MuVAN: A Multi-view Attention Network for Multivariate Temporal Data

Ye Yuan*, Guangxu Xun†, Fenglong Ma†, Yaqing Wang†, Nan Du‡, Kebin Jia*, Lu Su† and Aidong Zhang†
*College of Information and Communication Engineering, Beijing University of Technology, Beijing, China
Email: yuanye91@emails.bjut.edu.cn, kebinj@bjut.edu.cn
†Department of Computer Science and Engineering, State University of New York at Buffalo, NY, USA
Email: guangxux,fenglong,yaqingwa,lusu,azhang@buffalo.edu
‡Tencent Medical AI Lab, CA, USA Email: ndu@tencent.com

***Abstract*—Recent advances in attention networks have gained enormous interest in time series data mining. Various attention mechanisms are proposed to soft-select relevant timestamps from temporal data by assigning learnable attention scores. However, many real-world tasks involve complex multivariate time series that continuously measure target from multiple views. Different views may provide information of different levels of quality varied over time, and thus should be assigned with different attention scores as well. Unfortunately, the existing attention-based architectures cannot be directly used to jointly learn the attention scores in both time and view domains, due to the data structure complexity. Towards this end, we propose a novel multi-view attention network, namely MuVAN, to learn fine-grained attentional representations from multivariate temporal data. MuVAN is a unified deep learning model that can jointly calculate the two-dimensional attention scores to estimate the quality of information contributed by each view within different timestamps. By constructing a hybrid focus procedure, we are able to bring more diversity to attention, in order to fully utilize the multi-view information. To evaluate the performance of our model, we carry out experiments on three real-world benchmark datasets. Experimental results show that the proposed MuVAN model outperforms the state-of-the-art deep representation approaches in different real-world tasks. Analytical results through a case study demonstrate that MuVAN can discover discriminative and meaningful attention scores across views over time, which improves the feature representation of multivariate temporal data.**

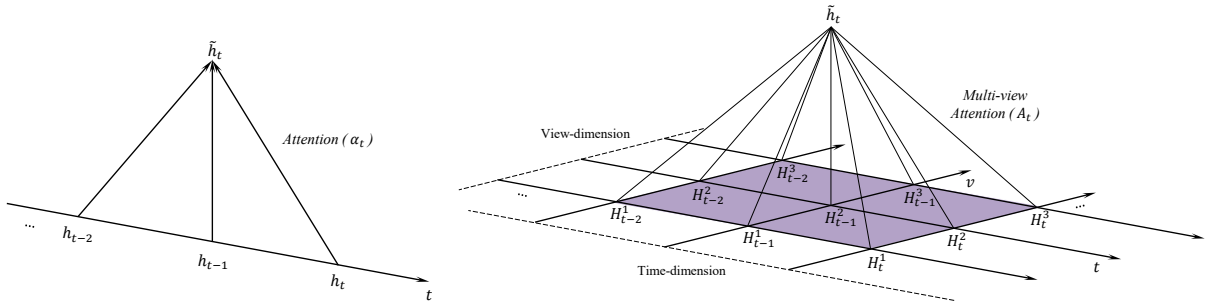
I. Introduction

Recently, attention-based neural networks have been successfully applied in a wide range of tasks, including neural machine translation [1], [2], speech recognition [3], [4], disease diagnosis [5]–[7], and risk prediction [8]. Among different applications, various attention mechanisms are proposed as a hidden layer to make soft-selection over several timestamps by assigning different attention scores from a categorical distribution [9], [10]. Combined with recurrent neural networks (RNN), such as long short-term memory (LSTM) and gated recurrent neural networks (GRU), attention mechanisms are able to focus on the most relevant hidden states of the sequence to conduct detection or prediction. These approaches have been proven to be useful for deep feature representation of temporal data.

However, the aforementioned existing attention mechanisms can only handle univariate temporal data. In practice, the recent advances of pervasive sensing make many real-world tasks involve complex multivariate temporal data that continuously measure target from multiple views (or different information sources) [11]–[13]. Different views may carry different amount of information varied over time, and thus should be assigned with different attention scores to make decisions. For instance, in healthcare setting, various physiological measurements, such as electrocardiogram (ECG) and electroencephalogram (EEG), provide complementary information on clinical observations and reflect patient's health condition from different perspectives, i.e., views. These raw temporal records are often used to diagnose diseases [14]–[16]. For the task of human activity recognition, multi-sensor data record synchronous movements (or actions) in different body areas, and each of which is monitored by a sensor serving as a 'view' of the activities. Intuitively, if we can estimate the importance of information contributed by each view within different timestamps, we will be able to enhance the feature representation of multivariate temporal data. Unfortunately, no attention mechanism can be directly used to jointly assign attention scores to both time and view dimensions. Due to the complex data structure, the existing attention-based architectures are not applicable to the modeling of a continuum of multi-view time series. Thus, it remains a difficult task to develop multi-view attention to learn meaningful representations from long and broad temporal inputs.

To develop a multi-view attention mechanism for multivariate temporal data, we must address several challenges. First, there exist inherent connections among views over time containing complicated observations that cannot be simply captured and interpreted. In practice, multivariate temporal data can be represented as a collection of heterogeneous continuous time series consisting of several non-uniformly sampled signals. It has a unique data structure where the record fragments of each view are temporally ordered but the views within a timestamp form an unordered set. Compared to the discrete setting that contains a single-dimensional fragment at each time step, e.g., binary diagnosis codes or one-hot words, multivariate temporal data comprises a two-dimensional heterogeneous fragment within one timestamp, e.g., multimodal biosignals with different sampling rates. This introduces a challenge of modeling complex structures with

978-1-5386-9159-5/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDM.2018.00087



(a) Structure of attention network

(b) Structure of multi-view attention network

Fig. 1. Deep model for attention-based architecture.

attention mechanism. Moreover, the hidden patterns of raw waveform records in decision-making vary significantly across individuals. For example, patients suffer from diseases to different extents, and the diagnosis is often regarded as a result dependent on physicians' experience with significant uncertainty [17].

To address the above challenges, we propose a novel **mu**lti-**v**iew **a**ttention **n**etwork (MuVAN) to learn fine-grained attentional representations from multivariate temporal data. The idea is to mimic the practical visual inspection that pays attention to the details jointly from both time and view dimensions. The learned attention scores are highly interpretable, since they can be used to explain how important each view at each time is to the goal of real-world tasks. Specifically, we first separately embed each view of the heterogeneous time series into a unified latent space through a view-wise recurrent encoder. Then we feed all the view representations into a new multi-view attention layer that jointly assigns attention scores to each view within different timestamps. In order to fully utilize the multi-view information, we construct a hybrid focus procedure that brings more diversity to attention. To preserve the spatial locality, the attentional representation is further aggregated through a spatial feature fusion layer. Finally, we adopt a softmax layer for the classification of task. We demonstrate that the proposed MuVAN model achieves better performance compared to the state-of-the-art deep representation learning approaches on three different real-world datasets. Moreover, we evaluate the interpretability of the learned attention scores through a case study.

