Cross-modal Memory Fusion Network for Multimodal Sequential Learning with Missing Values

Chen Lin(✉), Joyce C. Ho, and Eugene Agichtein

Emory University, Atlanta, GA 30322, USA
{chen.lin,joyce.c.ho,eugene.agichtein}@emory.edu

Abstract. Information in many real-world applications is inherently multi-modal, sequential and characterized by a variety of missing values. Existing imputation methods mainly focus on the recurrent dynamics in one modality while ignoring the complementary property from other modalities. In this paper, we propose a novel method called cross-modal memory fusion network (CMFN) that explicitly learns both modal-specific and cross-modal dynamics for imputing the missing values in multi-modal sequential learning tasks. Experiments on two datasets demonstrate that our method outperforms state-of-the-art methods and show its potential to better impute missing values in complex multi-modal datasets.

Keywords: Multi-modal information · Sequential learning · Missing value imputation · Recurrent neural networks

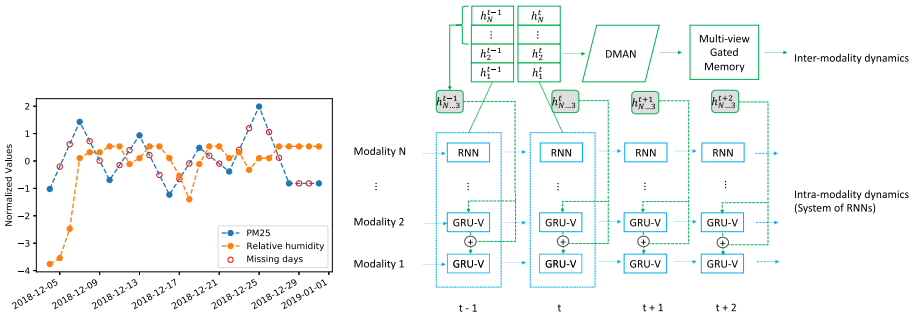
1 Introduction and Related Work

1.1 Introduction

In many real-world scenarios, information and data are multi-modal (e.g. heterogeneous features collected from multi-typed sensors for air quality surveillance [1,8,20]; and multi-modal perception for face-to-face communication [16,19]). In these scenarios, features from different modalities are seamlessly used together for classification/regression purposes. However, multi-modal sequential data is often incomplete due to various reasons, such as broken sensors, failed data transmission or low sampling rate. For example, Fig. 1a shows two time series of air quality data at Atlanta Fire Station #8, where two-thirds of fine particulate matter ($PM_{2.5}$) data is missing while relative humidity data is complete. Relative humidity data, as shown in Fig. 1a, is promising for improving daily $PM_{2.5}$ surveillance because of its high correlation and low missing rate. Many previous studies [2,3,13,15] have been developing models that could impute missing values in multivariate sequential data by either constructing local statistics or utilizing local and global recurrent dynamics. Although these methods have achieved remarkable success in multivariate sequential data of one modality, they can not

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12657, pp. 312–319, 2021.
https://doi.org/10.1007/978-3-030-72240-1_30

be naturally adapted to multi-modal sequential data. Specifically, they are not designed to incorporate the information from modalities with lower missing rates for imputing the missing values of modalities with higher missing rates.



(a) Two time series from Air Quality dataset (b) An illustration of CMFN

Fig. 1. Two time series from $PM_{2.5}$ monitoring station at Atlanta Fire Station #8 (left) and an illustraion of CMFN (right).

Previous studies [9,16,17] in multi-modal sequential learning have been proved successful in exploring intra-modality and inter-modality dynamics for more robust and accurate prediction. The strategies for multi-modal sequential learning can be classified into three categories. The first strategy is early fusion, which simply concatenates multi-modal features at the input level [10,12]. This fusion strategy could not efficiently model the intra-modality dynamics because the complex inter-modality dynamics can dominate the learning process or result in overfitting. The second strategy is late fusion, which trains unimodal classifiers independently and performs decision voting [14,19]. This strategy could lead to inefficient exploration of inter-modality dynamics by relying on the simple weighted averaging of multiple classifiers. The last strategy is to design models that could learn both the intra-modality and inter-modality end-to-end [9,16,17]. It has been shown that by exploring the consistency and complementary properties of different modalities, the third strategy is a more effective and promising way of multi-modal sequential learning. However, there is few studies examining the condition when there are missing values in one or more modalities and how to leverage the intra-modality and inter-modality dynamics for missing value imputation remains an under-explored problem.

