Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion

Paul Pu Liang[1], **Ziyin Liu**[2], **Amir Zadeh**[2], **Louis-Philippe Morency**[2]
[1]Machine Learning Department, [2]Language Technologies Institute
Carnegie Mellon University
{pliang,ziyinl,abagherz,morency}@cs.cmu.edu

Abstract

Computational modeling of human multimodal language is an emerging research application of spatiotemporal modeling spanning the language, visual and acoustic modalities. Comprehending multimodal language requires modeling not only the spatial interactions within each modality (intra-modal interactions) but more importantly the interactions between modalities (cross-modal interactions) from complex temporal data. We propose the Recurrent Multistage Fusion Network (RMFN) which decomposes the spatiotemporal fusion problem into multiple stages, each of them focused on a subset of multimodal signals for specialized, effective fusion. Spatial cross-modal interactions are modeled using this multistage fusion approach which builds upon intermediate representations of previous stages. Temporal and intra-modal interactions are modeled by integrating our proposed fusion approach with a system of recurrent neural networks. The RMFN displays state-of-the-art performance in modeling multimodal language across three tasks relating to multimodal sentiment analysis, emotion recognition, and speaker traits recognition. Experiments show that each stage of fusion focuses on a different subset of multimodal signals and learns increasingly discriminative representations.

1 Introduction

Computational modeling of human multimodal language is an emerging research application of spatiotemporal modeling. This area focuses on modeling tasks such as multimodal sentiment analysis [41], emotion recognition [9], and personality traits recognition [46]. The multimodal temporal signals include the language (spoken words), visual (facial expressions, gestures), and acoustic modalities (prosody, vocal expressions). At its core, these signals are highly structured with two prime forms of spatial interactions: intra-modal and cross-modal interactions [52]. Intra-modal interactions refer to information within a specific modality. For example, the arrangement of words in a sentence [15] or the sequence of facial muscle activations for a frown. Cross-modal interactions refer to interactions between modalities. For example, the simultaneous co-occurrence of a smile with a positive sentence or the delayed occurrence of a laughter after the end of a sentence. Modeling these spatiotemporal interactions lies at the heart of multimodal language analysis and has recently become a centric research direction in multimodal machine learning [38, 48, 11, 59, 23, 25, 31, 62, 58, 43].

Recent advances in cognitive neuroscience have demonstrated the existence of multistage aggregation across human cortical networks and functions [61], particularly during the integration of multisensory spatiotemporal information [45]. At later stages of cognitive processing, higher level semantic meaning is extracted from phrases, facial expressions, and tone of voice, eventually leading to the formation of higher level cross-modal concepts [45, 61]. Inspired by these discoveries, we hypothesize that the computational modeling of cross-modal interactions also requires a *multistage fusion* process. In this process, cross-modal representations can build upon the representations learned during earlier stages. This decreases the burden on each stage of spatiotemporal fusion and allows each stage of fusion to be performed in a more specialized and effective manner.

Workshop on Modeling and Decision-Making in the Spatiotemporal Domain, 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

In this paper, we propose the Recurrent Multistage Fusion Network (RMFN) which decomposes the spatiotemporal fusion problem into multiple stages. At each stage, a subset of multimodal signals is highlighted and fused with previous fusion representations (Figure 1). This divide-and-conquer approach decreases the burden on each fusion stage, allowing each stage to be performed in a more specialized and effective way. In contrast, conventional fusion approaches model interactions over multimodal signals in one step [6]. Temporal and intra-modal interactions are modeled by integrating our new multistage fusion process with a system of recurrent neural networks. Overall, RMFN jointly models intra-modal and cross-modal interactions for spatiotemporal fusion. RMFN achieves state-of-the-art performance on three tasks related to multimodal language: sentiment analysis, emotion recognition, and speaker traits recognition. Through a comprehensive set of ablation experiments and visualizations, we demonstrate the advantages of defining multiple stages for spatiotemporal fusion.

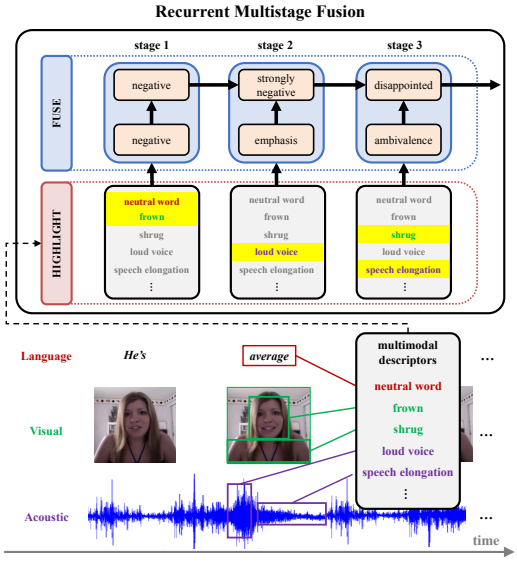

Figure 1: An illustrative example for Recurrent Multistage Fusion. At each stage, a subset of multimodal signals is highlighted and then fused with previous fusion representations. The first fusion stage selects the neutral word and frowning behaviors which create an intermediate representation reflecting negative emotion when fused together. The second stage selects the loud voice behavior which is locally interpreted as emphasis before being fused with previous stages into a strongly negative representation. The third stage selects the shrugging and speech elongation behaviors that reflect ambivalence and when fused with previous stages is interpreted as a representation for the disappointed emotion.

2 Related Work

Previous approaches in spatiotemporal modeling for multimodal language can be categorized as:

Non-temporal Models simplify the problem by averaging temporal information through time and using supervised learning methods [29, 12, 1, 44, 69, 41]. These approaches have trouble modeling long sequences since the average statistics do not accurately reflect temporal dynamics [64].

