GlobalFusion: A Global Attentional Deep Learning Framework for Multisensor Information Fusion

SHENGZHONG LIU*, University of Illinois at Urbana-Champaign
SHUOCHAO YAO*, University of Illinois at Urbana-Champaign
JINYANG LI, University of Illinois at Urbana-Champaign
DONGXIN LIU, University of Illinois at Urbana-Champaign
TIANSHI WANG, University of Illinois at Urbana-Champaign
HUAJIE SHAO, University of Illinois at Urbana-Champaign
TAREK ABDELZAHER, University of Illinois at Urbana-Champaign

The paper enhances deep-neural-network-based inference in sensing applications by introducing a lightweight attention mechanism called the *global attention module* for multi-sensor information fusion. This mechanism is capable of utilizing information collected from higher layers of the neural network to selectively amplify the influence of informative features and suppress unrelated noise at the fusion layer. We successfully integrate this mechanism into a new end-to-end learning framework, called *GlobalFusion*, where two global attention modules are deployed for spatial fusion and sensing modality fusion, respectively. Through an extensive evaluation on four public human activity recognition (HAR) datasets, we successfully demonstrate the effectiveness of GlobalFusion at improving information fusion quality. The new approach outperforms the state-of-the-art algorithms on all four datasets with a clear margin. We also show that the learned attention weights agree well with human intuition. We then validate the efficiency of GlobalFusion by testing its inference time and energy consumption on commodity IoT devices. Only a negligible overhead is induced by the global attention modules.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing theory, concepts and paradigms**; • **Computing methodologies → Machine learning**.

Additional Key Words and Phrases: Internet of Things (IoT), neural networks, multisensor information fusion.

ACM Reference Format:
Shengzhong Liu*, Shuochao Yao*, Jinyang Li, Dongxin Liu, Tianshi Wang, Huajie Shao, and Tarek Abdelzaher. 2020. GlobalFusion: A Global Attentional Deep Learning Framework for Multisensor Information Fusion. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 19 (March 2020), 27 pages. https://doi.org/10.1145/3380999

1 INTRODUCTION

In several multi-sensor application contexts [20, 21, 29, 30, 38, 46], deep learning algorithms have shown non-trivial accuracy improvements over conventional feature-engineering-based machine learning methods [43, 48],

[*] indicates equal contribution.

Authors' addresses: Shengzhong Liu*, University of Illinois at Urbana-Champaign, sl29@illinois.edu; Shuochao Yao*, University of Illinois at Urbana-Champaign, syao9@illinois.edu; Jinyang Li, University of Illinois at Urbana-Champaign, jinyang7@illinois.edu; Dongxin Liu, University of Illinois at Urbana-Champaign, dongxin3@illinois.edu; Tianshi Wang, University of Illinois at Urbana-Champaign, tianshi3@illinois.edu; Huajie Shao, University of Illinois at Urbana-Champaign, hshao5@illinois.edu; Tarek Abdelzaher, University of Illinois at Urbana-Champaign, zaher@illinois.edu.

2474-9567/2020/3-ART19 $15.00
https://doi.org/10.1145/3380999

motivating a closer look at the use of deep neural networks for multisensor data fusion [39]. Recent work developed novel neural network architectures for sensor data processing [25, 36, 40, 41], and neural network model reduction techniques for resource-constrained Internet-of-Things (IoT) devices [7, 18, 19, 44]. A key advantage of deep-learning-based solutions over the plethora of model-based approaches lies in reducing the human design burden. In a world increasingly dominated by data and computing power, trade-offs that replace human effort with machine-centered albeit data-intensive approaches are becoming increasingly attractive.

Personal devices, such as smart phones, smart watches, and fitness trackers, are typically equipped with multiple sensors that can be collaboratively utilized to capture user context, perform environmental measurements, and recognize body movements. Several devices may coexist in a typical body network. However, not all devices and modalities are equally useful for detecting a given class of outputs at all times. For example, when detecting walking, a wrist watch or a fitbit would usually be very helpful. However, when wrists are confined, say, by pushing a cart in a grocery store, a cell-phone in a back pocket might work better. How can one automatically decide, based on current measurement features, where to pay attention in a given situation to detect a given class of activity? This automatic attention guidance mechanism is the topic of the paper.

The work extends traditional ensemble methods by understanding global context in which given local sensor outputs may be "misleading". For example, the smart watch and the fitbit on the wrists of a person pushing the shopping cart might generate high confidence outputs saying the person is sitting in a slowly moving vehicle. If the only other device on the person correctly detects walking, it might be out-voted. An attention mechanism, in contrast, can attenuate the false claims (despite their associated high local confidence) because the global context suggests that these sensors are presently not in a position to yield accurate local results.

