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

[1]xyang35@cs.umd.edu, [2]Palghat.Ramesh@parc.com, [3]{Radha.Chitta,
Sriganesh.Madhvanath}@conduent.com, [4]bernalea@utrc.utc.com, [5]jluo@cs.rochester.edu

Abstract

*In recent years, Deep Learning has been successfully applied to multimodal learning problems, with the aim of learning useful joint representations in data fusion applications. When the available modalities consist of time series data such as video, audio and sensor signals, it becomes imperative to consider their temporal structure during the fusion process. In this paper, we propose the **Corr**elational **R**ecurrent Neural Network (CorrRNN), a novel temporal fusion model for fusing multiple input modalities that are inherently temporal in nature. Key features of our proposed model include: (i) simultaneous learning of the joint representation and temporal dependencies between modalities, (ii) use of multiple loss terms in the objective function, including a maximum correlation loss term to enhance learning of cross-modal information, and (iii) the use of an attention model to dynamically adjust the contribution of different input modalities to the joint representation. We validate our model via experimentation on two different tasks: video- and sensor-based activity classification, and audio-visual speech recognition. We empirically analyze the contributions of different components of the proposed CorrRNN model, and demonstrate its robustness, effectiveness and state-of-the-art performance on multiple datasets.*

1. Introduction

Automated decision-making in a wide range of real-world scenarios often involves acquisition and analysis of data from multiple sources. For instance, human activity may be more robustly monitored using a combination of video cameras and wearable motion sensors than with either
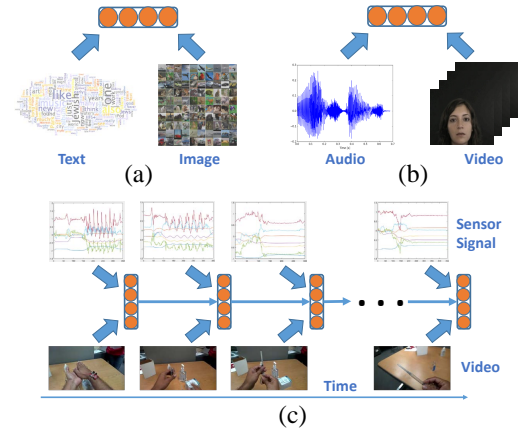

Figure 1. Different multimodal learning tasks. (a) Non-temporal model for non-temporal data [21]. (b) Non-temporal model for temporal data [13]. (c) Proposed CorrRNN model: temporal model for temporal data.

sensing modality by itself. When analyzing spontaneous socio-emotional behaviors, researchers can use multimodal cues from video, audio and physiological sensors such as electro-cardiograms (ECG) [17]. However, fusing information from different modalities is usually nontrivial due to the distinct statistical properties and highly non-linear relationships between low-level features [21] of the modalities. Prior work has shown that *multimodal learning* often provides better performance on tasks such as retrieval, classification and description [9, 13, 21, 12]. When the modalities being fused are temporal in nature, it becomes desirable to design a model for *temporal multimodal learning* (TML) that can simultaneously fuse the information from different sources, and capture temporal structure within the data.

In the past five years, several deep learning based approaches have been proposed for TML, in particular, for audio-visual data. Early models proposed for audiovi-

---
[*]Work carried out while at PARC, a Xerox Company

sual speech recognition (AVSR) were based on the use of non-temporal models such as deep multimodal *autoencoders* [13] or deep *Restricted Boltzmann Machines* (RBM) [21, 22] applied to concatenated data across a number of consecutive frames. More recent models have attempted to model the inherently sequential nature of temporal data, *e.g.*, Conditional RBMs [1], Recurrent Temporal Multimodal RBMs (RTMRBM) [7] for AVSR, and Multimodal Long-Short-Term Memory networks for speaker identification [16].

We believe that a good model for TML should simultaneously learn a joint representation of the multimodal input, and the temporal structure within the data. Moreover, the model should be able to dynamically weigh different input modalities to enable emphasis on the more useful signal(s) and to provide robustness to noise, a known weakness of AVSR [8]. Third, the model should be able to generalize to different kinds of multimodal temporal data, not just audio-visual data. Finally, the model should be tractable and efficient to train. In this paper, we introduce the ***Corr***elational ***Recurrent Neural Network*** (CorrRNN), a novel unsupervised model that satisfies the above desiderata.

An interesting characteristic of multimodal temporal data from many application scenarios is that the differences across modalities stem largely from the use of different sensors such as video cameras, motion sensors and audio recorders, to capture the same temporal phenomenon. In other words, modalities in multimodal temporal data are often different representations of the same phenomena, which is usually not the case with other multimodal data such as images and text, which are related because of their shared high-level semantics. Motivated by this observation, our CorrRNN attempts to explicitly capture the *correlation* between modalities through maximizing a correlation-based loss function, as well as minimizing a reconstruction-based loss for retaining information.