Temporal Graphical Models such as Hidden Markov Models [7], Conditional Random Fields (CRFs) [34], and Hidden Conditional Random Fields (HCRFs) [51] were shown to work well on modeling spatiotemporal data [40, 39, 27, 65]. Multimodal extensions have been proposed including multi-view HCRFs [56], multi-layered CRFs [56], and multi-view hierarchical models [57].

Temporal Neural Networks, such as Recurrent Neural Networks [20, 30] and Long-short Term Memory (LSTM) networks [26] have been used for spatiotemporal modeling [71, 55, 54, 21, 10, 24, 35, 52, 11]. Recently, more advanced models were proposed that use Bayesian ranking algorithms [37], external memory mechanisms [67], or low-rank tensors [38] for spatiotemporal fusion. These methods assume that fusion should be performed all at once rather than across multiple stages. Our empirical evaluations show the advantages of our spatiotemporal fusion approach.

3 Recurrent Multistage Fusion Network

In this section we describe the Recurrent Multistage Fusion Network (RMFN) for spatiotemporal fusion (Figure 2). Given a set of modalities $\{l(anguage), v(isual), a(coustic)\}$, each modality $m \in \{l, v, a\}$ is represented as a temporal sequence $\mathbf{X}^m = \{\mathbf{x}_1^m, \mathbf{x}_2^m, \mathbf{x}_3^m, \cdots, \mathbf{x}_T^m\}$, where $\mathbf{x}_t^m$ is the input at time t. Each sequence $\mathbf{X}^m$ is modeled with an intra-modal recurrent neural network. At time t, each recurrent network will output a unimodal representation $\mathbf{h}_t^m$. The Multistage Fusion Process uses a multistage approach to fuse all unimodal representations $\mathbf{h}_t^m$ into a cross-modal representation $\mathbf{z}_t$ which is then fed back into each intra-modal recurrent network.

Multistage Fusion Process (MFP) is a modular neural approach that performs multistage fusion to model cross-modal interactions. MFP has three modules: `HIGHLIGHT`, `FUSE` and `SUMMARIZE`. At each stage, `HIGHLIGHT` identifies a subset of multimodal signals from $[\mathbf{h}_t^l, \mathbf{h}_t^v, \mathbf{h}_t^a]$ that will be used for that stage of fusion. `FUSE` then performs two subtasks simultaneously: a local fusion of the highlighted features and integration with representations from previous stages. Both `HIGHLIGHT` and `FUSE` are realized using memory-based networks which enable coherence between stages and

storage of previously modeled representations. After all stages, SUMMARIZE takes the representation of the final stage and translates it into a cross-modal representation $\mathbf{z}_t$.

We now present the details of the three modules: HIGHLIGHT, FUSE and SUMMARIZE. Multistage fusion begins with the concatenation of intra-modal network outputs $\mathbf{h}_t = \bigoplus_{m \in M} \mathbf{h}_t^m$. We use superscript $^{[k]}$ to denote the indices of each stage $k = 1, \cdots, K$ during K total stages of multistage fusion. Θ denotes the network parameters across all modules.

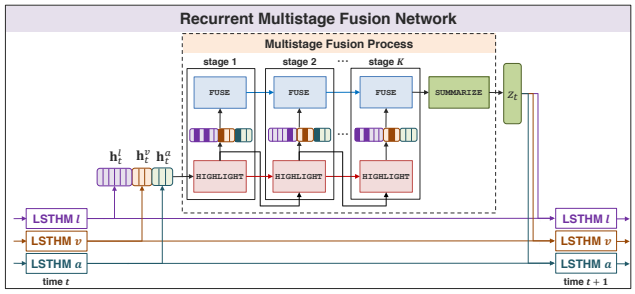

Figure 2: The Recurrent Multistage Fusion Network for spatiotemporal fusion. Multistage fusion begins with the concatenated intra-modal network outputs $\mathbf{h}_t^l, \mathbf{h}_t^v, \mathbf{h}_t^a$. At each stage, the HIGHLIGHT module identifies a subset of multimodal signals and the FUSE module performs local fusion before integration with previous fusion representations. The SUMMARIZE module translates the representation at the final stage into a cross-modal representation $\mathbf{z}_t$ to be fed into the intra-modal recurrent networks for temporal modeling.

