Collaborative Attention Mechanism for Multi-Modal Time Series Classification

Yue Bai* Zhiqiang Tao† Lichen Wang‡ Sheng Li§ Yu Yin¶ Yun Fu‖

Abstract

Multi-modal time series classification (MTC) uses complementary information from different modalities to improve the learning performance. Obtaining informative modality-specific representation plays an essential role in MTC. Attention mechanism has been widely adopted as an effective strategy for discovering discriminative cues underlying temporal data. However, most existing MTC methods only utilize attention to balance the feature weights within or cross modalities but ignore digging latent patterns from mutual-support information in attention space. Specifically, the attention distributions are different for multiple modalities which are supportive and instructional with each other. To this end, we propose a collaborative attention mechanism (CAM) for MTC based on a novel perspective to utilize attention module. CAM detects the attention differences among multi-modal time series, and adaptively integrates different attention information to benefit each other. We extend the long short-term memory (LSTM) to a Mutual-Aid RNN (MAR) for multi-modal collaboration. CAM takes advantages of modality-specific attention to guide another modality and discover potential information which is hard to be explored by itself. It paves a novel way of employing attention to enhance the capacity of multi-modal representations. Extensive experiments on four multi-modal time series datasets illustrate the CAM effectiveness to improve the single-modal and also boost multi-modal performances.

1 Introduction

Multi-model time series classification (MTC) has drawn more attention since the increasing usage of multi-modal sensors to improve classification performance in several data mining applications [3, 24]. Further, several algorithms are designed to explore multi-modal time series analysis [22, 1]. However, MTC is still a challenging task due to the difficulties: (1) how to represent modality-specific information, especially for temporal data with dynamic patterns; (2) how to utilize them for achieving better multi-modal performance.

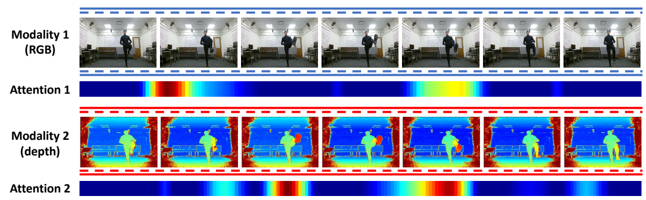

Figure 1: Illustration of the differences among multiple modalities in the attention space. We take a "kicking" motion sample with RGB and depth modalities as an example. RGB modality easily captures visible color changes between frames, and the depth modality is more sensitive to the changes in depth distance. In RGB, the visible changes are obvious during the lifting leg and drawing back the leg. However, in the middle of example sequence, the changes of RGB are tiny and hard to be discovered, whereas the depth modality illustrates significant changes during this period, because the changes are mainly in depth dimension when the leg is at the top position.

Subspace learning is widely used to seek a common subspace for multiple modalities [8, 9]. It aims to find consistent characteristics among multi-modal and derive robust representations. However, emphasizing the synchronous patterns may overlook the distinctive information of each modality. Besides, fusion mechanism is another popular way for multi-modal learning [18, 23]. Utilizing effective fusion takes advantage of the distinctive information from each modality and combine them for encouraging higher performance. However, some straightforward fusion methods (e.g., average, concatenation, and summation) may not fully exploit multi-modal data and hurt the final result. On the one hand, early fusion methods pay more attention on augmenting the capacity of each modality by borrowing information from the other modalities [23]. They integrate the multi-modal information in feature space. On the other hand, late fusion algorithms explore distinctive modality-specific decision in label space [17]. The mutual-support information across modalities are utilized by wisely fusing the predicted scores. Assisted by attention modules, some well-designed learnable weights are assigned to each modality individually [4] or cross-modal learning [7]. However, for both early and late

*bai.yue@northeastern.edu, Northeastern University
†ztao@scu.edu, Santa Clara University
‡wanglichenxj@gmail.com, Northeastern University
§sheng.li@uga.edu, University of Georgia
¶yin.yu1@northeastern.edu, Northeastern University
‖yunfu@ece.neu.edu, Northeastern University

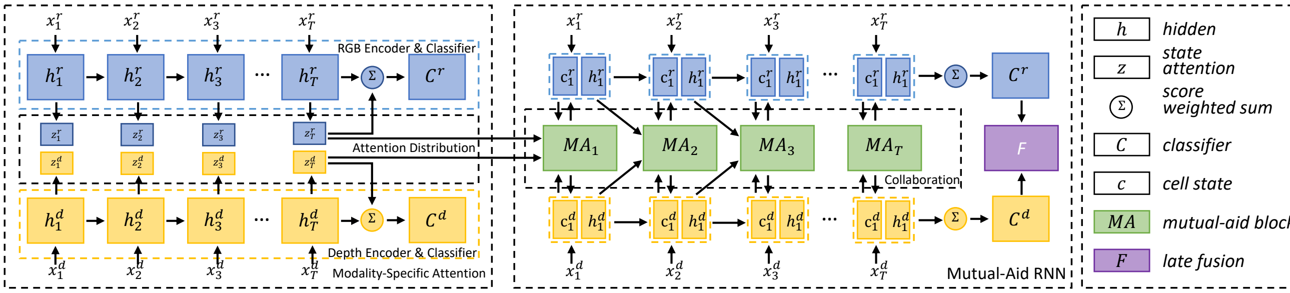

Figure 2: Illustration of our CAM framework. The Modality-Specific Attention is the first-stage. Two LSTM encoders make modality-specific classification, respectively. The attention score z_t of two modalities are fed into the second stage. Mutual-Aid RNN achieves the multi-modal collaboration in the second stage. Mutual-Aid block (MA_t) collaborates the multiple modalities step-by-step in temporal dimension. After multi-modal collaboration, we deploy an attention module for each modality and make a late fusion for multi-modal result.

fusion strategies, most existing attention-enhanced approaches benefit multi-modal learning only by exploiting the readily available information directly. For example, the late fusion uses attention to balance predicted scores from each modality for final decision. Similarly, the early fusion constructs a common feature space projected from different original modality-specific feature spaces adjusted by learnable attention weights. Although the attention module is used in different representation spaces, both of them utilize weighted summation to coordinate the contribution from each single modality. They ignore that how to excavate latent information from the attention distributions of different modalities, which can be seen as a guidance information to benefit multi-modal learning.