This observation regarding correlated inputs has motivated previous work in multi-view representation learning using the Deep Canonically Correlated Autoencoder (DCCAE) [25] and Correlational Neural Network [4]. Our model extends this work in two important ways. First, an RNN-based encoder-decoder framework that uses Gated Recurrent Units (GRU) [5] is introduced to capture the temporal structure, as well as long-term dependencies and correlation across modalities. Second, dynamic weighting is used while encoding input sequences to assign different weights to input modes based on their contribution to the fused representation.

The main contributions of this paper are as follows:

- We propose a novel generic model for temporal multimodal learning that combines an Encoder-Decoder RNN framework with Multimodal GRUs, a multi-aspect learning objective, and a dynamic weighting

mechanism;

- We show empirically that our model outperforms state-of-the-art methods on two different application tasks: video- and sensor-based activity classification and audio-visual speech recognition; and

- Our proposed approach is more tractable and efficient to train compared with RTMRBM and other probabilistic models designed for TML.

The remainder of this paper is organized as follows. In Sec. 2, we review the related work on multimodal learning. We describe the proposed CorrRNN model in Sec. 3. Sec. 4 introduces the two application tasks and datasets used in our experiments. In Secs. 4.1 and 4.2, we present empirical results demonstrating the robustness and effectiveness of the proposed model. The final section presents conclusions and future research directions.

2. Related work

In this section, we briefly review some related work on deep-learning-based multimodal learning and temporal data fusion. Generally speaking, and from the standpoint of dynamicity, fusion frameworks can be classified based on the type of data they support (*e.g.*, temporal vs. non-temporal data) and the type of model used to fuse the data (*e.g.*, temporal vs. non-temporal model) as illustrated in Fig. 1.

2.1. Multimodal Deep Learning

Within the context of data fusion applications, deep learning methods have been shown to be able to bridge the gap between different modalities and produce useful joint representations [13, 21]. Generally speaking, two main approaches have been used for deep-learning-based multimodal fusion. The first approach is based on common representation learning, which learns a joint representation from the input modalities. The second approach is based on *Canonical Correlation Analysis* (CCA) [6], which learns separate representations for the input modalities while maximizing their correlation.

An example of the first approach, the Multimodal Deep Autoencoder (MDAE) model [13], is capable of learning a joint representation that is predictive of either input modality. This is achieved by performing simultaneous self-reconstruction (within a modality) and cross-reconstruction (across modalities). Srivastava *et al.* [21] propose to learn a joint density model over the space of multimodal inputs using *Multimodal Deep Boltzmann Machines* (MDBM). Once trained, it is able to infer a missing modality through Gibbs sampling and obtain a joint representation even in the absence of some modalities. This model has been used to build a practical AVSR system [22]. Sohn *et al.* [19] propose a new learning objective to improve multimodal learn-

ing, and explicitly train their model to reason about missing modalities by minimizing the variation of information.

CCA-based methods, on the other hand, aim to learn separate features for the different modalities such that the correlation between them is mutually maximized. They are commonly used in multi-view learning tasks. In order to improve the flexibility of CCA, Deep CCA (DCCA) [2] was proposed to learn nonlinear projections using deep networks. Weirang *et al.* [25] extended this work by combining DCCA with the multimodal deep autoencoder learning objective [13]. The Correlational Neural Network model [4] is similar in that it integrates two types of learning objectives into a single model to learn a common representation. However, instead of optimizing the objective function under the hard CCA constraints, it only maximizes the empirical correlation of the learned projections.

2.2. Temporal Models for Multimodal Learning

In contrast to multimodal learning using non-temporal models, there is little literature on fusing temporal data using temporal models. Amer *et al.* [1] proposed a hybrid model for fusing audio-visual data in which a *Conditional Restricted Boltzmann Machines* (CRBM) is used to model short-term multimodal phenomena and a discriminative *Conditional Random Field* (CRF) is used to enhance the model. In more recent work [7], the *Recurrent Temporal Multimodal RBM* was proposed which learns joint representations and temporal structures. The model yields state-of-the-art performance on the ASVR datasets AVLetters and AVLetters2. A supervised multimodal LSTM was proposed in [16] for speaker identification using face and audio sequences. The method was shown to be robust to both distractors and image degradation by modeling long-term dependencies over multimodal high-level features.