In summary, the main contributions of this paper are as follows:

- We formalize the problem of multi-view attention learning for multivariate temporal data and identify its unique challenges resulting from structure complexity.
- We propose MuVAN, a unified multi-view attention-based deep learning model, to learn fine-grained attentional representations from multivariate temporal data. MuVAN can jointly calculate two-dimensional attention scores to estimate the quality of information contributed by each view within different timestamps.
- We empirically show that the proposed MuVAN outperforms existing deep representation learning methods on three real-world datasets. The results indicate that the learned attention scores can identify the influential concepts across views over time.

In the following sections, we first discuss the connection of the proposed approaches to related work in Section 2. Our proposed methodology is then described in Section 3. Section 4 presents and discusses the experimental results for our method. Finally, we conclude this work in Section 5.

II. Preliminaries and Related Work

In this section, we first give a brief introduction of attention-based neural networks, then review the existing work using deep representation learning for multivariate temporal data.

A. Attention-based Neural Networks

The basic idea of attention mechanism is to distinguish the task-related importance of different timestamps from a sequence $\boldsymbol{x}_{1:T}$. In practice, $\boldsymbol{x}_{1:T}$ is often processed by an encoder, e.g., RNN, which outputs a sequence of hidden vectors $\boldsymbol{h}_{1:T}$ that are more suitable for attention mechanism. Fig. 1a shows a basic structure of attention-based network following a self-attention strategy [9]. Intuitively, attention mechanism is trained to capture the dependencies by calculating a normalized energy score $\alpha_{t,i}$ corresponding to each input timestamp $i(1 \leq i \leq t)$ separately. More formally, the following three formulations describe a general attention procedure for the t-th timestamp:

$$e_{t,i} = \text{Energy}(\boldsymbol{h}_t, \boldsymbol{h}_i), \tag{1}$$

$$\boldsymbol{\alpha}_t = \text{Normalize}([e_{t,1}, e_{t,2}, ..., e_{t,t}]), \tag{2}$$

$$\tilde{\boldsymbol{h}}_t = \sum\nolimits_{i=1}^{t} \alpha_{t,i} \odot \boldsymbol{h}_i.$$

We can observe that the attentional vector $\tilde{\boldsymbol{h}}_t$ is fused by weighted sum of all hidden representations from timestamp 1 to t. Recently, researchers have attempted to adopt attention-based models with various energy functions based on Eq. (1), including location-based attention [6], [18], graph-based attention [7], concatenation-based attention [6], [18],

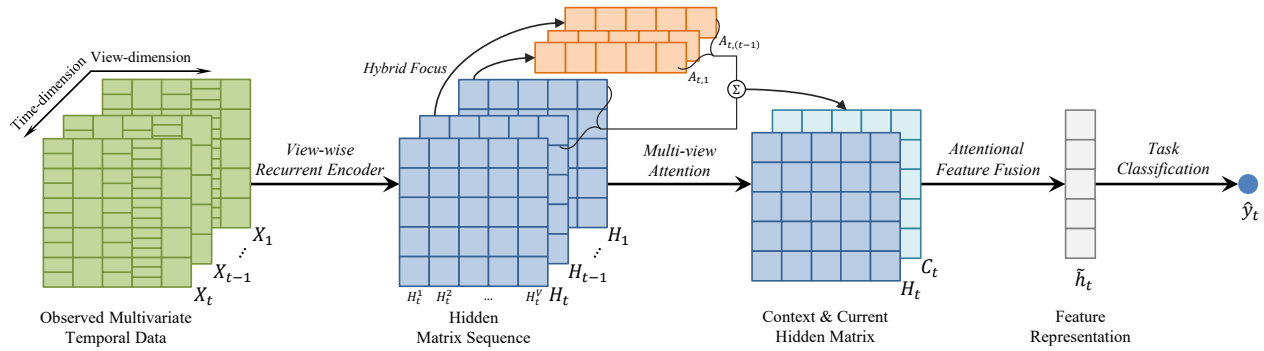


Fig. 2. Schematic illustration of the proposed multi-view attention model.

and channel-aware attention [19]. Different from existing work, we consider a more complex structure of the raw multivariate temporal data shown in Fig. 1b. The input contains multi-view information and thus is represented as a sequence of hidden matrices $\boldsymbol{H}_{1:T}^{1:V}$ rather than vectors. Compared to the conventional single-dimension attention-based networks, our multi-view attention mechanism should capture the two-dimension importance of different input data in both time and view domains, in order to unleash the power of feature representation for multivariate temporal data.

B. Deep Representation Learning for Multivariate Temporal Data

Deep representation learning for multivariate temporal data is one of the core research tasks in temporal data mining, as such multivariate time series refers to different observation views that can be continually measured and monitored for various real-world tasks. On one hand, some simplified deep representation learning approaches are designed to extract abstract features from single part (or a few parts) of views independently [20]–[22]. However, these methods may ignore information carried by the rest of views and hence obtain limited improvement in real-world tasks. On the other hand, in order to incorporate multi-view information, previous studies have validated that modifying deep learning structures can improve the performance in modeling multivariate temporal data. To extract hidden features from multivariate and multimodal Polysomnography (PSG) records, concatenated deep belief networks (DBN) are adopted for sleep stage classification [23], [24]. Multi-view stacked denoising autoencoders (SDAE) are employed to detect characteristic patterns of epileptic seizure in multi-channel EEG signals [25]. Some variants of convolutional neural networks (CNN) are used to learn common and modality-specific representations from multimodal temporal data for the task of human activity recognition (HAR) [26] and sleep stage classification [15]. A hybrid deep learning model that combines CNN with RNN is also proposed to extract both local and temporal relationships from different sensory modalities [27], [28]. In most of the aforementioned deep learning models, the joint features are extracted mainly using the parameter sharing architecture. In contrast, we propose to derive the combined representation by explicitly fusing the view representations according to the relative significance of each view over time. With the help of our multi-view attention mechanism, we can jointly capture dependencies from both time and view dimensions.

III. Methodology

In this section, we first introduce the structure of multivariate temporal data and some basic notations. Then we present an overview of the proposed MuVAN model, and discuss the details of the main components. Finally, we explain how to interpret the learned attention scores.

A. Basic Notations

In this part, we model the multivariate temporal data as a set of time-labeled heterogeneous sequences for a multi-class classification problem. Since the waveform pattern associated with practical meanings is related to an interval rather than a certain point [29], [30], in our model, we assume that there are M multivariate temporal records, and the m-th record has $T^{(m)}$ timestamps with $V^{(m)}$ views. Then, the record can be further represented by a sequence of fragments $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_{T^{(m)}}\}$. Each fragment $\boldsymbol{X}_t$ consists of a set of waveform vectors $\{\boldsymbol{x}_t^1, \boldsymbol{x}_t^2, \cdots, \boldsymbol{x}_t^{V^{(m)}}\}$ where $\boldsymbol{x}_t^v \in \mathbb{R}^{n^{(v)}}$. Each fragment $\boldsymbol{X}_t$ also has a corresponding coarse-grained category label $\boldsymbol{y}_t \in \{0,1\}^{|\mathcal{C}|}$ where $|\mathcal{C}|$ is the unique number of categories related to different real-world tasks. Moreover, in signal processing, the sampled waveform data expressed in time-frequency domain are more meaningful than time domain [30]. In the context of our model, the input vector $\boldsymbol{x}_t^v$ refers to a scalogram vector using wavelet transform.