To discover the latent cues from attention distributions across multiple modalities, we propose a Collaborative Attention Mechanism (CAM) model for MTC as shown in Fig. 2. Attention effectively enhances the representation learning accompanied with the capacity of interpreting model and providing intuitions of data. Inspired by the interpretability of temporal attentions, we instantiate CAM based on the observations from multi-modal time series: *different modalities have different attention distributions* (see Fig. 1). Specifically, taking human motion data as an example, the RGB modality pays attention to certain video frames; the depth modality values more contributions from some other frames. Each modality has its own concentrations, yet ignoring the frames that are hard to explored by itself. However, the ignored time steps reserving valuable patterns also deserves to be investigated. To disclose the overlooked information, we propose a Mutual-Aid RNN (MAR) cell to collaboratively guide multi-modal representation learning. Specifically, one modality utilizes the attention differences and selectively directs the other modality to focus on certain temporal steps containing obscure information. In this way, the previous overlooked temporal steps of one modality can be revisited helped from the other modality and its extracted feature will be enhanced. Please note that these clues are still from the modality itself instead of borrowing from other modality, but discovered based on guidance from the other modality. Leveraging on this mechanism, single-modal performance is improved and the multi-modal performance is also boosted. Different from conventionally using attention to adjust fusion for multi-modal data, we fully exploit the multi-modal attention distributions to achieve multi-modal collaboration. It is motivated by the interpretability of attention and naturally developed to expand a new way to analyze multi-modal data. To the best of our knowledge, we are the first to go deeply into the attention differences and explore multi-modal time series analysis from this novel perspective. We summarize our contribution as below:

- We propose a collaborative attention mechanism (CAM) framework to improve the multi-modal time series classification (MTC) performance. It effectively utilizes the attention information across different modalities to mutually enhance multi-modal learning, which boosts the single-modal and multi-modal performance simultaneously.
- A novel Mutual-Aid RNN (MAR) cell is proposed for multi-modal time series. It relies on attention distribution to capture the latent patterns and adaptively enhance the temporal representation of each modality.
- We provide a new perspective to reacquaint multi-modal learning by leveraging the interpretability of attention mechanism to guide learning process. Extensive experiments on four multi-modal time series datasets illustrate the effectiveness of the proposed CAM.

2 Methodology

Let $X^1 \in \mathbb{R}^{T \times d^1}$ and $X^2 \in \mathbb{R}^{T \times d^2}$ are multi-modal feature inputs. T represents the length of time series. d^1 and d^2 are feature dimensions of two modalities. $Y \in \mathbb{R}^C$ is the one-hot label vector, where C is the number of classes. The first phase contains modality-specific encoders and classifiers. We use LSTM with self-attention to encode the sequence input and obtain the attention information. The training is supervised by the label information. In the second phase, our CAM utilizes the modality-specific attention distributions from the first phase to achieve the multi-modal collaboration process. In this way, the modality-specific representation is enhanced to obtain higher single-modal performance. After that, we use a correlative late fusion to obtain multi-modal result which is boosted by the enhanced single-modal representations.

2.1 Attention for Time Series Given an time series sample and the corresponding label, the temporal attention model aims to encode the sequential input and optimize the following objective:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{(\boldsymbol{X},\boldsymbol{y})} \log p(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{\theta}), \tag{2.1}$$

where $\boldsymbol{\theta}$ is the set of parameters of model. $\boldsymbol{X} = \{\boldsymbol{x_1}, ..., \boldsymbol{x_t}\}$ is the multiple steps of one time series sample, and $\boldsymbol{y}$ is the corresponding label. The dynamic information is the key factor for classification. Thus, wisely choosing temporal encoder is decisive for temporal feature extraction. In our work, we deploy long short-term memory (LSTM) [6] to model sequential data. Each input step $\boldsymbol{x_t}$ is encoded as a hidden representation $\boldsymbol{h_t}$, and the cell state $\boldsymbol{c_t}$ is updated correspondingly. The LSTM update processes are given by

$$\begin{aligned}
\boldsymbol{f_t} &= \sigma_g(W_f \boldsymbol{x_t} + U_f \boldsymbol{h_{t-1}} + b_f),\\
\boldsymbol{i_t} &= \sigma_g(W_i \boldsymbol{x_t} + U_i \boldsymbol{h_{t-1}} + b_i),\\
\boldsymbol{o_t} &= \sigma_g(W_o \boldsymbol{x_t} + U_o \boldsymbol{h_{t-1}} + b_o),\\
\boldsymbol{c_t} &= \boldsymbol{f_t} \circ \boldsymbol{c_{t-1}} + \boldsymbol{i_t} \circ \sigma_c(W_c \boldsymbol{h_t} + U_c \boldsymbol{h_{t-1}} + b_c),\\
\boldsymbol{h_t} &= \boldsymbol{o_t} \circ \sigma_h(\boldsymbol{c_t}),
\end{aligned} \tag{2.2}$$

where $\boldsymbol{f_t}$, $\boldsymbol{i_t}$, $\boldsymbol{o_t}$, $\boldsymbol{c_t}$, and $\boldsymbol{h_t}$ represent forget gate, input gate, output gate, cell state, and hidden state at time t, respectively. $\boldsymbol{c_{t-1}}$ and $\boldsymbol{h_{t-1}}$ are cell and hidden states at time $t-1$. σ_g, σ_c, and σ_h are activation functions. $\circ$ represents the element-wise product. W, U, and b are learnable parameters.

