INTERPRETABLE FUSION MECHANISHMS FOR MULTIMODAL REPRESENTATION LEARNING

Akella Ravi Tej[*♣] *Hardik Chauhan*[*♣] *Arunav Pratap Shandeelya*[*♠]
Asif Ekbal[♥] *Pushpak Bhattacharyya*[♥]

[♥] Indian Institute of Technology Patna
[♣] Indian Institute of Technology Roorkee
[♠] International Institute of Information Technology Bhubaneswar
{arunavshandilya96, chauhanhardik23, ravitej.akella}@gmail.com, {asif, pb}@iitp.ac.in

ABSTRACT

Multimodal representation learning extends the paradigm of traditional representation learning to multiple input modalities. A crucial research problem in this field is to effectively model the fine connections between the unimodal representations. This is particularly challenging since it is infeasible to learn the complete multilinear interaction across all the input modalities. In this paper, we propose a *Multilinear Superdiagonal Fusion* strategy to approximate the multilinear fusion tensor with a block-superdiagonal tensor decomposition. This interpretable decomposition allows to better capture inter-modality dynamics while working with a tractable number of learnable parameters. Additionally, we introduce a parameter sharing procedure to further limit the number of parameters in the decomposition without considerably affecting the accuracy. We demonstrate the effectiveness of the proposed approach on two multimodal tasks, *viz.* speaker trait analysis and emotion recognition.

Index Terms— Multimodal representational learning, block-superdiagonal tensor decomposition, classification, emotion recognition, speaker trait analysis

1. BACKGROUND

The performance of a learning algorithm is significantly influenced by the features it employs to represent the input data. In several real-world tasks, this input data comes from multiple information sources that are complementary to each other. While each input modality only represents a partial observation of the input state, their effective fusion can potentially provide a more complete information of the input state. In simpler words, input modalities are like the pieces of a jigsaw puzzle in a way that the right combination of these pieces conveys more information than the individual pieces themselves.

An important open problem in multimodal representation learning is to design an effective fusion mechanism that captures the fine interactions between unimodal representations. The efficiency of multimodal fusion is heavily dependent on the inductive bias of the fusion layer. A natural extension of linearity in multimodal settings would be to model a multilinear relationship across the unimodal representations. However, the growth in the parameters of a multilinear tensor is exponential with the number of input modalities and strictly super-linear in input-dimensions. In [1], the authors limit the computation to a linear growth in input modalities and input dimensions by performing a low-rank (Candecomp/PARAFAC (CP)) decomposition [2, 3] of the multilinear fusion tensor. While this approach is computationally efficient, there is no clear interpretation of the constraining nature of the rank of a tensor on the multilinear interactions it can model. A similar issue has been addressed in BLOCK [4] for the bilinear fusion of two modalities in visual question answering (VQA). Further, this approach [4] provides an interpretable bimodal fusion strategy that allows to trade-off the expressivity and complexity of the fusion mechanism.

2. MAIN CONTRIBUTIONS

1. This paper introduces a reparameterizion trick that reduces the complexity of a full-rank multilinear fusion operation to a linear growth in the number of input modalities and a quadratic growth in input-dimensions.

* contributed equally as part of their internship at IIT Patna, India.

A short preview of our work. Full manuscript is under review for ICASSP 2020 conference track.