With the aforementioned notations, the inputs of the proposed multi-view attention model are the set of time-ordered heterogeneous sequences $\{\boldsymbol{X}_1^{(m)}, \boldsymbol{X}_2^{(m)}, \cdots, \boldsymbol{X}_T^{(m)}\}_{m=1}^M$ with a set of corresponding labels $\{\boldsymbol{y}_1^{(m)}, \boldsymbol{y}_2^{(m)}, \cdots, \boldsymbol{y}_T^{(m)}\}_{m=1}^M$.

B. Model Architecture

Attention networks aim at performing a soft-selection procedure over sequential inputs using an internal inference step. In this work, we present MuVAN to mimic the practical visual inspection that focuses on several details jointly from both time

and view dimensions. For example, physicians diagnose diseases by exploring the subject's monitoring records, and they always pay attention to some influential clinical observations in different aspects (i.e., views) happened at different time, such as fast heart beat, abnormal brain activity, and irregular body motion. Fig. 2 depicts the high-level overview of our proposed model. Given a multivariate temporal record from timestamp 1 to t, the i-th input vector from the v-th view $\boldsymbol{x}_t^v$ is fed into a view-wise recurrent encoder, which outputs a hidden vector $\boldsymbol{h}_t^v$. Along with the set of integrated hidden matrix $\{\boldsymbol{H}_i\}_{i=1}^{t-1}$, we are able to compute a score matrix $\boldsymbol{A}_t$ for the current timestamp t. Subsequently, a context matrix $\boldsymbol{C}_t$ is computed from $\boldsymbol{A}_t$ and $\{\boldsymbol{H}_i\}_{i=1}^{t-1}$. The procedure is presented as multi-view attention mechanism combining with a hybrid focus module, which will be detailed in the rest of this section. From the context matrix $\boldsymbol{C}_t$ and the current hidden matrix $\boldsymbol{H}_t$, we can further obtain an attentional hidden representation $\tilde{\boldsymbol{h}}_t$ through a spatial feature fusion module to predict the label, i.e., $\boldsymbol{y}_t$. The proposed network can be trained end-to-end.

C. View-wise Recurrent Encoder

In the task of learning deep representations from multivariate temporal data, simply concatenating raw input features, namely global feature learning, may not be enough to achieve accurate and robust performance, especially with the presence of a large number of views with different modalities. The effectiveness of extracting features from different views, referred to view-wise feature learning, has been proven for different tasks [19], [25], [27]. Inspired by these studies, we further extend this strategy by adopting view-wise RNN to learn latent representations from multi-view time series. Given the v-th view sequence from $\boldsymbol{x}_1^v$ to $\boldsymbol{x}_T^v$, we can obtain their hidden representations $\boldsymbol{h}_t^v \in \mathbb{R}^{2p}$ through a 2-layer stacked Bidirectional Gated Recurrent Units (BGRU) [31] as follows:

$$\boldsymbol{h}_{1:T}^v = \text{BGRU}(\boldsymbol{x}_{1:T}^v; \boldsymbol{\theta}_r), \tag{3}$$

where $\boldsymbol{\theta}_r$ denotes all the parameters of BGRU. The obtained $\boldsymbol{h}_t^v$ is the concatenation of both forward and backward hidden vectors, denoted as $\overrightarrow{\boldsymbol{h}}_t^v, \overleftarrow{\boldsymbol{h}}_t^v \in \mathbb{R}^p$, respectively.

In our model, we choose BGRU since it takes advantages of all the available sequence information from two directions [6]. Moreover, compared to LSTM, BGRU shows similar performance but with a more concise expression [32], which reduces network complexity for our view-wise settings. The hidden features extracted from different views are further integrated into a hidden matrix $\boldsymbol{H}_t = \{\boldsymbol{h}_t^1, \boldsymbol{h}_t^2, ..., \boldsymbol{h}_t^V\} \in \mathbb{R}^{2p \times V}$. In this way, we unify the feature dimension from the heterogeneous inputs, and the unique characteristics of each view are preserved for the attention mechanism.

D. Multi-view Attention Mechanism

To distinguish the importance of different views over time, we propose two multi-view attention mechanisms, namely location-based attention and context-based attention, respectively, to jointly assign attention energy to both view and time domains.

• *Location-based Attention*. An easy way to calculate the attention energy $e_{t,i}^v \in \mathbb{R}$ is scoring solely from each view-wise representation $\boldsymbol{H}_i^v$, as follows:

$$e_{t,i}^v = \text{Energy}(\boldsymbol{H}_i^v) = \boldsymbol{W}_e^\top \boldsymbol{H}_i^v + b_e, \tag{4}$$

where $\boldsymbol{W}_e \in \mathbb{R}^{2p}$ and $b_e \in \mathbb{R}$ are the weight vector and bias value, respectively. Note that the location-based multi-view attention mechanism does not capture any relationships among views or timestamps, as it only considers individual information. Moreover, it is time-consuming to directly measure all pairwise correlations of every two view-wise representations, since the computational complexity may grow exponentially with the increase in the number of correlated views. Thus, to fully characterize the complicated view relationships, we propose a novel context-based multi-view attention mechanism in the proposed MuVAN.

• *Context-based Attention*. We use a multi-layer perceptron (MLP) [2] to calculate the energy based on four context information sources: 1) local self-context of $\boldsymbol{H}_i$, 2) target self-context of $\boldsymbol{H}_t$, 3) cross-context between $\boldsymbol{H}_t$ and $\boldsymbol{H}_i$, and 4) previous score information from $\boldsymbol{\alpha}_{t,(i-1)}$. Specifically, for self-context expression, we learn a weighted sum vector $\bar{\boldsymbol{h}}_t \in \mathbb{R}^k$ by convolving $\boldsymbol{H}_t$ with a matrix $\boldsymbol{W}_{sc} \in \mathbb{R}^{V \times (2p-k+1)}$, as follows:

$$\bar{\boldsymbol{h}}_t = \sum_{v=1}^{V} \boldsymbol{H}_t^v * \boldsymbol{W}_{sc}^v, \tag{5}$$

where $*$ denotes the convolution operator. Similarly, we derive a vector $\bar{\boldsymbol{h}}_{t,i} \in \mathbb{R}^k$ to express the cross-context between representations $\boldsymbol{H}_t$ and $\boldsymbol{H}_i$, defined as:

$$\bar{\boldsymbol{h}}_{t,i} = \boldsymbol{H}_t * \boldsymbol{W}_{cc}^1 + \boldsymbol{H}_i * \boldsymbol{W}_{cc}^2, \tag{6}$$

