Learning Representations from Imperfect Time Series Data via Tensor Rank Regularization

Paul Pu Liang[♣*], **Zhun Liu**[♢*], **Yao-Hung Hubert Tsai**[♣]
Qibin Zhao[♡], **Ruslan Salakhutdinov**[♣], **Louis-Philippe Morency**[♢]
[♣]Machine Learning Department, Carnegie Mellon University, USA
[♢]Language Technologies Institute, Carnegie Mellon University, USA
[♡]Tensor Learning Unit, RIKEN Center for Artificial Intelligence Project, Japan
{pliang,zhunl,yaohungt,rsalakhu,morency}@cs.cmu.edu
qibin.zhao@riken.jp

Abstract

There has been an increased interest in multimodal language processing including multimodal dialog, question answering, sentiment analysis, and speech recognition. However, naturally occurring multimodal data is often *imperfect* as a result of imperfect modalities, missing entries or noise corruption. To address these concerns, we present a regularization method based on *tensor rank minimization*. Our method is based on the observation that high-dimensional multimodal time series data often exhibit correlations across time and modalities which leads to low-rank tensor representations. However, the presence of noise or incomplete values breaks these correlations and results in tensor representations of higher rank. We design a model to learn such tensor representations and effectively regularize their rank. Experiments on multimodal language data show that our model achieves good results across various levels of imperfection.

1 Introduction

Analyzing multimodal language sequences spans various fields including multimodal dialog (Das et al., 2017; Rudnicky, 2005), question answering (Antol et al., 2015; Tapaswi et al., 2015; Das et al., 2018), sentiment analysis (Morency et al., 2011), and speech recognition (Palaskar et al., 2018). Generally, these multimodal sequences contain heterogeneous sources of information across the language, visual and acoustic modalities. For example, when instructing robots, these machines have to comprehend our verbal instructions and interpret our nonverbal behaviors while grounding these inputs in their visual sensors (Schmerling et al., 2017; Iba et al., 2005). Likewise, comprehending human intentions requires integrating human language, speech, facial behaviors, and body postures (Mihalcea, 2012; Rossiter, 2011). However, as much as more modalities are required for improved performance, we now face a challenge of *imperfect* data where data might be 1) incomplete due to mismatched modalities or sensor failure, or 2) corrupted with random or structured noise. As a result, an important research question involves learning robust representations from imperfect multimodal data.

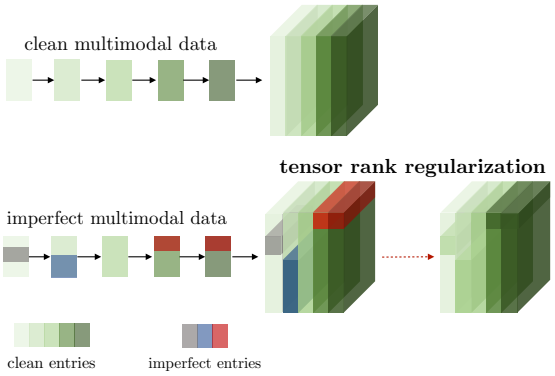


Figure 1: Clean multimodal time series data (in shades of green) exhibits correlations across time and across modalities, leading to redundancy in low rank tensor representations. On the other hand, the presence of imperfect entries (in gray, blue, and red) breaks these correlations and leads to higher rank tensors. In these scenarios, we use *tensor rank regularization* to learn tensors that more accurately represent the true correlations and latent structures in multimodal data.

Recent research in both unimodal and multimodal learning has investigated the use of tensors for representation learning (Anandkumar et al., 2014). Given representations $\mathbf{h}_1, ..., \mathbf{h}_M$ from M modalities, the order-M outer product tensor $\mathcal{T} = \mathbf{h}_1 \otimes \mathbf{h}_2 \otimes ... \otimes \mathbf{h}_M$ is a natural representation for all possible interactions between the modality dimensions (Liu et al., 2018). In this paper, we propose a model called the Temporal Tensor Fusion Network (T2FN) that builds tensor representations from multimodal time series data. T2FN

