARISON OF VARIOUS MULTIMODAL FUSION SCHEMES

Fusion Schemes	Inter Modal	Intra Modal	Attention Reliance	Convolution	Multimode Representation
TFN	✓	×	×	×	✓
LMF	✓	×	×	×	×
DeepCU	✓	×	×	✓	✓
MFN	✓	✓	✓	×	×
MARN	✓	✓	✓	×	×
<i>Deep-HOSeq</i>	✓	✓	×	✓	✓

product. The elements within this multi-mode tensor reflect the strength of inter-modal interactions as correlations [9], and this rich inter-modal information is finally captured by utilizing convolution kernels followed by fully connected layers. On the other hand, we also design a unique network for leveraging the temporal-granularity among multimodal sequences. This is achieved by first obtaining latent features from each unimodality by utilizing feed-forward layers to increase their discriminative power. We then obtain higher-order interactions within the modalities at each temporal-step followed by feature extraction with convolution layers and fully connected layers (as in the common network). We finally unify the information from all the temporal-steps with a pooling operation, which encapsulates the temporal-granularity as the unique information in *Deep-HOSeq*. Although one may argue that our choice of unification scheme is not sophisticated but this scheme efficiently captures the temporal-dynamics within multimodal sequences and is demonstrated in the results section.

We finally integrate both these kinds of information with a fusion layer to perform multimodal sentiment prediction and call our novel multimodal fusion scheme as *Deep-HOSeq* (Deep Higher-Order Sequence Fusion). An important characteristic of our *Deep-HOSeq* is that it does not rely on attention-based schemes and hence does not face the same critiques as state of the art (SOTA) techniques such as MARN and MFN. Its superiority lies in simple but careful design choices that enable joint discovery and utilization of all-essential information to perform multimodal fusion. To aid the understanding of our technique, we summarize the similarities and differences between *Deep-HOSeq* and SOTA techniques in Table. I. Besides, our major contributions in this work are summarized as below:

- 1) We design a common network to extract both intra-modal and inter-modal information in a cascaded framework for multimodal fusion. Conceptually, the information obtained by our common network is more expressive than the SOTA as we utilize convolution on multi-mode tensors, which efficiently captures all-essential inter-modal interactions. Besides, the use of basic LSTMs efficiently discovers the underlying intra-modal dynamics and does not require strenuous efforts for parameter optimization.
- 2) We design a unique network that encapsulates the temporal-granularity from multimodal sequences. This

¹Visual-Gesture occurring at the end of the spoken words.

²It should be noted that the terms ‘common’ and ‘unique’ information are also utilized in DeepCU [4] to refer to a different but related concept. The concept of common information in DeepCU is limited to inter-modal information, whereas in *Deep-HOSeq*, the common information is comprised of both inter-modal and intra-modal information. Besides, the unique information in the DeepCU is comprised of factorized information from unimodality’ integrated by late fusion. In contrast, the unique information in *Deep-HOSeq* refers to the information present via asynchronous and synchronous temporal occurrence among modalities.

enhances the *Deep-HOSeq*'s robustness with multimodal synchronous and asynchronous interactions.

- 3) We design a deep consolidated network for joint discovery and utilization of both common and unique information from multimodal temporal sequences, which we call as *Deep-HOSeq*.
- 4) We perform comprehensive experiments on multimodal CMU-MOSEI and CMU-MOSI datasets and demonstrate the effectiveness of utilizing both common and unique information in comparison to other techniques.

The rest of the paper is organized in the following sections: Sec. II presents literature review of existing multimodal fusion techniques followed by details of our proposed *Deep-HOSeq* in Sec. III. Experimental setup and results are described in Sec. IV and Sec. V, respectively. We finally conclude our work and discuss its possible future directions in Sec. VI.

II. RELATED WORK

We focus our review on techniques performing neural-based fusion of multimodal sequences where arguably the simplest deep architecture performing fusion of heterogeneous data is Deep Multimodal Fusion (DMF) [10]. The DMF is a successor to the Early Fusion (EF) [11], which is one of the most utilized non-neural technique performing multimodal data fusion. The DMF is developed to perform both a) EF: combine raw (or latent) features by concatenating them; b) late fusion: process each modality with a deep network and then synthesize their decisions. Although powerful, the DMF (and EF) is a basic technique and assumes that a modality (for example, visual) does not share any relevant information within itself. In other words, it can not leverage the intra-modal information a particular modality might offer. Hence, it is limited to express only the inter-modal interactions and thus faces the same limitation as in EF [6]. We now review SOTA that leverages both inter-modal and intra-modal relationships while performing multimodal fusion.

a) *Memory Fusion Network (MFN)*: [6] is a recurrent model that consists of three sub-modules a) System of LSTMs to obtain intra-modal dynamics from each unimodality; b) Delta-memory Attention Network which discovers inter-modal dynamics; and c) Multi-view Gated Memory responsible for integrating the intra-modal and inter-modal dynamics. The final input for fusion in MFN comprises of concatenated intra-modal information from LSTMs and the final state of the Multi-View Gated Memory and hence can be assumed as a sophisticated EF system. Albeit powerful, the MFN relies entirely on the attention network and the Multi-View Gated Memory to obtain inter-modal dynamics while performing multimodal fusion. This complete reliance on attention mechanism is problematic as the MFN assumes synchronous inputs which is hard to achieve in real-world scenarios, and more importantly, the reliability of attention-memory to discover inter-modal interactions is questionable as shown with deceptive attention masks in [7].

b) *Multi-attention Recurrent Network (MARN)*: [5] is also a recurrent model and consists of two sub-modules a)

Long-short Term Hybrid Memory (LSTHM) that amalgamates intra-modal dynamics inter-modal temporal-dynamics by explicitly augmenting LSTM with a hybrid memory; and b) Multi-attention Block (MAB) which discovers the inter-modal dynamics and successively updates the hybrid memory of LSTHMs. Similar to MFN, the MARN also completely relies on the attention scheme to obtain the inter-modal dynamics. The key difference between the two is that the earlier utilizes basic LSTMs [12], whereas the latter augments a hybrid memory within the LSTMs. Besides, the MARN attributes usage of multiple-attention for gains in the predictive performance, but it is obscure whether it is from the discovery of inter-modal information or the training dynamics of the MAB (and LSTHM, which are much strenuous than MFN) [8].