where $\boldsymbol{W}_{cc} \in \mathbb{R}^{2 \times V \times (2p-k+1)}$ is the parameter to be learned. Based on Eq. (5) and Eq. (6), the attention energy vector $\boldsymbol{e}_{t,i} \in \mathbb{R}^V$ can be calculated as follows:

$$\begin{aligned} \boldsymbol{e}_{t,i} &= \text{Energy}(\boldsymbol{H}_t, \boldsymbol{H}_i, \boldsymbol{\alpha}_{t,(i-1)}) \\ &= \tanh\left(\boldsymbol{W}_a \bar{\boldsymbol{h}}_t + \boldsymbol{W}_b \bar{\boldsymbol{h}}_i + \boldsymbol{W}_c \bar{\boldsymbol{h}}_{t,i} + \boldsymbol{W}_d \boldsymbol{\alpha}_{t,(i-1)} + \boldsymbol{b}_e\right), \end{aligned} \tag{7}$$

where $\boldsymbol{W}_a \in \mathbb{R}^{V \times k}$, $\boldsymbol{W}_b \in \mathbb{R}^{V \times k}$, $\boldsymbol{W}_c \in \mathbb{R}^{V \times k}$, $\boldsymbol{W}_d \in \mathbb{R}^{V \times V}$, $\boldsymbol{b}_e \in \mathbb{R}^V$ are the parameters to be learned. Different from the location-based attention, context-based attention captures hidden connections of $\boldsymbol{H}_t$ and $\boldsymbol{H}_i$ by considering the surrounding context information from both time and view domains, and thus can generate informative representations from multi-view data. Based on Eq. (4) or Eq. (7), we can obtain an energy matrix $\boldsymbol{E}_t \in \mathbb{R}^{V \times (t-1)}$ of current timestamp t for score normalization.

E. Hybrid Focus Procedure

According to Eq. (2), to obtain a normalized attention score matrix $\boldsymbol{A}_t \in \mathbb{R}^{V \times (t-1)}$ from the energy matrix $\boldsymbol{E}_t$, the softmax function can be directly used, as follows:

$$\boldsymbol{A}_{t,i}^v = \exp\left(\boldsymbol{E}_{t,i}^v\right) \Big/ \sum_{i=1}^{t-1} \sum_{v=1}^{V} \exp\left(\boldsymbol{E}_{t,i}^v\right). \tag{8}$$

The conventional softmax-based normalization is based on the assumption that only a few elements are related to the task goal. As a result, it tends to assign close-to-zero scores to most of the elements due to the unbounded exponential function in Eq. (8). However, in the scenario of our multi-view attention, such assumption does not hold, since a significant portion of the views may provide informative observations for decision making.

To address this issue, we design a new score assignment strategy, named hybrid focus module, to fully utilize the multi-view information and bring more diversity to attention. In particular, we propose to separately enlarge the details in view domain within each timestamp, while preserving the global energy distribution in time domain. More formally, given the energy matrix $\boldsymbol{E}_t$, we first adopt view-aware smoothing for each timestamp i, as follows:

$$\hat{e}^v_{t,i} = \beta_{t,i}\sigma(e^v_{t,i}) \Big/ \sum_{v=1}^{V} \sigma(e^v_{t,i}) \ ,$$

where $\sigma(\cdot)$ is the bounded sigmoid function used to narrow the energy distance among views within each timestamp, and $\beta_{t,i}$ denotes the coefficient of global energy distribution at timestamp i, defined as:

$$\beta_{t,i} = \frac{\sum_{i,v\in U^v_i} e^{v+}_{t,i}}{\sum_{i=1}^{t-1}\sum_{i,v\in U^v_i} e^{v+}_{t,i}},$$

where U^v_i identifies the group of the v-th view at the i-th timestamp. Here $e^{v+}_{t,i}$ means that only positive values are counted in order to avoid energy offset. Given the smoothed energy matrix $\hat{\boldsymbol{E}}_t$, we then adopt time-aware sharpening for all the timestamps, to obtain the final score matrix $\boldsymbol{A}_t$, as follows:

$$\alpha^v_{t,i} = \exp\left(\gamma\hat{e}^v_{t,i}\right) \Big/ \sum_{i=1}^{t-1}\sum_{v=1}^{V} \exp\left(\gamma\hat{e}^v_{t,i}\right) \ ,$$

where γ is the sharpening factor [4] to prevent aggregating multiple focus. In this way, the proposed hybrid module would focus on the topology of energy distribution rather than scattered points, and thus can help MuVAN preserve more useful information from the learned attention energy.

F. Attentional Feature Fusion

In order to preserve the spatial locality of view-wise characteristics during feature fusion, CNN can be employed. The benefit of adopting CNN is to utilize the layers with non-linear filters to share weights among all the locations in the input, which has shown its superior capability for several content-related tasks, such as image analysis [33] and language modeling [34]. In our model, we obtain the context matrix $\boldsymbol{C}_t \in \mathbb{R}^{2p\times V}$ according to the attention score matrix $\boldsymbol{A}_t$ and the hidden matrix from $\boldsymbol{H}_1$ to $\boldsymbol{H}_{t-1}$, as follows:

$$\boldsymbol{C}_t = \sum_{i=1}^{t-1} \boldsymbol{A}_{t,i} \odot \boldsymbol{H}_i.$$

Given the context matrix, we combine it with the current hidden matrix $\boldsymbol{H}_t$ to generate a 3D-tensor composed of two input planes. The attentional hidden representation can be further obtained using CNN, defined as:

$$\tilde{\boldsymbol{h}}_t = \mathrm{CNN}_{\mathrm{2D}}([\boldsymbol{H}_t \oplus \boldsymbol{C}_t]; \boldsymbol{\theta}_c), \tag{9}$$

where $\oplus$ is the combination operator, and $\mathrm{CNN}_{\mathrm{2D}}$ denotes a series of 2D convolutional-nonlinear-pooling cells with the parameter $\boldsymbol{\theta}_c$. All the features extracted by CNN are then flattened to represent the attentional vector $\tilde{\boldsymbol{h}}_t \in \mathbb{R}^r$. The advantage of the proposed CNN-based attentional feature fusion module is that it not only keeps the content across views, but also extracts the correlations between context and hidden matrix, which further helps MuVAN to enhance the capability of feature representation. Note that the dimension r relies on the input size and the structure of CNN, which are both given in Section 4.2.

Finally, the attentional vector $\tilde{\boldsymbol{h}}_t$ is fed to a softmax layer for classification, as follows:

$$\hat{\boldsymbol{y}}_t = \mathrm{Softmax}(\boldsymbol{W}_s\tilde{\boldsymbol{h}}_t + \boldsymbol{b}_s), \tag{10}$$

where $\boldsymbol{W}_s \in \mathbb{R}^{|\mathcal{C}|\times r}$ and $\boldsymbol{b}_s \in \mathbb{R}^{|\mathcal{C}|}$ are the parameters to be learned.