Original temporal sequence $\boldsymbol{X}$ is encoded as $\boldsymbol{H} = \{\boldsymbol{h_1}, ..., \boldsymbol{h_T}\}$. Commonly, we pick the last hidden state $\boldsymbol{h_T}$ to represent the whole sequence. However, it may lose temporal information to some degree. A reasonable way is using the weighted summation of $\boldsymbol{h_t}$. The weights are calculated based on the importance of each temporal step by attention mechanism. Here, we adopt a self-attention variant [21] which is proposed for document classification. It can be easily utilized for modeling temporal data and given by

$$\begin{aligned}
\boldsymbol{u_t} &= tanh(W_w \boldsymbol{h_t} + b_w),\\
\boldsymbol{z_t} &= \frac{exp(\boldsymbol{u_t}^T \boldsymbol{u_w})}{\sum_t exp(\boldsymbol{u_t}^T \boldsymbol{u_w})},\\
\boldsymbol{r} &= \sum_t \boldsymbol{z_t} \boldsymbol{h_t},
\end{aligned} \tag{2.3}$$

where $\boldsymbol{u_t}$ denotes the attention vector derived from $\boldsymbol{h_t}$. W_w and b_w are learnable parameters. $\boldsymbol{u_w}$ is the context vector, which is random initialized and updated through the optimization procedure. It depicts the global meaning of the temporal sequence itself. $\boldsymbol{z_t}$ means the degree of importance for each $\boldsymbol{u_t}$ among the whole temporal context $\boldsymbol{u_w}$ by using softmax activation. $\boldsymbol{r}$ is the weighted summation of $\boldsymbol{h_t}$.

To introduce our CAM clearly, we go deeper to provide more insights about LSTM. The key factor of LSTM cell is the $\boldsymbol{c_t}$. It reflects memory states of the whole sequence. $\boldsymbol{f_t}$ and $\boldsymbol{i_t}$ update the $\boldsymbol{c_t}$ internally through the *forget* and *input* procedures. The contents of *forget*/*input* are derived from current input $\boldsymbol{x_t}$ and last hidden state $\boldsymbol{h_{t-1}}$. The content of current hidden state $\boldsymbol{h_t}$ is also extracted from $\boldsymbol{x_t}/\boldsymbol{h_{t-1}}$, then filtered by $\boldsymbol{c_t}$. All information flows cross several control gates center on the $\boldsymbol{c_t}$. As the memory state, $\boldsymbol{c_t}$ only records the temporal dynamic characteristic instead of specific domain knowledge. To this end, we conclude that fully exploiting the cell state c_t is decisive for informative temporal encoding. We introduce our framework starting from the temporal attention and cell state c_t.

2.2 Modality-Specific Attention Multi-modal time series contain mutual-support information for each other, however, each modality has its own distinctive patterns. To fully exploit the distinctive information from each modality, we utilize the modality-specific attention as follows:

$$\begin{aligned}
\boldsymbol{H^v} &= E^v(\boldsymbol{X^v}, \phi_E^v),\\
\boldsymbol{r^v} &= Q^v(\boldsymbol{H^v}, \phi_Q^v),\\
\hat{Y}_a^v &= C^v(\boldsymbol{r^v}, \phi_C^v),
\end{aligned} \tag{2.4}$$

where superscript v represents the modality v. E is LSTM module (Eq. 2.2), encoding sequence $\boldsymbol{X}$ into hid-

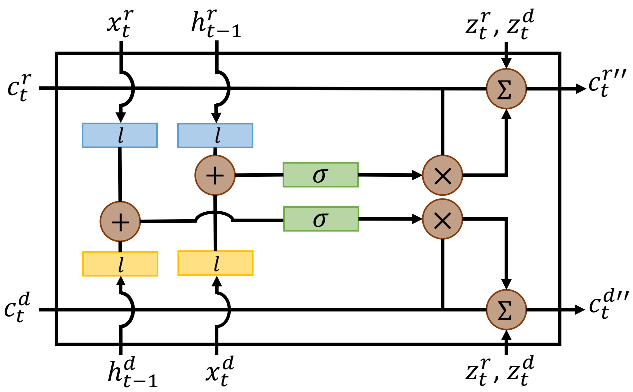

Figure 3: Illustration of the proposed Mutual-Aid RNN (MAR) cell. $\boldsymbol{x_t}$ and $\boldsymbol{h_{t-1}}$ of two modalities are set as input to integrate information. They collaboratively update the $\boldsymbol{c_t}$ of two modalities, respectively. Attention distributions $\boldsymbol{z_t}$ of two modalities are involved in weighted summation to adaptively integrate the multi-modal knowledge.

den sequence $\boldsymbol{H}$. Q is the attention module (Eq. 2.3), transferring $\boldsymbol{H}$ into weighted summation vector $\boldsymbol{r}$. C is the modality-specific linear classifier, resulting in the predicted label $\hat{Y}_a$. ϕ_E^v, ϕ_Q^v, and ϕ_C^v are learnable parameters. They are optimized using following objective:

$$L_a^v = \ell(Y, \hat{Y}_a^v), \tag{2.5}$$

where ℓ represents cross-entropy loss, Y is the ground truth. The goal of modality-specific attention aims to derive the $\boldsymbol{Z^v} = \{\boldsymbol{z_1^v}, ..., \boldsymbol{z_T^v}\}$ which is the intermediate product of Eq. 2.3 and preserves the modality-specific dynamic patterns. We regard the modality-specific attention as our first stage model. Multi-modal attention distributions $\boldsymbol{Z^v}$ are reserved for the second stage.