Differently from the above, few notable multimodal fusion techniques which do utilize sophisticated attention mechanisms are Tensor Fusion Networks (TFN) [13], Low-rank Multimodal Fusion (LMF) [14], and DeepCU [4]. These techniques perform multimodal fusion by utilizing the summarized information within visual (and acoustic) modality as its average. Although this leads to the loss of sequential information present in the form of visual and acoustic interactions. These techniques compensate for this information loss by modelling multiple combinations of inter-modal interactions, either as tensors or its low-rank factorized representation.

Our proposed *Deep-HOSeq* is similar to the above as it also aims to exploit both the inter-modality and intra-modality relationship while performing multimodal fusion, but substantially differs from them due to the following:

- I. The common network in *Deep-HOSeq* extracts information from inter-modal tensors obtained via modelling the intra-modal information in multimodal sequences. Since the elements of this tensor signify the correlation strength between the fusion modalities, the information obtained is not obscure or deceptive as in MARN and MFN.
- II. Obtaining inter-modal temporal-granularity independently with unique network is a distinctive characteristic of *Deep-HOSeq*, and inclusion of this information enhances the *Deep-HOSeq*'s capability while dealing with asynchronous (and synchronous) multimodal sequences.
- III. The fusion layer integrates both the common and the unique information to perform multimodal sentiment analysis. It is worth mentioning that this layer uses averaging and hence does not introduce extra model parameters. More importantly, it also refrains the common network to influence the parameters of the unique sub-network and vice-versa. This restriction allows the sub-networks to obtain complementary information and hence increase the diversity during fusion.

Although, all the techniques mentioned above are fundamentally different from proposed Deep-HOSeq; one must not consider the equality of Deep-HOSeq in particular to DeepCU – based on the terms common and unique. The concept of common and unique information in both techniques is disparate and explained in detail in the footnote². Furthermore, the fea-

ture dissection process is also distinct in the technique where the earlier is proposed to perform multimodal fusion from asynchronous (and synchronous) interactions within temporal sequences whereas the latter is proposed to perform fusion of independent data units.

III. PROPOSED METHODOLOGY

We aim to utilize intra-modal and inter-modal dynamics by amalgamating them as common information and the dynamics of the temporal-granularity as unique information for multimodal fusion. To achieve this, we propose two sub-networks, i.e., 1) common sub-network which extracts intra-modal and inter-modal information in a cascaded manner as described in Sec. III-A, and 2) unique sub-network for encapsulating the temporal-granularity as detailed in Sec. III-B. The information obtained by both networks is then integrated via a fusion layer to perform multimodal sentiment prediction in Sec. III-C. To aid the understanding of our sub-networks, we illustrate the workflow of *Deep-HOSeq* in Fig. 2.

We begin with raw feature vectors from a spoken utterance in the form of acoustic, visual, and language modalities denoted as $z_a \in \mathbb{R}^{1 \times t_k \times d_a}$, $z_v \in \mathbb{R}^{1 \times t_k \times d_v}$, and $z_l \in \mathbb{R}^{1 \times t_k \times d_l}$ respectively, where d_a , d_v , and d_l represents the dimensionality of the feature vectors and t_k represents the sequence length. These feature vectors are then independently processed with basic LSTMs in the common network and with feed-forward layers in the unique network which constrains both the sub-networks to obtain unshared latent representation. This restriction allows the sub-networks to obtain complementary feature representations at lower layers as the latent space of unique sub-network remains unaffected by the gradient from the common sub-network and vice-versa for the common sub-network. Furthermore, optimizing unshared latent space also enhances the expressiveness of *Deep-HOSeq* and is empirically shown beneficial in the works [15], [16].

A. Common Network

The common network first obtains intra-modal dynamics from the individual modality by processing them with basic uni-directional LSTMs; we chose basic LSTMs as they are simple-yet-powerful enough to discover the relevant parts within a modality. We then process the latent features obtained with all the unimodal LSTMs (the final state of LSTMs) with fully-connected layers to increase their discriminatory strength, followed by an outer product to obtain multi-mode tensors. These tensors represent the amalgamated intra-modal and inter-modal information within the multimodal sequences.

$$\begin{aligned} h_V &= \sigma(LSTM(z_v) \times W_v + \mathbf{b}_v) \\ h_A &= \sigma(LSTM(z_a) \times W_a + \mathbf{b}_a) \\ h_L &= \sigma(LSTM(z_l) \times W_l + \mathbf{b}_l) \\ T_{VAL} &= h_V \otimes h_A \otimes h_L \end{aligned} \quad (1)$$

The elements of tensor T_{VAL} signifies the inter-modal interactions as correlation strengths [9], which can be efficiently

derived by processing the tensor with a series of convolution and fully-connected layers as detailed in (2). The highly discriminative information available after dissecting T_{VAL} denoted as h_{com} represents the amalgamated intra-modal and inter-modal information perceived as the common information in *Deep-HOSeq*. Although this common information i.e. h_{com} can still be utilized to perform multimodal prediction as in (3); it is still not as effective as utilizing both common and unique information, and a relative comparison of utilizing both common and information vs. only common information is presented in the experiments section. It should be noted that our common network can be extended with sophisticated convolutions schemes such as ResNet [17] etc. to boost its generalization power.