[*]first two authors contributed equally

learns a tensor representation that captures multimodal interactions across time. A key observation is that clean data exhibits tensors that are *low-rank* since high-dimensional real-world data is often generated from lower dimensional latent structures (Lakshmanan et al., 2015). Furthermore, clean multimodal time series data exhibits correlations across time and across modalities (Yang et al., 2017; Hidaka and Yu, 2010). This leads to redundancy in these overparametrized tensors which explains their low rank (Figure 1). On the other hand, the presence of noise or incomplete values breaks these natural correlations and leads to higher rank tensor representations. As a result, we can use *tensor rank minimization* to learn tensors that more accurately represent the true correlations and latent structures in multimodal data, thereby alleviating imperfection in the input. With these insights, we show how to integrate tensor rank minimization as a simple regularizer for training in the presence of imperfect data. As compared to previous work on imperfect data (Sohn et al., 2014; Srivastava and Salakhutdinov, 2014; Pham et al., 2019), our model does not need to know which of the entries or modalities are imperfect beforehand. Our model combines the strength of temporal non-linear transformations of multimodal data with a simple regularization technique on tensor structures. We perform experiments on multimodal video data consisting of humans expressing their opinions using a combination of language and nonverbal behaviors. Our results back up our intuitions that imperfect data increases tensor rank. Finally, we show that our model achieves good results across various levels of imperfection.

2 Related Work

Tensor Methods: Tensor representations have been used for learning discriminative representations in unimodal and multimodal tasks. Tensors are powerful because they can capture important higher order interactions across time, feature dimensions, and multiple modalities (Kossaifi et al., 2017). For unimodal tasks, tensors have been used for part-of-speech tagging (Srikumar and Manning, 2014), dependency parsing (Lei et al., 2014), word segmentation (Pei et al., 2014), question answering (Qiu and Huang, 2015), and machine translation (Setiawan et al., 2015). For multimodal tasks, Huang et al. (2017) used tensor products between images and text features for image captioning. A similar approach was proposed to learn representations across text, visual, and acoustic features to infer speaker sentiment (Liu et al., 2018; Zadeh et al., 2017). Other applications include multimodal machine translation (Delbrouck and Dupont, 2017), audio-visual speech recognition (Zhang et al., 2017), and video semantic analysis (Wu et al., 2009; Gao et al., 2009).

Imperfect Data: In order to account for imperfect data, several works have proposed generative approaches for multimodal data (Sohn et al., 2014; Srivastava and Salakhutdinov, 2014). Recently, neural models such as cascaded residual autoencoders (Tran et al., 2017), deep adversarial learning (Cai et al., 2018), or translation-based learning (Pham et al., 2019) have also been proposed. However, these methods often require knowing which of the entries or modalities are imperfect beforehand. While there has been some work on using low-rank tensor representations for imperfect data (Chang et al., 2017; Fan et al., 2017; Chen et al., 2017; Long et al., 2018; Nimishakavi et al., 2018), our approach is the first to integrate rank minimization with neural networks for multimodal language data, thereby combining the strength of non-linear transformations with the mathematical foundations of tensor structures.

3 Proposed Method

In this section, we present our method for learning representations from imperfect human language across the language, visual, and acoustic modalities. In §3.1, we discuss some background on tensor ranks. We outline our method for learning tensor representations via a model called Temporal Tensor Fusion Network (T2FN) in §3.2. In §3.3, we investigate the relationship between tensor rank and imperfect data. Finally, in §3.4, we show how to regularize our model using tensor rank minimization.

We use lowercase letters $x \in \mathbb{R}$ to denote scalars, boldface lowercase letters $\mathbf{x} \in \mathbb{R}^d$ to denote vectors, and boldface capital letters $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ to denote matrices. Tensors, which we denote by calligraphic letters $\mathcal{X}$, are generalizations of matrices to multidimensional arrays. An order-M tensor has M dimensions, $\mathcal{X} \in \mathbb{R}^{d_1 \times \ldots \times d_M}$. We use $\otimes$ to denote outer product between vectors.