2.3 Multi-Modal Collaboration by Mutual-Aid RNN

To substantially take advantage of multi-modal data, in our second stage, we propose the multi-modal collaboration mechanism which is achieved by the well-designed Mutual-Aid Rnn (MAR) cell. It mutually supports each single-modal representation in multi-modal scenario. Note that unlike some data augmentation strategies cross multi-modal (e.g., representation mapping, feature fusion, and generative model), our goal is referring to attention information of the other modality to help the target modality discovering more clues by itself, instead of transferring or fusing information from others directly. To convey our insight clearly, we elaborate it as follow: Modality-specific attention provides the distinctive temporal patterns by attention distributions. It is extracted through optimizing single-modal classifier individually and reflects the modality-specific characteristics. Particularly, for the same temporal sample, the attention distribution of one modality focuses on certain time steps. On the other hand, that of the other modality focuses on different steps. This is caused by the inherent attribute of each modality. For example, in human motion temporal data, RGB modality could easily capture the color changes to recognize human motion and depth may be more sensitive to the distance variations. As a result, the effective temporal steps for two modalities could be different. However, the differences are not opposite but complementary for each other. Some steps are ignored by certain modality, due to its inherent attribute, still restore valuable information. This information may be easily discovered by the other modality. To this end, we propose the multi-modal collaboration mechanism. It encourages multi-modal data to help with each other by guiding other modality to focus on implicit but effective information by itself. We first encode the multi-modal temporal sequence $\boldsymbol{X^v}$ with LSTM (Eq. 2.2) abbreviated as follows:

$$\begin{aligned} \boldsymbol{c_t} &= \boldsymbol{f_t} \circ \boldsymbol{c_{t-1}} + \boldsymbol{i_t} \circ \sigma_c(W_c \boldsymbol{h_t} + U_c \boldsymbol{h_{t-1}} + b_c), \\ \boldsymbol{h_t} &= \boldsymbol{o_t} \circ \sigma_h(\boldsymbol{c_t}), \end{aligned} \tag{2.6}$$

where $\boldsymbol{c_t}$ and $\boldsymbol{h_t}$ are the cell state and hidden state for time t, deriving information from $\boldsymbol{c_{t-1}}$, h_{t-1}, and $\boldsymbol{x_t}$, respectively. We extend the LSTM to our proposed MAR by designing a novel recurrent cell (see Fig. 3). Instead of setting the $\boldsymbol{c_t}$ as the cell state for next time step directly, MAR guides the target modality to dig more latent information by leveraging the temporal dynamic characteristic from the other modality. Next, we formulate our proposed MAR step-by-step.

Cross-Modal Collaborator is proposed to integrate multi-modal information and prepared for following collaborative learning. It is formulated as follows:

$$\begin{aligned} \boldsymbol{G_{r\to d}} &= \sigma(W_{rd}\boldsymbol{x_t^r} + W_d \boldsymbol{h_{t-1}^d}), \\ \boldsymbol{G_{d\to r}} &= \sigma(W_{dr}\boldsymbol{x_t^d} + W_r \boldsymbol{h_{t-1}^r}), \end{aligned} \tag{2.7}$$

where W_* is learnable parameters. $\boldsymbol{G_*}$ extract information from current input $\boldsymbol{x_t^*}$ and collaborate with last hidden state $\boldsymbol{h_{t-1}^*}$ from the other modality. σ represents the sigmoid activation. The knowledge of each time step from the other modality is reserved in $\boldsymbol{G_*}$.

Mutual Filtering is designed based on cross-modal collaborator above. Cell state $\boldsymbol{c_t^*}$ contains temporal dynamic patterns for each modality. It could be updated internally in LSTM (Eq. 2.6). However,

$\boldsymbol{c_t^*}$ will only contain the memory information from single modality and cannot take advantage of temporal patterns of the other modality. Mutual filtering helps model update the $\boldsymbol{c_t^*}$ using cross-modal collaborator to derive knowledge from the other one, which is given by

$$\begin{aligned} \boldsymbol{c_t^{r\prime}} &= \boldsymbol{G_{d\rightarrow r}} \circ \boldsymbol{c_t^r}, \\ \boldsymbol{c_t^{d\prime}} &= \boldsymbol{G_{r\rightarrow d}} \circ \boldsymbol{c_t^d}, \end{aligned} \tag{2.8}$$

where $\circ$ is the point-wise product. $\boldsymbol{c_t^{*\prime}}$ are the enhanced cell states containing mutual-support temporal information from the other modality.

Mutual Collaboration is finally achieved by combining the attention distributions and two proposed modules above. Attention distributions $\boldsymbol{z_t^*}$ reflect the importance of each time step for each single-modality. Further, it also decides the information importance during updating $\boldsymbol{c_t^*}$ in multi-modal collaborative learning. We first normalize the attention scores by

$$\begin{aligned} \boldsymbol{z_t^{r\prime}} &= \frac{\boldsymbol{z_t^r}}{\boldsymbol{z_t^r} + \boldsymbol{z_t^d}}, \\ \boldsymbol{z_t^{d\prime}} &= \frac{\boldsymbol{z_t^d}}{\boldsymbol{z_t^r} + \boldsymbol{z_t^d}}. \end{aligned} \tag{2.9}$$