$$\begin{aligned} \mathbf{g}_{VAL} &= \sigma(Conv(T_{VAL})) \\ h_1 &= \sigma(\mathbf{g}_{VAL} \times W_1 + \mathbf{b}_1) \\ h_2 &= \sigma(h_1 \times W_2 + \mathbf{b}_2) \\ &\dots \\ h_n &= \sigma(h_{(n-1)} \times W_{(n-1)} + \mathbf{b}_{(n-1)}) \\ h_{com} &= \sigma(h_{(n)} \times W_{(com)} + \mathbf{b}_{com}) \end{aligned} \quad (2)$$

where \mathbf{g}_{VAL} is obtained by flattening \mathbf{G}_{VAL} to process the convolution output with fully connected layers. Multimodal sentiment prediction with common information can be obtained by utilizing h_{com} as in (3).

$$\hat{y}_{com} = (\mathbf{h}_n \times \mathbf{w}_{0_c}) + b_{0_c} \quad (3)$$

B. Unique Network

The unique network first obtains latent representations of raw unimodal features by processing with a sequence of feed-forward layers and then utilizes the obtained discriminative representations to capture cross-categorical correlations in the form of multi-mode tensors. These tensors accommodate complex dynamical factors such as inter-region spatial correlations between the temporal sequence. Thus we process them with convolution and fully-connected layers (as in the common network) to extract the concealed unique information in them. The whole feature extraction process is mathematically described as in (4), where $k = 1, 2, \dots, t_k$ is the sequence length.

$$\begin{aligned} h_{V_k} &= \sigma(z_{v_k} \times W_{v_k} + \mathbf{b}_{v_k}) \\ h_{A_k} &= \sigma(z_{a_k} \times W_{a_k} + \mathbf{b}_{a_k}) \\ h_{L_k} &= \sigma(z_{l_k} \times W_{l_k} + \mathbf{b}_{l_k}) \\ T_{VAL_k} &= h_{V_k} \otimes h_{A_k} \otimes h_{L_k} \\ \mathbf{h}_k &= \sigma\left(\sigma(Conv(T_{VAL_k})) \times W_{val_k} + \mathbf{b}_{val_k}\right) \end{aligned} \quad (4)$$

The discovery of such complex factors is essential as they promote collaboration among temporal and semantic views of the data and thus enhance classifiers' predictive performance by

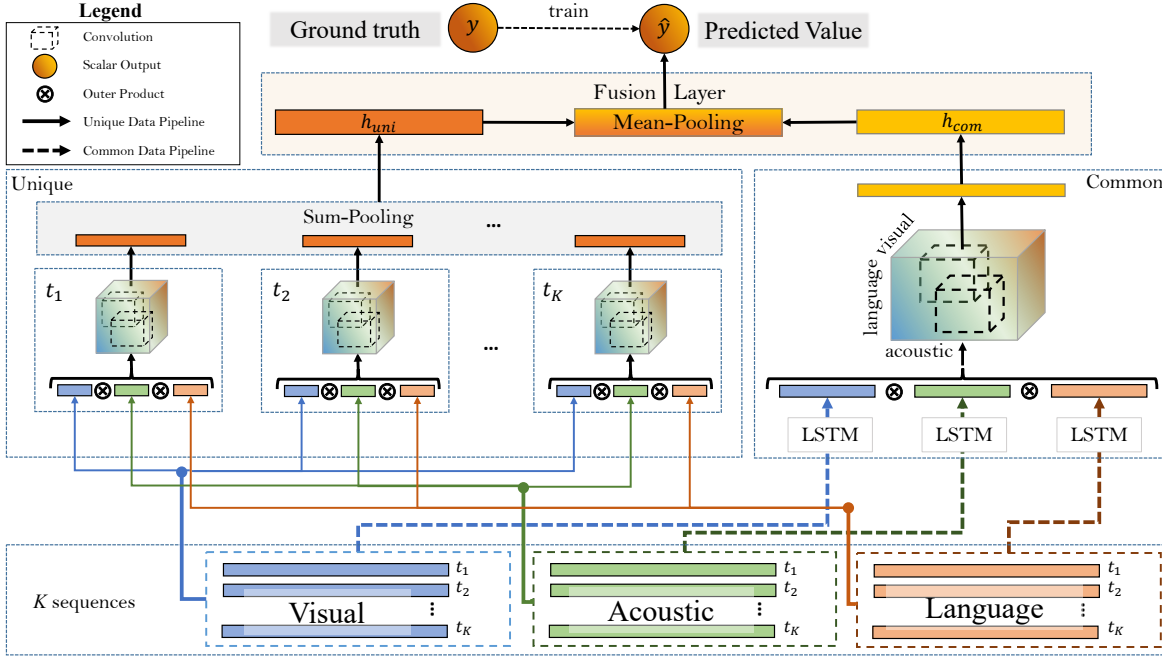


Fig. 2. Workflow of the proposed Deep Higher-Order Sequence Fusion network *Deep-HOSeq*.

learning the dynamical relationship within multiple modalities [18], [19]. We finally unify the sequential information at each temporal step with a pooling operation in (5). We arguably again utilize a simple-yet-effective operation to summarize the most discriminate pattern from the intermediate layer. Importantly, this unification does not add any extra parameters and provides further opportunities to increase the discriminatory strength of the unique features.

$$h_{pool} = \sum_{k=1}^{t_k} h_k \quad (5)$$

$$h_{uni} = \sigma(h_{pool} \times W_{pool} + b_{pool})$$

These pooled representations denoted as h_{pool} encapsulates the temporal-granularity from the modalities interactions and are perceived as the unique information in this paper. Similar to common information, one can also utilize unique information to perform multimodal sentiment prediction, as in (6).