3. Proposed Model

In this section, we describe the proposed CorrRNN model. We start by formulating the temporal multimodal learning problem mathematically. For simplicity, and without loss of generality, we consider the problem of fusing two modalities X and Y; it should be noted, however, that the model seamlessly extends to more than two modalities. We then present an overview of the model architecture, which consists of two components: the multimodal encoder and the multimodal decoder. We describe the multimodal encoder, which extracts the joint data representation, in Sec. 3.3, and the multimodal decoder, which attempts to reconstruct the individual modalities from the joint representation in Sec. 3.4.

3.1. Temporal Multimodal Learning

Let us denote the two temporal modalities as sequences of length T, namely $\mathcal{X} = (x_1^m, x_2^m, ..., x_T^m)$ and $\mathcal{Y} =$
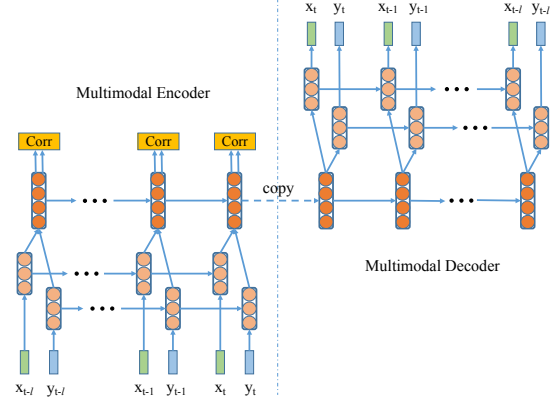

Figure 2. Basic architecture of the proposed model

$(y_1^n, y_2^n, ..., y_T^n)$, where x_t^m denotes the m dimensional feature of modality X at time t. For simplicity, we omit the superscripts m and n in most of the following discussion.

In order to achieve temporal multimodal learning, we fuse the two modalities at time t by considering both their current state and history. Specifically, at time t we append the recent per-modality history to the current samples x_t and y_t to obtain extended representations $\tilde{x}_t = \{x_{t-l}, ..., x_{t-1}, x_t\}$ and $\tilde{y}_t = \{y_{t-l}, ..., y_{t-1}, y_t\}$, where l denotes the scope of the history taken into account. Given pairs of multimodal data sequences $\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^N$, our goal is to train a feature learning model $\mathcal{M}$ that learns a d-dimensional joint representation $\left\{\tilde{h}_i\right\}_{i=1}^N$ which simultaneously fuses information from both modalities and captures underlying temporal structures.

3.2. Model Overview

We first describe the basic model architecture, as shown in Fig. 2. We implement an *Encoder-Decoder* framework, which enables sequence-to-sequence learning [23] and learning of sequence representations in an unsupervised fashion [20]. Specifically, our model consists of two recurrent neural nets: the *multimodal encoder* and the *multimodal decoder*. The multimodal encoder is trained to map the two input sequences into a joint representation, *i.e.*, a common space. The multimodal decoder attempts to reconstruct two input sequences from the joint representation obtained by the encoder. During the training process, the model learns a joint representation that retains as much information as possible from both modalities.

In our model, both the encoder and decoder are two-layer networks. The multimodal inputs are first mapped to separate hidden layers before being fed to a common layer called the *fusion layer*. Similarly, the joint representation is first decoded to separate hidden layers before reconstruction of the multimodal inputs takes place.

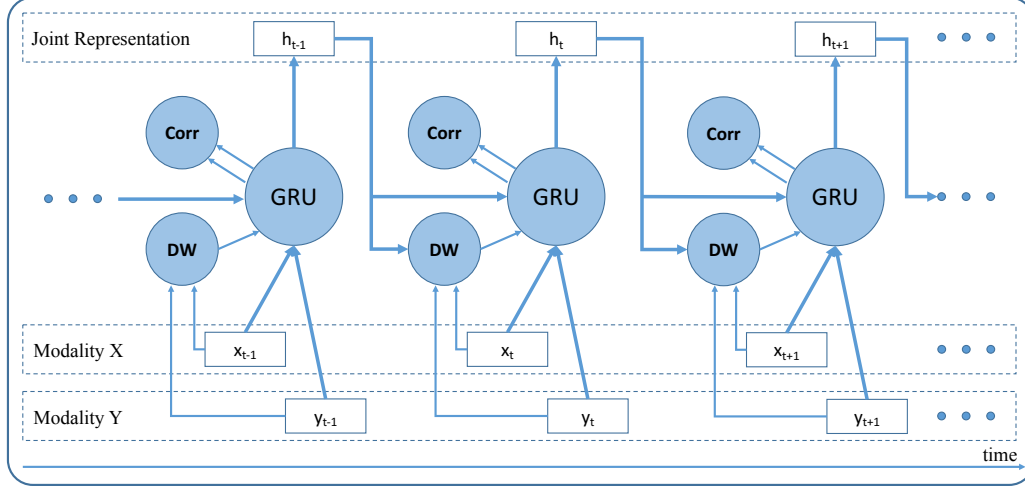The standard Encoder-Decoder framework relies on the

(reconstruction) loss function only in the decoder. As mentioned in Section 1, in order to obtain a better joint representation for temporal multimodal learning, we introduce two important components into the multimodal encoder, one that explicitly captures the correlation between the modalities, and another that performs dynamic weighting across modality representations. We also consider different types of reconstruction losses to enhance the capture of information within and between modalities.