G. Unified Training Procedure

To train a unified model, we adopt cross-entropy to measure the loss between the ground truth $\boldsymbol{y}_t$ and the $\hat{\boldsymbol{y}}_t$ obtained by Eq. 10. Formally, the final cost function of our end-to-end MuVAN model is defined as:

$$\begin{aligned} &J_{\mathsf{MuVAN}}(\boldsymbol{X}^{(1)}_1, \cdots, \boldsymbol{X}^{(1)}_{T^{(1)}}, \cdots, \boldsymbol{X}^{(M)}_1, \cdots, \boldsymbol{X}^{(M)}_{T^{(M)}}) \\ &= -\frac{1}{M}\sum_{i=1}^{M}\frac{1}{T^{(i)}}\sum_{t=1}^{T^{(i)}}\left[\boldsymbol{y}_t^\top \log\hat{\boldsymbol{y}}_t + (\boldsymbol{1}-\boldsymbol{y}_t)^\top \log(\boldsymbol{1}-\hat{\boldsymbol{y}}_t)\right]. \end{aligned}$$

H. Interpretation

For various real-world applications, interpreting the learned representations is important to understand the practical meanings. We focus on analyzing the interpretability of each view over time, in order to discover which ones are crucial to the real-world task. Since the proposed model is based on multi-view attention mechanism, it is easy to find relevant inputs by analyzing the attention score matrix $\boldsymbol{A}_t$. For the t-th fragment $\boldsymbol{X}_t$, if the attention score $\alpha^v_{t,i}$ is large, then the information of the v-th view at the i-th timestamp has high probability to be related to the current label. Detailed examples and analysis are given in Section 4.4.

IV. Experiments

In this section, we experimentally evaluate the performance of the proposed MuVAN model on three benchmark datasets, compare its performance with other state-of-the-art deep representation learning models, and conduct a case study to show the benefit of the proposed multi-view attention mechanisms in real-world tasks.

A. Dataset Description

In the experiments, three benchmark multivariate temporal datasets from different real-world tasks are used to validate our proposed MuVAN model. The datasets are the CHI-MIT dataset, the UCD dataset, and the MHEALTH dataset, respectively.

The CHB-MIT Dataset
The CHB-MIT dataset is a publicly available multi-channel EEG dataset collected at the Children's Hospital Boston [35]. In this dataset, the EEG signals contain 23 channels (i.e., views) recorded from different brain areas at 256 Hz. The beginning and end of intractable seizures are both annotated in the ground truth. We follow the preprocessing steps adopted in previous work [30]. Finally, the total number of input vectors we generated is $252,862$ by sliding a fix-length window through the entire signals of 23 subjects parameterized by two predefined parameters: window length $l = 3sec$ and step length $s = 1sec$. We use this dataset to show that our model can learn meaningful representations for seizure detection task.

The UCD Dataset
The UCD dataset is a multi-modal PSG dataset provided by St. Vincent's University Hospital and University College Dublin, which can be downloaded from PhysioNet [36]. The PSG time series consist of 14 views from adult subjects, including EEG at 128 Hz, electrooculography (EOG) at 64Hz, electromyography (EMG) at 64Hz, and other signals related to patient movement, posture and breathing. In addition, each 30-second fragment is labeled as in one of the five sleep stages by experts. Different from previous work that select specific views [20], [24] or subject groups [23] using prior knowledge, we generate $287,840$ input vectors from all the 25 subjects and feed all the views into our model. We perform sleep stage classification on this dataset.

The MHEALTH Dataset
MHEALTH [37] is a HAR benchmark dataset, which contains 23 body sensor views while performing 12 physical activities. Multiple sensors are placed on chest, right wrist, and left ankle, i.e., accelerometer, gyroscope, magnetometer, and ECG. All the sensing views are recorded at a sampling rate of 50Hz. We generate $137,494$ input vectors from all the 10 subjects based on the segmentation experience gained in [37], [38], where the window length and step length are set as $l = 5.12sec$ and $s = 1sec$, respectively. Similarly, we employ all the views for evaluation.

Table I provides detailed statistics of each multivariate temporal dataset used in our experiments. We can observe that different datasets have different data formats in terms of sample rates, sensor type, and modalities, which can comprehensively evaluate our model in different situations. Note that all the datasets are imbalanced.

TABLE I
STATISTICS OF EACH EXPERIMENTAL DATASET: TASKS INCLUDE SEIZURE DETECTION (SD), SLEEP STAGE CLASSIFICATION (SSC), AND HUMAN ACTIVITY RECOGNITION (HAR).

Dataset	CHB-MIT	UCD	MHEALTH
# of fragments	252,862	287,840	137,494
# of classes	2	5	12
# of views	23	14	23
- # of modalities	1	12	4
- # of sensors	23	14	8
Sample rate(s) (Hz)	256	128, 64, 8, 4	50
Task	SD	SSC	HAR

B. Experiment setup

In this subsection, we first introduce the state-of-the-art deep representation learning approaches which are used as baselines, and then outline the criteria used for evaluation. Finally, we describe the implementation details.

Baseline Approaches
To validate the performance of the proposed model for different real-world tasks, we compare it with several state-of-the-art models. We select the following eight existing approaches as baselines:

• RNN. RNN is a commonly used baseline for global feature learning. We first concatenate all the inputs into a vector space, and then feed it to the BGRU. The hidden representations produced by the BGRU are directly used for task-related training using softmax.

• RNNAtt. We incorporate attention mechanism into RNN. After the BGRU outputs the hidden vectors $\boldsymbol{h}_{1:t}$, RNNAtt adopts attention module to obtain a context vector $\boldsymbol{c}_t$. Then, RNNAtt concatenates both $\boldsymbol{c}_t$ and $\boldsymbol{h}_t$ as an attentional representation for final training. For the sake of fairness, two existing strategies, namely location-based and concatenation-based attention [6], are employed, denoted as RNNAtt$_l$ and RNNAtt$_c$, respectively.

• vRNN. vRNN is an RNN variant considering view-specific characteristics, which is widely used in several view-related tasks [25], [27], [38]. We first feed each view to the BGRU based on Eq. (3), and then concatenate the hidden representations of all the views into a vector space to train an end-to-end model using softmax.

• vRNNAtt. vRNNAtt employs attention mechanism on each view after the feature extraction of vRNN. Similarly, we perform the same process as RNNAtt, denoted as vRNNAtt$_l$ and vRNNAtt$_c$, respectively.

• CNN. We first integrate the inputs from all views as a matrix, and then extract features through a plain convolutional architecture with the same structure in Eq. (9). The learned hidden representations are directly used for task-related training using softmax.

• ChannelAtt [19]. ChannelAtt focuses on soft-selecting critical views from multivariate signals. Compared to RNNAtt, ChannelAtt adopts a new global attention mechanism in the view domain instead of the time domain.

Our Approaches
We show the performance of the following three approaches

TABLE II
CNN STRUCTURE OF THE ATTENTIONAL FEATURE FUSION MODULE IN MUVAN.

Cell No.	Conv	Non-linear	Pooling
1	$5 \times 5 \times 16$	ReLU	2×2
2	$5 \times 5 \times 32$	ReLU	2×2

in the experiments.