HIGHLIGHT: At each stage k, a subset of the multimodal signals $\mathbf{h}_t$ will be highlighted for fusion. This module is defined by the process function f_H: $\mathbf{a}_t^{[k]} = f_H(\mathbf{h}_t \; ; \; \mathbf{a}_t^{[1:k-1]}, \Theta)$ where at stage k, $\mathbf{a}_t^{[k]}$ is a set of attention weights which are inferred based on the previously assigned attention weights $\mathbf{a}_t^{[1:k-1]}$. As a result, the highlights at a specific stage k will be dependent on previous highlights. To fully encapsulate these dependencies, the attention assignment process is performed in a recurrent manner using a LSTM which we call the HIGHLIGHT LSTM. The initial HIGHLIGHT LSTM memory at stage 0, $\mathbf{c}_t^{\texttt{HIGHLIGHT}[0]}$, is initialized using a network $\mathcal{M}$ that maps $\mathbf{h}_t$ into LSTM memory space $\mathbf{c}_t^{\texttt{HIGHLIGHT}[0]} = \mathcal{M}(\mathbf{h}_t \; ; \; \Theta)$. This allows the memory of the HIGHLIGHT LSTM to dynamically adjust to the intra-modal representations $\mathbf{h}_t$. The output of the HIGHLIGHT LSTM $\mathbf{h}_t^{\texttt{HIGHLIGHT}[k]}$ is softmax activated to produce attention weights $\mathbf{a}_t^{[k]}$ at every stage k of the multistage fusion process: $\mathbf{a}_t^{[k]}{}_j = \exp\left(\mathbf{h}_t^{\texttt{HIGHLIGHT}[k]}{}_j\right)/Z$, $Z = \sum_{d=1}^{|\mathbf{h}_t^{\texttt{HIGHLIGHT}[k]}|} \exp\left(\mathbf{h}_t^{\texttt{HIGHLIGHT}[k]}{}_d\right)$ and $\mathbf{a}_t^{[k]}$ is fed as input into the HIGHLIGHT LSTM at stage $k+1$. Therefore, the HIGHLIGHT LSTM functions as a decoder LSTM [60, 14] in order to capture the dependencies on previous attention assignments. Highlighting is performed by $\tilde{\mathbf{h}}_t^{[k]} = \mathbf{h}_t \odot \mathbf{a}_t^{[k]}$, where $\odot$ denotes the Hadamard product and $\tilde{\mathbf{h}}_t^{[k]}$ are the attended multimodal signals that will be used for the fusion at stage k.

FUSE: The highlighted multimodal signals are simultaneously fused in a local fusion and then integrated with fusion representations from previous stages. This module is defined by the process function f_F: $\mathbf{s}_t^{[k]} = f_F(\tilde{\mathbf{h}}_t^{[k]} \; ; \; \mathbf{s}_t^{[1:k-1]}, \Theta)$ where $\mathbf{s}_t^{[k]}$ denotes the integrated fusion representations at stage k. We employ a FUSE LSTM to simultaneously perform the local fusion and the integration with previous fusion representations. The FUSE LSTM input gate enables a local fusion while the FUSE LSTM forget and output gates enable integration with previous fusion results. The initial FUSE LSTM memory at stage 0, $\mathbf{c}_t^{\texttt{FUSE}[0]}$, is initialized using random orthogonal matrices [5, 36].

SUMMARIZE: After completing K stages, SUMMARIZE generates a cross-modal representation using all fusion representations $\mathbf{s}_t^{[1:K]}$. This operation is defined as: $\mathbf{z}_t = \mathcal{S}(\mathbf{s}_t^{[1:K]} \; ; \; \Theta)$ where $\mathbf{z}_t$ is the final output of the multistage fusion process and represents all cross-modal interactions discovered at time t. The summarized representation is fed into the intra-modal recurrent networks.

Intra-model Recurrent Networks: To integrate $\mathbf{z}_t$ with the temporal intra-modal representations, we employ a system of Long Short-term Hybrid Memories (LSTHMs) [68]. The LSTHM extends the LSTM formulation to include $\mathbf{z}_t$ in a hybrid memory component. The hybrid memory contains both intra-modal interactions from individual modalities $\mathbf{x}_t^m$ as well as the cross-modal interactions captured in $\mathbf{z}_t$. Multimodal prediction is performed using a representation $\mathcal{E}$ which integrates (1) the last outputs from the LSTHMs and (2) the last cross-modal representation $\mathbf{z}_T$. $\mathcal{E}$ is defined as $\mathcal{E} = (\bigoplus_{m \in M} \mathbf{h}_T^m) \oplus \mathbf{z}_T$ where $\oplus$ denotes vector concatenation. $\mathcal{E}$ summarizes all intra-modal and cross-modal representations from multimodal spatiotemporal data. RMFN is differentiable end-to-end which allows network parameters Θ to be learned using gradient descent approaches.

Dataset	CMU-MOSI					IEMOCAP Emotions							
Task	Sentiment					Happy		Sad		Angry		Neutral	
Metric	A2 ↑	F1 ↑	A7 ↑	MAE ↓	Corr ↑	A2 ↑	F1 ↑	A2 ↑	F1 ↑	A2 ↑	F1 ↑	A2 ↑	F1 ↑
SOTA3	76.5$^{\diamond}$	74.5†	33.2$^{\#}$	0.968§	0.622$^{\flat}$	86.1$^{\times}$	83.6§	83.2$^{\bullet}$	81.7$^{\bullet}$	85.0*	84.2§	68.2$^{\flat}$	66.7$^{\#}$
SOTA2	77.1§	77.0§	34.1*	0.965*	0.625§	86.5*	84.0*	83.4†	82.1*	85.1$^{\#}$	84.3$^{\#}$	68.8$^{\flat}$	68.5$^{\flat}$
SOTA1	77.4*	77.3*	34.7§	0.955$^{\diamond}$	0.632*	86.7§	84.2$^{\flat}$	83.5*	82.8†	**85.2**$^{\flat}$	84.5$^{\flat}$	**69.6***	**69.2***
RMFN	**78.4**	**78.0**	**38.3**	**0.922**	**0.681**	**87.5**	**85.8**	**83.8**	**82.9**	85.1	**84.6**	69.5	69.1

Table 1: Results on CMU-MOSI and IEMOCAP. Best results in bold. Symbols denote baseline model which achieves the reported performance: MFN: $\star$, MARN: §, GME-LSTM(A): $\diamond$, BC-LSTM: $\bullet$, TFN: †, MV-LSTM: #, EF-LSTM: $\flat$, SVM: $\times$. RMFN achieves state-of-the-art or competitive performance for all metrics.