The original cell state $\boldsymbol{c_t^*}$ are updated by single-modal information, while $\boldsymbol{c_t^{*\prime}}$ are updated by the cross-modal collaborator $\boldsymbol{G_*}$. $\boldsymbol{z_t^{*\prime}}$ represent the importance of dynamic knowledge from different modalities. We integrate the multi-modal information for updating cell states via the weighted summation:

$$\begin{aligned} \boldsymbol{c_t^{r\prime\prime}} &= \boldsymbol{z_t^{r\prime}} \boldsymbol{c_t^r} + \boldsymbol{z_t^{d\prime}} \boldsymbol{c_t^{r\prime}}, \\ \boldsymbol{c_t^{d\prime\prime}} &= \boldsymbol{z_t^{d\prime}} \boldsymbol{c_t^d} + \boldsymbol{z_t^{r\prime}} \boldsymbol{c_t^{d\prime}}, \end{aligned} \tag{2.10}$$

where $\boldsymbol{c_t^{*\prime\prime}}$ are the final cell states containing the dynamic knowledge from multi-modal data. Through being the inputs for next time step, they bring the knowledge from the other modality to overcome the inherent drawback of each single modality. In this way, some implicit information could be discovered by each single modality via the guidance from mutual collaboration.

So far, we have introduced the proposed multi-modal collaboration via our MAR cell. Its input and output are multi-modal time series samples and sequential representations, respectively. In order to fully utilize the discovered information via our collaboration mechanism, we reuse the self-attention (Eq. 2.3) to obtain the final representation and make the modality-specific classification again similar to Eq. 2.4. We briefly formulate these steps by

$$\begin{aligned} \boldsymbol{H_M^v} &= E_M^v(\boldsymbol{X^v}, \phi_{E_M}^v), \\ \boldsymbol{r_M^v} &= Q_M^v(\boldsymbol{H_M^v}, \phi_{Q_M}^v), \\ \hat{Y}_M^v &= C_M^v(\boldsymbol{r_M^v}, \phi_M^v), \end{aligned} \tag{2.11}$$

where all the terms with subscript $\boldsymbol{M}$ represents the similar meanings with Eq. 2.4 under our multi-modal collaboration mechanism. We obtain another attention distribution $\boldsymbol{Z_M^v}$ and the predicted label $\hat{Y}_M^v$ for multi-modal results. The learnable parameters are optimized by minimizing following loss:

$$L_M^v = \ell(Y, \hat{Y}_M^v). \tag{2.12}$$

The modality-specific attention (first stage) and the multi-modal collaboration (second stage) constitute our whole framework Collaborative Attention Mechanism (CAM). It exploits the knowledge from multi-modal attention distributions to guide the multi-modal information discovering and enhance the learning process. More implicit but valuable patterns could be discovered for performance boosting. After obtaining the $\hat{Y}_M^v$ from each single modality, we use a correlative late fusion to evaluate final multi-modal performance.

2.4 Correlative Late Fusion Our CAM discovers more clues to enhance the single-modal representation. We deploy a correlative late fusion [17] for multi-modal evaluation, which is given by

$$D = \hat{Y}_M^r \cdot \hat{Y}_M^{d\top}, \tag{2.13}$$

where $\hat{Y}_M^r \in \mathbb{R}^{d^l \times 1}$ and $\hat{Y}_M^{d\top} \in \mathbb{R}^{1 \times d^l}$ are the predicted label from multiple modalities. $D \in \mathbb{R}^{d^l \times d^l}$ is the correlative matrix constructed by the multiplication of multi-modal predicted labels. D is flatten into a $d^l \times d^l$ dimension vector as input of the final classifier $C^f : \mathbb{R}^{d^l \times d^l} \rightarrow \mathbb{R}^{d^l}$. C^f is parameterized by ϕ_{C^f} and updated by minimizing following loss:

$$L_f = \ell(Y, C^f(D, \phi_{C^f})), \tag{2.14}$$

where Y is the ground truth, ℓ is the cross-entropy loss. L_f represents the final multi-modal loss.

As a summary, our model consists of the modality-specific attention and the multi-modal collaboration, followed by a late fusion model for multi-modal learning performance. The modality-specific attention aims to capture the differences among multiple modalities, especially focusing on the attention distribution. These differences are leveraged as guidance information for multi-modal collaboration. A novel MAR cell is proposed for extracting cross-modal knowledge and updating memory cell effectively. A concise late fusion is deployed to evaluate multi-modal performance. More implicit yet valuable information could be discovered by each single modality to enhance the modality-specific representation, thus to improve both single-modal and multi-modal performances.

3 Experiments

3.1 Multi-Modal Time Series Datasets Four datasets are used for evaluation: **EV-Action** [19] is a novel large-scale multi-modal human motion dataset. It contains 20 common human actions. We use the first 53 subjects with RGB and depth modalities for our experiments. Each subject performs each action 5 times and we have 5300 samples in total. We choose the first 40 subjects as training set and the rest 13 subjects as test set. **NTU RGB+D (NTU)** [16] is a popular large-scale multi-modal action dataset. It contains 56000 action clips in 60 action classes performed by 40 subjects. We use RGB and depth modalities in our evaluation. We use the cross-subject evaluation strategy in the original dataset paper, which contains 40320 samples for training and 16560 samples for test. **UWA3D Multimodal Activity II (UWA3D II)** [14, 15] contains 30 human actions performed by 10 subjects. We use RGB and depth recorded from front for evaluation. There are totally 270 samples and we randomly choose 150 for training and 120 for test. **Depth-included Human Action Dataset (DHA)** [11] is a multi-modal dataset with RGB and depth modalities. It contains 23 classes performed by 21 subjects. There are 483 samples in total. We randomly choose 240 samples for training and the rest 243 samples for test.