Attention mechanisms are an emerging technique for dynamically adjusting neural network's focus by scaling the features using corresponding weights computed by a separate attention module. Given an appropriate attention design, the network can automatically amplify the influence of informative features and suppress unrelated noise. In conventional attention mechanisms for regression problems, the extraction of features weights are often computed from matching features from the input signal and features from the output signal. The features used to estimate input features importance are called the *query*. For example, in neural machine translation problems, the word embeddings in the output sentence are used as the query to find alignment with each input sentence word. How to design a query (that can accurately identify informative features) in classification problems with no output signal but just an output label is a key challenge in attention mechanism research.

In this paper, we follow an idea called *self attention* proposed by Vaswani *et al.* [32] where the query is also extracted from the input features. Our insight is that we try to match the neural network node outputs (at certain layers) against inputs at earlier layers with the idea that inputs that are more correlated with outputs deserve more attention. We propose a *global attention module* that (i) uses *high-level node features* (i.e., features of nodes that are closer to the output layer) to estimate the contribution (and hence, scaling factor) of each input feature vector in the low-level fusion layer, then (ii) adds these scaled local features to the output (high-level node) features. The design literally superimposes selected local context (scaled low-level features) and global context (high-level node features), allowing subsequent nodes to consider the combination of the two. By adding local context and global context, more informed decisions can be made based on the combination. For example, the system might learn that when a hip-mounted device detects features compatible with walking, while multiple wrist-mounted devices detect features compatible with sitting or standing, the latter devices should be suppressed, as the ground truth is usually more correlated with the former. The attention module is an automated mechanism to learn such global feature weighting policies.

The idea of using information at higher layers to guide attention comes from the commonly observed fact [45] that, in a deep neural network, lower level features are local to the input and general to the task, whereas higher-level features are global to the input but specific to the particular output class. In other words, the high-level features contain more information related to specific output classes. Our attention module design is

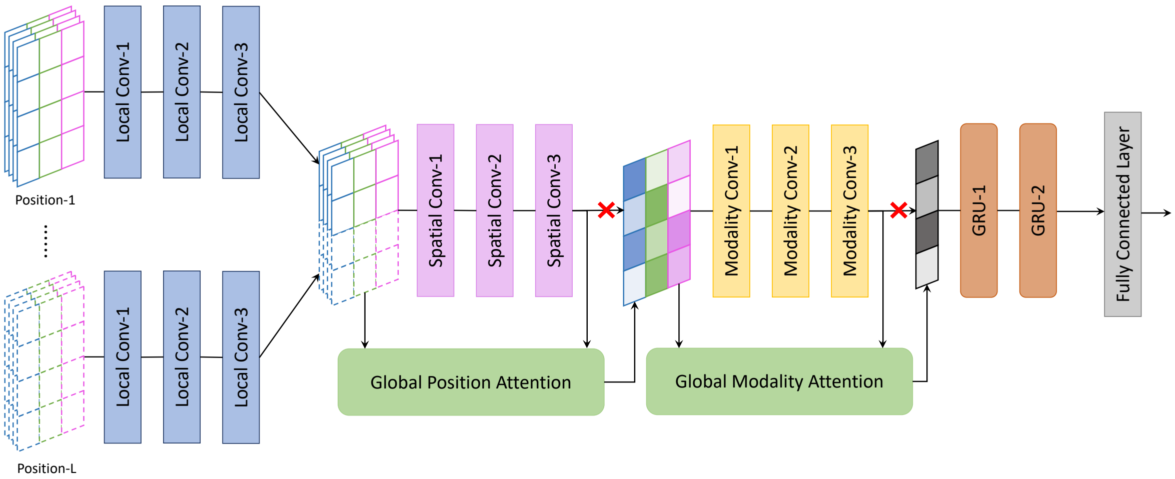


Fig. 1. GlobalFusion framework overview. (We use different dash lines to represent features from different sensor positions, and use different colors to represent features from different sensing modalities.)

successfully integrated into an end-to-end learning framework, called GlobalFusion, where two global attention modules, named global position attention and global modality attention, are deployed to fuse information from heterogeneous sensing modalities and diverse sensor positions, respectively. Figure 1 provides an overview of the proposed GlobalFusion framework. Our backbone network is based on DeepSense proposed by Yao *et al.* [40]. Actually, global attention is a flexible and configurable module that can be used as building block in most state-of-the-art sensing data processing frameworks when fusing heterogeneous information. The global attention module is implemented by a lightweight shallow convolution module, so that its incurred computational overhead is low.

Our design is targeted at general IoT applications because we do not rely on any application-specific insight. Rather, we use a pure data-driven approach. However, due to the limitations of current publicly available sensing datasets, we demonstrate the effectiveness of GlobalFusion on four public human activity recognition (HAR) datasets: 1) PAMAP2 [27], 2) Realworld-HAR [31], 3) DSADS [4], and 4) DG [5]. In these datasets, we are given multiple types of sensor readings collected from several body positions to infer the human activities. We compare GlobalFusion to both the state-of-the-art non-attentional DeepSense [40] framework, and attentional frameworks, including SADeepSense [42], BANet [34], and attnLSTM [47]. GlobalFusion is able to consistently outperform these algorithms with a clear margin on all datasets. Through a qualitative analysis of attention weight distribution, we also demonstrate the interpretability of global attention design. Finally, we test the inference time and energy consumption of GlobalFusion on two commodity IoT devices, Nexus 5 and Raspberry Pi 3 Model B. The results show that the overhead of our global attention module is negligible compared to the backbone DeepSense network.