• MuVAN−. MuVAN− is a reduced model that only uses the hidden representations generated by Eq. (3), i.e., without employing any multi-view attention mechanisms.

• MuVAN$_{loc}$. It is based on location-based multi-view attention mechanism with softmax-based normalization in the unified model.

• MuVAN$_{ctx}$. This model uses context-based multi-view attention and hybrid focus procedure when calculating the score matrix.

Evaluation criteria

Since the evaluation tasks belong to classification problem, we use F1-score and Accuracy to validate our model. Moreover, the area-under-the-curve of receiver operator characteristic (AUCROC) and precision-recall (AUCPR) scores are also employed to numerically evaluate the quality of each method. Regarding the tasks related to multi-class classification problem, here we show both the Macro-F1 and Micro-F1 scores. Macro-F1 score biases the metric towards the least populated labels, while Micro-F1 score biases towards to the most populated labels. Note that the AUCROC and AUCPR scores are both based on Macro metric.

Implementation Details

We implement all the approaches with Pytorch [39]. In order to conduct subject-independent evaluations, we randomly split each dataset by subjects into the training, validation and testing sets with a $0.7 : 0.1 : 0.2$ ratio (i.e., the model is never trained on data from both the validation and testing subjects). The validation set is used to determine the best values of parameters within 100 iterations. We repeat experiments with different data combination for 5 times (or folds) and report the average test performance for each method. Note that, to fairly compare the performance, we adopt the same data combination for all models at each fold. During the whole training step, we use Adadelta [40] with mini-batch to minimize the cost function. We also use momentum ($\rho = 0.95$), weight decay ($L2$ penalty with the coefficient 0.001), and dropout strategies (the dropout rate is 0.5) for all the approaches. Furthermore, the CNN structure of our attentional feature fusion module is shown in Table II, and we set the same $p = 128$ and $k = 128$ for baselines and our models.

C. Performance on Real-world Tasks

In this subsection, we investigate the effectiveness of MuVAN compared to other models in different real-world tasks, including seizure detection, sleep stage classification, and human activity recognition, respectively.

TABLE III
DETECTION PERFORMANCE COMPARISONS ON THE CHB-MIT DATASET.

	CHB-MIT Dataset			
Method	AUCROC	AUCPR	F1-score	Accuracy
RNN	0.9343	0.8274	0.6201	0.8775
RNNAtt$_l$	0.9251	0.7794	0.7149	0.8958
RNNAtt$_c$	0.9289	0.7780	0.6517	0.9021
vRNN	0.9414	0.8467	0.7529	0.8840
vRNNAtt$_l$	0.9484	0.8941	0.7924	0.9255
vRNNAtt$_c$	0.9522	0.9056	0.8199	0.9268
CNN	0.9263	0.8702	0.7959	0.9088
ChannelAtt	0.9556	0.9119	0.8675	0.9506
MuVAN−	0.9480	0.8645	0.7561	0.9189
MuVAN$_{loc}$	0.9749	0.9233	0.8916	0.9566
MuVAN$_{ctx}$	**0.9832**	**0.9654**	**0.9238**	**0.9705**

Results of Seizure Detection

In this set of experiments, we evaluate the performance of various models for detecting epileptic seizure onset on the CHB-MIT dataset, as shown in Table III. Given the results of baselines, we can observe that the attention-based RNN models perform better than plain RNN methods both in global and view-wise strategies. This is because attention mechanism can help model to learn reasonable parameters and hence make correct detection. We can also see that the results of view-wise attention models perform on par with those of global-based attention models. The reason is that all the views belong to the same modality, i.e., EEG, which makes both strategies easy to identify critical seizure patterns. Moreover, the ChannelAtt model adopting view-wise attention performs better than the time-wise attention models. This results from the fact that brain related activities are often associated with several different brain areas. There exist more hidden connections across views related to the EEG seizures. This observation can also be found from the performance comparison between RNN and MuVAN−, where MuVAN− achieves better results since the spatial information is remained instead of simply adopting concatenation.

From the results, both our MuVAN$_{loc}$ and MuVAN$_{ctx}$ models outperform the baselines on all five different evaluation measurements. Specifically, MuVAN$_{ctx}$ significantly outperforms all other models. For example, MuVAN$_{ctx}$ obtains the best of 0.9238 on F1-score compared with 0.8916 and 0.8675 achieved by our MuVAN$_{loc}$ model and the ChannelAtt baseline, respectively. This means that MuVAN$_{ctx}$ is able to enhance feature representation in imbalanced classes. Based on the overall performance comparisons on the CHB-MIT dataset, we can demonstrate that our proposed multi-view attention mechanism can improve the performance for seizure detection task in healthcare.

Results of Sleep Stage Classification

In this set of experiments, we evaluate the performance of various models for the purpose of sleep stage classification. Table IV shows the classification results on the UCD dataset.

TABLE IV
CLASSIFICATION PERFORMANCE COMPARISONS ON THE UCD DATASET.

Method	UCD Dataset				
	AUCROC	AUCPR	MacroF1	MicroF1	Accuracy
RNN	0.6228	0.3350	0.2663	0.3970	0.5091
RNNAtt$_l$	0.6172	0.3305	0.2457	0.3734	0.5002
RNNAtt$_c$	0.6234	0.3335	0.2554	0.3712	0.5010
vRNN	0.8489	0.5923	0.5523	0.6287	0.6538
vRNNAtt$_l$	0.8433	0.5858	0.5579	0.6342	0.6619
vRNNAtt$_c$	0.8529	0.5970	0.5845	0.6530	0.6717
CNN	0.8732	0.6725	0.5925	0.6492	0.6590
ChannelAtt	0.8662	0.6458	0.6137	0.6773	0.6859
MuVAN−	0.8474	0.6023	0.6022	0.6952	0.7180
MuVAN$_{loc}$	0.8538	0.6105	0.6200	0.7068	0.7193
MuVAN$_{ctx}$	**0.8722**	**0.6611**	**0.6510**	**0.7313**	**0.7430**

TABLE V
RECOGNITION PERFORMANCE COMPARISONS ON THE MHEALTH DATASET.

Method	MHEALTH Dataset				
	AUCROC	AUCPR	MacroF1	MicroF1	Accuracy
RNN	0.9835	0.9459	0.8684	0.8582	0.8915
RNNAtt$_l$	0.9765	0.9196	0.8858	0.8658	0.8931
RNNAtt$_c$	0.9720	0.9163	0.9026	0.8949	0.9106
vRNN	0.9849	0.9204	0.9080	0.9062	0.9091
vRNNAtt$_l$	0.9868	0.9267	0.9197	0.9191	0.9266
vRNNAtt$_c$	0.9833	0.9136	0.9400	0.9384	0.9399
CNN	0.9821	0.8992	0.9134	0.9122	0.9167
ChannelAtt	0.9848	0.9336	0.9296	0.9276	0.9310
MuVAN−	0.9825	0.9334	0.9358	0.9328	0.9352
MuVAN$_{loc}$	0.9846	0.9287	0.9262	0.9252	0.9328
MuVAN$_{ctx}$	**0.9915**	**0.9576**	**0.9600**	**0.9678**	**0.9698**

In the task of PSG-based sleep stage classification, the variability of rhythmic patterns in different modalities makes it hard to train an effective model with simple design. This can be observed from the bad performance of all the global-based models, which shows that directly concatenating raw features into model may not work on multimodal data. However, taking the view-wise strategy into consideration, the performance enormously increases.

