Automatic Recognition of Bipolar Disorder from Multimodal Data

Ziheng Zhang (zz362@cam.ac.uk)

21st June, 2019

Table of Contents

Table of Contents

Bipolar Disorder

Bipolar Disorder is a serious mental disorder.

- BD is associated with significant mortality risk.
- 3.9% of US population are affected by BD in some point of their lives.

Automatic recognition systems of BD can help early detection of BD symptoms and reduce the treatment resistance. Moreover, it could assist psychologists during the face-to-face interviews.

Bipolar Disorder Corpus

The Turkish Audio-Visual Bipolar Disorder Corpus

- was introduced in 2018
- consists of audio-visual recordings of interview sessions
- aims to help develop automatic recognition system

Audio/Visual Emotion Challenge (AVEC) 2018 introduces a challenge on the BD recognition from multimodal data based on the BD corpus.

Table of Contents

Proposed Framework

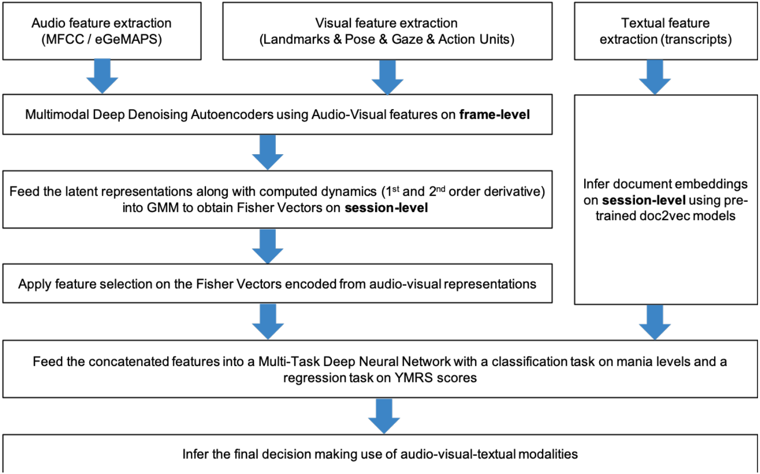


Figure: Pipeline of proposed architecture

Table of Contents

Unimodal Deep Denoising Autoencoders

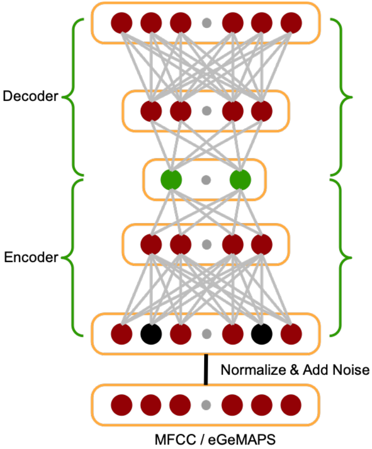


Figure: Unimodal Deep Denoising Autoencoders

Bimodal Deep Denoising Autoencoders

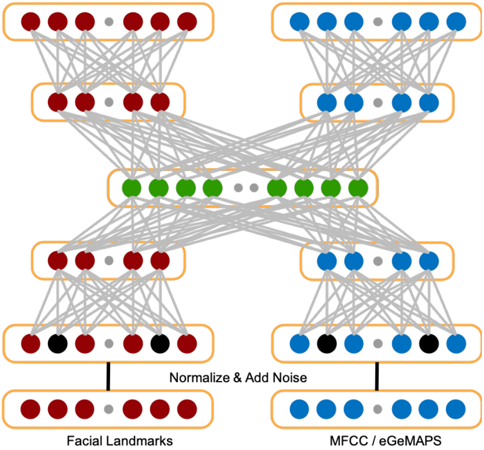


Figure: Bimodal Deep Denoising Autoencoders

Two acoustic features are investigated: MFCC or eGeMAPS features.

Multimodal Deep Denoising Autoencoders

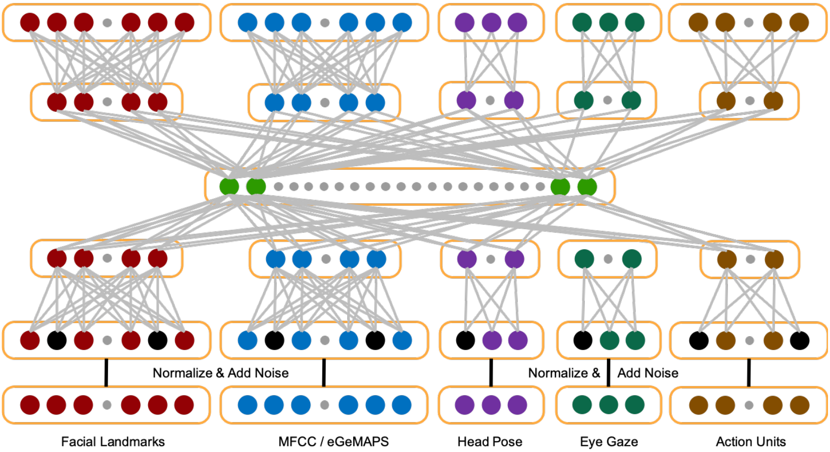
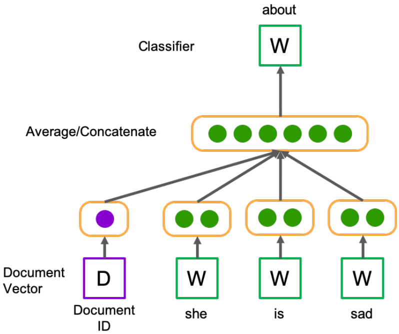