Once the model is trained using a pair of multimodal inputs, the multimodal encoder plays the role of a feature extractor. Specifically, the activations of the fusion layer in the encoder at the last time step is output as the sequence feature representation. Two types of feature representation may be obtained depending on the model inputs: if both input modalities are present, we obtain their joint representation; on the other hand, if only one of the modalities is present, we obtain an "enhanced" unimodal representation. The model may be extended to more than two modalities by maximizing the sum of correlations between all pairs of modalities. This can be implemented by adding more correlation modules to the multimodal encoder.

3.3. Multimodal Encoder

The multimodal encoder is designed to fuse the input modality sequences into a common representation such that a coherent input is given greater importance, and the correlation between the inputs is maximized. Accordingly, three main modules are used by the multimodal encoder at each time step.

- Dynamic Weighting module (DW): Dynamically assigns weights to the two modalities by evaluating the coherence of the incoming signal with recent past history.

- GRU module (GRU): Fuses the input modalities to generate the fused representation. The module also captures the temporal structure of the sequence using *forget* and *update* gates.

- Correlation module (Corr): Takes the intermediate states generated by the GRU module as inputs to compute the correlation-based loss.

The structure of the multimodal encoder and the relationships among the three modules are illustrated in Fig. 3. We now describe the implementation of these modules in detail.

The **Dynamic Weighting module** assigns a weight to each modality input at a given time step according to an evaluation of its coherence over time. With reference to recent work on attention models [3], our approach may be characterized as a soft attention mechanism that enables the model to focus on the modality with the more useful signal when, for example, the other is corrupted with noise. The dynamic weights assigned to the input modalities are based on the agreement between their current input and the fused data representation from the previous time step. This is based on the intuition that an input corrupted by noise would be less in agreement with the fused representation from the previous time step when compared with a "clean" input. We use bilinear functions to evaluate the coherence scores α_t^1 and α_t^2 of the two modalities, namely:

$$\alpha_t^1 = x_t A_1 h_{t-1}^T, \quad \alpha_t^2 = y_t A_2 h_{t-1}^T,$$

where $A_1 \in \mathbb{R}^{m \times d}, A_2 \in \mathbb{R}^{n \times d}$ are parameters learned during the training of the module. The weights of the two modalities is obtained by normalizing the scores using Laplace smoothing:

$$w_t^i = \frac{1 + \exp(\alpha_t^i)}{2 + \sum_k \exp(\alpha_t^k)}, \quad i = 1, 2$$
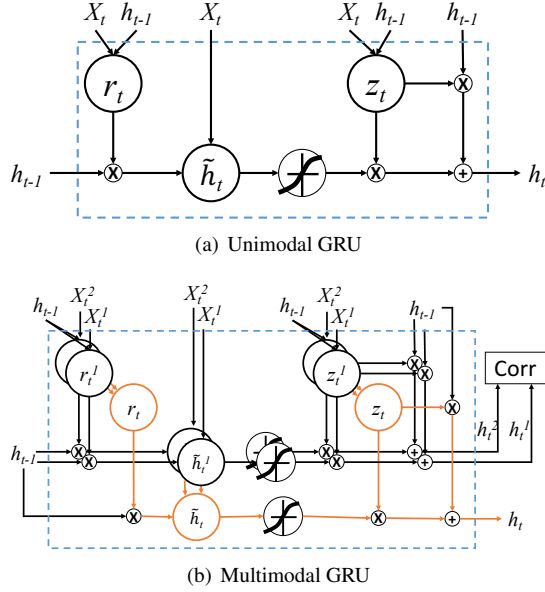
(a) Unimodal GRU

(b) Multimodal GRU

Figure 4. Block diagram illustrations of unimodal and multimodal GRU modules.