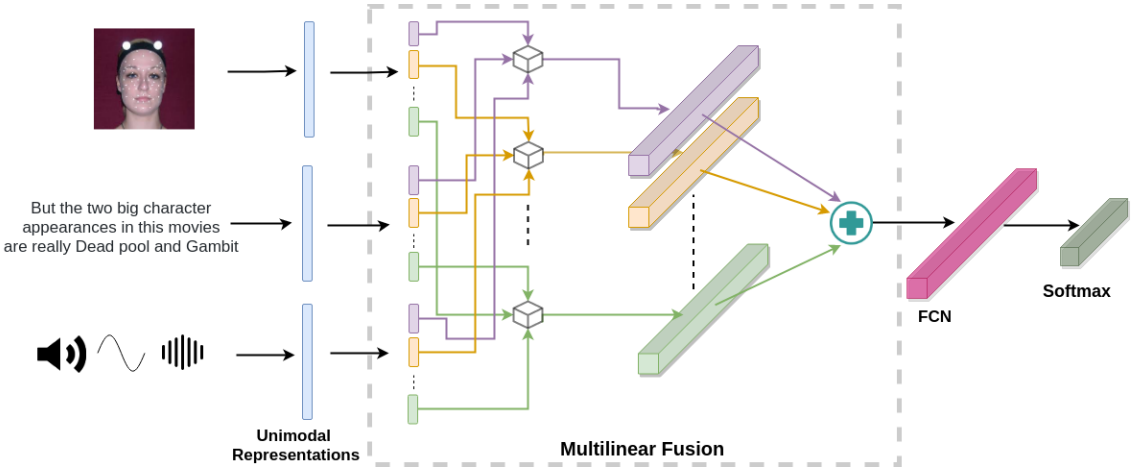

Fig. 1. The diagram illustrates the *Multilinear Superdiagonal Fusion* mechanism for an example with three input modalities.

2. We then impose a super-diagonal approximation that provides an low-dimensional full-rank fusion strategy with a linear growth rate. Further, we explain how this decomposition brings the best of Candecomp/PARAFAC (CP)) decomposition [2, 3] and Tucker decomposition [5].

3. We also highlight that applying a low-rank constraint on the blocks of a superdiagonal tensor, a common practice in tensor fusion literature, has a detrimental impact on capturing the inter-modality fusion dynamics.

4. To further restrict the parameters in the model, we propose an aggressive parameter sharing scheme that maintains the rank of the multilinear tensor and is supported by most automatic differentiation packages.

The overall architecture is illustrated in Figure 1.

Model		**POM**			**IEMOCAP**			
		MAE	***Corr***	***Acc***	***F1-Happy***	***F1-Sad***	***F1-Angry***	***F1-Neutral***
Non-Fusion	***SVM*** [6]	*0.887*	*0.104*	*33.9*	*81.5*	*78.8*	*82.4*	*64.9*
	DF [7]	*0.869*	*0.144*	*34.1*	*81*	*81.2*	*65.4*	*44.0*
	BC-LSTM [8]	*0.840*	*0.278*	*34.8*	*81.7*	*81.7*	*84.2*	*64.1*
	MV-LSTM [9]	*0.891*	*0.270*	*34.6*	*81.3*	*74.0*	*84.3*	*66.7*
	MARN [10]	-	-	-	*83.6*	*81.2*	*84.2*	*65.9*
	MFN [10]	*0.805*	*0.349*	*41.7*	*84*	*82.1*	*83.7*	*69.2*
Fusion	***TFN*** [8]	*0.886*	*0.093*	*31.6*	*83.6*	*82.8*	*84.2*	*65.4*
	LMF [1]	***0.796***	***0.396***	***42.8***	*85.8*	*85.9*	***89.0***	*71.7*
Proposed	***mBLOCK***	*0.845*	*0.29*	*36.0*	*85.4*	*85.8*	*88*	*71.4*
	mBLOCK$_{share}$	0.823	0.31	32.3	***85.9***	***86.2***	*88.2*	***72.6***

Table 1. Results on IEMOCAP & POM dataset based on *Multilinear Superdiagonal Fusion* mechanism

References

[1] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe 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)*, Melbourne, Australia, July 2018, pp. 2247–2256, Association for Computational Linguistics.

[2] J. Douglas Carroll and Jih-Jie Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, Sep 1970.

[3] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[4] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord, "Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," in *Proceedings of the 33st Conference on Artificial Intelligence (AAAI)*, 2019.

[5] Ledyard R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, Sep 1966.

[6] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep 1995.

[7] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, New York, NY, USA, 2016, ICMI '16, pp. 284–288, ACM.

[8] 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*, Copenhagen, Denmark, Sept. 2017, pp. 1103–1114, Association for Computational Linguistics.

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

[10] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Memory fusion network for multi-view sequential learning," in *AAAI*, 2018.