3.1 Background: Tensor Rank

The rank of a tensor measures how many vectors are required to reconstruct the tensor. Simple tensors that can be represented as outer products of

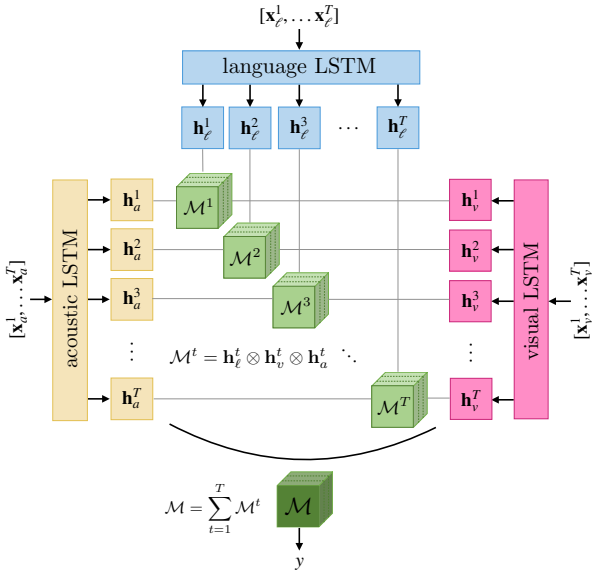


Figure 2: The Temporal Tensor Fusion Network (T2FN) creates a tensor $\mathcal{M}$ from temporal data. The rank of $\mathcal{M}$ increases with imperfection in data so we regularize our model by minimizing an upper bound on the rank of $\mathcal{M}$.

vectors have lower rank, while complex tensors have higher rank. To be more precise, we define the rank of a tensor using Canonical Polyadic (CP)-decomposition (Carroll and Chang, 1970). For an order-M tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times \ldots \times d_M}$, there exists an exact decomposition into vectors $\mathbf{w}$:

$$\mathcal{X} = \sum_{i=1}^{r} \bigotimes_{m=1}^{M} \mathbf{w}_m^i. \tag{1}$$

The minimal r for exact decomposition is called the *rank* of the tensor. The vectors $\{\{\mathbf{w}_m^i\}_{m=1}^{M}\}_{i=1}^{r}$ are called the rank r decomposition factors of $\mathcal{X}$.

3.2 Multimodal Tensor Representations

Our model for creating tensor representations is called the Temporal Tensor Fusion Network (T2FN), which extends the model in Zadeh et al. (2017) to include a temporal component. We show that T2FN increases the capacity of TFN to capture high-rank tensor representations, which itself leads to improved prediction performance. More importantly, our knowledge about tensor rank properties allows us to regularize our model effectively for imperfect data.

We begin with time series data from the language, visual and acoustic modalities, denoted as $[\mathbf{x}_\ell^1, ..., \mathbf{x}_\ell^T]$, $[\mathbf{x}_v^1, ..., \mathbf{x}_v^T]$, and $[\mathbf{x}_a^1, ..., \mathbf{x}_a^T]$ respectively. We first use Long Short-term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) to encode the temporal information from each modality, resulting in a sequence of hidden representations $[\mathbf{h}_\ell^1, ..., \mathbf{h}_\ell^T]$, $[\mathbf{h}_v^1, ..., \mathbf{h}_v^T]$, and $[\mathbf{h}_a^1, ..., \mathbf{h}_a^T]$. Similar to prior work which found tensor representations to capture higher-order interactions from multimodal data (Liu et al., 2018; Zadeh et al., 2017; Fukui et al., 2016), we form tensors via outer products of the individual representations through time (as shown in Figure 2):

$$\mathcal{M} = \sum_{t=1}^{T} \begin{bmatrix} \mathbf{h}_\ell^t \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{h}_v^t \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{h}_a^t \\ 1 \end{bmatrix} \tag{2}$$

where we append a 1 so that unimodal, bimodal, and trimodal interactions are all captured as described in Zadeh et al. (2017). $\mathcal{M}$ is our multimodal representation which can then be used to predict the label y using a fully connected layer. Observe how the construction of $\mathcal{M}$ closely resembles equation (1) as the sum of vector outer products. As compared to TFN which uses a single outer product to obtain a multimodal tensor of rank one, T2FN creates a tensor of high rank (upper bounded by T). As a result, the notion of *rank* naturally emerges when we reason about the properties of $\mathcal{M}$.