To address the aforementioned problems, we propose a novel cross-modal memory fusion network (CMFN) for multi-modal sequential learning with missing values. CMFN extends the memory fusion network [17], where recurrent neural networks (RNNs) are leveraged for learning intra-modality dynamics and attention-based modules are leveraged for learning inter-modality dynamics. Since the original RNN is unable to handle incomplete input, we introduced a novel variant of gated recurrent units (GRU) [5] called GRU-V to impute the

missing values by leveraging modal-specific and cross-modal dynamics. The main contributions of the paper are:

- We study a new problem of multi-modal sequential learning with missing values by leveraging intra-modality and inter-modality dynamics.
- We propose a novel framework CMFN, with a GRU-V module to impute missing values in multi-modal sequential learning.
- We conduct experiments on both real-world datasets and synthetic datasets to validate the proposed approach.

1.2 Related Work

We now briefly review related work to place our contribution in context.

Multivariate Sequential Learning with Missing Values. A variety of imputation methods such as statistical imputation (e.g., mean, median), EM-based imputation [11], K-nearest neighborhood [6] and tensor factorization [4] have been applied to estimate missing values. However, these approaches fail to model the sequential pattern of data and are independent of the training process, which often leads to sub-optimal results. To tackle this issue, recent studies [2,3,13] propose end-to-end frameworks that jointly estimate missing values and make the prediction. For example, Che *et al.* [3] introduced the GRU-D model to impute missing values in a single modality using the linear combination of statistical features, which is under strong assumptions that missing values could be learned by assigning weights between the last observed value and statistical mean value.

Multi-modal Sequential Learning. Previous studies dealing with multi-modal sequential data have largely focused on three major types of models as mentioned in Sect. 1.1. The third category of models [9,17,18] relies on collapsing the time dimension from sequences by learning a temporal representation for each of the different modalities. Memory fusion network (MFN) [17] is one of these models, which uses a special attention mechanism called the Delta-memory Attention Network (DMAN) and a Multi-view Gated Memory to identify the cross-modal interactions. Experiments show that these models [16–18] achieve remarkable success on a variety of tasks, including multi-modal sentiment analysis and emotion recognition; however, none of them can handle input with missing values in one or more modalities.

2 Methodology

In this section, we first define the problem setting, and then we present the model architecture in detail.

2.1 Problem Formulation

The input is multi-modal sequential data with $N \geq 2$ modalities. For those N modalities, we order them from high missing rate to low missing rate as modality 1, modality 2, ..., modality N. For each modality k, the input data is denoted as $X_k = \left[x_k^t : t \leq T, x_k^t \in R^{d_{x_k}}\right]$, where d_{x_k} is the input dimensionality of modality k. We also input the masking matrix $M_k = \{m_1, m_2, \ldots, m_t\}, m_i \in \{0,1\}^{d_{x_k}}$ to denote missing status ($m = 0$ means missing) and the time interval matrix $D_k = \{\delta_1, \delta_2, \ldots, \delta_t\}, \delta_i \in R^{d_{x_k}}$ to denote the number of time steps since last observation.

2.2 Model Architecture

The Cross-modal Memory Fusion Network (CMFN) is a recurrent model for multi-modal sequential learning with missing values, which consists of two main components: 1) A system of RNNs consisting of multiple RNNs for learning intra-modality dynamics. 2) DMAN and Multi-view Gated Memory [17] for learning inter-modality dynamics. As shown in Fig. 1b, RNNs such as GRU and long short-term memory (LSTM) [7] are applied for modalities without missing values, GRU-V is applied for imputing the missing values with intra-modality and inter-modality dynamics for modalities with missing values.

GRU-V is inspired by the structure of GRU-D proposed by Che *et al.* [3]. To explain the procedure of missing value imputation, we assume that the input for modality 1 is feature matrix X_1, masking matrix M_1 and time interval matrix D_1. As shown in Fig. 1b, at time step t, for the $N-1$ modalities with lower missing values, we concatenate their hidden outputs $\{h_2^t, h_3^t, \ldots, h_{N-1}^t\}$ as $h_{N\ldots2}^t$ to represent cross-modal dynamics. For modality 1, we have the hidden output h_1^{t-1} at last time step to represent modal-specific dynamics. We then concatenate the cross-modal and modal-specific dynamics, denoted as $c^{[h_1^{t-1}, h_{N\ldots2}^t]}$, and pass the concatenated tensor to a neural network $\mathcal{D}_v : R^{d_c} \mapsto R^{d_{x_1}}$ to infer the variance of the missing values from its empirical mean $\tilde{X_1}$ in modality 1 as:

$$V_{X_1}^t = \mathcal{D}_v\left(c^{[h_1^{t-1}, h_{N\ldots2}^t]]}\right) \tag{1}$$