Although adopting view-wise learning significantly improves the classification performance, given the results of vRNNAtt$_c$ using complicated attention mechanism, only relatively minor gains are achieved in terms of Accuracy and F1 scores, compared with vRNN and CNN. This observation can also be found from the performance of ChannelAtt in terms of AUCROC and AUCPR, which illustrates that in sleep PSGs data, the inherent dependencies shift across views over time, and hence simply using multilayer perceptrons on each domain cannot work well.

Our MuVAN model still performs better than the baselines, which demonstrates that the proposed multi-view attention mechanism can help the models to enhance the ability of feature representation. Furthermore, MuVAN− does not use any attention mechanism, but the performance is higher than other baselines, which shows that keeping spacial information can improve the recognition performance. Thus, it is reasonable to employ matrix-based representation for feature learning. The limited improvement of MuVAN$_{loc}$, compared with MuVAN−, demonstrates that the learned context matrix using location-based attention does not provide much useful information. Our proposed MuVAN$_{ctx}$ model, on the other hand, yields better results than the other methods, which confirms that the proposed context-based attention mechanism can focus on more useful information from both view and time domains and hence help learn better representations.

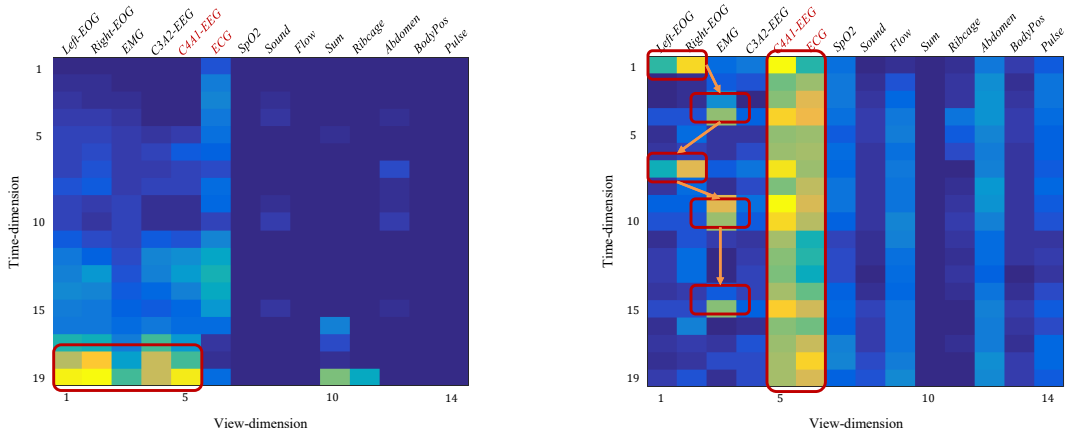
Results of Human Activity Recognition

We also evaluate the performance for human activity recognition. The experimental results on the benchmark MHEALTH dataset are listed in Table V. Among all the approaches, similarly, the results of global-based models still perform worse than those of view-wise models, under the influence of multiple modalities. We observe that vRNNAtt$_c$ outperforms all other baselines in terms of Accuracy and F1 scores, including ChannelAtt, and even performs better than the proposed MuVAN− and MuVAN$_{loc}$ on this dataset. This is because of the fact that, in human related activities, there exist more distinctive rhythmic patterns within views than those across views, which illustrates the advantages of time-wise attention mechanism over other attention strategies. However, as we mentioned before, vRNNAtt$_c$ fails to perform well across different datasets and tasks. On the other hand, our MuVAN$_{ctx}$ model consistently achieves better results than others. It not only achieves better Accuracy and F1 scores, but also obtains higher AUC-ROC and AUC-PR than baselines. Moreover, the comparison between MuVAN$_{ctx}$ and vRNNAtt$_c$ indicates that the context-based attention mechanism provides complementary information carried by multiple views and hence learns more meaningful representations that are helpful for human activity recognition.

Based on all the above analysis, we can conclude that the conventional single-dimension attention mechanisms may lose critical information, and hence do not work well dealing with complex data structure. The context-based multi-view attention mechanism achieves better results on all the datasets compared with the location-based multi-view attention mechanism. We arrive at a conclusion that our proposed multi-view attention model indeed learns informative representations to improve the performance in different real-world tasks.

D. Case Study

To demonstrate the benefit of adopting proposed multi-view attention mechanisms in real-world tasks, in this part, we analyze the attention scores learned from both of our MuVAN$_{loc}$ and MuVAN$_{ctx}$ models. In addition, we interpret how the multi-view attention works using different mechanisms to affect the quality of learned feature representations. Fig. 3 shows a case study for PSG sleep stage classification on the UCD dataset



(a) Score matrix leaned from MuVAN$_{loc}$

(b) Score matrix leaned from MuVAN$_{ctx}$

Fig. 3. A case study for sleep stage classification of a patient using different multi-view mechanisms from MuVAN on the UCD dataset.

where a patient woke up suffering from central sleep apnea. We choose this task because of the complex PSG data structure and relationships among views and timestamps. Specifically, the score matrix $\boldsymbol{A}_t$ for the 20-th timestamp is calculated according to previous representations from 1 to 19 timestamps with 14 views. Thus, in Fig. 3, X-axis represents views and Y-axis denotes the previous timestamps. We can observe that the score matrices learned by the proposed two multi-view attention mechanisms are different since they adopt different focus strategies. This is reasonable in visual inspection, since physicians may make the same diagnosis based on different observations [41], [42], which also reflects the uncertainty and variability in decision making.

Analyzing the score matrix of MuVAN$_{loc}$ from Fig. 3a, which utilizes location-based attention mechanism with softmax-based normalization, we observe that the recent two timestamps with the first five views significantly contribute to the current sleep stage. This demonstrates that the previous eye movements with salient brain activities cause the patient to wake up. Different from MuVAN$_{loc}$, the score matrix learned from MuVAN$_{ctx}$ in Fig. 3b, using context-based attention mechanism with hybrid focus normalization, provides an in-depth interpretation. We can observe that C3-A1 EEG and ECG are the two most active views, and the contributions of EOG and EMG appear in turns. There is a strong probability that the wake-up is related to the nervous system irregularities which trigger the heart abnormalities and muscles movements. According to [43], the second interpretation conforms more to the pathology of central sleep apnea. To sum up, between two multi-view attention mechanisms, MuVAN$_{loc}$ can focus on surface phenomena that are close to the current timestamp, while MuVAN$_{ctx}$ tends to incorporate topological correlations to derive explicit explanations with more practical meanings. In addition, as shown in Table IV, the meaningful representations support the effectiveness of our model to yield good performance in sleep stage classification.