The **GRU module** (see Fig. 4(b)) is a multimodal extension of the standard GRU (see Fig. 4(a)), and contains different gating units that modulate the flow of information inside the module. The GRU module takes x_t and y_t as input at time step t and keeps track of three quantities, namely the fused representation h_t, and modality-specific representations h_t^1, h_t^2. The fused representation h_t constitutes a single representation of historical multimodal input that propagates along the time axis to maintain a consistent concept and learn its temporal structure. The modality-specific representations h_t^1, h_t^2 may be thought of as projections of the modality inputs which are maintained so that a measure of their correlation can be computed. The computation within this module may be formally expressed as follows:

$$r_t^i = \sigma\left(\mathbf{W}_r^i X_t^i + \mathbf{U}_r h_{t-1} + b_r^i\right), \qquad i = 1, 2 \quad (1)$$

$$z_t^i = \sigma\left(\mathbf{W}_z^i X_t^i + \mathbf{U}_z h_{t-1} + b_z^i\right), \qquad i = 1, 2 \quad (2)$$

$$\tilde{h}_t^i = \varphi\left(\mathbf{W}_h^i X_t^i + \mathbf{U}_h(r_t^i \odot h_{t-1}) + b_h^i\right), \quad i = 1, 2 \quad (3)$$

$$r_t = \sigma\left(\sum_{i=1}^{2} w_t^i \left(\mathbf{W}_r^i X_t^i + b_r^i\right) + \mathbf{U}_r h_{t-1}\right) \quad (4)$$

$$z_t = \sigma\left(\sum_{i=1}^{2} w_t^i \left(\mathbf{W}_z^i X_t^i + b_z^i\right) + \mathbf{U}_z h_{t-1}\right) \quad (5)$$

$$\tilde{h}_t = \varphi\left(\sum_{i=1}^{2} w_t^i \left(\mathbf{W}_h^i X_t^i + b_h^i\right) + \mathbf{U}_h(r_t \odot h_{t-1})\right) \quad (6)$$

$$h_t^i = (1 - z_t^i) \odot h_{t-1} + z_t^i \odot \tilde{h}_t^i, \qquad i = 1, 2 \quad (7)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (8)$$

where σ is the logistic sigmoid function and φ is the hyperbolic tangent function, r and z are the input to the reset and update gates, and h and $\tilde{h}$ represent the activation and candidate activation, respectively, of the standard GRU [5]. Note that our model uses separate weights for the different inputs X and Y, which differs from the approach proposed in [16]. However, as we enforce an explicit correlation-based loss term in the fusing process, our model in principle can capture both the correlation across modalities, and specific aspects of each modality.

The **Correlation module** computes the correlation between the projections of the modality inputs h_t^1 and h_t^2 obtained from the GRU module. Formally, given N mappings of two modalities $H_t^1 = \{h_{ti}^1\}_{i=1}^N$ and $H_t^2 = \{h_{ti}^2\}_{i=1}^N$ at time t, the correlation is calculated as follows:

$$corr(H_t^1, H_t^2) = \frac{\sum_{i=1}^{N}(h_{ti}^1 - \overline{H_t^1})(h_{ti}^2 - \overline{H_t^2})}{\sqrt{\sum_{i=1}^{N}(h_{ti}^1 - \overline{H_t^1})^2 \sum_{i=1}^{N}(h_{ti}^2 - \overline{H_t^2})^2}}$$

where $\overline{H_t^1} = \frac{1}{N}\sum_i^N h_{ti}^1$ and $\overline{H_t^2} = \frac{1}{N}\sum_i^N h_{ti}^2$. We denote the correlation-based loss function as $L_{\text{corr}} = corr(H_t^1, H_t^2)$ and maximize the correlation between two modalities by maximizing this function. In practice, the empirical correlation is computed within a mini-batch of size N.

3.4. Multimodal Decoder

The multimodal decoder attempts to reconstruct the individual modality input sequences X and Y simultaneously, from the joint representation h_t computed by the multimodal encoder described above. By minimizing the reconstruction loss at training, the resulting joint representation retains as much information as possible from both modalities. In order to better share information across the modalities, we introduce two additional reconstruction loss terms into the multimodal decoder: *cross-reconstruction* and *self-reconstruction*. These two terms not only benefit the joint representation, but also improve the performance of the model in cases when only one of the modalities is present, as shown in Section 4.1. In all, our multimodal decoder includes three reconstruction losses:

- **Fused-reconstruction loss**. The error in reconstructing $\tilde{x}_i$ and $\tilde{y}_i$ from joint representation $\tilde{h}_i = f(\tilde{x}_i, \tilde{y}_i)$.

$$L_{\text{fused}} = L(g(f(\tilde{x}_i, \tilde{y}_i)), \tilde{x}_i) + \beta L(g(f(\tilde{x}_i, \tilde{y}_i)), \tilde{y}_i)$$

- **Self-reconstruction loss**. The error in reconstructing $\tilde{x}_i$ from $\tilde{x}_i$, and $\tilde{y}_i$ from $\tilde{y}_i$.

$$L_{\text{self}} = L(g(f(\tilde{x}_i)), \tilde{x}_i) + \beta L(g(f(\tilde{y}_i)), \tilde{y}_i)$$