$$\hat{y}_{uni} = (h_{uni} \times w_{0_u}) + b_{0_u} \quad (6)$$

C. Fusion Layer

The feature vectors from the last hidden layer of common and the unique sub-networks are integrated by performing a mean-pooling operation followed by feed-forward layer to derive the combined prediction with common and unique information in *Deep-HOSeq* in (7). Our motivation for applying mean-pooling instead of any other pooling (to integrate common and unique information) is that the mean-pooling will enforce equal learning within both common and unique sub-networks as it allows equal gradient flow in both the

sub-networks while training. This scenario, however, is not guaranteed in case max-Pooling (or any other pooling layer) is employed, as there is a possibility that a single network might be dominant while training and thus will result in the absence of either kind of information. The validity of this hypothesis is shown in Fig. 3 by comparing the performance of different kinds of information.

$$h_{combined} = average_pooling(h_{com}, h_{uni}) \quad (7)$$

$$\hat{y}(x) = h_{combined} \times W + b$$

The weights of our proposed *Deep-CUSeq* is optimized via minimizing the mean square error (MSE) loss in (8), 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 *Deep-HOSeq*.

$$L = \frac{1}{n} \sum_{\forall x \in \chi} (\hat{y}(x) - y(x))^2 \quad (8)$$

D. Complexity Analysis

The paramount computation complexity in *Deep-HOSeq* arises with obtaining unshared latent features in the unique network. This is because we utilize feed-forward layers for obtaining latent features from each temporal-sequence within modalities, and this accumulates to approximately 43% of the total trainable neurons in *Deep-HOSeq*.

While a direct comparison of running time is not possible as the baselines are customised in Pytorch and *Deep-HOSeq* is written in Tensorflow, but the number of parameters in the optimized MFN model is equal to 2.34×10^5 , and for MARN it is equal to 4.58×10^5 . Whereas, the number of trainable parameters in optimized *Deep-HOSeq* is equal to 2.61×10^5 .

TABLE II
THE SPEAKER INDEPENDENT SPLITS FOR TRAINING, VALIDATION, AND
TEST SETS FROM CMU-MOSEI AND CMU-MOSI DATASETS.

Dataset	CMU-MOSI	CMU-MOSEI
#Training Instances	1284	15290
#Validation Instances	229	2291
#Testing Instances	686	4832

Additionally, both MFN and *Deep-HOSeq* took less than 30 epochs to converge on the CMU-MOSEI dataset while the MARN did not converge with 1000 epochs.

IV. EXPERIMENTAL SETUP

A. Dataset.

We perform experiments on the CMU-MOSI [20] and CMU-MOSEI datasets [21] where both the datasets consist of opinion videos collected from YouTube with only a single person in front of the camera expressing his opinion. The CMU-MOSI dataset consists of reviews from 93 distinct speakers where each video consists of multiple opinion segments with a total of 2199 utterances (segments) in the whole dataset. On the other hand, the CMU-MOSEI dataset consists of reviews from 1000 distinct speakers with a total of 23,453 utterances. Each utterance in the video is annotated with the sentiment in the range $[-3, 3]$. Here -3 indicates highly negative and $+3$ indicates highly positive sentiment.

1) *Features*: We accessed the language, visual, and acoustic features provided by the authors [20] at their official publicly available repository³. The modality specific features are provided after word alignment using P2FA [22] aligning them at the word granularity.

a) *Language*: Pre-trained 300-dimensional Glove word embeddings [23] were utilized to encode each sequence of transcribed word into a sequence of word vectors.

b) *Visual*: The library Facet⁴ is used to extract visual features for each frame (sampled at 30Hz). Extracted features consists of 20 facial action units, 68 facial landmarks, head pose estimates, gaze tracking and HOG features [24].

c) *Acoustic*: COVAREP acoustic framework [25] is utilized to extract features including 12 MFCCs, pitch, glottal source, peak, slope, voiced/unvoiced segmentation, and maxima dispersion quotient.

The feature vectors for each modality is publicly available via CMU-MultimodalDataSDK. Also, in order to evaluate the generalization capability of models the training, testing, and validation splits of datasets are speaker-independent and pre-defined in the CMU-MultimodalDataSDK and the number of instances for both datasets are reported in Table. II.

B. Baselines.

We extensively evaluate the performance of our proposed *Deep-HOSeq* against neural-based and non-neural based

schemes available for multimodal sentiment analysis. Thus we trained our *Deep-HOSeq* and also the baselines with MSE loss in Eq. 8. The details of the baselines are described as below:

1) *Early Fusion, Non-Neural Approaches*: We first collapsed the sequence dimension in all the modalities by taking their average and then trained them for a regression task by concatenating the average features. The baselines thus reported are Support Vector Machines (SVM) and Random Forest (RF).

2) *Joint Representation, Neural Approaches*: For baselines under joint representation, we followed the protocol as in [21] and thus trained basic LSTMs on individual modalities and treated their final state as the latent features available for fusion. We then concatenated these features and trained a deep network for regression reported as EF_{LSTM} [26]. We also trained RF and SVM on these latent features reported as RF-MD [20] and SVM-MD [20], respectively. We also trained an extreme learning machine (ELM) classifier on these latent features as utilized to predict multimodal sentiment in [27].

3) *Deep Networks without temporal information*: As a requirement for baselines under this such as TFN [13], LMF [14], and DeepCU [4], we collapsed the sequence dimension of acoustic and visual modalities and performed a grid search to optimize to network's hyperparameters.

4) *State of the art deep neural networks*: Under this we have two state of the art techniques performing multimodal fusion with temporal sequences.

MARN (Multi Attention Recurrent Networks) *SOTA1* [5] is described in Sec. II. The source code of MARN is publicly available and a grid search is performed to optimize the network's hyperparameters on MOSEI dataset.

MFN (Memory Fusion Networks) *SOTA2* [6] is another state of the art as described in Sec. II. The source code is publicly available and a grid search is performed to optimize the network's hyperparameters on MOSEI dataset.