3.3 How Does Imperfection Affect Rank?

We first state several observations about the rank of multimodal representation $\mathcal{M}$:

1) r_{noisy}: The rank of $\mathcal{M}$ is maximized when data entries are sampled from i.i.d. noise (e.g. Gaussian distributions). This is because this setting leads to no redundancy at all between the feature dimensions across time steps.

2) $r_{clean} < r_{noisy}$: Clean real-world data is often generated from lower dimensional latent structures (Lakshmanan et al., 2015). Furthermore, multimodal time series data exhibits correlations across time and across modalities (Yang et al., 2017; Hidaka and Yu, 2010). This redundancy leads to low-rank tensor representations.

3) $r_{clean} < r_{imperfect} < r_{noisy}$: If the data is imperfect, the presence of noise or incomplete values breaks these natural correlations and leads to higher rank tensor representations.

These intuitions are also backed up by several experimental results which are presented in §4.2.

3.4 Tensor Rank Regularization

Given our intuitions above, it would then seem natural to augment the discriminative objective function with a term to minimize the rank of $\mathcal{M}$.

In practice, the rank of an order-M tensor is computed using the nuclear norm $\|\mathcal{X}\|_*$ which is defined as (Friedland and Lim, 2014),

$$\|\mathcal{X}\|_* = \inf\left\{\sum_{i=1}^{r}|\lambda_i| : \mathcal{X} = \sum_{i=1}^{r}\lambda_i\left(\bigotimes_{m=1}^{M}\mathbf{w}_m^i\right), \|\mathbf{w}_m^i\| = 1, r \in \mathbb{N}\right\}. \tag{3}$$

When $M = 2$, this reduces to the matrix nuclear norm (sum of singular values). However, computing the rank of a tensor or its nuclear norm is NP-hard for tensors of order ≥ 3 (Friedland and Lim, 2014). Fortunately, there exist efficiently computable upper bounds on the nuclear norm and minimizing these upper bounds would also minimize the nuclear norm $\|\mathcal{M}\|_*$. We choose the upper bound as presented in Hu (2014), which upper bounds the nuclear norm with the tensor Frobenius norm scaled by the tensor dimensions:

$$\|\mathcal{M}\|_* \leq \sqrt{\frac{\prod_{i=1}^{M} d_i}{\max\{d_1, ..., d_M\}}}\|\mathcal{M}\|_F, \tag{4}$$

where the Frobenius norm $\|\mathcal{M}\|_F$ is defined as the sum of squared entries in $\mathcal{M}$ which is easily computable and convex. Since $\|\mathcal{M}\|_F$ is easily computable and convex, including this term adds negligible computational cost to the model. We will use this upper bound as a surrogate for the nuclear norm in our objective function. Our objective function is therefore a weighted combination of the prediction loss and the tensor rank regularizer in equation (4).

4 Experiments

Our experiments are designed with two research questions in mind: 1) What is the effect of various levels of imperfect data on tensor rank in T2FN? 2) Does T2FN with rank regularization perform well on prediction with imperfect data? We answer these questions in §4.2 and §4.3 respectively.

4.1 Datasets

We experiment with real video data consisting of humans expressing their opinions using a combination of language and nonverbal behaviors. We use the CMU-MOSI dataset which contains 2199 videos annotated for sentiment in the range $[-3, +3]$ (Zadeh et al., 2016). CMU-MOSI and related multimodal language datasets have been studied in the NLP community (Gu et al., 2018; Liu et al., 2018; Liang et al., 2018) from fully supervised settings but not from the perspective of supervised learning with imperfect data. We use 52 segments for training, 10 for validation and 31 for testing. GloVe word embeddings (Pennington et al., 2014), Facet (iMotions, 2017), and COVAREP (Degottex et al., 2014) features are extracted for the language, visual and acoustic modalities respectively. Forced alignment is performed using P2FA (Yuan and Liberman, 2008) to align visual and acoustic features to each word, resulting in a multimodal sequence. Our data splits, features, alignment, and preprocessing steps are consistent with prior work on the CMU-MOSI dataset (Liu et al., 2018).