The rest of this paper is organized as follows. Section 2 briefly reviews recent literature in multisensor information fusion for IoT applications. The design details of the global attention module and the GlobalFusion framework are explained in Section 3. We describe the evaluation results in Section 4. After a brief discussion about existing issues and future work in Section 5, we conclude the paper in Section 6.

2 RELATED WORK

Textbooks have been written on classical multisensor data fusion [12] prior to the recent resurrection of neural networks research. These traditional techniques fuse carefully-designed features from each sensor thus calling for a human feature engineering effort. We shall not consider these approaches further as we seek a fully automated machine learning solution.

Deep learning revolutionized data fusion as it obviates feature engineering, instead ingesting raw sensor measurements only. Several deep-learning-based fusion methods have recently been proposed [22, 24, 26, 33, 39, 40]. For example, Yao *et al.* [40] concatenate different modality representations and use a convolution operation to fuse their information. Others explicitly model sensor interactions and correlations for better fusion quality [22, 24, 26, 33, 39]. While these approaches compute global context from local data, the "wiring" of global context as a function of local data is fixed. In contrast, an attention mechanism allows selective retrieval of some local data to add to the global context for further processing, in essence allowing for different weighting of the same local feature in different global contexts.

Indeed, a key challenge in information fusion is to dynamically understand which inputs or features are more important when. This is akin to attention mechanisms in neural networks [6, 8, 23, 37]. It is an emerging technique for dynamically adjusting neural network model's focus by multiplying the features of each sensor with a corresponding weight, where the weight is dynamically estimated by an independent module based on the sensor inputs. Different solutions vary in their choice of weight calculation methods for fusion inputs [25, 34, 42, 47].

The attention mechanisms used in multisensor fusion are generally divided into two categories, called *additive attention* and *multiplicative attention*, respectively. In additive attention designs [34, 47], a small fully connected neural network is utilized to learn the weight of sensors / positions directly based on their inputs. In multiplicative attention, the weight of each sensor or position is decided by the compatibility between its features and a special feature vector, called *query*. The compatibility function is typically defined as a dot product of two vectors. In the sensor fusion problem, multiplicative attention is more intuitive because we can define the relevance of sensor features through a corresponding query design, unlike the black-box implementation of additive attention. The choice of the query directly decides the attention weight received by each sensor/position. For example, Yao *et al.* [42] use the mean of all sensor feature vectors as the query to estimate the attention weight of local features from each sensor component. They rely on the assumption that sensing information is highly correlated while the noise is not. However, this solution does not address different sensing modalities well because the information contained in fusion inputs are probably dissimilar but complementary to each other.

In contrast, in our design, we propose to use aggregated information from *higher layers* of the neural network to estimate the importance of local sensor features. We show that such an approach outperforms others because it is able to choose weights based on more advanced features, not available (i.e., not yet computed) at lower layers of the neural network. A similar idea of using aggregated global information to selectively emphasize informative local features originated in recent efforts on applying convolutional neural networks (CNN) to image recognition [9, 14, 15, 35]. In Squeeze-and-Excitation (SE) block [14]. Hu *et al.* explore channel relationships (i.e., RGB channels) in an image by stacking an information gathering stage (i.e., squeeze block) with a following information distribution stage (i.e, excitation block) to adaptively recalibrate channel features. In [15], Jetley *et al.* leverage global image representation fed to the last classification layer as the query, to estimate weights of local area features at intermediate convolution layers, after which the highlighted local features are output directly for classification. Both of them have observed significant improvement in image recognition accuracy.

To the best of our knowledge, we are the first to apply *global information based attention design* to *multisensor fusion* for IoT applications. Compared to the channels or local areas within an image, differences in the nature of information obtained from different sensors or spatial positions is a key challenge that hinders the direct application of global information based attention mechanism here. We tackle this challenge as follows. First, compared to [14], we explore a different method in information gathering stage. Instead of using a fully connected layer, we leverage the three-layer convolution module in DeepSese [40] to gather global information from sensors/spatial positions, which is known to be both effective and efficient in extracting representative features from multi-channel input. The gathered global information is used as the *global query* to recalibrate the *local features*. Second, compared to [15], we divide the information fusion into a hierarchy of two stages, named *modality fusion* and *position fusion*, respectively. Position fusion is performed before sensor fusion. The global information gathered from sensors/positions are immediately sent back to recalibrate the local features. We maximally preserve the flexibility in attention distributions to tackle information heterogeneity: Different sensors correspond to their own position attention distributions. Similarly, considering the time-varying sensing quality, the sensor fusion at different time intervals is independently performed with no information interference from other intervals.

