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 across 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-modal 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.

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

Copyright © 20XX by SIAM
Unauthorized reproduction of this article is prohibited
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 summations to coordinate the contribution from each single modality. They ignore that how to excavate latent information to coordinate the contribution from each single modality.

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 preserving valuable patterns also deserve 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 follow:

- 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 multimodal 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:

$$\theta^* = \arg\max_{\theta} \sum_{(X,y)} \log p(y|X; \theta),$$  \hspace{1cm} (2.1)

where $\theta$ is the set of parameters of model. $X = \{x_1, ..., x_t\}$ is the multiple steps of one time series sample, and $y$ 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 $x_t$ is encoded as a hidden representation $h_t$, and the cell state $c_t$ is updated correspondingly. The LSTM update processes are given by

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f),$$
$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i),$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o),$$
$$c_t = f_t \odot c_{t-1} + i_t \odot o_c(W_c h_t + U_c h_{t-1} + b_c),$$
$$h_t = o_t \odot \sigma_h(c_t),$$  \hspace{1cm} (2.2)

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

Original temporal sequence $X$ is encoded as $H = \{h_1, ..., h_T\}$. Commonly, we pick the last hidden state $h_T$ to represent the whole sequence. However, it may lose temporal information to some degree. A reasonable way is using the weighted summation of $h_t$. 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

$$u_t = \tanh(W_w h_t + b_w),$$
$$z_t = \frac{\exp(u^T w)}{\sum_t \exp(u^T w)},$$
$$r = \sum_t z_t h_t,$$  \hspace{1cm} (2.3)

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

To introduce our CAM clearly, we go deeper to provide more insights about LSTM. The key factor of LSTM cell is the $c_t$. It reflects memory states of the whole sequence. $f_t$ and $i_t$ update the $c_t$ internally through the forget and input procedures. The contents of forget/input are derived from current input $x_t$ and last hidden state $h_{t-1}$. The content of current hidden state $h_t$ is also extracted from $x_t/h_{t-1}$, then filtered by $c_t$. All information flows cross several control gates center on the $c_t$. As the memory state, $c_t$ 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:

$$H^v = E^v(X^v, \phi_E^v),$$
$$r^v = Q^v(H^v, \phi_Q^v),$$
$$Y^v = C^v(r^v, \phi_C^v),$$  \hspace{1cm} (2.4)

where superscript $v$ represents the modality $v$. $E$ is LSTM module (Eq. 2.2), encoding sequence $X$ into hid-
2.3 Multi-Modal Collaboration by Mutual-Aid RNN

To substantially take advantage of multi-modal information and prepared for following collaborative learning. It is formulated as follows:

\[ G_{r \rightarrow d} = \sigma(W_r x_t^r + W_d h_{t-1}^d), \]
\[ G_{d \rightarrow r} = \sigma(W_d x_t^d + W_r h_{t-1}^r), \]

where \( W_r \) is learnable parameters. \( G_r \) extract information from current input \( x_t^r \) and collaborate with last hidden state \( h_{t-1}^r \) from the other modality. \( \sigma \) represents the sigmoid activation. The knowledge of each time step from the other modality is reserved in \( G_d \).

**Mutual Filtering** is designed based on cross-modal collaborator above. Cell state \( c_t^c \) contains temporal dynamic patterns for each modality. It could be updated internally in LSTM (Eq. 2.6). However,
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 \( c^r_t \) using cross-modal collaborator to derive knowledge from the other one, which is given by

\[
\begin{align*}
c^{r'}_t &= G_{d \rightarrow r} \circ c^r_t, \\
c^{d'}_t &= G_{r \rightarrow d} \circ c^d_t,
\end{align*}
\]

(2.8)

where \( \circ \) is the point-wise product. \( c^{r'}_t \) 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 \( z^r_t \) reflect the importance of each time step for each single-modality. Further, it also decides the information importance during updating \( c^r_t \) in multi-modal collaborative learning. We first normalize the attention scores by

\[
\begin{align*}
z^{r'}_t &= \frac{z^r_t}{z^r_t + z^d_t}, \\
z^{d'}_t &= \frac{z^d_t}{z^r_t + z^d_t}.
\end{align*}
\]

(2.9)

The original cell state \( c^r_t \) are updated by single-modal information, while \( c^{r'}_t \) are updated by the cross-modal collaborator \( G_* \). \( z^{r'}_t \) represent the importance of dynamic knowledge from different modalities. We integrate the multi-modal information for updating cell states via the weighted summation:

\[
\begin{align*}
c^{r''}_t &= z^{r'}_t c^r_t + z^{d'}_t c^{d'}_t, \\
c^{d''}_t &= z^{d'}_t c^d_t + z^{r'}_t c^{r'}_t,
\end{align*}
\]

(2.10)

where \( c^{r''}_t \) 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

\[
\begin{align*}
H^r_M &= E^v_M(X^v, \phi^v_{E_M}), \\
\tau^r_M &= Q^v_M(H^r_M, \phi^v_{Q_M}), \\
\hat{Y}^v_M &= C^v_M(r^v_M, \phi^v_{C_M}),
\end{align*}
\]

(2.11)

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

\[
L^v_M = \ell(Y, \hat{Y}^v_M).
\]

(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 \( \hat{Y}^v_M \) 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

\[
D = \hat{Y}^r_M \cdot \hat{Y}^d_M^\top,
\]

(2.13)

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

\[
L_f = \ell(Y, C^f(D, \phi_{C^f})),
\]

(2.14)

where \( Y \) is the ground truth, \( \ell \) 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. 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 Multi-modal 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 [13] 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 BNIception 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 $Z^v$ 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 $Z^v$ from the first-stage. The MAR model is conducted with the additional input $Z^v$. 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.
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.

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
3.6 Attention Visualization We visualize and compare the changes between the $Z^v$ and $Z^v_M$, 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 scenarios. A detailed ablation study is also provided to validate the effectiveness of each model component.

References