4 Results and Discussion

Experimental Setup: To evaluate the performance of RMFN, three domains of multimodal language were selected: multimodal sentiment analysis on **CMU-MOSI** [69], emotion recognition on **IEMOCAP** [9], and speaker traits recognition on **POM** [46]. All datasets consist of monologue videos. GloVe word embeddings [47], Facet [28] and COVAREP [17] are extracted for the language, visual and acoustic modalities respectively[1]. For classification, we report accuracy Ac across c classes and F1 score. For regression, we report Mean Absolute Error (MAE) and Pearson's correlation (Corr).

Performance: Table 1 shows results on CMU-MOSI and IEMOCAP[2]. We achieve state-of-the-art or competitive results for all metrics, highlighting RMFN's capability in spatiotemporal fusion.

Dataset	CMU-MOSI Sentiment				
Metric	A2 ↑	F1 ↑	A7 ↑	MAE ↓	Corr ↑
RMFN-R1	75.5	75.5	35.1	0.997	0.653
RMFN-R2	76.4	76.4	34.5	0.967	0.642
RMFN-R3	**78.4**	**78.0**	**38.3**	**0.922**	**0.681**
RMFN-R4	76.0	76.0	36.0	0.999	0.640

Dataset	CMU-MOSI Sentiment				
Metric	A2 ↑	F1 ↑	A7 ↑	MAE ↓	Corr ↑
MARN	77.1	77.0	34.7	0.968	0.625
RMFN (no MFP)	76.5	76.5	30.8	0.998	0.582
RMFN (no HIGHLIGHT)	77.9	77.9	35.9	0.952	0.666
RMFN	**78.4**	**78.0**	**38.3**	**0.922**	**0.681**

Table 2: Left: Effect of varying the number of stages on CMU-MOSI performance. Multistage fusion improves performance as compared to single stage fusion. Right: Comparison studies of RMFN on CMU-MOSI. Modeling cross-modal interactions using multistage fusion and attention weights are crucial for spatiotemporal fusion.

Analysis: To achieve a deeper understanding of the multistage fusion process, we study four research questions. (Q1): the effect of the number of stages K during multistage fusion on performance. (Q2): the comparison between multistage and independent modeling of cross-modal interactions. (Q3): whether modeling cross-modal interactions are helpful. (Q4): whether attention weights from the HIGHLIGHT module are required for modeling cross-modal interactions.

Q1: We test the baseline RMFN-RK which performs fusion K stages of fusion. From Table 2, we observe that RMFN-R1 (single fusion stage) underperforms as compared to RMFN which performs multistage fusion, and increasing the number of stages K increases the model's capability to model cross-modal interactions up to a certain point (K = 3) in our experiments. Further increases led to decreases in performance and we hypothesize this is due to overfitting on the dataset.

Q2: We pay close attention to the performance comparison with respect to MARN which models multiple cross-modal interactions all at once (see Table 2). RMFN shows improved performance, indicating that multistage fusion is both effective and efficient for spatiotemporal modeling.

Q3: RMFN (no MFP) represents a system of LSTHMs without the integration of $\mathbf{z}_t$ from the MFP to model cross-modal interactions. From Table 2, RMFN (no MFP) is outperformed by RMFN, confirming that modeling cross-modal interactions is crucial for spatiotemporal fusion.

Q4: RMFN (no HIGHLIGHT) removes the HIGHLIGHT module from MFP during multistage fusion. From Table 2, RMFN (no HIGHLIGHT) underperforms, indicating that highlighting multimodal representations using attention weights are important for modeling cross-modal interactions.

5 Conclusion

In conclusion, this paper proposed the Recurrent Multistage Fusion Network (RMFN) which decomposes the spatiotemporal fusion problem into multiple stages, each focused on a subset of multimodal signals. Extensive experiments across three spatiotemporal datasets reveal that RMFN is highly effective in modeling multimodal language. Our visualizations also reveal that the stages coordinate to capture both synchronous and asynchronous spatiotemporal interactions.

[1]Details on datasets, feature extraction and baseline models are in supplementary.

[2]State-of-the-art (SOTA)1/2/3 represent the three best performing baseline models on each dataset. Results for POM, individual baseline models, and visualizations of the trained model are in supplementary.

References

[1] Harika Abburi, Rajendra Prasath, Manish Shrivastava, and Suryakanth V Gangashetty. Multimodal sentiment analysis using deep neural networks. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 58–65. Springer, 2016.

[2] Paavo Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication*, 11(2-3):109–118, 1992.

[3] Paavo Alku, Tom Bäckström, and Erkki Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *the Journal of the Acoustical Society of America*, 2002.

[4] Paavo Alku, Helmer Strik, and Erkki Vilkman. Parabolic spectral parameterâa new method for quantification of the glottal flow. *Speech Communication*, 22(1):67–79, 1997.

[5] Martín Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. *CoRR*, abs/1511.06464, 2015.

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *arXiv preprint arXiv:1705.09406*, 2017.

[7] Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

[8] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[9] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 2008.

[10] E. Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107, Mar 2016.

[11] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI, 2017.

[12] Xilun Chen, Ben Athiwaratkun, Yu Sun, Kilian Q. Weinberger, and Claire Cardie. Adversarial deep averaging networks for cross-lingual sentiment classification. *CoRR*, abs/1606.01614, 2016.

[13] Donald G Childers and CK Lee. Vocal quality factors: Analysis, synthesis, and perception. *the Journal of the Acoustical Society of America*, 90(5):2394–2410, 1991.

[14] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014.

[15] Noam Chomsky. *Syntactic Structures*. Mouton and Co., The Hague, 1957.

[16] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[17] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarepâa collaborative voice analysis repository for speech technologies. In *ICASSP*. IEEE, 2014.

[18] Thomas Drugman and Abeer Alwan. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Interspeech*, pages 1973–1976, 2011.

