Multimodal Language Analysis with Recurrent Multistage Fusion: Supplementary Material

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

1 Experiment Details

1.1 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 (Pennington et al., 2014). 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 (iMotions, 2017) is used to extract a set of visual features including facial action units, facial landmarks, head pose, gaze tracking and HOG features (Zhu et al., 2006). 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 (Degottex et al., 2014) is used to extract acoustic features including 12 Mel-frequency cepstral coefficients, pitch tracking and voiced/unvoiced segmenting features (Drugman and Alwan, 2011), glottal source parameters (Childers and Lee, 1991; Drugman et al., 2012; Alku, 1992; Alku et al., 1997, 2002), peak slope parameters and maxima dispersion quotients (Kane and Gobl, 2013). 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.

1.2 Multimodal Alignment

We perform forced alignment using P2FA (Yuan and Liberman, 2008) 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 (Zadeh et al., 2018).

2 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.

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**
Δ_{SOTA}	↑ **1.0**	↑ **0.7**	↑ **3.6**	↓ **0.033**	↑ **0.049**
Human	85.7	87.5	53.9	0.710	0.820

Table 1: Sentiment prediction results on CMU-MOSI test set. The best results are highlighted in bold and Δ_{SOTA} shows the change in performance over previous state of the art (SOTA). Improvements are highlighted in green. The RMFN significantly outperforms the current SOTA 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
Δ_{SOTA}	↑ **0.8**	↑ **1.6**	↑ **0.3**	↑ **0.1**	↓ **0.1**	↑ **0.1**	↓ **0.1**	↓ **0.1**

Table 2: Emotion recognition results on IEMOCAP test set. The best results are highlighted in bold and Δ_{SOTA} shows the change in performance over previous SOTA. Improvements are highlighted in green. The RMFN achieves state-of-the-art or competitive performance across all evaluation metrics.

Dataset	POM Speaker Personality Traits											
Task	Con	Pas	Voi	Cre	Viv	Exp	Res	Rel	Tho	Ner	Per	Hum
Metric	A^7	A^7	A^7	A^7	A^7	A^7	A^5	A^5	A^5	A^5	A^7	A^5
Majority	19.2	20.2	30.5	21.7	25.6	26.1	29.6	39.4	31.0	24.1	20.7	6.9
SVM	20.6	20.7	32.0	25.1	29.1	26.6	34.0	49.8	39.9	41.4	28.1	36.0
RF	26.6	27.1	29.6	23.2	23.6	26.6	34.0	40.9	37.4	36.0	25.6	40.4
THMM	24.1	15.3	19.2	27.6	26.1	18.7	22.7	31.5	30.0	27.1	17.2	24.6
DF	25.6	24.1	33.0	26.1	32.0	26.6	30.0	50.2	37.9	42.4	26.6	34.5
EF-LSTM	20.7	27.6	31.5	25.1	31.0	25.1	30.0	48.3	42.4	40.4	25.6	36.0
EF-SLSTM	22.2	28.6	30.5	27.1	32.0	27.6	32.5	46.8	39.9	41.9	22.7	35.0
EF-BLSTM	25.1	26.1	34.0	29.6	31.0	25.6	30.0	46.3	41.9	42.9	25.6	39.4
EF-SBLSTM	23.2	30.5	29.1	27.6	32.5	31.0	33.5	47.8	39.4	44.8	25.6	38.9
MV-LSTM	25.6	28.6	28.1	25.6	32.5	29.6	33.0	50.7	37.9	42.4	26.1	38.9
BC-LSTM	26.6	26.6	31.0	27.6	36.5	30.5	33.0	47.3	45.8	36.0	27.1	36.5
TFN	24.1	31.0	31.5	24.6	25.6	27.6	30.5	35.5	33.0	42.4	27.6	33.0
MARN	29.1	34.0	34.5	31.5	35.0	31.5	36.9	52.2	46.8	47.3	31.0	44.8
MFN	34.5	35.5	37.4	34.5	36.9	36.0	38.4	53.2	47.3	47.8	34.0	**47.3**
RMFN	**37.4**	**38.4**	**37.4**	**37.4**	**38.9**	**38.9**	**39.4**	**53.7**	**48.3**	**48.3**	**35.0**	46.8
Δ_{SOTA}	↑ **2.9**	↑ **2.9**	↑ **0.0**	↑ **2.9**	↑ **2.0**	↑ **2.9**	↑ **1.0**	↑ **0.5**	↑ **1.0**	↑ **0.5**	↑ **1.0**	↑ **0.5**