3 GLOBALFUSION FRAMEWORK

Our idea is motivated by the commonly observed fact [45] that in a deep neural network, the lower level features are local to the input and general to the task, while higher level features are global to the input and specific to particular classes. We call the feature vectors at lower layers as *local features*, while call the feature vectors at higher layers as *global features*. Therefore, juxtaposing local features and global information can help improve classification based on the combination. In this section, we first describe the general architecture of GlobalFusion, which is an end-to-end deep learning framework designed for multisensor information fusion. Walking through the GlobalFusion architecture, we point out the positions in network where we need a global attention module and outline requirement for its correct functionality. Next, we explain the technical details of global attention module design for multisensor information fusion, especially how to utilize high level global features to compute low level attention weights.

Before diving into the details, we first introduce the notations used in the rest of this paper. All vectors are denoted by bold lower-case letters (e.g., $\mathbf{x}$ and $\mathbf{y}$), while matrices and tensors are represented by bold upper-case letters (e.g., $\mathbf{X}$ and $\mathbf{Y}$). For a vector $\mathbf{x}$, the j^{th} element is denoted by $x_{[j]}$. For a tensor $\mathbf{X}$, the t^{th} matrix along the first axis is denoted by $X_{[t \cdot \cdot]}$, and other slicing denotations are defined similarly. Assume we have L spatial positions and S distinct sensor types deployed at each position, $\mathbf{X}^{sl}$ means the s-th sensor reading at l-th position. For each layer k, we use $\mathbf{X}^{(k)}$ to represent the input to this layer, and use $\mathbf{Y}^{(k)}$ to denote corresponding output. For any tensor $\mathbf{X}$, $|\mathbf{X}|$ denotes the size of $\mathbf{X}$.

3.1 GlobalFusion Architecture

Before introducing the technical design of global attention, we first give an end-to-end overview of the GlobalFusion framework for multisensor information fusion. GlobalFusion is based on the state-of-the-art DeepSense[40] framework as back-bone network. It exploits the power of both Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN) in time-series sensing data processing. In the vanilla DeepSense model, fusion of musitisensor inputs is performed by a three-layer convolution module after concatenating inputs from all fusing components. By doing so, the model does not take the heterogeneity among input sources into consideration and pays equal attention to each fusing component. Thus, the fusion layer cannot maximize the information extracted from all information sources (i.e., sensing modalities and body positions). To solve this problem, we cut the direct connection between output of information fusion convolution module and the input of the next layer, and add one global attention module between them to first enhance fusion output with complementary local features before feeding it into next layer. The overall architecture of GlobalFusion is presented in Figure 1. In this subsection, we temporarily regard the global attention module as a black-box implementation which is able to

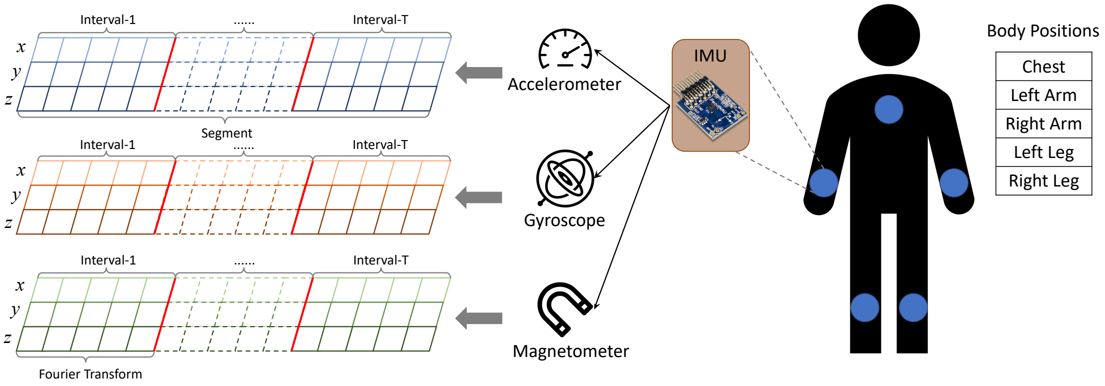


Fig. 2. An example illustrating our data preprocessing and input format.

automatically extract complementary information from each input local feature vector compared to the output global feature vector. The information fusion at sensing modalities and diverse body positions are both included.