[19] Thomas Drugman, Mark Thomas, Jon Gudnason, Patrick Naylor, and Thierry Dutoit. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):994–1006, 2012.

[20] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[21] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 2011.

[22] A. Graves, A. r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*, pages 6645–6649, May 2013.

[23] Abhinav Gupta, Yajie Miao, Leonardo Neves, and Florian Metze. Visual features for context-aware speech recognition. *CoRR*, abs/1712.00489, 2017.

[24] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. September 2014.

[25] David F. Harwath and James R. Glass. Learning word-like units from joint audio-visual analysis. *CoRR*, abs/1701.07481, 2017.

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

[27] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *CoRR*, abs/1508.01991, 2015.

[28] iMotions. Facial expression analysis, 2017.

[29] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *ACL*, 2015.

[30] L. C. Jain and L. R. Medsker. *Recurrent Neural Networks: Design and Applications*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition, 1999.

[31] Herman Kamper, Shane Settle, Gregory Shakhnarovich, and Karen Livescu. Visually grounded learning of keyword prediction from untranscribed speech. *CoRR*, abs/1703.08136, 2017.

[32] John Kane and Christer Gobl. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013.

[33] Patricia K. Kuhl. A new view of language acquisition. *Proceedings of the National Academy of Sciences*, 97(22):11850–11857, 2000.

[34] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, 2001.

[35] Egor Lakomkin, Cornelius Weber, Sven Magg, and Stefan Wermter. Reusing neural speech representations for auditory emotion recognition. *CoRR*, abs/1803.11508, 2018.

[36] Quoc V. Le, Navdeep Jaitly, and Geoffrey E. Hinton. A simple way to initialize recurrent networks of rectified linear units. *CoRR*, abs/1504.00941, 2015.

[37] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Multimodal local-global ranking fusion for emotion recognition. ICMI, 2018.

[38] 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 *ACL*, 2018.

[39] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf, 2016. ACL 2016.

[40] Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. Character-based bidirectional lstm-crf with words and characters for japanese named entity recognition. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*. ACL, 2017.

[41] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *ICMI*. ACM, 2011.

[42] Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *CVPR*. IEEE, 2007.

[43] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In Lise Getoor and Tobias Scheffer, editors, *ICML*, pages 689–696. Omnipress, 2011.

[44] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. Deep multimodal fusion for persuasiveness prediction. In *ICMI*, New York, NY, USA, 2016. ACM.

[45] German I. Parisi, Jun Tani, Cornelius Weber, and Stefan Wermter. Emergence of multimodal action representations from neural network self-organization. *Cognitive Systems Research*, 43:208 – 221, 2017.

[46] Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 50–57, New York, NY, USA, 2014. ACM.

[47] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[48] Hai Pham, Thomas Manzini, Paul Pu Liang, and Barnabas Poczos. Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*. ACL, 2018.

[49] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *EMNLP*, 2015.

[50] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *ACL*, 2017.

[51] Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell. Hidden conditional random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007.

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

[53] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681, November 1997.

[54] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013.

[55] Hagen Soltau, Hank Liao, and Hasim Sak. Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition. *CoRR*, abs/1610.09975, 2016.

[56] Yale Song, Louis-Philippe Morency, and Randall Davis. Multi-view latent variable discriminative models for action recognition. In *CVPR*. IEEE, 2012.

[57] Yale Song, Louis-Philippe Morency, and Randall Davis. Action recognition by hierarchical sequence summarization. In *CVPR*, 2013.

[58] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*. 2012.

[59] F. Sun, D. Harwath, and J. Glass. Look, listen, and decode: Multimodal speech recognition with images. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016.

[60] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014.

[61] P. Taylor, J. N. Hobbs, J. Burroni, and H. T. Siegelmann. The global landscape of cognition: hierarchical aggregation as an organizational principle of human cortical networks and functions. *Scientific Reports*, 5:18112 EP –, 12 2015.

[62] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

[63] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis. *arXiv preprint arXiv:1609.05244*, 2016.

[64] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

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

[66] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, 2017.

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

[68] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. *AAAI*, 2018.

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

[70] Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, and Shai Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *CVPR*. IEEE, 2006.

[71] Julian Georg Zilly, Rupesh Kumar Srivastava, Jan Koutník, and Jürgen Schmidhuber. Recurrent Highway Networks. *arXiv preprint arXiv:1607.03474*, 2016.

A Experiment Details

A .1 Multimodal Tasks and Datasets

Multimodal Sentiment Analysis involves analyzing speaker sentiment based on video content. Multimodal sentiment analysis extends conventional language-based sentiment analysis to a multimodal setup where both verbal and non-verbal signals contribute to the expression of sentiment. We use **CMU-MOSI** [69] which consists of 2199 opinion segments from online videos each annotated with sentiment in the range [-3,3].

Multimodal Emotion Recognition involves identifying speaker emotions based on both verbal and nonverbal behaviors. We perform experiments on the **IEMOCAP** dataset [9] which consists of 7318 segments of recorded dyadic dialogues annotated for the presence of human emotions happiness, sadness, anger and neutral.

Multimodal Speaker Traits Recognition involves recognizing speaker traits based on multimodal communicative behaviors. **POM** [46] contains 903 movie review videos each annotated for 12 speaker traits: confident (con), passionate (pas), voice pleasant (voi), credible (cre), vivid (viv), expertise (exp), reserved (res), trusting (tru), relaxed (rel), thorough (tho), nervous (ner), persuasive (per) and humorous (hum).

A .2 Multimodal Features

Here we present extra details on feature extraction for the language, visual and acoustic modalities.