V. Conclusions

In this paper, we propose a novel multi-view attention network, named MuVAN, to address the challenges of modeling multivariate temporal data. MuVAN is a unified model that consists of several components: (1) view-wise recurrent encoder (2) multi-view energy assignment, (3) hybrid focus, and (4) spatial attentional feature fusion. Two multi-view attention mechanisms are developed to jointly learn two-dimensional attention scores to estimate the quality of information contributed by each view over time. Experimental results on three benchmark datasets justify the effectiveness of our proposed MuVAN model in real-world tasks. Analytical results through a case study demonstrate that MuVAN can discover meaningful attention score matrices to provide complementary information to enhance feature representation. As the proposed multi-view attention mechanism is end-to-end and task-oriented, it is applicable to other applications with similar data structure, especially in pervasive sensing where interpretable broad learning is still a major challenge.

Acknowledgment

This paper is supported by the National Natural Science Foundation of China (61672064), the Beijing Natural Science Foundation (4172001), the Science and Technology Project of Beijing Municipal Education Commission (KZ201610005007), Beijing Laboratory of Advanced Information Networks (040000546617002), and the China Scholarship Council Fund (201606540008). The authors would like to thank the anonymous reviewers, and NVIDIA Corporation for the donation of the Titan Xp GPU.

References

[1] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[3] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4960–4964.

[4] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[5] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao, "Kame: Knowledge-based attention model for diagnosis prediction in healthcare," in *Proceedings of CIKM*. ACM, 2018.

[6] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 1903–1911.

[7] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "Gram: Graph-based attention model for healthcare representation learning," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 787–795.

[8] F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, and A. Zhang, "Risk prediction on electronic health records with prior medical knowledge," in *Proceedings of SIGKDD*. ACM, 2018, pp. 1910–1919.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010.

[10] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[11] R. Moskovitch and Y. Shahar, "Classification-driven temporal discretization of multivariate time series," *Data Mining and Knowledge Discovery*, vol. 29, no. 4, pp. 871–913, 2015.

[12] R. Rawassizadeh, E. Momeni, C. Dobbins, J. Gharibshah, and M. Pazzani, "Scalable daily human behavioral pattern mining from multivariate temporal data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 11, pp. 3098–3112, 2016.

[13] I. Batal, G. F. Cooper, D. Fradkin, J. Harrison, F. Moerchen, and M. Hauskrecht, "An efficient pattern mining approach for event detection in multivariate temporal data," *Knowledge and information systems*, vol. 46, no. 1, pp. 115–150, 2016.

[14] M. Ghassemi, M. A. Pimentel, T. Naumann, T. Brennan, D. A. Clifton, P. Szolovits, and M. Feng, "A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data." in *AAAI*, 2015, pp. 446–453.

[15] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2018.

[16] Y. Yuan, G. Xun, K. Jia, and A. Zhang, "A novel wavelet-based model for eeg epileptic seizure detection using multi-context learning," in *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*. IEEE, 2017.

[17] S. Saria, D. Koller, and A. Penn, "Learning individual and population level traits from clinical temporal data," in *Proc. Neural Information Processing Systems (NIPS), Predictive Models in Personalized Medicine workshop*. Citeseer, 2010.

[18] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems*, 2016, pp. 3504–3512.

[19] Y. Yuan, G. Xun, F. Ma, Q. Suo, H. Xue, K. Jia, and A. Zhang, "A novel channel-aware attention framework for multi-channel eeg seizure detection via multi-view deep learning," in *Biomedical & Health Informatics (BHI), 2018 IEEE EMBS International Conference on*. IEEE, 2018.

[20] M. Manzano, A. Guillén, I. Rojas, and L. J. Herrera, "Deep learning using eeg data in time and frequency domains for sleep stage classification," in *International Work-Conference on Artificial Neural Networks*. Springer, 2017, pp. 132–141.

[21] A. Supratak, H. Dong, C. Wu, and Y. Guo, "Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.

[22] L. F. Polania, L. K. Mestha, D. T. Huang, and J.-P. Couderc, "Method for classifying cardiac arrhythmias using photoplethysmography," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 6574–6577.

[23] J. Zhang, Y. Wu, J. Bai, and F. Chen, "Automatic sleep stage classification based on sparse deep belief net and combination of multiple classifiers," *Transactions of the Institute of Measurement and Control*, vol. 38, no. 4, pp. 435–451, 2016.

[24] M. Längkvist, L. Karlsson, and A. Loutfi, "Sleep stage classification using unsupervised feature learning," *Advances in Artificial Neural Systems*, vol. 2012, p. 5, 2012.

[25] Y. Yuan, G. Xun, K. Jia, and A. Zhang, "A multi-view deep learning method for epileptic seizure detection using short-time fourier transform," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2017, pp. 213–222.

[26] S. Ha and S. Choi, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, pp. 381–388.

[27] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 351–360.

[28] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate lstm-fcns for time series classification," *arXiv preprint arXiv:1801.04503*, 2018.

[29] R. Moskovitch and Y. Shahar, "Classification of multivariate time series via temporal abstraction and time intervals mining," *Knowledge and Information Systems*, vol. 45, no. 1, pp. 35–74, 2015.

[30] Y. Yuan, G. Xun, Q. Suo, K. Jia, and A. Zhang, "Wave2vec: Learning deep representations for biosignals," in *Data Mining (ICDM), 2017 IEEE 16th International Conference on*. IEEE, 2017.

[31] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[32] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[33] W. Zhang, R. Li, T. Zeng, Q. Sun, S. Kumar, J. Ye, and S. Ji, "Deep model based transfer and multi-task learning for biological image analysis," *IEEE Transactions on Big Data*, 2016.

[34] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[35] A. H. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[36] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[37] O. Banos, R. Garcia, J. A. Holgado-Terriza, M. Damas, H. Pomares, I. Rojas, A. Saez, and C. Villalonga, "mhealthdroid: a novel framework for agile development of mobile health applications," in *International Workshop on Ambient Assisted Living*. Springer, 2014, pp. 91–98.

[38] H. Guo, L. Chen, L. Peng, and G. Chen, "Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 1112–1123.

[39] A. Paszke, S. Gross, and S. Chintala, "Pytorch," https://github.com/pytorch/pytorch, 2017.

[40] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[41] Q. Suo, F. Ma, Y. Yuan, M. Huai, W. Zhong, J. Gao, and A. Zhang, "Deep patient similarity learning for personalized healthcare," *IEEE Transactions on NanoBioscience*, 2018.

[42] Q. Suo, W. Zhong, F. Ma, Y. Yuan, M. Huai, and A. Zhang, "Multi-task sparse metric learning on monitoring patient similarity progression," in *Data Mining (ICDM), 2018 IEEE International Conference on*. IEEE, 2018.

[43] C. Guilleminault, A. Tilkian, and W. C. Dement, "The sleep apnea syndromes," *Annual review of medicine*, vol. 27, no. 1, pp. 465–484, 1976.