C. Parameter Setting in Deep-HOSeq

We implemented *Deep-HOSeq* in TensorFlow⁵ and optimized it by minimizing the loss in Eq. 8 with Adam Optimizer [28]. The learning rate was set to 6×10^{-3} with a mini batch size of 256. To avoid over-fitting we applied dropout [29] in our model and tune the dropout probability from $[0.05, 0.8]$ with a step size of 0.05. The optimal dimensions of latent space in each sub-network was searched in $[5, 10, 15, 20, 25, 30]$, while the number of convolution filters were set between $[1, 3]$. We also applied batch-normalization [30] to the convolution layers to speed up the training of *Deep-HOSeq*. Besides, we utilized basic uni-directional LSTM-cell⁶ for obtaining the intra-modal dynamics. Lastly, we employed early stopping as in [6], where the training is terminated if the MAE on the validation-set did not improved in 5 consecutive epochs.

⁵Link to our *Deep-HOSeq*'s source code repository is available at: <https://github.com/sverma88/Deep-HOSeq--ICDM-2020>

⁶https://www.tensorflow.org/api_docs/python/tf/nn/rnn_cell/BasicLSTMCell

³<https://github.com/A2Zadeh/CMU-MultimodalSDK>, SDK Version 1.0.1

⁴<https://imotions.com/>

D. Evaluation Metrics

We evaluate the performance of the baselines and *Deep-HOSeq* for regression, binary classification (positive and negative sentiments), and multi-class classification (7 sentiments). In this regard, we report Mean Absolute Error (MAE) and Pearson’s Correlation (Correlation) for regression, and in the case of binary classification we report accuracy and F1-score. Whereas for multi-class classification we only report accuracy. Note that, for all metrics, higher value is better except for MAE where a lower value is better.

Besides, while calculating the binary and multi-class accuracies, we followed the protocol in [4], [14] and map the predicted sentiment (and the true sentiment) to integer values.

V. RESULTS AND DISCUSSIONS

The key contribution of this work is utilization of both common and unique information for multimodal data fusion. Therefore, in order to study the significance of proposed *Deep-HOSeq* and its relative sub-components we performed the experiments as per the following research questions:

A. Does the integration of both common and unique information enhance the generalization performance in Deep-HOSeq or does it deteriorate the performance?

To evaluate the effectiveness of integrating both the common and unique information in *Deep-HOSeq*, we studied the performance of multimodal sentiment prediction by considering information from a) unique sub-network; b) common sub-network; and c) *Deep-HOSeq*. In this regard, we obtained the performance of these different kinds of information by performing a grid search on all the hyper-parameter settings as in Sec. IV-C and present their optimized MAE on CMU-MOSEI test dataset with a box-plot in Fig. 3.

First, the plot clearly suggests that the integration of both the common and the unique information is beneficial for performing multimodal sentiment analysis. We argue that this is because the *Deep-HOSeq* efficiently leverages the advantages of both the amalgamated inter-modal and inter-modal information as common information and temporal-dynamics as unique information.

Second, the plot also shows that the common information achieves lower MAE compared to the unique information suggesting that the information obtained by amalgamating the intra-modal and inter-modal dynamics is more important than the temporal-granularity of multimodal sequences. Although, this might depend on the data as there may be fewer instances with high temporal-variability in the testing dataset.

B. How does the hyper-parameters affect the performance of unique, common, and Deep-HOSeq?

To answer the above question, we present a comprehensive study to understand the effects of various hyper-parameters on the performance of *Deep-HOSeq*. In this regard, we split the discussion into three subsections and first discuss the effects of hyper-parameters on common and unique networks. We then discuss ways to enhance the feature discriminability

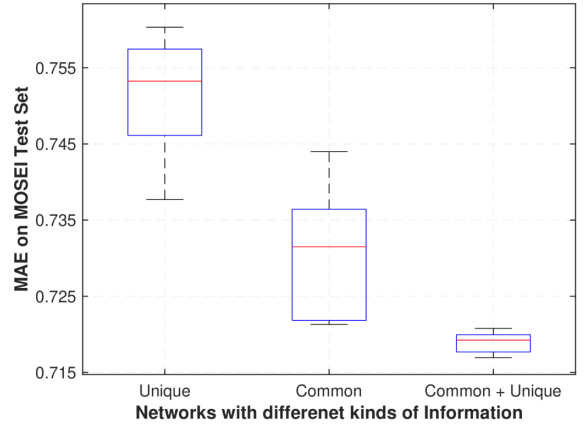


Fig. 3. Performance comparison of *Deep-HOSeq* vs. only common and unique information on the CMU-MOSEI test dataset.

of common networks by utilizing sophisticated convolution networks such as ResNet. We finally utilize these optimized hyperparameters and study the effect of varying the dimensionality of the latent space (multi-mode tensor) on *Deep-HOSeq*.

1) Analysis of common and unique sub-network: We first analyze the effects of various hyperparameters in the common and unique subnetworks, in particular the effect of activation functions in the lower layers. Since activation functions such as *relu* can lead to a sparse multi-mode tensor as compared to *identity* or *sigmoid* activation functions. We are thus interested in understanding whether extracting features from a full or a sparse tensor has significant effects on the networks’ prediction performance. In this regard, we plot the MAE achieved by concurrently optimizing the dropout ratio and the size of convolution kernel with constant size of the latent dimensions as 10 in the common and unique sub-networks in Fig. 4. The *x-axis* in this figure represents the choice of different activation functions in the lower layers. Besides, we have repeated the same experiment by applying batch normalization referred to as *BatchNorm* and without batch normalization referred to as *W/O-BatchNorm* in the legend of Fig 4.

The performance comparison of different activation functions in the two sub-networks clearly dictates that *relu* as an activation function performs equally good as any other activation function. However, we empirically found that the two networks converge faster with *relu* activation than others. Moreover, these performances also suggest that applying batch normalization has a negligible effect on the performance of the sub-networks. However, empirically we found that the networks with batch normalization layer converged much faster than their counterparts. Besides, both the sub-networks perform marginally better with smaller kernel size as this might be due to the increase in overlapping regions between segments in the multi-mode tensors.