$V_{X_1}^t$ are softmax activated scores, which is then used to infer the missing values as:

$$\mathcal{X}_1^t = \tilde{X_1} + 2K \cdot (V_{X_1}^t - 0.5) \tag{2}$$

$\mathcal{X}_1^t$ are the inferred values, and we rescale $V_{X_1}^t$ from $[0,1]$ to $[-K, K]$ using rescale parameter K. Because all the input values are normalized, we set $K = 3$ to represent the variance of input values. Following GRU-D, we then use a weight decay function $\mathit{\Gamma}_{D_1^t}$ to assign weights between the last observed value $X_1^{t'}$ and the inferred value $\mathcal{X}_1^t$ and get final imputed value $\hat{X}_1^t$ as:

$$\mathit{\Gamma}_{D_1^t} = \exp\left\{-\max\left(\tilde{\Gamma}, W_{\mathit{\Gamma}} D_1^t + b_{\mathit{\Gamma}}\right)\right\} \tag{3}$$

$$\hat{X}_1^t = \Gamma_{D_1^t} X_1^{t'} + (1 - \Gamma_{D_1^t}) \cdot \mathcal{X}_1^t \tag{4}$$

where W_Γ and b_Γ are model parameters that we train jointly with other parameters of the GRU. $\tilde{\Gamma}$ is the default weight decay, which is set as a hyper-parameter in range $[0, 1]$.

3 Experiments

In this section, we describe experiments in four parts. First, we describe the datasets. Second, we present the baseline models. Then we describe the experimental setup. Last, we summarize experimental results comparing with state-of-the-art baselines.

3.1 Datasets

Air Quality Dataset. Air Quality dataset is time series of daily measurement of $PM_{2.5}$ and meteorological data (i.e. relative humidity and temperature) in Atlanta Fire Station #8 monitoring site from Jan 1, 2011 to Dec 31, 2018. This dataset consists of two modalities and it facilitates a regression task of predicting $PM_{2.5}$ concentration based on data of the past 7 days.

CMU-MOSI Dataset. Multimodal Opinion Sentiment Intensity (CMU-MOSI) dataset [19] is a collection of 93 opinion videos from online sharing websites with three modalities: language, vision, and acoustic. Each video consists of multiple opinion segments and each segment is annotated with sentiment in the range $[-3, 3]$. This benchmark dataset facilitates three prediction tasks: 1) Binary Sentiment classification 2) Seven-Class sentiment classification 3) Sentiment regression in range $[-3, 3]$. This dataset contains no missing values, so we synthetically introduce missing values by randomly masking 50% percent of the values in acoustic modality. We construct the synthetic datasets in two ways to test our model under different conditions. Synthetic Dataset #1: For 5 features in acoustic modality, We randomly mask values separately, which means this modality is partly masked when selected. Synthetic Dataset #2: We mask values for all 5 features randomly, which means this modality is masked totally when selected.

3.2 Baseline Models

Here, we use the following models for baselines and ablation studies.

- EFLSTM: LSTM model using early fusion strategy. The missing values are simply imputed by the last observed values and all modalities are concatenated into a single modality at the input level.
- MFN: State-of-the-art multi-modal learning model that learns the temporal representation for each modality using an RNN. The missing values are simply imputed by the last observed values.

– GRU-D: Baseline for multivariate sequential learning with missing values. All modalities are concatenated into a single modality using early fusion method at the input level.
– MFN-GRUD: This model is proposed for the ablation study and the RNNs in MFN are replaced with the GRU-D. Thus, it is a multi-modal learning architecture that imputes the missing values based only on intra-modality dynamics.

3.3 Experimental Setup

For the Air Quality dataset, we split the training (2011–2016), validation (2017) and testing (2018) sets chronologically. For the CMU-MOSI dataset, there are 1284, 229, and 686 samples in the training, validation, and testing sets respectively. We implement our models using Pytorch[1]. For all the experiments, the batch size is set to be 32 and all the parameters are tuned by the validation dataset.

3.4 Performance Comparison

Table 1. Comparison with state-of-the-art approaches for multi-modal sequential learning with missing values.

Task	Air Quality		CMU-MOSI Dataset #1					CMU-MOSI Dataset #2				
Metric	MAE	MSE	BA	F1	MA(7)	MAE	r	BA	F1	MA(7)	MAE	r
ELLSTM	3.19	15.5	0.726	0.725	0.325	1.051	0.584	0.739	0.735	0.343	1.021	0.623
MFN	3.17	15.35	0.739	0.735	0.322	1.012	0.618	0.749	0.745	0.327	1.008	0.616
GRUD	3.13	15.22	0.739	0.738	0.294	1.037	0.620	0.755	0.750	0.331	**0.957**	0.652
MFN-GRUD	3.07	14.8	0.736	0.729	0.321	**0.996**	**0.621**	0.755	0.753	**0.354**	0.987	0.626
CMFN	**3.04**	**14.21**	**0.755**	**0.751**	**0.354**	1.007	0.615	**0.767**	**0.759**	0.353	0.958	**0.660**