Suppose the sensing data $\mathbf{X}$ comes from L spatial locations, where N sensors are deployed simultaneously at each position. This models the general scenario of intelligent human activity recognition where multiple integrated sensing units such as Inertial Measurement Units (IMUs) are deployed at a set of diverse body positions to carry out the sensing task collaboratively. Each sensing unit consists of multiple sensors. For example, most IMUs usually contain accelerometer, gyroscope, and magnetometer so that the moving speed and orientation can be simultaneously tracked. For a given sensing modality n and spatial location l, its sensor readings are divided into fixed-length but non-overlapped time segments. One segment aggregated from all spatial positions and sensors constitutes one data sample in our model. Each sensor also has d dimensions (i.e., x, y, z axises). Number of dimensions can be different among sensors, but we assume the same dimension just for notation simplicity. Different sensor readings are upsampled and downsampled into a unified sampling rate. Each data segment is further divided into T non-overlapped time intervals. In data preprocessing step, we perform a Fourier transform to each interval to extract their frequency domain representations, which have been proved to be more informative than pure time domain representations [28]. By utilizing the time-frequency input, both the time domain order information and frequency domain pattern information are well preserved. To help illustrate our data segmentation and pre-processing procedure, we show an example of data input to our model in Figure 2. After preprocessing, the input fed into the model should have a shape of $\mathbf{X} \in \mathbb{R}^{T \times N \times L \times d \times 2f}$, where T represents time intervals, N denotes sensors, L denotes spatial locations, d is the sensor dimension, and $2f$ is spectral samples with f frequency magnitude and phase pairs within each interval. The order of dimensions in $\mathbf{X}$ is determined by the order of information fusion in GlobalFusion.

In general, GlobalFusion is divided into four stacked sub-modules: individual sensor convolution module, spatial fusion module, modality fusion module, and time recurrent module. We will introduce these sub-modules one by one from bottom to top.

Individual Convolution Module. The processed sensing data from each (sensor, position, interval) combination is first separately fed into the individual convolution module. The input data is $\mathbf{X}^{sl}_{[t \cdot \cdot]}$, where s denotes sensor, l denotes position, t denotes time interval. No sensor or body position interaction is considered in this module. Convolution layer [17] has been successful in aggregating information from local area, i.e., adjacent frequencies in our problem. In GlobalFusion, convolution layers are used to gradually extract frequency pattern features within the spectrum of each time interval. We call the output of this module as *(sensor, position, interval)* *features*, which contain the information contained in given sensor at a specific body position and time interval. They are also the input of spatial fusion module.

Spatial Fusion Module. In this module, we fuse the obtained local (sensor, position, interval) features for same sensing modality across different spatial positions, into *(sensor, interval) features.* (sensor, interval) features represent the information contained in given sensor across all body positions at specific interval. Regarding the information fusion order, we profiled the performance improvement of spatial-fusion-first strategy and modality-fusion-first strategy. It turns out that spatial-fusion-first models consistently outperform modality-fusion-first models, so we choose to first perform spatial fusion in GlobalFusion. We will give more explanation on this design choice from the heterogeneity level in Section 4.6. We preserve the three-layer convolution module in DeepSense [40] to extract preliminary global information from different body positions. However, all body positions are homogeneously convolved by now so that some local information is inevitably missed while some noises are included. After that, we stack a global position attention module to allow the incomplete global information to absorb more complementary information from input features at each spatial position. These two sub-modules together constitute our spatial fusion module. The input to this module is $\mathbf{X}_{[t\cdot\cdot]}^{sl(4)}$ from sensor s across every spatial position l, where upper subscript (4) means input to the 4-th layer. The output is fed into next-level modality fusion module for further information aggregation across all sensing modalities.

Modality Fusion Module. In this module, we fuse (sensor, interval) features $\mathbf{X}_{[t\cdot\cdot]}^{s(7)}$ collected from all sensing modalities s into an *interval feature* vector, which is a general feature representation of time interval t. Specifically, we still use a three-layer convolution module to extract preliminary global information from all sensing modalities, the output of which contains incomplete interval features. We use another global attention module to help incomplete interval features absorb more complementary information from input (sensor, interval) features. Deploying global fusion module here is more beneficial than deploying it at spatial fusion level for two reasons: first, the heterogeneity level is higher among different sensing modalities compared to different spatial positions of the same sensing modality; second, modality fusion layer is closer to the output layer, which means the global features at this level is more class-specific and "mature" according to the observations in [15]. After the post-processing by global attention, we feed the flattened vector representation of each time interval into next level time recurrent module.

Time Recurrent Module. After obtaining each interval feature vector, the activity recognition problem turns into a typical sequence classification problem. Here we use a stacked two-layer Gated Recurrent Unit (GRU) to sequentially encode the temporal pattern information. The input to this module is the interval features $\mathbf{X}_{[t\cdot\cdot]}^{(10)}$ across all time intervals t. The order information across time intervals is extracted by the recurrent neural network. We do not use any attention mechanism here for two reasons: first, the step-by-step recurrent structure has already been excellent enough to learn the global temporal patterns; second, it is difficult to decide attention weight of input at each time interval since the information contained in hidden state of each interval is an aggregation of all previous intervals. A more detailed discussion about our design choice here will be give in Section 5. We simply take the average of hidden states at all time intervals as the output, and feed it into the last fully connected layer for ultimate class prediction.

After obtaining a high-level understanding of GlobalFusion architecture and the corresponding deploying position of global attention modules, we are going to discuss more technical details about the global attention design in next subsection.