3.2 Comparison Methods We use seven methods for comparisons (first five are for EV-Action and NTU, and the last four for UWA3D II and DHA). **MLSTM-FCN** [10] is a novel deep framework proposed for handling multivariate temporal data. It contains a two-pathway structure (CNN and LSTM) to encode temporal data. Comprehensive patterns are captured for classification. **RC Classifier** [2] proposes a reservoir computing (RC) approach to model temporal data as vectorial representations in an unsupervised fashion. **MFN** [23] designs a memory fusion mechanism for multi-modal learning based on temporal data. It proposes an early fusion strategy to integrate multi-modal information in the feature space and improve the multi-modal performance. **GMVAR** [18] utilizes the generative strategy to mutually augment the multi-modal representations. It boosts the multi-modal learning performance significantly and improves the model robustness simultaneously. **TSN** [20] is an effective benchmark model for temporal action data. It utilizes an efficient sampling method and a two-stream structure to effectively collect valuable patterns and achieve promising performance. **AMGL** [13] is a novel multi-modal classification method based on graph learning. It aims to optimize weights for each graph automatically in a parameter-free fashion. **MLAN** [12] proposes an adaptive graph-based algorithm. It achieves the local structure and semi-supervised learning at the same time for multi-modal learning.

3.3 Implementation We use the same strategy to preprocess the raw data for four datasets. Specifically, we use TSN [20] to extract features for RGB modality using the BNInception network as backbone. Each RGB frame is extracted into 1024 dimension feature vector. The depth is transferred into RGB format first using HHA encoding algorithm [5]. Then, we use the exactly the same TSN framework to extract depth features. We arrange the length of samples with a unified number for each dataset via the cutting and repeating strategies. Concretely, for longer samples, we pick the first certain time steps and cut the rest off; for shorter samples, we repeat the whole temporal sequence several times until it reaches the target number. We set the lengths as 60, 60, 60, and 40 for EV-Action, NTU, UWA3D II, and DHA, respectively.

We concatenate the multi-modal data in feature dimension to conduct the MLSTM-FCN and RC classifier methods. The MFN and GMVAR are for multi-modal learning which fit our input data appropriately. TSN is conducted for each single-modal individually. We adopt the AMGL and MLAN to fit our multi-modal temporal classification setting and make evaluation.

As shown on Fig. 2, the modality-specific attention is first trained individually. The input is multi-modal time series data. The attention distributions $\boldsymbol{Z^v}$ are derived through optimizing Eq. 2.5 during first-stage model. Next, the same input data is set as input for the multi-modal collaboration (second-stage) with $\boldsymbol{Z^v}$ from the first-stage. The MAR model is conducted with the additional input $\boldsymbol{Z}^v$. Single-modal results from MAR model are fed into the final late fusion model to obtain the multi-modal performance. We set 128 batch size for EV-Action and NTU, and 32 for DHA and UWA3D II datasets. The hidden dimensions for both temporal encoders (first and second stages) and attention are 128. The learning rates are 0.0005 and 0.001 for first-stage and second-stage. Our model is implemented by PyTorch with GPU acceleration.

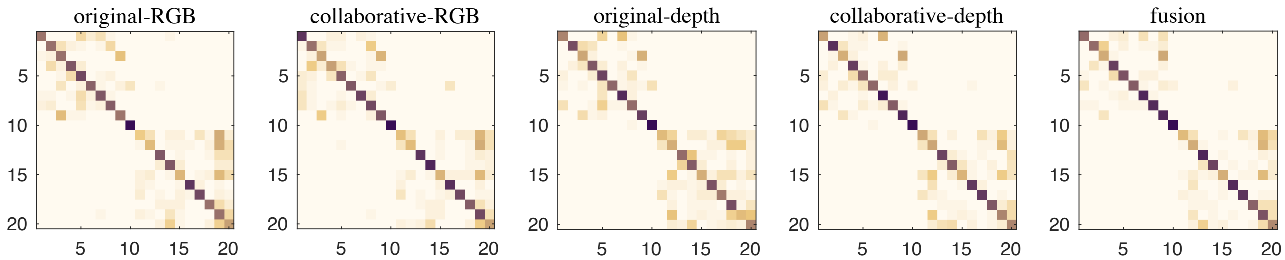

Figure 4: Visualization results of confusion matrices on EV-Action dataset.

Datasets	Methods	RGB	Depth	Fusion
EV-Action	TSN [20]	0.6855	0.6723	-
	RC Classifier [2]	0.5992	0.5790	0.6213
	MFN [23]	0.5743	0.4082	0.6423
	MLSTM-FCN [10]	0.6804	0.6926	0.7014
	GMVAR [18]	0.6792	0.6739	0.7088
	CAM (ours)	**0.7022**	**0.7123**	**0.7359**
NTU	TSN [20]	0.7517	0.7691	-
	RC Classifier [2]	0.7683	0.8014	0.8258
	MFN [23]	0.7089	0.8062	0.8125
	MLSTM-FCN [10]	0.7662	0.7941	0.8217
	GMVAR [18]	0.7545	0.7702	0.8018
	CAM (ours)	**0.7720**	**0.8134**	**0.8408**
DHA	TSN [20]	0.6785	0.8324	-
	AMGL [13]	0.6461	0.7284	0.7489
	MLAN [12]	0.6791	0.7296	0.7613
	GMVAR [18]	0.6972	0.8348	**0.8872**
	CAM (ours)	**0.7407**	**0.8642**	0.8724
UWA3D II	TSN [20]	0.4833	0.5936	-
	AMGL [13]	0.3067	0.3667	0.3933
	MLAN [12]	0.2933	0.2867	0.3800
	GMVAR [18]	0.4917	0.5846	0.6035
	CAM (ours)	**0.5083**	**0.6073**	**0.6314**

Table 1: Single-modal and multi-modal evaluation on four multi-modal time series datasets.