4.2 Rank Analysis

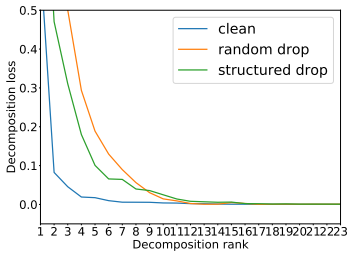
We first study the effect of imperfect data on the rank of tensor $\mathcal{M}$. We introduce the following types of noises parametrized by `noise_level` $= [0.0, 0.1, ..., 1.0]$. Higher noise levels implies more imperfection: 1) **clean**: no imperfection, 2) **random drop**: each entry is dropped independently with probability $p \in$ `noise_level`, and 3) **structured drop**: independently for each modality, each time step is chosen with probability $p \in$ `noise_level`. If a time step is chosen, all feature dimensions at that time step are dropped. For all imperfect settings, features are dropped during both training and testing.

We would like to show how the tensor ranks vary under different imperfection settings. However, as is mentioned above, determining the exact rank of a tensor is an NP-hard problem (Friedland and Lim, 2014). In order to analyze the effect of imperfections on tensor rank, we perform CP decomposition (equation (5)) on the tensor representations under different rank settings r and measure the reconstruction error ϵ,

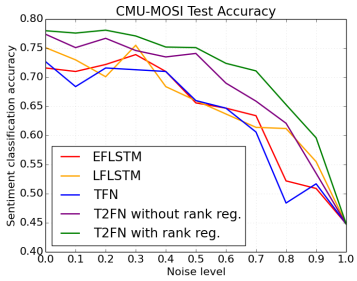
$$\epsilon = \min_{\mathbf{w}_m^i}\left\|\left(\sum_{i=1}^{r}\bigotimes_{m=1}^{M}\mathbf{w}_m^i\right) - \mathcal{X}\right\|_F. \tag{5}$$

Given the true rank r^*, ϵ will be high at ranks $r < r^*$, while ϵ will be approximately zero at ranks $r \geq r^*$ (for example, a rank 3 tensor would display a large reconstruction error with CP decomposition at rank 1, but would show almost zero error with CP decomposition at rank 3). By analyzing the effect of r on ϵ, we are then able to derive a surrogate $\tilde{r}$ to the true rank r^*.

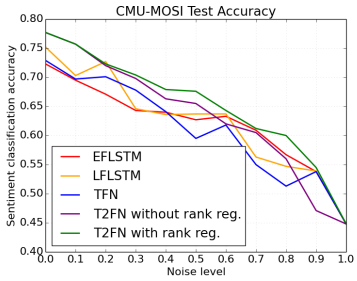
Using this approach, we experimented on CMU-MOSI and the results are shown in Figure 3(a). We observe that imperfection leads to an increase in (approximate) tensor rank as measured



(a) CP decomposition error of $\mathcal{M}$ under random and structured dropping of features. Imperfect data leads to an increase in decomposition error and an increase in (approximate) tensor rank.



(b) Sentiment classification accuracy under random drop (i.e. dropping entries randomly with probability $p \in$ `noise_level`). T2FN with rank regularization (green) performs well.



(c) Sentiment classification accuracy under structured drop (dropping entire time steps randomly with probability $p \in$ `noise_level`). T2FN with rank regularization (green) performs well.

Figure 3: (a) Effect of imperfect data on tensor rank. (b) and (c): CMU-MOSI test accuracy under imperfect data.

by reconstruction error (the graph shifts outwards and to the right), supporting our hypothesis that imperfect data increases tensor rank (§3.3).