Dataset	POM Speaker Personality Traits											
Task	Con	Pas	Voi	Cre	Viv	Exp	Res	Rel	Tho	Ner	Per	Hum
Metric	MAE											
Majority	1.483	1.268	1.089	1.260	1.158	1.164	1.166	0.753	0.939	1.181	1.328	1.774
SVM	1.071	1.159	0.938	1.036	1.043	1.031	0.877	0.594	0.728	0.714	1.110	0.801
DF	1.033	1.083	0.899	1.039	0.997	1.027	0.884	0.591	0.732	0.695	1.086	0.768
EF-LSTM	1.035	1.067	0.911	1.022	0.981	0.990	0.880	0.594	0.712	0.706	1.084	0.768
EF-SLSTM	1.047	1.069	0.924	1.040	0.990	1.005	0.872	0.597	0.726	0.686	1.098	0.787
EF-BLSTM	1.103	1.105	0.975	1.053	1.018	1.069	0.882	0.607	0.762	0.705	1.085	0.762
EF-SBLSTM	1.062	1.055	0.926	1.027	1.020	0.991	0.877	0.594	0.746	0.697	1.102	0.779
MV-LSTM	1.029	1.098	0.971	1.082	0.976	1.012	0.877	0.625	0.792	0.687	1.163	0.770
BC-LSTM	1.016	1.008	0.914	0.942	**0.905**	0.906	0.888	0.630	0.680	0.705	1.025	0.767
TFN	1.049	1.104	0.927	1.058	1.000	1.029	0.900	0.621	0.743	0.727	1.132	0.770
MARN	0.993	1.033	0.886	0.945	0.950	0.945	0.860	0.576	0.685	0.677	1.036	0.749
MFN	0.952	0.993	0.882	**0.903**	0.908	**0.886**	0.821	0.566	**0.665**	0.654	**0.981**	**0.727**
RMFN	**0.933**	**0.938**	**0.868**	0.967	0.914	0.887	**0.781**	**0.566**	0.666	**0.640**	1.014	0.728
Δ_{SOTA}	↓ **0.019**	↓ **0.055**	↓ **0.014**	↓ **0.064**	↓ **0.001**	↓ **0.020**	↓ **0.040**	↓ **0.0**	↑ **0.001**	↓ **0.014**	↑ **0.033**	↓ **0.001**

Dataset	POM Speaker Personality Traits											
Task	Con	Pas	Voi	Cre	Viv	Exp	Res	Rel	Tho	Ner	Per	Hum
Metric	r											
Majority	-0.041	-0.029	-0.104	-0.122	-0.044	-0.065	0.006	-0.024	-0.130	0.097	-0.127	-0.069
SVM	0.063	0.086	-0.004	0.113	0.076	0.134	0.166	0.104	0.134	0.068	0.064	0.147
DF	0.240	0.273	0.017	0.112	0.173	0.118	0.148	0.019	0.041	0.136	0.168	0.259
EF-LSTM	0.200	0.302	0.031	0.170	0.244	0.265	0.142	0.083	0.260	0.105	0.217	0.227
EF-SLSTM	0.221	0.327	0.042	0.177	0.239	0.268	0.204	0.092	0.252	0.159	0.218	0.196
EF-BLSTM	0.162	0.289	0.034	0.191	0.279	0.274	0.184	0.093	0.245	0.166	0.243	0.272
EF-SBLSTM	0.174	0.310	0.021	0.170	0.224	0.261	0.155	0.097	0.215	0.121	0.216	0.171
MV-LSTM	0.358	0.416	0.131	0.280	0.347	0.323	0.295	0.119	0.284	0.258	0.239	0.317
BC-LSTM	0.359	0.425	0.081	0.358	0.417	0.450	0.293	0.075	0.363	0.184	0.344	0.319
TFN	0.089	0.201	0.030	0.124	0.204	0.171	-0.051	0.114	0.048	-0.002	0.106	0.213
MARN	0.340	0.410	0.166	0.340	0.374	0.406	0.282	0.215	0.348	0.235	0.303	0.287
MFN	0.395	0.428	0.193	**0.367**	**0.431**	0.452	0.333	0.255	**0.381**	0.318	**0.377**	**0.386**
RMFN	**0.441**	**0.502**	**0.247**	0.355	0.414	**0.470**	**0.446**	**0.291**	0.376	**0.379**	0.327	0.346
Δ_{SOTA}	↑ **0.046**	↑ **0.074**	↑ **0.054**	↓ **0.012**	↓ **0.017**	↓ **0.018**	↑ **0.113**	↑ **0.036**	↓ **0.005**	↑ **0.061**	↓ **0.050**	↓ **0.040**

Table 3: Results for personality trait recognition on the POM dataset. The best results are highlighted in bold and Δ_{SOTA} shows the change in performance over previous state of the art. Improvements are highlighted in green. The RMFN achieves state-of-the-art or competitive performance across all evaluation metrics.

References

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

Paavo Alku, Tom Bäckström, and Erkki Vilkman. 2002. Normalized amplitude quotient for parametrization of the glottal flow. *the Journal of the Acoustical Society of America* 112(2):701–710.

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

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

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarepa collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, pages 960–964.

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

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

iMotions. 2017. Facial expression analysis. goo.gl/1rh1JN.

John Kane and Christer Gobl. 2013. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing* 21(6):1170–1179.

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

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

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. *arXiv preprint arXiv:1802.00927* .

Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, and Shai Avidan. 2006. Fast human detection using a cascade of histograms of oriented gradients. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, volume 2, pages 1491–1498.