Table 1 summarizes the comparison between CMFN and proposed baselines for all the multi-modal sequential learning tasks. For the regression tasks, we report mean absolute error (MAE), mean squared error (MSE) and Pearson's correlation r. For binary classification, we report binary accuracy (BA) and binary F1 score. For multiclass classification, we report multiclass accuracy MA(k) where k denotes the number of classes. The results show that CMFN outperforms all the baseline methods in 8/12 tasks. For the CMU-MOSI dataset, when the features in acoustic modality are either partly missing (Dataset #1) or completely missing (Dataset #2), CMFN can robustly impute the missing values and outperform the compared methods. For the ablation study, the difference between CMFN and MFN-GRUD is that the latter only uses intra-modality dynamics for missing value imputation. The results show that CMFN outperforms MFN-GRUD in 9/12 tasks, which suggests that cross-modal dynamics can improve the missing value imputation performance.

[1] https://pytorch.org.

4 Conclusion

In this paper, we investigate a novel problem of exploring intra-modality and inter-modality dynamics for multi-modal sequential learning with missing values. We propose a new framework CMFN, which adopts modality-specific and cross-modal information for imputing missing values. To validate the framework, we instantiated a setup incorporating real-world data and synthetic data on benchmark multi-modal learning data. Our result outperforms existing state-of-the-arts models, with ablation studies to show architectural advantages.

References

1. Cabaneros, S.M.S., Calautit, J.K., Hughes, B.R.: A review of artificial neural network models for ambient air pollution prediction. Environ. Modell. Software **119**, 285–304 (2019). https://doi.org/10.1016/j.envsoft.2019.06.014
2. Cao, W., Wang, D., Li, J., Zhou, H., Li, L., Li, Y.: BRITS: Bidirectional Recurrent Imputation for Time Series. arXiv (2018)
3. Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y.: Recurrent neural networks for multivariate time series with missing values. Sci. Rep. **8**(1), 6085 (2018). https://doi.org/10.1038/s41598-018-24271-9
4. Chen, X., He, Z., Chen, Y., Lu, Y., Wang, J.: Missing traffic data imputation and pattern discovery with a Bayesian augmented tensor factorization model. Transp. Res. Part C: Emerging Technol. **104**, 66–77 (2019). https://doi.org/10.1016/j.trc.2019.03.003
5. Cho, K., Merrienboer, B.v., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches. arXiv (2014)
6. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning, vol. 1. Springer, New York (2001). https://doi.org/10.1007/978-0-387-21606-5
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
8. Li, V.O.K., Lam, J.C.K., Chen, Y., Gu, J.: deep learning model to estimate air pollution using M-BP to fill in missing proxy urban data. In: GLOBECOM 2017–2017 IEEE Global Communications Conference, pp. 1–6 (2017). https://doi.org/10.1109/glocom.2017.8255004
9. Liang, P.P., Liu, Z., Tsai, Y.H.H., Zhao, Q., Salakhutdinov, R., Morency, L.P.: Learning representations from imperfect time series data via tensor rank regularization. arXiv (2019)
10. Morency, L.P., Mihalcea, R., Doshi, P.: Towards multimodal sentiment analysis: harvesting opinions from the web, pp. 169–176 (2011). https://doi.org/10.1145/2070481.2070509
11. Nelwamondo, F.V., Mohamed, S., Marwala, T.: Missing data: a comparison of neural network and expectation maximization techniques. Current Sci. 1514–1521 (2007)
12. Poria, S., Chaturvedi, I., Cambria, E., Hussain, A.: Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 439–448 (2016). https://doi.org/10.1109/icdm.2016.0055

13. Tang, X., Yao, H., Sun, Y., Aggarwal, C., Mitra, P., Wang, S.: Joint modeling of local and global temporal dynamics for multivariate time series forecasting with missing values (2019)
14. Wang, H., Meghawat, A., Morency, L.P., Xing, E.P.: Select-additive learning: improving generalization in multimodal sentiment analysis. arXiv (2016)
15. Yi, X., Zheng, Y., Zhang, J., Li, T.: St-mvl: filling missing values in geo-sensory time series data (2016)
16. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. arXiv (2017)
17. Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E., Morency, L.P.: Memory fusion network for multi-view sequential learning. arXiv (2018)
18. Zadeh, A., Mao, C., Shi, K., Zhang, Y., Liang, P.P., Poria, S., Morency, L.P.: Factorized multimodal transformer for multimodal sequential learning. arXiv (2019)
19. Zadeh, A., Zellers, R., Pincus, E., Morency, L.P.: MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv (2016)
20. Zhao, X., Zhang, R., Wu, J.L., Chang, P.C.: A deep recurrent neural network for air quality classification. J. Inf. Hiding Multimed. Sig. Proc. **9**, 346–354 (2018)