Language: We used 300 dimensional Glove word embeddings trained on 840 billion tokens from the common crawl dataset [47]. These word embeddings were used to embed a sequence of individual words from video segment transcripts into a sequence of word vectors that represent spoken text.

Visual: The library Facet [28] is used to extract a set of visual features including facial action units, facial landmarks, head pose, gaze tracking and HOG features [70]. These visual features are extracted from the full video segment at 30Hz to form a sequence of facial gesture measures throughout time.

Acoustic: The software COVAREP [17] is used to extract acoustic features including 12 Mel-frequency cepstral coefficients, pitch tracking and voiced/unvoiced segmenting features [18], glottal source parameters [13, 19, 2, 4, 3], peak slope parameters and maxima dispersion quotients [32]. These visual features are extracted from the full audio clip of each segment at 100Hz to form a sequence that represent variations in tone of voice over an audio segment.

A .3 Multimodal Alignment

We perform forced alignment using P2FA [65] to obtain the exact utterance time-stamp of each word. This allows us to align the three modalities together. Since words are considered the basic units of language we use the interval duration of each word utterance as one time-step. We acquire the aligned video and audio features by computing the expectation of their modality feature values over the word utterance time interval [67].

A .4 Baseline Models

We compare to the following models for multimodal machine learning: **MFN** [67] synchronizes multimodal sequences using a multi-view gated memory. It is the current state of the art on CMU-MOSI and POM. **MARN** [68] models intra-modal and cross-modal interactions using multiple attention coefficients and hybrid LSTM memory components. **GME-LSTM(A)** [11] learns binary gating mechanisms to remove noisy modalities that are contradictory or redundant for prediction. **TFN** [66] models unimodal, bimodal and trimodal interactions using tensor products. **BC-LSTM** [50] performs context-dependent sentiment analysis and emotion recognition, currently state of the art on IEMOCAP. **EF-LSTM** concatenates the multimodal inputs and uses that as input to a single LSTM [26]. We also implement the Stacked, (**EF-SLSTM**) [22] Bidirectional (**EF-BLSTM**) [53] and Stacked Bidirectional (**EF-SBLSTM**) LSTMs. The best model is reported as **EF-**$(\star)$**LSTM**. **EF-HCRF:** (Hidden Conditional Random Field) [51] uses a HCRF to learn latent variables conditioned on the concatenated input. We also implement the following variations: **EF-LDHCRF:** (Latent Discriminative HCRFs) [42], **MV-HCRF** (Multi-view HCRF) [56], **MV-LDHCRF**, **EF-HSSHCRF** (Hierarchical Sequence Summarization HCRF) [57] and **MV-HSSHCRF**. The best performing early fusion model is reported as **EF-**$(\star)$**HCRF** while the best multi-view model is reported as **MV-**$(\star)$**HCRF**. For descriptions of the remaining baselines, we refer the reader to **EF-HCRF** [51], **EF/MV-LDHCRF** [42], **MV-HCRF** [56], **EF/MV-HSSHCRF** [57], **MV-LSTM** [52], **DF** [44], **SAL-CNN** [63], **C-MKL** [49], **THMM** [41], **SVM** [16, 46] and **RF** [8].

B Additional Results

Here we record the complete set of results for all the baseline models across all the datasets, tasks and metrics. Table 1 summarizes the complete results for sentiment analysis on the CMU-MOSI dataset. Table 2 presents the complete results for emotion recognition on the IEMOCAP dataset and Table 3 presents the complete results for personality traits prediction on the POM dataset. For experiments on the POM dataset we report additional results on MAE and correlation metrics for personality traits regression. We achieve significant improvement over state-of-the-art multi-view and dataset specific approaches across all these datasets, highlighting the RMFN's capability in analyzing sentiment, emotions and speaker traits from human multimodal language.

C Visualizations

Using an attention assignment mechanism during the `HIGHLIGHT` process gives interpretability to the model since it allows us to visualize the attended multimodal signals at each stage and time step (see Figure 3). Using RMFN trained on the CMU-MOSI dataset, we plot the attention weights across the multistage fusion process for three videos in CMU-MOSI. Based on these visualizations we first draw the following general observations on spatiotemporal fusion:

Across stages (Spatio): Attention weights change their behaviors across the multiple stages of fusion. Some features are highlighted by earlier stages while other features are used in later stages. This supports our hypothesis that RMFN learns to specialize in different stages of the spatiotemporal fusion process.

Across time (Temporal): Attention weights vary over time and adapt to the multimodal inputs. We observe that the attention weights are similar if the input contains no new information. As soon as new multimodal information comes in, the highlighting mechanism in RMFN adapts to these new inputs.

Priors: Based on the distribution of attention weights, we observe that the language and acoustic modalities seem the most commonly highlighted. This represents a prior over the expression of sentiment in human multimodal language and is closely related to the strong connections between language and speech in human communication [33].

Inactivity: Some attention coefficients are not active (always orange) throughout time. We hypothesize that these corresponding dimensions carry only intra-modal dynamics and are not involved in the formation of cross-modal interactions.

C .1 Qualitative Analysis

In addition to the general observations above, Figure 3 shows three examples where multistage fusion learns cross-modal representations across three different scenarios.