3.4 Performance Analysis The classification performances of four datasets are shown in Table 1. For EV-Action and NTU datasets, our method outperforms all other approaches on both single-modal and multi-modal scenarios. MLSTM-FCN is an effective model for single-modal temporal data which achieves competitive results. However, the fusion result is lower than ours. More importantly, our single-modal performances are also higher than MLSTM-FCN which demonstrates our MAR model works well to discover more valuable information for each single modality. GMVAR is another competitive multi-modal temporal data classification algorithm based on a generative model. However, it suffers from the difficulties of training generative model and cannot obtain promising performance on these two large-scale datasets. Our CAM obtains the highest classification performances for both single-modal and multi-modal evaluation. For DHA and UWA3D II datasets, GMVAR achieves better performances on these two small-scale datasets. Its generative strategy improves the multi-modal learning performance and model robustness. However, our method still generally outperforms it especially on single-modal scenario. We visualize the confusion matrices on EV-Action dataset using the single-modal results before/after our collaborative (first-stage/second-stage) and the multi-modal fusion result in Fig. 4.

Datasets	Method	RGB	Depth	Fusion
EV-Action	LSTM (baseline)	0.6878	0.6772	-
	CAM w/o MAR	0.6894	0.6796	0.7154
	CAM w/o RGB	0.6978	0.6802	0.7285
	CAM w/o Depth	0.6874	0.7084	0.7255
	CAM (ours)	**0.7022**	**0.7123**	**0.7359**
NTU	LSTM (baseline)	0.7393	0.7974	-
	CAM w/o MAR	0.7401	0.7986	0.8262
	CAM w/o RGB	0.7425	0.8115	0.8342
	CAM w/o Depth	0.7680	0.8059	0.8316
	CAM (ours)	**0.7720**	**0.8134**	**0.8408**

Table 2: Ablation Study on EV-Action and NTU.

3.5 Ablation Study We provide a detailed ablation study on the EV-Action and NTU datasets to prove the necessity of each model component. The results are shown in Table 2. Particularly, we compare with four ablated models as follows: 1) `LSTM (baseline)` indicates single-modal performance without multi-modal collaboration learning and late fusion. 2) `CAM w/o MAR` means we train the multi-modal data synchronously and add late fusion without collaborative learning using the MAR cell. 3) `CAM w/o Depth` and 4) `CAM w/o RGB` denote we only deploy collaboration learning to update the cell state of each single modality individually. We conclude deploying collaborative learning on each single-modal enhances the representations and improves the performance correspondingly. Our complete model

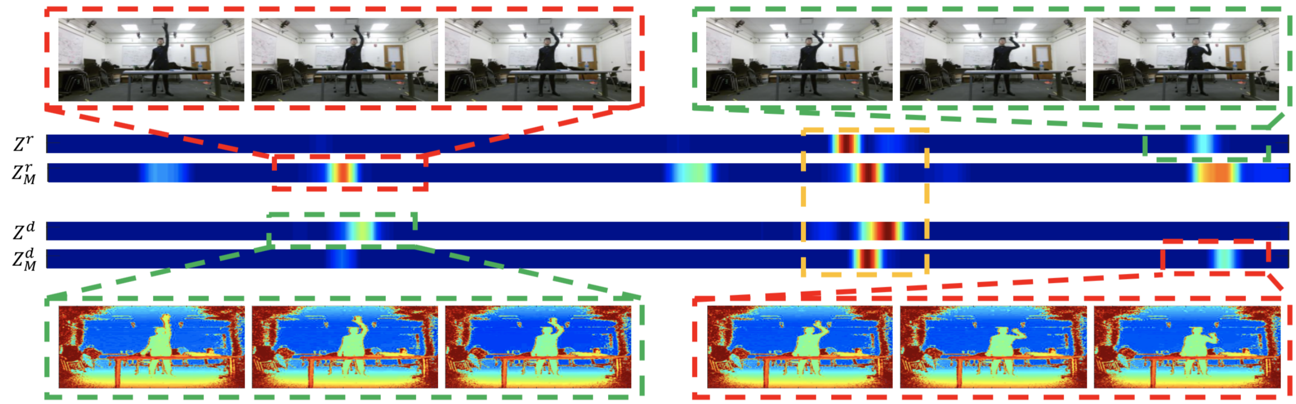

Figure 5: The colorbars represent the attention distribution scores of Z^v and Z^v_M (before/after multi-modal collaboration). The dash boxes indicate the temporal locations and show its corresponding frames. Green indicates frames originally attended by single modality itself. Red represents frames being attended after our collaborative learning. Yellow means frames attended by both two modalities.

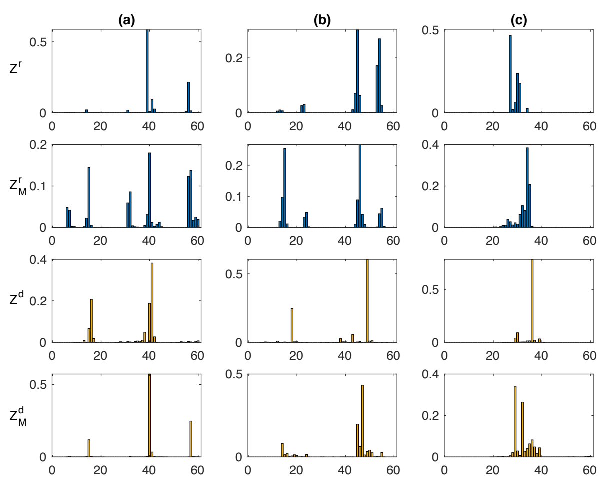

Figure 6: Visualization of attention score changes for two modalities. Each column represents one sample. X-axis/Y-axis are the time step and score values.

deploys the multi-modal collaboration to achieve high performance on each single modality. Further the late fusion leverages the enhanced single-modal representations and obtains the best multi-modal performance.