3.2 Global Attention Module

We explain the specific design details of our global attention module in this subsection. Suppose there is already a feature extraction module in the backbone network that can take input from multiple input sources and output a feature vector containing information aggregated from all sources (i.e., sensing modalities or body positions). For

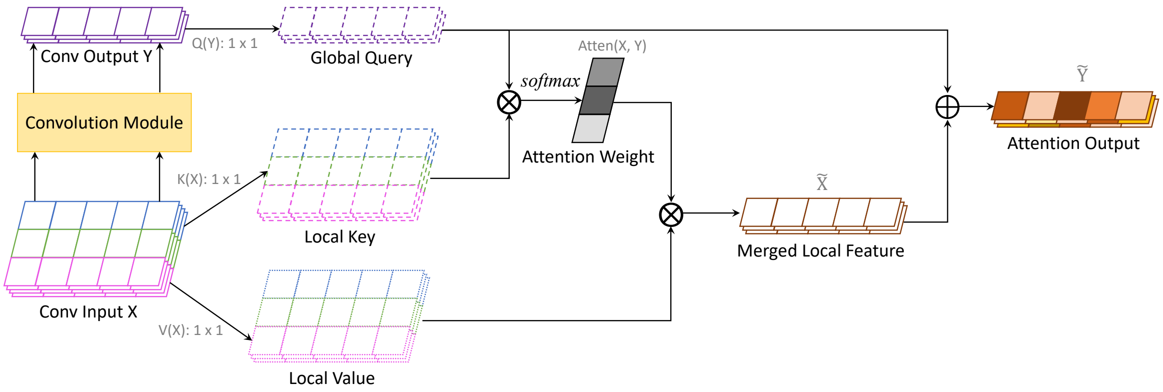


Fig. 3. Global attention module design. (We use different colors to represent input features from each fusing component. $\otimes$ represents dot product operation, while $\oplus$ represents element-wise addition operation.)

example, we use the three-layer convolution module in DeepSense [40] to serve this purpose. However, in these modules, the heterogeneity among sources is not necessarily addressed and their information is homogeneously merged. We ultimately want to see the features from informative sensing modalities / body positions being amplified, while the unrelated noises being suppressed. To narrow down the gap between our learning objective and limitation of backbone feature extraction modules, the global attention mechanism is accordingly designed. In global attention module, the complementary information is extracted from input local features and added to global output features before feeding them into next layer.

It has been widely accepted in machine learning community that for a deep neural network, the features of lower layers are specific to the input but general to the classes, while the features of higher layers are general to the input but specific to the classes. Let's use *local features* and *global features* to refer to input and output of the feature extraction module respectively. The global features possess more class-related information than the local features. Therefore, we use the global features as the *global query* to estimate the importance, i.e., attention weight, of each sensing modality / body position. By doing so, the class-specific information in global features become the standard to evaluate the informativeness of local features. During the back-propagation, the class-related information is propagated from the output layer to the attention module, further back to the global / local feature extraction part. Different from the design in [42], where the mean of all input features is used as the query, our global attention design handles the heterogeneity among input sources better because we do not assume that the information from all sources is similar. The residual connection between global query and highlighted local features can help the global features absorb more complementary local information. A graphical illustration of our proposed global attention module is given in Figure 3.

As mentioned before, in GlbalFusion, the preliminary global feature extraction module refers to a three-layer convolution network (i.e., the yellow box in Figure 3, but it can be configured as other well-behaved computation units). Suppose the input and output of this unit are $\mathbf{X} \in \mathbb{R}^{N \times W \times C_1}$ and $\mathbf{Y} \in \mathbb{R}^{1 \times W \times C_2}$ respectively, where N is the number of components to be fused (we denote it as height of features in the figure), W is the width of features, C_1 and C_2 represent input channels and output channels separately. $\mathbf{X}$ and $\mathbf{Y}$ can be of any shape, we use vectorized local features here just to simplify the notation and make it visually easy understanding. What we need is an attention function $Atten : (\mathbf{X}, \mathbf{Y}) \longrightarrow w \in \mathbb{R}^{1 \times N}$ that takes $\mathbf{X}$ and $\mathbf{Y}$ as input to calculate the attention weight of each input source automatically so that we can fuse the multisensor inputs correspondingly as follows:

$$\widetilde{\mathbf{X}} = \sum_{i=1}^{N} Atten(\mathbf{X_i}, \mathbf{Y}) \cdot \mathbf{X_i}, \quad \text{s.t.} \quad \sum_{i=1}^{N} Atten(\mathbf{X_i}, \mathbf{Y}) = 1. \tag{1}$$