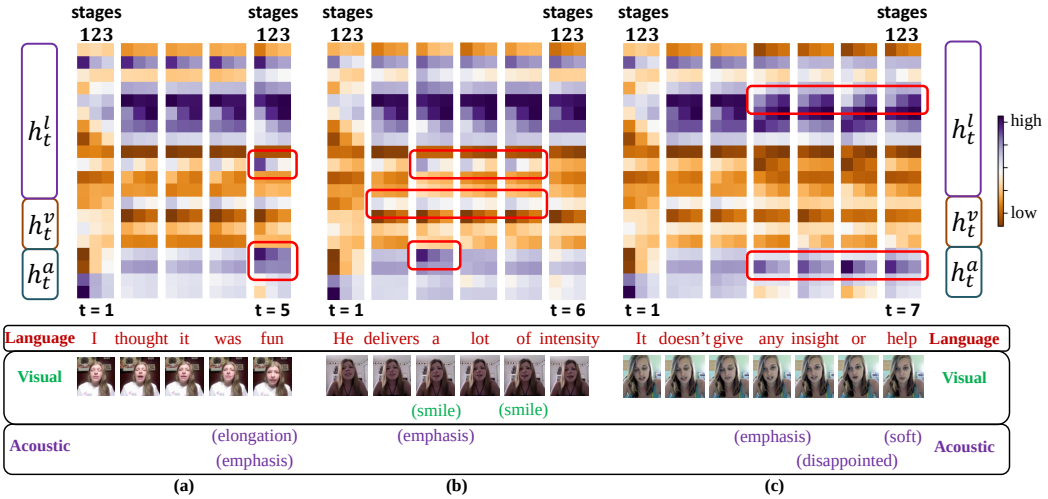

Figure 3: Visualization of learned attention weights across stages 1,2 and 3 of the multistage fusion process and across time of the multimodal sequence. We observe that the attention weights are diverse and evolve across stages and time. In these three examples, the red boxes emphasize specific moments of interest. (a) Synchronized interactions: the positive word "fun" and the acoustic behaviors of emphasis and elongation ($t = 5$) are synchronized in both attention weights for language and acoustic features. (b) Asynchronous trimodal interactions: the asynchronous presence of a smile ($t = 2 : 5$) and emphasis ($t = 3$) help to disambiguate the language modality. (c) Bimodal interactions: the interactions between the language and acoustic modalities are highlighted by alternating stages of fusion ($t = 4 : 7$).

Synchronized Interactions: In Figure 3(a), the language features are highlighted corresponding to the utterance of the word "fun" that is highly indicative of sentiment ($t = 5$). This sudden change is also accompanied by a synchronized highlighting of the acoustic features. We also notice that the highlighting of the acoustic features lasts longer across the 3 stages since it may take multiple stages to interpret all the new acoustic behaviors (elongated tone of voice and phonological emphasis).

Asynchronous Trimodal Interactions: In Figure 3(b), the language modality displays ambiguous sentiment: "delivers a lot of intensity" can be inferred as both positive or negative. We observe that the circled attention units in the visual and acoustic features correspond to the asynchronous presence of a smile ($t = 2 : 5$) and phonological emphasis ($t = 3$) respectively. These nonverbal behaviors resolve ambiguity in language and result in an overall display of positive sentiment. We further note the coupling of attention weights that highlight the language, visual and acoustic features across stages ($t = 3 : 5$), further emphasizing the coordination of all three modalities during multistage fusion despite their asynchronous occurrences.

Bimodal Interactions: In Figure 3(c), the language modality is better interpreted in the context of acoustic behaviors. The disappointed tone and soft voice provide the nonverbal information useful for sentiment inference. This example highlights the bimodal interactions ($t = 4 : 7$) in alternating stages: the acoustic features are highlighted more in earlier stages while the language features are highlighted increasingly in later stages.

Dataset	CMU-MOSI				
Task	Sentiment				
Metric	A^2	F1	A^7	MAE	Corr
Majority	50.2	50.1	17.5	1.864	0.057
RF	56.4	56.3	21.3	-	-
SVM-MD	71.6	72.3	26.5	1.100	0.559
THMM	50.7	45.4	17.8	-	-
SAL-CNN	73.0	-	-	-	-
C-MKL	72.3	72.0	30.2	-	-
EF-HCRF	65.3	65.4	24.6	-	-
EF-LDHCRF	64.0	64.0	24.6	-	-
MV-HCRF	44.8	27.7	22.6	-	-
MV-LDHCRF	64.0	64.0	24.6	-	-
CMV-HCRF	44.8	27.7	22.3	-	-
CMV-LDHCRF	63.6	63.6	24.6	-	-
EF-HSSHCRF	63.3	63.4	24.6	-	-
MV-HSSHCRF	65.6	65.7	24.6	-	-
DF	72.3	72.1	26.8	1.143	0.518
EF-LSTM	74.3	74.3	32.4	1.023	0.622
EF-SLSTM	72.7	72.8	29.3	1.081	0.600
EF-BLSTM	72.0	72.0	28.9	1.080	0.577
EF-SBLSTM	73.3	73.2	26.8	1.037	0.619
MV-LSTM	73.9	74.0	33.2	1.019	0.601
BC-LSTM	73.9	73.9	28.7	1.079	0.581
TFN	74.6	74.5	28.7	1.040	0.587
GME-LSTM(A)	76.5	73.4	-	0.955	-
MARN	77.1	77.0	34.7	0.968	0.625
MFN	77.4	77.3	34.1	0.965	0.632
RMFN	**78.4**	**78.0**	**38.3**	**0.922**	**0.681**
Human	85.7	87.5	53.9	0.710	0.820

Table 3: Sentiment prediction results on CMU-MOSI test set. The best results are highlighted in bold. RMFN outperforms the current state-of-the-art across all evaluation metrics.