- **Cross-reconstruction loss**. The error in reconstructing $\tilde{x}_i$ from $\tilde{y}_i$, and $\tilde{y}_i$ from $\tilde{x}_i$.

$$L_{\text{cross}} = L(g(f(\tilde{y}_i), \tilde{x}_i) + \beta L(g(f(\tilde{x}_i)), \tilde{y}_i)$$

where β is a hyperparameter used to balance the relative scale of the loss function values of the two input modalities, and f, g denote the functional mappings implemented by the multimodal encoder and decoder, respectively. The objective function used to train our model may thus be expressed as:

$$\mathcal{L} = \sum_{i=1}^{N} \left(L_{\text{fused}} + L_{\text{cross}} + L_{\text{self}} \right) - \lambda L_{\text{corr}}$$

where λ is a hyperparameter used to scale the contribution of the correlation loss term, and N is the mini-batch size used in the training stage. The objective function thus combines different forms of reconstruction losses computed by the decoder, with the correlation loss computed as part of the encoding process. We use a stochastic gradient descent algorithm with an adaptive learning rate to optimize the objective function above.

4. Empirical Analysis

In the following sections, we describe experiments to demonstrate the effectiveness of CorrRNN at modeling temporal multimodal data. We demonstrate its general applicability to multimodal learning problems by evaluating it on multiple datasets, covering two different types of multimodal data (video-sensor and audio-video) and two different application tasks (activity classification and audio-visual speech recognition). We also evaluate our model in three *multimodal learning settings* [13] for each task. We review these settings in Table 1.

	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 Representation Learning	$X + Y$	X	Y
	$X + Y$	Y	X

Table 1. Multimodal Learning settings, where X and Y are different input modalities

For each application task and dataset, the CorrRNN model is first trained in an unsupervised manner using both the input modalities and the composite loss function described. The trained model is then used to extract the fused representation and the modality-specific representations of

the data. Each of the multimodal learning settings is then implemented as a supervised classification task using a classifier, either an SVM or a logistic-regression classifier (in order to maintain consistency, the choice of classifier depends on the method involved in the benchmarking implemented).

4.1. Experiments on Video-Sensor Data

In this section, we apply the CorrRNN model to the task of human activity classification. For this purpose, we use the **ISI dataset** [10], a multimodal dataset in which 11 subjects perform seven actions related to an insulin self-injection activity. The dataset includes egocentric video data acquired using a Google Glass wearable camera, and motion data acquired using an Invensense motion wrist sensor. Each subject's video and motion data is manually labeled and segmented into seven videos corresponding to the seven actions in the self-injection procedure. Each of these videos are further segmented into short video clips of fixed length.

4.1.1 Implementation Details

We first temporally synchronize the video and motion sensor data with the same sampling rate of 30 fps. We compute a 1024-dimensional CNN feature representation for each video frame using GoogLeNet [24]. Raw motion sensor signals are smoothed by applying an averaging filter of width 4. Sensor features are obtained by computing the output of the last convolutional layer (layer 5) of a Deep Convolutional and LSTM (DCL) Network [14] pre-trained on the OPPORTUNITY dataset [18] to the smoothed sensor data input. The extracted features are a temporal sequence of 448-dimensional elements.

We build sequences from the video and sensor data, using a sliding window of 8 frames with a stride of 2, sampled from a duration of 2 seconds, resulting in 13, 456 sequences. These video and motion sequences are used to train the CorrRNN model, using stochastic gradient descent with the mini-batch size set to 256. The values of β and λ were set to 1 and 0.1, respectively; these values were optimized using grid search methods.

4.1.2 Results

Figure 5 shows the activity recognition accuracy of the proposed CorrRNN model. We evaluate the contribution of each component in our model under the various multimodal learning settings listed in Table 1. In order to understand the contribution of different aspects of the CorrRNN design, we also evaluate different model configurations summarized in Table 2. The baseline results are obtained by first training a single layer GRU recurrent neural network with 512 hidden units, separately for each modality. The 512-dimensional

Config	Description
Baseline	Single-layer GRU RNN per modality
Fused	Objective uses only L_{fused} term
Self	Objective uses L_{fused} & L_{self}
Cross	Objective uses L_{fused} & L_{cross}
All	Objective uses L_{fused}, L_{self} & L_{cross}
Corr	Objective uses all loss terms
Corr-DW	Objective uses all loss terms & dyn. weights