2) Analysis of ResNet layer in common network: The performance study of applying ResNet on the common sub-network does not yield any new insights. However, it confirms that applying batch normalization does not have a significant

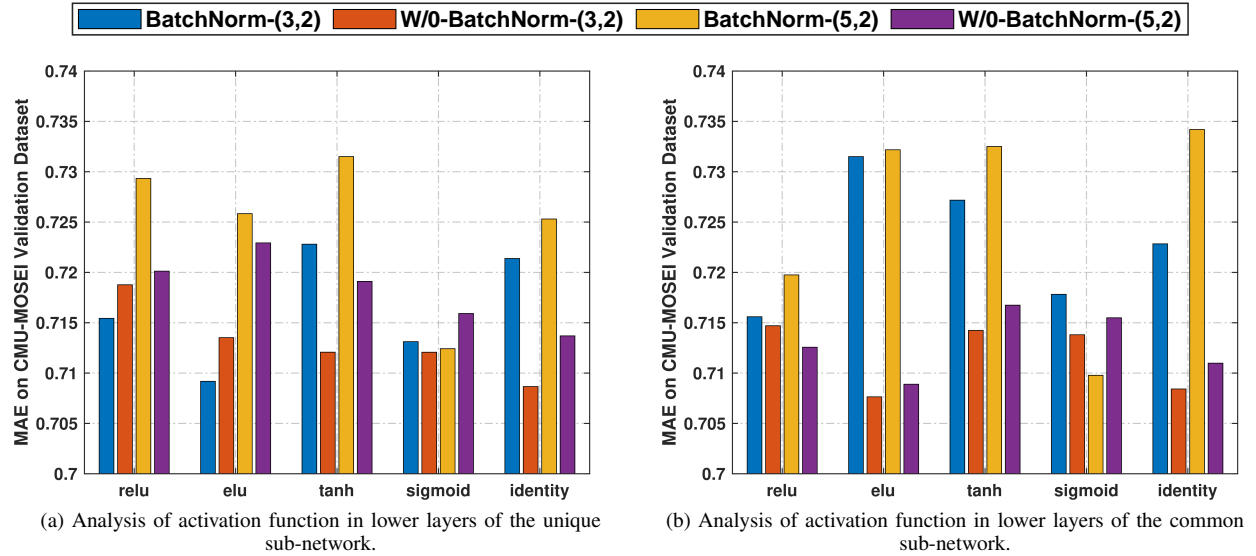


Fig. 4. Hyperparameter study of common and unique sub-networks on the CMU-MOSEI validation set. In the legend, integers x, y represents the kernel size and stride size in the convolution layers.

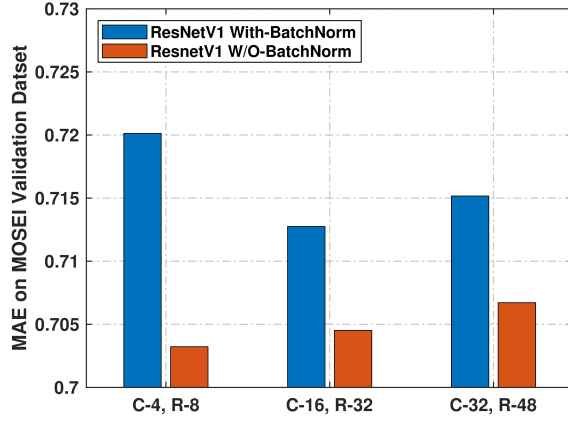


Fig. 5. Performance of ResNet-V1 by varying the number of convolution kernels in the common sub-network on the CMU-MOSEI dataset. The integers in the legend represents the number of kernels in the previous convolution layer and the ResNet layer respectively.

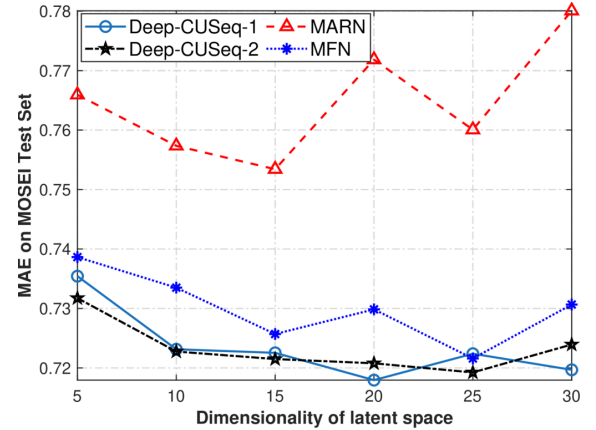


Fig. 6. Performance of *Deep-HOSeq*, MARN, and MFN by varying hyperparameters on the CMU-MOSEI dataset. The integer x in *Deep-HOSeq*- x represents the number of convolution kernels.

advantage to our model. This might be justifiable as contradictory to the image recognition the inputs to our network are features extracted by pre-trained networks. Therefore, the network does not face a covariate shift during prediction. However, the networks did converge faster with batch normalization.

We decide to drop ResNet as a layer from our model as the predictive performance of *Deep-HOSeq* with a simple convolution layer is as good as with a sophisticated ResNet layer. Moreover, the number of trainable parameters increases with the ResNet layer and encourages us to proceed with a simple convolution layer in our common sub-network.

The failure of ResNet scheme in our model might be due to less amount of training data or the depth of the network, which is at most 2 in our case. We would like to perform some statistical analysis in the future to investigate the role of

ResNet in such kind of similar schemes.