Dataset	IEMOCAP Emotions							
Task	Happy		Sad		Angry		Neutral	
Metric	A^2	F1	A^2	F1	A^2	F1	A^2	F1
Majority	85.6	79.0	79.4	70.3	75.8	65.4	59.1	44.0
SVM	86.1	81.5	81.1	78.8	82.5	82.4	65.2	64.9
RF	85.5	80.7	80.1	76.5	81.9	82.0	63.2	57.3
THMM	85.6	79.2	79.5	79.8	79.3	73.0	58.6	46.4
EF-HCRF	85.7	79.2	79.4	70.3	75.8	65.4	59.1	44.0
EF-LDHCRF	85.8	79.5	79.4	70.3	75.8	65.4	59.1	44.0
MV-HCRF	15.0	4.9	79.4	70.3	24.2	9.4	59.1	44.0
MV-LDHCRF	85.7	79.2	79.4	70.3	75.8	65.4	59.1	44.0
CMV-HCRF	14.4	3.6	79.4	70.3	24.2	9.4	59.1	44.0
CMV-LDHCRF	85.8	79.5	79.4	70.3	75.8	65.4	59.1	44.0
EF-HSSHCRF	85.8	79.5	79.4	70.3	75.8	65.4	59.1	44.0
MV-HSSHCRF	85.8	79.5	79.4	70.3	75.8	65.4	59.1	44.0
DF	86.0	81.0	81.8	81.2	75.8	65.4	59.1	44.0
EF-LSTM	85.2	83.3	82.1	81.1	84.5	84.3	68.2	67.1
EF-SLSTM	85.6	79.0	80.7	80.2	82.8	82.2	68.8	68.5
EF-BLSTM	85.0	83.7	81.8	81.6	84.2	83.3	67.1	66.6
EF-SBLSTM	86.0	84.2	80.2	80.5	**85.2**	84.5	67.8	67.1
MV-LSTM	85.9	81.3	80.4	74.0	85.1	84.3	67.0	66.7
BC-LSTM	84.9	81.7	83.2	81.7	83.5	84.2	67.5	64.1
TFN	84.8	83.6	83.4	82.8	83.4	84.2	67.5	65.4
MARN	86.7	83.6	82.0	81.2	84.6	84.2	66.8	65.9
MFN	86.5	84.0	83.5	82.1	85.0	83.7	**69.6**	**69.2**
RMFN	**87.5**	**85.8**	**83.8**	**82.9**	85.1	**84.6**	69.5	69.1

Table 4: Emotion recognition results on IEMOCAP test set. The best results are highlighted in bold. RMFN achieves state-of-the-art or competitive performance across all evaluation metrics.

Dataset	POM Speaker Personality Traits															
Task	Con	Pas	Voi	Dom	Cre	Viv	Exp	Ent	Res	Tru	Rel	Out	Tho	Ner	Per	Hum
Metric	A^7	A^7	A^7	A^7	A^7	A^7	A^7	A^7	A^5	A^5	A^5	A^5	A^5	A^5	A^7	A^5
Majority	19.2	20.2	30.5	18.2	21.7	25.6	26.1	19.7	29.6	44.3	39.4	36.0	31.0	24.1	20.7	6.9
SVM	20.6	20.7	32.0	35.0	25.1	29.1	26.6	31.5	34.0	50.2	49.8	42.9	39.9	41.4	28.1	36.0
RF	26.6	27.1	29.6	26.1	23.2	23.6	26.6	26.1	34.0	53.2	40.9	32.5	37.4	36.0	25.6	40.4
THMM	24.1	15.3	19.2	29.1	27.6	26.1	18.7	12.3	22.7	31.0	31.5	30.0	30.0	27.1	17.2	24.6
DF	25.6	24.1	33.0	34.0	26.1	32.0	26.6	29.6	30.0	53.7	50.2	39.4	37.9	42.4	26.6	34.5
EF-LSTM	20.7	27.6	31.5	35.0	25.1	31.0	25.1	29.1	30.0	48.3	48.3	38.4	42.4	40.4	25.6	36.0
EF-SLSTM	22.2	28.6	30.5	36.9	27.1	32.0	27.6	27.6	32.5	49.3	46.8	40.4	39.9	41.9	22.7	35.0
EF-BLSTM	25.1	26.1	34.0	32.0	29.6	31.0	25.6	33.5	30.0	52.2	46.3	34.0	41.9	42.9	25.6	39.4
EF-SBLSTM	23.2	30.5	29.1	31.0	27.6	32.5	31.0	25.1	33.5	52.7	47.8	38.4	39.4	44.8	25.6	38.9
MV-LSTM	25.6	28.6	28.1	34.5	25.6	32.5	29.6	29.1	33.0	52.2	50.7	38.4	37.9	42.4	26.1	38.9
BC-LSTM	26.6	26.6	31.0	33.0	27.6	36.5	30.5	29.6	33.0	52.2	47.3	37.9	45.8	36.0	27.1	36.5
TFN	24.1	31.0	31.5	34.5	24.6	25.6	27.6	29.1	30.5	38.9	35.5	37.4	33.0	42.4	27.6	33.0
MARN	29.1	33.0	-	38.4	31.5	-	-	33.5	36.9	55.7	52.2	-	-	47.3	31.0	44.8
MFN	34.5	35.5	37.4	**41.9**	34.5	36.9	36.0	**37.9**	38.4	**57.1**	53.2	**46.8**	47.3	47.8	34.0	**47.3**
RMFN	**37.4**	**38.4**	**37.4**	39.4	**37.4**	**38.9**	**38.9**	36.9	**39.4**	56.7	**53.7**	46.3	**48.3**	**48.3**	**35.0**	46.8

Table 5: Results for personality trait recognition on the POM dataset. The best results are highlighted in bold. The MFP outperforms the current state of the art across most evaluation metrics.