3.6 Attention Visualization We visualize and compare the changes between the $\boldsymbol{Z^v}$ and $\boldsymbol{Z^v_M}$, which are the temporal attention distributions (scores) before and after our multi-modal collaboration. It illustrates the collaborative learning process and provides the intuition about our model insight. Fig. 6 shows three samples from EV-Action dataset with different collaborative learning cases. Each column represents one sample. In (a), each modality captures specific attention patterns and guides the other modality correspondingly. In (b), depth exerts an influence on RGB, while RGB has little impact on depth. In (c), two modalities roughly pay attention to the same location, however, they still adjust their attention scores in a small-scale through the collaborative learning process.

Further, we provide more details of sample (a) with corresponding frames in Fig. 5. The colorbars in the middle are the temporal attention scores of Z^v and Z^v_M. Being lighter means higher value. The green dash boxes indicate the frames have been noticed by each single modality itself. The red boxes represent the frames gained attention after our collaborative learning, which is hard to be discovered by single modality itself. The yellow boxes denote the frames noticed by both two modalities simultaneously. In this case, the action class is “throwing a ball”. The process of *hands up* and *hands down* are easily captured by RGB. However, the *throwing* when hands at the highest point is easily noticed by depth due to its motion changes in depth direction. The collaborative learning process takes advantages of characteristics of each modality to guide the other modality obtaining more implicit patterns and enhancing the learned representations.

4 Conclusions

In this paper, we propose a Collaborative Attention Mechanism (CAM) for the multi-modal time series classification (MTC). A modality-specific attention is first utilized for capturing multi-modal attention distributions. Then, the multi-modal collaboration is achieved with the proposed Mutual-Aid RNN (MAR) cell. In this way, each modality is guided by the knowledge from the other modalities and enhanced to discover more latent information by itself. The proposed CAM provides a novel perspective to leverage the attention mechanism for exploring multi-modal temporal learning. The interpretability of attention is appropriately exploited to guide the learning process. Taking advantage of the collaboration strategy, the proposed CAM outperforms state-of-the-art methods on four public multi-modal time series datasets in both single and multi-modal scenarios. A detailed ablation study is also provided to validate the effectiveness of each model component.

References

[1] Mohiuddin Ahmad and Seong-Whan Lee. Hmm-based human action recognition using multiview image sequences. In *ICPR*, pages 263–266. IEEE, 2006.

[2] Filippo Maria Bianchi, Simone Scardapane, Sigurd Løkse, and Robert Jenssen. Reservoir computing approaches for representation and classification of multivariate time series. *arXiv preprint arXiv:1803.07870*, 2018.

[3] Stanislas Chambon, Mathieu N Galtier, Pierrick J Arnal, Gilles Wainrib, and Alexandre Gramfort. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4):758–769, 2018.

[4] Shengdong Du, Tianrui Li, Yan Yang, and Shi-Jinn Horng. Multivariate time series forecasting via attention-based encoder–decoder framework. *Neurocomputing*, 388:269–279, 2020.

[5] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*, pages 345–360. Springer, 2014.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202, 2017.

[8] Chengcheng Jia and Yun Fu. Low-rank tensor subspace learning for rgb-d action recognition. *TIP*, 25(10):4641–4652, 2016.

[9] Chengcheng Jia, Yu Kong, Zhengming Ding, and Yun Fu. Latent tensor transfer learning for rgb-d action recognition. In *ACM MM*, pages 87–96, 2014.

[10] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Samuel Harford. Multivariate lstm-fcns for time series classification. *Neural Networks*, 116:237–245, 2019.

[11] Yan-Ching Lin, Min-Chun Hu, Wen-Huang Cheng, Yung-Huan Hsieh, and Hong-Ming Chen. Human action recognition and retrieval using sole depth information. In *ACM MM*, pages 1053–1056, 2012.

[12] Feiping Nie, Guohao Cai, and Xuelong Li. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *AAAI*, 2017.

[13] Feiping Nie, Jing Li, Xuelong Li, et al. Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In *IJCAI*, pages 1881–1887, 2016.

[14] Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. Histogram of oriented principal components for cross-view action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2430–2443, 2016.

[15] Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In *ECCV*, pages 742–757. Springer, 2014.

[16] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[17] Lichen Wang, Zhengming Ding, Seungju Han, Jae-Joon Han, Changkyu Choi, and Yun Fu. Generative correlation discovery network for multi-label learning. In *ICDM*, pages 588–597. IEEE, 2019.

[18] Lichen Wang, Zhengming Ding, Zhiqiang Tao, Yunyu Liu, and Yun Fu. Generative multi-view human action recognition. In *ICCV*, pages 6212–6221, 2019.

[19] Lichen Wang, Bin Sun, Joseph Robinson, Taotao Jing, and Yun Fu. Ev-action: Electromyography-vision multi-modal action dataset. *arXiv preprint arXiv:1904.12602*, 2019.

[20] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36. Springer, 2016.

[21] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *NAACL*, pages 1480–1489, 2016.

[22] Hui Yin, Shuiqiao Yang, Xiangyu Song, Wei Liu, and Jianxin Li. Deep fusion of multimodal features for social media retweet time prediction. *World Wide Web*, pages 1–18, 2020.

[23] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *AAAI*, 2018.

[24] Zhihua Zhang, Khelifi Zhang, and A Khelifi. *Multivariate time series analysis in climate and environmental research*. Springer, 2018.