3) *Analysis of Deep-HOSeq*: We now plot the mean MAE obtained by varying the dimensionality of the latent space in *Deep-HOSeq*'s sub-networks shown as the x -axis in Fig. 6, and illustrate different colors to represent the number of convolution filters. Besides, as sanity checks, we also plot the optimized MAE obtained from MARN and MFN on the same latent dimensions. A clear trend is visible in performance curves of *Deep-HOSeq* in Fig. 6 where the MAE improves significantly by increasing the latent dimensions with noticeable improvements beyond latent space of 5. This might be due to the size of the convolution filter which happens to be equal to the size of the multi-mode tensors, and hence applying convolutions does not prove much beneficial to *Deep-HOSeq*. However, the performance gradually improves with

TABLE III

PERFORMANCE COMPARISON OF *Deep-HOSeq* VS. OTHER FUSION TECHNIQUES ON CMU-MOSEI DATASET. EACH PERFORMANCE METRIC IS EXECUTED FOR 5 TIMES AND THEIR MEAN AND STANDARD DEVIATION ARE REPORTED. NOTE FOR ALL THE METRICS A HIGHER VALUE IS BETTER BUT FOR MAE.

MOSEI Dataset	Regression		Binary		7-class
	MAE (lower is better)	Correlation	Accuracy	F1	Accuracy
<i>RF</i>	$0.7794 \pm 3.54 \times 10^{-3}$	$0.4229 \pm 7.40 \times 10^{-3}$	$70.63 \pm 3.16 \times 10^{-3}$	$70.98 \pm 3.12 \times 10^{-3}$	$39.23 \pm 3.53 \times 10^{-3}$
<i>SVR</i>	$0.7758 \pm 1.66 \times 10^{-3}$	$0.4348 \pm 1.82 \times 10^{-3}$	$70.53 \pm 5.96 \times 10^{-4}$	$71.06 \pm 5.39 \times 10^{-4}$	$37.64 \pm 6.68 \times 10^{-3}$
<i>EF_{LSTM}</i> [26]	$0.7861 \pm 1.03 \times 10^{-2}$	$0.3815 \pm 3.78 \times 10^{-2}$	$71.03 \pm 2.76 \times 10^{-3}$	$71.80 \pm 9.01 \times 10^{-3}$	$40.34 \pm 5.86 \times 10^{-3}$
<i>RF-MD</i>	$0.7995 \pm 1.93 \times 10^{-2}$	$0.3879 \pm 2.86 \times 10^{-2}$	$72.08 \pm 3.06 \times 10^{-3}$	$72.17 \pm 4.76 \times 10^{-3}$	$38.81 \pm 1.18 \times 10^{-2}$
<i>SVM-MD</i> [20]	$0.7886 \pm 2.34 \times 10^{-2}$	$0.3906 \pm 4.46 \times 10^{-2}$	$72.03 \pm 3.09 \times 10^{-3}$	$71.71 \pm 4.74 \times 10^{-3}$	$38.87 \pm 1.44 \times 10^{-2}$
<i>ELM</i> [27]	$0.7699 \pm 3.37 \times 10^{-3}$	$0.4439 \pm 6.28 \times 10^{-3}$	$69.06 \pm 2.04 \times 10^{-3}$	$70.41 \pm 2.19 \times 10^{-3}$	$39.62 \pm 5.06 \times 10^{-3}$
<i>TFN</i> [13]	$0.7483 \pm 1.06 \times 10^{-2}$	$0.5005 \pm 6.62 \times 10^{-3}$	$69.08 \pm 1.74 \times 10^{-2}$	$69.34 \pm 1.08 \times 10^{-2}$	$40.88 \pm 2.03 \times 10^{-2}$
<i>LMF</i> [14]	$0.7417 \pm 1.19 \times 10^{-2}$	$0.5058 \pm 1.09 \times 10^{-2}$	$71.04 \pm 1.04 \times 10^{-2}$	$71.32 \pm 6.23 \times 10^{-3}$	$40.64 \pm 1.60 \times 10^{-3}$
<i>DeepCU</i> [4]	$0.7331 \pm 4.32 \times 10^{-3}$	$0.5125 \pm 3.54 \times 10^{-3}$	$71.82 \pm 8.96 \times 10^{-3}$	$70.87 \pm 7.50 \times 10^{-3}$	$41.30 \pm 4.29 \times 10^{-3}$
<i>MARN</i> (SOTA 1) [5]	$0.7532 \pm 4.46 \times 10^{-2}$	$0.4828 \pm 4.19 \times 10^{-2}$	$68.12 \pm 3.03 \times 10^{-2}$	$69.10 \pm 2.82 \times 10^{-2}$	$39.39 \pm 8.75 \times 10^{-3}$
<i>MFN</i> (SOTA 2) [6]	$0.7270 \pm 7.47 \times 10^{-3}$	$0.5243 \pm 3.54 \times 10^{-3}$	$72.47 \pm 8.09 \times 10^{-3}$	$73.11 \pm 7.06 \times 10^{-2}$	$42.69 \pm 3.97 \times 10^{-3}$
<i>Deep-HOSeq</i> (proposed)	$0.7189 \pm 1.15 \times 10^{-3}$	$0.5438 \pm 2.24 \times 10^{-3}$	$74.32 \pm 8.00 \times 10^{-3}$	$75.12 \pm 3.35 \times 10^{-2}$	$44.17 \pm 2.60 \times 10^{-3}$

TABLE IV

PERFORMANCE COMPARISON OF *Deep-HOSeq* VS. OTHER FUSION TECHNIQUES ON CMU-MOSI DATASET.

MOSI Dataset	Regression		7-class
	MAE (lower is better)	Correlation	Accuracy
<i>TFN</i>	1.1111 ± 0.0003	0.5341 ± 0.0010	31.98 ± 1.1321
<i>LMF</i>	1.0960 ± 0.0021	0.5455 ± 0.0032	30.76 ± 0.0339
<i>DeepCU</i>	1.0595 ± 0.0007	0.5506 ± 0.0076	33.14 ± 0.0639
<i>MARN</i>	1.1215 ± 0.0481	0.5116 ± 0.0323	30.54 ± 0.0661
<i>MFN</i>	1.0406 ± 0.0568	0.5461 ± 0.0291	34.14 ± 0.0219
<i>Deep-HOSeq</i>	1.0201 ± 0.0218	0.5676 ± 0.0166	35.87 ± 0.0332

the increase in the latent dimension supporting the learning requirement of the convolution kernels.