4.3 Prediction Results

Our next experiment tests the ability of our model to learn robust representations despite data imperfections. We use the tensor $\mathcal{M}$ for prediction and report binary classification accuracy on CMU-MOSI test set. We compare to several baselines: Early Fusion (EF)-LSTM, Late Fusion (LF)-LSTM, TFN, and T2FN *without* rank regularization. These results are shown in Figure 3(b) for random drop and Figure 3(c) for structured drop. T2FN *with* rank regularization maintains good performance despite imperfections in data. We also observe that our model's improvement is more significant on random drop settings, which results in a higher tensor rank as compared to structured drop settings (from Figure 3(a)). This supports our hypothesis that our model learns robust representations when imperfections that increase tensor rank are introduced. On the other hand, the existing baselines suffer in the presence of imperfect data.

5 Discussion and Future Work

We acknowledge that there are other alternative methods to upper bound the true rank of a tensor (Alexeev et al., 2011; Atkinson and Lloyd, 1980; Ballico, 2014). From a theoretical perspective, there exists a trade-off between the cost of computation and the tightness of approximation. In addition, the tensor rank can (far) exceed the maximum dimension, and a low-rank approximation for tensors may not even exist (de Silva and Lim, 2008). While our tensor rank regularization method seems to work well empirically, there is definitely room for a more thorough theoretical analysis of constructing and regularizing tensor representations for multimodal learning.

6 Conclusion

This paper presented a regularization method based on *tensor rank minimization*. We observe that clean multimodal sequences often exhibit correlations across time and modalities which leads to low-rank tensors, while the presence of imperfect data breaks these correlations and results in tensors of higher rank. We designed a model, the Temporal Tensor Fusion Network, to learn such tensor representations and effectively regularize their rank. Experiments on multimodal language data show that our model achieves good results across various levels of imperfections. We hope to inspire future work on regularizing tensor representations of multimodal data for robust prediction in the presence of imperfect data.

Acknowledgements

PPL, ZL, and LM are partially supported by the National Science Foundation (Award #1750439 and #1722822) and Samsung. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of Samsung and NSF, and no official endorsement should be inferred. YHT and RS are supported in part by DARPA HR00111990016, AFRL FA8750-18-C-0014, NSF IIS1763562, Apple, and Google focused award. QZ is supported by JSPS KAKENHI (Grant No. 17K00326). We also acknowledge NVIDIA's GPU support and the anonymous reviewers for their constructive comments.

References

Boris Alexeev, Michael A. Forbes, and Jacob Tsimerman. 2011. Tensor rank: Some lower and upper bounds. *CoRR*, abs/1102.0072.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. 2014. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

M.D. Atkinson and S. Lloyd. 1980. Bounds on the ranks of some 3-tensors. *Linear Algebra and its Applications*, 31:19 – 31.

E. Ballico. 2014. An upper bound for the real tensor rank and the real symmetric tensor rank in terms of the complex ranks. *Linear and Multilinear Algebra*, 62(11):1546–1552.

Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. 2018. Deep adversarial learning for multi-modality missing data completion. In *KDD '18*, pages 1158–1166.

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

Yi Chang, Luxin Yan, Houzhang Fang, Sheng Zhong, and Zhijun Zhang. 2017. Weighted low-rank tensor recovery for hyperspectral image restoration. *CoRR*, abs/1709.00192.

Xiai Chen, Zhi Han, Yao Wang, Qian Zhao, Deyu Meng, Lin Lin, and Yandong Tang. 2017. A general model for robust tensor factorization with unknown noise. *CoRR*, abs/1705.06755.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep - a collaborative voice analysis repository for speech technologies. In *ICASSP*. IEEE.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. Multimodal compact bilinear pooling for multimodal neural machine translation. *CoRR*, abs/1703.08084.

Haiyan Fan, Yunjin Chen, Yulan Guo, Hongyan Zhang, and Gangyao Kuang. 2017. Hyperspectral image restoration using low-rank tensor recovery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP:1–16.

Shmuel Friedland and Lek-Heng Lim. 2014. Computational complexity of tensor nuclear norm. *CoRR*, abs/1410.6072.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Xinbo Gao, Yimin Yang, Dacheng Tao, and Xuelong Li. 2009. Discriminative optical flow tensor for video semantic analysis. *Computer Vision and Image Understanding*, 113(3):372 – 383. Special Issue on Video Analysis.

Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *ACL*.

Shohei Hidaka and Chen Yu. 2010. Analyzing multimodal time series as dynamical systems. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ICMI-MLMI '10, pages 53:1–53:8, New York, NY, USA. ACM.