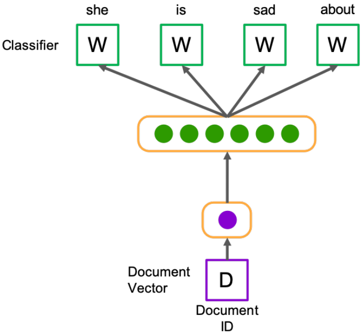
Figure: Multimodal Deep Denoising Autoencoders

Table of Contents

Document Embeddings



(a) doc2vec-DM



(b) doc2vec-DBOW

Figure: Paragraph Vector architectures

The multimodal framework also utilizes the textual modality, transcripts of interview sessions.

Table of Contents

Baseline

Table: Baseline systems with Random Forest classifiers on baseline features

Metric	MFCC	eGeMAPS	BoAW	FAU	BoVW
UAR (F)	0.414	0.396	0.443	-	0.452
UAR (S)	0.413	0.455	0.489	0.481	0.452
UAP	0.410	0.370	0.439	0.528	0.445
F1	0.411	0.408	0.463	0.503	0.448

The baseline systems do not take into account temporary information and the correlation across modality.

Multimodal Feature Learning

Table: Comparison of proposed Multimodal DDAE architectures (selected experimental results)

Acoustic feature	Hidden ratio	Noise	GMM kernel	UAR	UAP	F1
MFCC	0.4	0.1	32	0.656	0.678	0.667
eGeMAPS	0.5	0.1	32	0.622	0.665	0.642
Baseline (BoAW)				0.489	0.439	0.463
Baseline (FAU)				0.481	0.528	0.503
Best Unimodal DDAE (Landmarks)				0.624	0.692	0.656
Best Unimodal DDAE (MFCC)				0.587	0.611	0.599
Best Unimodal DDAE (eGeMAPS)				0.632	0.654	0.637
Best Bimodal DDAE (MFCC)				0.656	0.677	0.666
Best Bimodal DDAE (eGeMAPS)				0.566	0.611	0.587

Document Embeddings

Table: Comparison of proposed document embeddings on the transcripts (selected experimental results)

Model	Vector size	Window size	Negative words	UAR	UAP	F1
PV-DM	50	10	5	0.492	0.481	0.486
PV-DBOW	50	-	5	0.505	0.544	0.524
Baseline (BoAW)				0.489	0.439	0.463
Baseline (FAU)				0.481	0.528	0.503

Visualization

(a) multi-DDAE on MFCC

(b) multi-DDAE on eGeMAPS

(c) PV-DM

(d) PV-DBOW

Figure: Visualization of Fisher vectors and document embeddings in 2D space using t-SNE algorithm

Comparison with competing frameworks in AVEC2018

After feature fusion, a Multi-Taks neural network is implemented for classification, which makes use of regression task to address the overfitting issues.

Table: Comparison with proposed frameworks in AVEC2018

Framework	UAR (dev)	Accuracy (dev)
Yang *et al.* 2018	0.714	0.717
Du *et al.* 2018	0.651	0.650
Xing *et al.* 2018	0.868	NA
Syed *et al.* 2018	0.635	NA
This work	0.709	0.717

Generalization

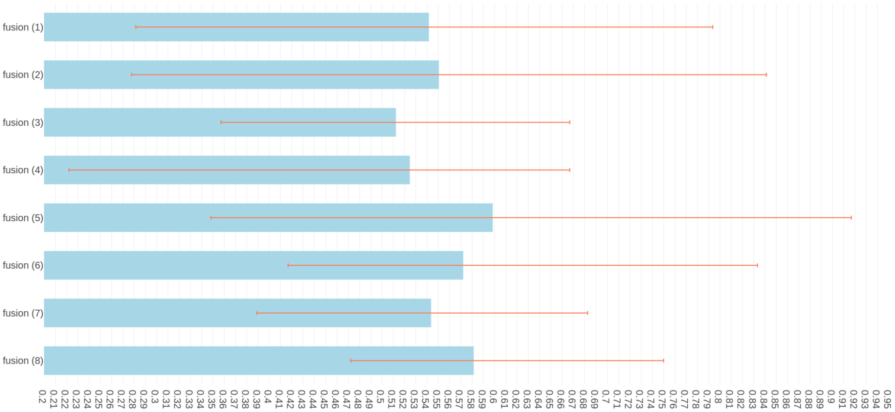


Figure: Barchart of the averaged UAR and variance via a 10-fold cross-validation on different fusion frameworks

Table of Contents

Conclusion

- The proposed multimodal framework demonstrates effective in learning the shared representations across modalities while managing the discrepancy.
- It achieves the state-of-the-art performance when compared with competing frameworks in the BD recognition task proposed in AVEC2018.

Future Work

- To introduce more layers in Deep Denoising Autoencoders to capture the spatial information (like Convolutional layers).
- To correlate and decouple different modalities via a semantic interface to obtain more robust representations.
- To evaluate the performance of the proposed framework on other similar problems, like the recognition of human state-of-mind proposed in AVEC2019.

Thank you for your attention