A second noticeable trend from the plot is that the performance of all the fusion schemes generally improves until the latent dimensions of 15 and then deteriorating sporadically indicating over-fitting, in particular at latent dimension of 30. Although, the *Deep-HOSeq* does not face a significant performance degradation as compared to MARN (and MFN) and this might be due to 1) less number of parameters required by convolution kernels, and 2) the ability of convolutions to efficiently capture utmost expressiveness concealed in multi-modal tensors.

C. Does Deep-HOSeq provide a better multi-modal fusion technique compared to SOTA such as MARN and MFN? Besides, are convolutions effective in obtaining inter-modal dynamics from multi-mode tensors, and whether inclusion of this information necessary?

To address this requirement, we compare the performance of *Deep-HOSeq* and baselines on the CMU-MOSEI and CMU-MOSI datasets. The performance evaluations are reported in Table. III and Table. IV, respectively. On the CMU-MOSEI dataset we improve the state of the art by **3.60%** for correlation, **2.55%** for binary class, and **3.46%** on multi-class accuracy. On the CMU-MOSI dataset, the *Deep-HOSeq*

improves the correlation by **3.94%**, and **5.07%** on multi-class accuracy compared to SOTA approaches.

The above results validate our hypothesis regarding a) utilizing both the common and unique information obtained with unshared latent space; b) the use of convolutions to capture utmost expressiveness offered by multi-mode representation; and c) basic LSTMs to effectively obtain intra-modal dynamics, and d) incorporating temporal-granularity from multimodal sequences with *Deep-HOSeq*.

Furthermore, we can easily see that the performance achieved with the common network in Fig. 3 is equally competitive to the performance achieved with state of the art MFN. This indicates that our common network is able to capture inter-modal dynamics with the utilization of convolutions efficiently. It is worth mentioning that the discovery of this information does not require any attention-mechanism whose information discovery is powerful but obscure.

Besides, on comparing the MAE achieved with *EF_{LSTM}* [26] against that our proposed *Deep-HOSeq*. Our proposed *Deep-HOSeq* reduces the MAE by **6.48%** as it consists of both the inter-modal and intra-modal information, while *EF_{LSTM}* only consists of intra-modal information. This demonstrates that inter-modal interactions play a significant role in multi-modal fusion and hence should not be neglected.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have introduced *Deep-HOSeq* to perform multimodal fusion from temporal sequences. The *Deep-HOSeq* integrates two kinds of information 1) common: amalgamated inter-modal and intra-modal information from multimodal temporal sequences, and 2) unique: temporal-granularity from multimodal interactions. We then demonstrated that both these two kinds of information are essential for multimodal fusion and integrated them via a fusion layer in *Deep-HOSeq*. The parameters of our consolidated *Deep-HOSeq* are optimized by back-propagation on the target loss function. The superiority of the combined information obtained with *Deep-HOSeq* is demonstrated by performing sentiment prediction on multiple

benchmark datasets where the proposed *Deep-HOSeq* outperformed state of the art and other baseline approaches. This enhancement in *Deep-HOSeq* is attributable to the expressiveness from all-types of information obtained by learning complementary information from the two sub-networks. Comprehensive experiments demonstrated the effectiveness of our proposed *Deep-HOSeq* for multimodal data fusion. In the future, we plan to reduce the computational complexity of the unique sub-network by designing factorized representations. Besides, generalizing *Deep-HOSeq* to perform multiple tasks is another interesting research direction for this work.

REFERENCES

- [1] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.
- [2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.
- [3] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intelligent Systems*, vol. 33, no. 6, pp. 17–25, 2018.
- [4] S. Verma, C. Wang, L. Zhu, and W. Liu, "Deepcu: Integrating both common and unique latent information for multimodal sentiment analysis," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 3627–3634.
- [5] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *32 AAAI Conference on Artificial Intelligence*, 2018.
- [6] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *32 AAAI Conference on Artificial Intelligence*, 2018.
- [7] D. Pruthi, M. Gupta, B. Dhingra, G. Neubig, and Z. C. Lipton, "Learning to deceive with attention-based explanations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 2020, pp. 4782–4793.
- [8] P. Michel, O. Levy, and G. Neubig, "Are sixteen heads really better than one?" in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., 2019, pp. 14 014–14 024.
- [9] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, "Deep multimodal multilinear fusion with high-order polynomial pooling," in *Advances in Neural Information Processing Systems*, 2019, pp. 12 113–12 122.
- [10] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 284–288.
- [11] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*. ACM, 2011, pp. 169–176.
- [12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [13] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.
- [14] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2247–2256.
- [15] X. He and T.-S. Chua, "Neural factorization machines for sparse predictive analytics," in *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2017, pp. 355–364.
- [16] C.-T. Lu, L. He, H. Ding, B. Cao, and P. S. Yu, "Learning from multi-view multi-way data via structural factorization machines," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2018, pp. 1593–1602.
- [17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [18] C. Huang, C. Zhang, J. Zhao, X. Wu, D. Yin, and N. Chawla, "Mist: A multiview and multimodal spatial-temporal learning framework for city-wide abnormal event forecasting," in *The World Wide Web Conference*, 2019, pp. 717–728.
- [19] X. Wu, B. Shi, Y. Dong, C. Huang, and N. V. Chawla, "Neural tensor factorization for temporal interaction learning," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 2019, pp. 537–545.
- [20] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.
- [21] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1)*, 2018, pp. 2236–2246.
- [22] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.
- [23] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [24] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Computer Vision and Pattern Recognition, 2006 IEEE Conference on*, vol. 2. IEEE, 2006, pp. 1491–1498.
- [25] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 960–964.
- [26] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1)*, 2013, pp. 973–982.
- [27] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016.
- [28] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.
- [29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [30] S. Ioffe and C. 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, pp. 448–456.