For the design of attention module, we adopt the standard multiplicative attention mechanism [32], where the dot product between the query and keys are used to estimate the attention weight for each sensor. Instead of proposing a new attention computation paradigm, the key innovation lies in the choice of the query. Instead of computing the compatibility score between each local feature vector $\mathbf{X_i}$ and global representation $\mathbf{Y}$ directly, we first let them go through a transformation independently. There are two reasons for this pre-transformation: first, the local features and global features belong to different latent spaces, so a non-linear projection can transform them into the same latent space to make them compatible in semantic level; second, the dimensions of local features and global features can be different in general case so that they are mathematically incompatible, while this transformation can unify their dimensions by needs. Adopting the concepts in [32], we call the transformation for global features and local features as *query function* $Q(\mathbf{Y})$ and *key function* $K(\mathbf{X})$ respectively. What differs from previous works is that, instead of extracting both keys and query from local features, we choose to use global output features to extract query, which is closer to the output layer to provide more global information. Both functions are chosen as 1×1 convolution with *relu* activation followed by a flatten operation. After the transformation, both local keys and global query have been projected into a h dimension latent space, i.e., $K(\mathbf{X}) \in \mathbb{R}^{n \times h}$ and $Q(\mathbf{Y}) \in \mathbb{R}^{1 \times h}$. h is a hyper-parameter that we can adjust during training. According to our experience, model performance is not sensitive to the value of h, so this value is empirically fixed at 64, in all 4 datasets during evaluation.. Next, we use projected keys and query to calculate their compatibility score, i.t., attention weight for each fusing component:

$$Atten(\mathbf{X}_i, \mathbf{Y}) = \text{softmax}\left(\frac{K(\mathbf{X_i})^T Q(\mathbf{Y})}{\sqrt{h}}\right). \tag{2}$$

Here we choose the dot product similarity (i.e., multiplicative attention), instead of a feed-forward neural network (i.e., additive attention), as the compatibility function to compute the attention weight, because it is more intuitive and computational efficient. We scale the product by a factor of $\frac{1}{\sqrt{h}}$ to prevent it from becoming too large to get into regions with extremely small gradients after applying outlier *softmax* [32]. The outlier *softmax* is used to emphasize informative features, and normalize the attention weight distribution. By far, each input source corresponds to a normalized attention weight. Next, we let the local features go through another transformation, called *value function* $V(\mathbf{X})$, to match their dimensions with the global query. The local values are multiplied with their attention weights and sumed up. The weighted sum is called *merged local features.* After that, we combine the merged local features with the global query $Q(\mathbf{Y})$ through a residual connection later. The *value function* $V(\mathbf{X}) \in \mathbb{R}^{n \times h}$ consists of a 1×1 convolution operation with *relu* activation and a flatten operation. The merged local features are computed by:

$$\widetilde{\mathbf{X}} = \sum_{i=1}^{N} Atten(\mathbf{X_i}, \mathbf{Y}) \cdot V(\mathbf{X_i}), \tag{3}$$

where $\cdot$ means that the scalar attention weight is propagated to each element of local feature vector. This merged local feature vector provides a good complement to global feature (not replacement), so we combine them up through a residual connection:

$$\widetilde{\mathbf{Y}} = Q(\mathbf{Y}) + \widetilde{\mathbf{X}}, \tag{4}$$

which becomes the output of our global attention module. To fit it back into original network dimension, we can let the output go through another 1×1 convolution for dimension extension or suppression if necessary.

Table 1. Statistical Summary of Selected Datasets.

Dataset	Activities	Subjects	Sensors	Positions	Segment	Intervals	Spectral Samples
PAMAP2	18	9	3	3	2 sec	10	20
RealWorld-HAR	8	15	4	5	2 sec	10	10
DSADS	19	8	3	5	5 sec	5	25
DS	2	10	1	3	2 sec	8	16

From a different perspective of view, our global attention can be regarded as an innovative residual connection design [13] for sensing data processing. Instead of directly connecting the input and output of a processing module, we first utilize the calculated output to search and highlight local features in input before merging them. The heterogeneity in local features is well addressed while at the same time the residual property is preserved. So far, we have introduced all technical details related to GlobalFusion framework and global attention module design. Next, we validate the effectiveness of GlobalFusion especially the contribution of global attention modules through experiments on four realworld human activity recognition (HAR) datasets.

4 EXPERIMENTS

In this section, we compare GlobalFusion to other state-or-the-art deep learning frameworks using four publicly available human activity recognition (HAR) datasets. We first introduce the experimental setup, datasets used, data preprocessing steps, and baseline algorithms we are comparing with. We then show the evaluation results for each dataset, make qualitative observations, and discuss insights obtained from attention weight distributions. Finally, we present time and energy efficiency comparisons on commodity IoT devices.

4.1 Experimental Setup

All the models evaluated in this paper are trained with Tensorflow 1.14 [3] on a workstation equipped with an Intel i9-9960X processor, 64GB memory, and four NVIDIA RTX 2080 Ti GPU. For training the model, we adopt a standard cross entropy loss for classification, along with L2 normalization. The normalization factor is set as 5e-4. The model is optimized by the ADAM algorithm [16] with a learning rate of 1e-4, while $\beta_1 = 0.5$ and $\beta_2 = 0.9$. We add a batch normalization layer and a dropout layer after each convolutional layer to stabilize the training process and prevent overfitting. Training batch size is set as 64.