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

Shenglong Hu. 2014. Relations of the Nuclear Norms of a Tensor and its Matrix Flattenings. *arXiv e-prints*, page arXiv:1412.2443.

Qiuyuan Huang, Paul Smolensky, Xiaodong He, Li Deng, and Dapeng Oliver Wu. 2017. Tensor product generation networks. *CoRR*, abs/1709.09118.

Soshi Iba, Christiaan J. J. Paredis, and Pradeep K. Khosla. 2005. Interactive multimodal robot programming. *The International Journal of Robotics Research*, 24(1):83–104.

iMotions. 2017. Facial expression analysis.

Jean Kossaifi, Zachary C. Lipton, Aran Khanna, Tommaso Furlanello, and Anima Anandkumar. 2017. Tensor regression networks. *CoRR*, abs/1707.08308.

Karthik Lakshmanan, Patrick T. Sadtler, Elizabeth C. Tyler-Kabara, Aaron P. Batista, and Byron M. Yu. 2015. Extracting low-dimensional latent structure from time series in the presence of delays. *Neural Computation*, 27:1825–1856.

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

Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. *EMNLP*.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*.

Zhen Long, Yipeng Liu, Longxi Chen, and Ce Zhu. 2018. Low rank tensor completion for multiway visual data. *CoRR*, abs/1805.03967.

Rada Mihalcea. 2012. Multimodal sentiment analysis. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 1–1, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Madhav Nimishakavi, Pratik Kumar Jawanpuria, and Bamdev Mishra. 2018. A dual framework for low-rank tensor completion. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5484–5495. Curran Associates, Inc.

Shruti Palaskar, Ramon Sanabria, and Florian Metze. 2018. End-to-end multimodal speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Wenzhe Pei, Tao Ge, and Baobao Chang. 2014. Max-margin tensor neural network for chinese word segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 293–303. Association for Computational Linguistics.

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

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabas Poczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. *AAAI*.

Xipeng Qiu and Xuanjing Huang. 2015. Convolutional neural tensor network architecture for community-based question answering. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 1305–1311. AAAI Press.

James Rossiter. 2011. Multimodal intent recognition for natural human-robotic interaction.

Alexander I. Rudnicky. 2005. *Multimodal Dialogue Systems*, pages 3–11. Springer Netherlands, Dordrecht.

Edward Schmerling, Karen Leung, Wolf Vollprecht, and Marco Pavone. 2017. Multimodal probabilistic model-based planning for human-robot interaction. *CoRR*, abs/1710.09483.

Hendra Setiawan, Zhongqiang Huang, Jacob Devlin, Thomas Lamar, Rabih Zbib, Richard Schwartz, and John Makhoul. 2015. Statistical machine translation features with multitask tensor networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 31–41. Association for Computational Linguistics.

Vin de Silva and Lek-Heng Lim. 2008. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM J. Matrix Anal. Appl.*, 30(3):1084–1127.

Kihyuk Sohn, Wenling Shang, and Honglak Lee. 2014. Improved multimodal deep learning with variation of information. In *NIPS*.

Vivek Srikumar and Christopher D Manning. 2014. Learning distributed representations for structured output prediction. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3266–3274. Curran Associates, Inc.

Nitish Srivastava and Ruslan Salakhutdinov. 2014. Multimodal learning with deep boltzmann machines. *JMLR*, 15.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Movieqa: Understanding stories in movies through question-answering. *CoRR*, abs/1512.02902.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In *CVPR*.

F. Wu, Y. Liu, and Y. Zhuang. 2009. Tensor-based transductive learning for multimodality video semantic concept detection. *IEEE Transactions on Multimedia*, 11(5):868–878.

Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C. Lee Giles. 2017. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, pages 1114–1125.

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

Qingchen Zhang, Laurence T. Yang, Xingang Liu, Zhikui Chen, and Peng Li. 2017. A tucker deep computation model for mobile multimedia feature learning. *ACM Trans. Multimedia Comput. Commun. Appl.*, 13(3s):39:1–39:18.