Deep Multimodal Representation Learning from Temporal Data

Xitong Yang[1], Palghat Ramesh[2], Radha Chitta[3], Sriganesh Madhvanath[3], Edgar A. Bernal[4] and Jiebo Luo[5]

[1] University of Maryland, College Park [2] PARC [3] Conduent Labs US

[4] United Technologies Research Center [5] University of Rochester

IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR July 21-26 2017 HONOLULU

Introduction

- We proposed the ***Corr**elational **R**ecurrent **N**eural **N**etwork (**CorrRNN**)*, a novel temporal fusion model for fusing multiple input modalities that are inherently temporal in nature.
- Characteristics of multimodal temporal data
  - **Temporal consistency**
    → Multimodal GRU Module (GRU)
  - **High correlation**
    → Correlation Module (Corr)
  - **Dynamic noise-level changes**
    → Dynamic Weighting Module (DW)
- Multimodal learning setup (unsupervised)

Tasks	Feature Learning	Supervised Training	Testing
Multimodal Fusion	$X+Y$	$X+Y$	$X+Y$
Cross Modality Learning	$X+Y$	X	X
	$X+Y$	Y	Y
Shared Represe-ntation Learning	$X+Y$	X	Y
	$X+Y$	Y	X

Proposed Model

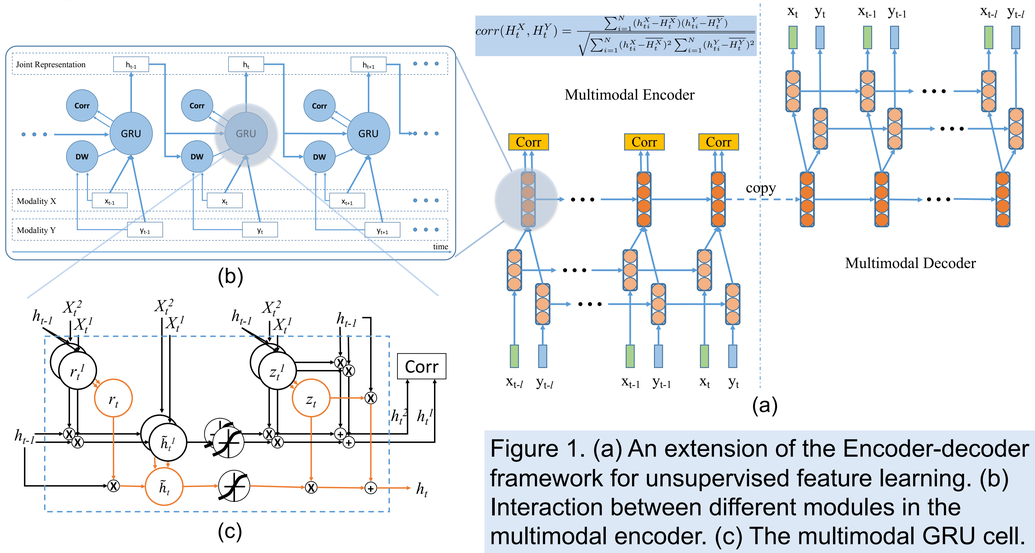


Figure 1. (a) An extension of the Encoder-decoder framework for unsupervised feature learning. (b) Interaction between different modules in the multimodal encoder. (c) The multimodal GRU cell.

Experiments on Audio-Video Data

- We test our model on two classic datasets.
  - **AVLetters** includes audio and video of 10 speakers uttering the English alphabet three times each.
  - **CUAVE** consists of videos of 36 speakers pronouncing the digits 0-9.
- For fair comparison, we perform the same data pre-processing approaches as those for RTMRBM (2016).

Method (years)	Accuracy	
	AVLetters	CUAVE
MDAE (2011)	62.04	66.70
MDBN (2012)	63.2	67.20
MDBM (2012)	64.7	69.00
RTMRBM (2016)	66.04	-
CRBM (2014)	67.10	69.10
CorrRNN (ours)	**83.40**	**95.9**

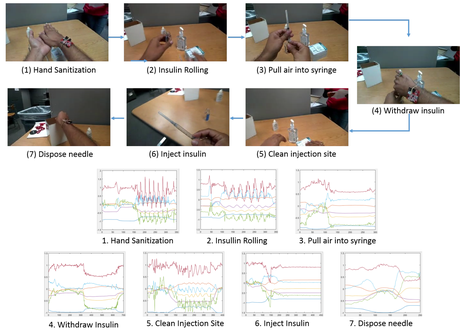
Table 1. Classification performance for audio-visual speech recognition compared to the best published results in literature, using the fused representation of the two modalities.

Method	Accuracy	
	Clean Audio	Noisy Audio
MDAE	94.4	77.3
Audio RBM	95.8	75.8
MDAE + Audio RBM	94.4	82.2
CorrRNN	**96.11**	**90.88**

Table 2. Our proposed model is more robust to noise, compared with other multimodal learning methods.

Experiments on Video-Sensor Data

- The Insulin Self-Injection (**ISI**) dataset
- 11 subjects perform seven actions related to an insulin self-injection activity.
- Egocentric video data acquired using a Google Glass wearable camera, and motion data acquired using an Invensense motion wrist sensor.
- Sensor data includes Quaternion rotation, Linear acceleration and Gravity compensated acceleration.



- We perform extensive experiments for different model configurations.
- Each loss component contributes to better performance, especially in the settings of cross-modality learning and shared representation learning.

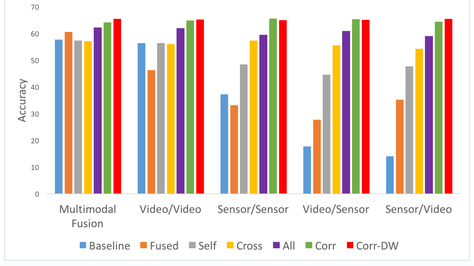


Figure 2. Classification accuracy on the ISI dataset for different model configurations.