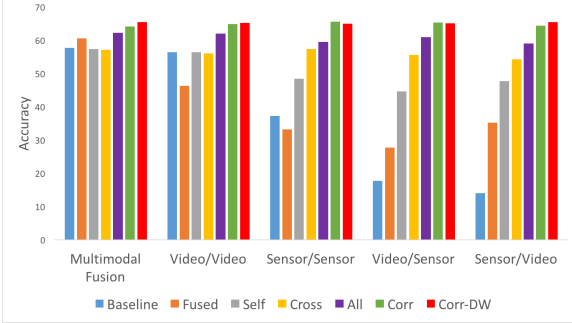Table 2. CorrRNN model configurations evaluated

Figure 5. Classification accuracy on the ISI dataset for different model configurations

hidden layer representations obtained from each network are then reduced to 256 dimensions using PCA, and concatenated to obtain a 512-dimensional fused representation. We observe that the fused representation obtained using CorrRNN significantly improves over this baseline fused representation.

Each loss component contributes to better performance, especially in the settings of cross-modality learning and shared representation learning. Performance in the presence of poor fidelity or noisy modality (for instance, the motion sensor modality) is boosted by the inclusion of the other modality, due to the cross reconstruction loss component. Inclusion of the correlation loss and dynamic weighting further improves the accuracy.

In Table 3, we compare the correlation between the projections of the modality inputs for different model configurations. This measure of correlation is computed as the mean encoder loss over the training data in the final training epoch, divided by the number of hidden units in the fusion layer. These values demonstrate that the use of the correlation-based loss term maximizes the correlation between the two projections, leading to a richer joint and shared representations.

4.2. Experiments on Audio-Video Data

The task of audio-visual speech classification using multimodal deep learning has been well studied in the literature [7, 13]. In this section, we focus on comparing the

Configuration	Correlation
Fused	0.46
Self	0.67
Cross	0.76
Corr	0.95
Corr-DW	0.93

Table 3. Normalized correlation for different model configurations

performance of the proposed model with other published methods on the AVLetters and CUAVE datasets:

- **AVLetters** [11] includes audio and video of 10 speakers uttering the English alphabet three times each. We use the videos corresponding to the first two times for training (520 videos) and the third time for testing (260 videos). This dataset provides pre-extracted lip regions scaled to 60×80 pixels for each video frame and 26-dimensional *Mel-Frequency Cepstrum Coefficient* (MFCC) features for the audio.

- **CUAVE** [15] consists of videos of 36 speakers pronouncing the digits 0-9. Following the protocol in [13], we use the first part of each video, containing the frontal facing speakers pronouncing each digit 5 times. The even-numbered speakers are used for training, and the odd-numbered speakers are used for testing. The training dataset contains 890 videos and the test data contains 899 videos. We pre-processed the video frames to extract only the region of interest containing the mouth, and rescaled each image to 60×60 pixels. The audio is represented using 26-dimensional MFCC features.

4.2.1 Implementation Details

We reduced the dimensionality of the video features of both the datasets to 100 using PCA whitening, and concatenated the features representing every 3 consecutive audio samples, in order to align the audio and the video data. In order to train the CorrRNN model, we generated sequences with length 8 using a stride of 2. Training was performed using stochastic gradient descent with the size of the mini-batch set to 32. The number of hidden units in the hidden layers was set to 512. After training the model in an unsupervised manner, the joint representation generated by CorrRNN is treated as the fused feature. Similar to [7], we first break down the fused features of each speaking example into one and three equal slices and perform mean-pooling over each slice. The mean-pooled features for each slice are then concatenated and used to train a linear SVM classifier in a supervised manner.

4.2.2 Results

Table 4 showcases the classification performance of the proposed CorrRNN model using the Corr-DW configuration on the AVLetters and the CUAVE datasets. The fused representation of the audio-video data generated using the CorrRNN model is used to train and test an SVM classifier. We observe that the CorrRNN representation leads to more accurate classification than the representation generated by non-temporal models such as Multimodal deep autoencoder (MDAE), multimodal deep belief networks (MDBN), and the multimodal deep Boltzmann machines (MDBM). This is because the CorrRNN model is able to learn the temporal dependencies between the two modalities. CorrRNN also outperforms conditional RBM (CRBM), and RTM-RBM models due to the incorporation of the correlational loss and the dynamic weighting mechanism.

The CorrRNN model also produces rich representations for each modality, as demonstrated in the cross-modality and shared representation learning experimental results in Table 5. Indeed, there is a significant improvement in accuracy from using CorrRNN features relative to the scenarios where only the raw features for both audio and video modalities are used, and this improvement holds for both the datasets. For instance, the accuracy improves by more than two times on the CUAVE dataset by learning the video features with both audio and video, compared to learning only with the video features. In the shared representation learning experiments, we learn the feature representation using both the audio and video modalities, but the supervised training and testing are performed using different modalities. The results show that the CorrRNN model captures the correlation between the modalities very well.

In order to evaluate the robustness of the CorrRNN model to noise, we added white Gaussian noise at 0dB SNR to the original audio signal in the CUAVE dataset. Unlike prior models whose performance degrades significantly ($12 - 20\%$) due to presence of noise , there is only a minor decrease of about 5% in the accuracy of the CorrRNN model, as shown in Table 6. This may be ascribed to the richness of the cross-modal information embedded in the fused representation learned by CorrRNN.

5. Conclusions

In this paper, we have proposed CorrRNN, a new model for multimodal fusion of temporal inputs such as audio, video and sensor data. The model, based on an Encoder-Decoder framework, learns joint representations of the multimodal input by exploiting correlations across modalities. The model is trained in an unsupervised manner (*i.e.*, by minimizing an input-output reconstruction loss term and maximizing a cross-modality-based correlation term) which obviates the need for labeled data, and incorporates GRUs

Method	Accuracy	
	AVLetters	CUAVE
MDAE [13]	62.04	66.70
MDBN [21]	63.2	67.20
MDBM [21]	64.7	69.00
RTMRBM [7]	66.04	-
CRBM [1]	67.10	69.10
CorrRNN	**83.40**	**95.9**

Table 4. Classification performance for audio-visual speech recognition on the AVLetters and CUAVE datasets, compared to the best published results in literature, using the fused representation of the two modalities.

	Train /Test	Method	Accuracy	
			AVLetters	CUAVE
Cross-modality learning	Video /Video	Raw	38.08	42.05
		CorrRNN	81.85	96.22
	Audio /Audio	Raw	57.31	88.32
		CorrRNN	85.33	96.11
Shared representation learning	Video /Audio	MDAE	-	24.30
		CorrRNN	85.33	96.77
	Audio /Video	MDAE	-	30.70
		CorrRNN	81.85	96.33

Table 5. Classification accuracy for the cross-modality and shared representation learning settings. MDAE results from [13].

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 6. Classification accuracy for audio-visual speech recognition on the CUAVE dataset, under clean and noisy audio conditions. White Gaussian noise is added to the audio signal at 0dB SNR. Baseline results from [13].

to capture long-term dependencies and temporal structure in the input. We also introduced a dynamic weighting mechanism that allows the encoder to dynamically modify the contribution of each modality to the feature representation being computed. We have demonstrated that the CorrRNN model achieves state-of-the-art accuracy in a variety of temporal fusion applications. In the future, we plan to apply the model to a wider variety of multimodal learning scenarios. We also plan to extend the model to seamlessly ingest asynchronous inputs.

References

[1] M. R. Amer, B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney. Multimodal fusion using dynamic hybrid models. In *IEEE Winter Conference on Applications of Computer Vision*, pages 556–563. IEEE, 2014.

[2] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1247–1255, 2013.

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *in ICLR 2015*, abs/1409.0473, 2014.

[4] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran. Correlational neural networks. *Neural computation*, 2015.

[5] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734, 2014.

[6] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[7] D. Hu, X. Li, et al. Temporal multimodal learning in audiovisual speech recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3574–3582, 2016.

[8] A. K. Katsaggelos, S. Bahaadini, and R. Molina. Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE*, 103(9):1635–1653, 2015.

[9] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 595–603, 2014.

[10] J. Kumar, Q. Li, S. Kyal, E. Bernal, and R. Bala. On-the-fly hand detection training with application in egocentric action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–27, 2015.

[11] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):198–213, 2002.

[12] N. Neverova, C. Wolf, G. Taylor, and F. Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016.

[13] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[14] F. J. Ordóñez and D. Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[15] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy. Cuave: A new audio-visual database for multimodal human-computer interface research. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II–2017. IEEE, 2002.

[16] J. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan. Look, listen and learn-a multimodal lstm for speaker identification. *arXiv preprint arXiv:1602.04364*, 2016.

[17] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. The av+ ec 2015 multimodal affect recognition challenge: Bridging across audio, video, and physiological data. In *Proceedings of the 5rd ACM International Workshop on Audio/Visual Emotion Challenge. ACM*, 2015.

[18] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, et al. Collecting complex activity datasets in highly rich networked sensor environments. In *Networked Sensing Systems (INSS), 2010 Seventh International Conference on*, pages 233–240. IEEE, 2010.

[19] K. Sohn, W. Shang, and H. Lee. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*, pages 2141–2149, 2014.

[20] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. *arXiv preprint arXiv:1502.04681*, 2015.

[21] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

[22] C. Sui, M. Bennamoun, and R. Togneri. Listening with your eyes: Towards a practical visual speech recognition system using deep boltzmann machines. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 154–162, 2015.

[23] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[25] W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1083–1092, 2015.