4.2 Datasets

We evaluate inference accuracy on human activity recognition tasks that use multiple sensors as input. All models are evaluated under a leave-one-user-out scenario with k-fold cross validation. Specifically, one subject is chosen as the test user each time, while the activity traces of all remaining subjects are used for training. The test user is then rotated until all users have been exchusted. A statistical summary of each dataset is listed in Table 1.

PAMAP2 Physical Activity Monitoring Data Set (PAMAP2). [27] This dataset contains data of 18 different physical activities (e.g., walking, cycling, playing soccer, etc) performed by 9 subjects using 3 inertial measurement units (IMUs) that are put at the chest, wrist (of dominant arm), and dominant side's ankle respectively. Each IMU records readings from a 3-axis accelerometer, gyroscope and magnetometer. The sampling rates of all sensors are 100 Hz. We divide data into segment of $2s$ where each segment is further divided into $T = 10$ fixed-length and non-overlapped time intervals. Each interval contains 20 spectral samples. Sensor readings in each interval are sent through a Fourier transform as pre-processing. "subject109" is excluded for testing because s/he has too few contributed data samples.

RealWorld Human Activity Recognition (RealWorld-HAR). [31] This dataset covers 8 activities (climbing stairs down and up, jumping, lying, standing, sitting, running/jogging, and walking) from 15 subjects on 7 body positions. Sensing modalities include acceleration, GPS, gyroscope, light, magnetic field, and sound level. We choose 5 body positions out of 7 (i.e., head, chest, forearm, waist, and shin), and use four sensor types only (i.e., accelerometer, gyroscope, magnetometer, and light). Upper arm and thigh data are not used because we want to make chosen body positions more diverse. GPS and sound level readings are not used here because their sampling rates are too low. GPS is sampled at 0.08 Hz and sound level is sampled at 2 Hz, while all remaining sensors have a ~50 Hz sampling rate. The sampling rate of all selected sensors is interpolated to 50 Hz by up-sampling and down-sampling. We still use segment of $2s$ consisting of 10 non-overlapped intervals, where each interval contains 10 spectral samples. Every subject is selected once as test user in cross validation.

Daily and Sports Activities Data Set (DSADS). [4] In this dataset, each of 19 activities (e.g., sitting, standing, ascending and descending stairs, exercise on stepper, playing basketball, etc) is performed by 8 subjects (4 female and 4 male). The contained sensors are still accelerometer, magnetometer, and gyroscope on 5 body positions (torso, right arm, left arm, right leg, and left leg). The data sampling rate is 25 Hz, so we choose the segment length as $5s$ with 5 intervals within each segment. Therefore, we have 25 spectral samples in each time interval. Every subject is selected once as test user in cross validation.

Daphnet Gait (DG). [5] The daphnet freezing of gait dataset is devised to benchmark automatic methods to recognize gait freeze from wearable acceleration sensors placed on legs and hip. It is a binary classification problem (i.e., freeze or not freeze). Only accelerometer data at three body positions (i.e., ankle, upper leg, trunk) is provided. Here we do a minor adjustment on our GlobalFusion model, where the modality fusion module is removed. It is collected from 10 Parkinson's disease patients. The sampling rate is 64 Hz. We choose a $2s$ segment and divide it into 8 time intervals, so that each interval has 16 spectral samples. Similarly, every subject is selected once as test user in cross validation.

The purpose of using the first three datasets is two-fold. First, we want to compare our model to the baselines on multiple multisensor datasets to reach more broadly substantiated conclusions. Second, we want to find common observations in attention weight distributions to check if attention allocation meets intuition. Through the DG dataset, we want to see whether proper attention mechanisms can overcome the unbalanced class distribution problem that is heavily represented in that dataset.

4.3 Baselines

Before presenting the results, we briefly review state-of-the-art deep learning frameworks for heterogeneous information fusion that are selected as baselines in our experiments. We pay special attention to attention-mechanism-based designs.

GlobalFusion-Single: To show the effectiveness of our global attention design and give a straightforward understanding of contribution by each attention module, we choose a variant of GlobalFusion here. In this model, only global modality attention is used before time recurrent layer. Comparing performance of this model with DeepSense could help understand the effectiveness of global modality attention module, while comparison with GlobalFusion could help show the contribution of global position attention module.

DeepSense [40]: This is the back-bone network for our GlobalFusion framework, which is one of the commonly used sensing data processing frameworks for IoT applications. It's generally based on the combination of convolutional modules (i.e., within a time interval) and a recurrent module (i.e., across time intervals) for temporal pattern extraction. To emphasize the effectiveness of our global attention module, we use the same backbone architecture as GlobalFusion here excluding global attention modules.

SADeepSense [42]: This is a recent self-attention based DeepSense framework. They use a self-attention (SA) module for heterogeneous sensor information fusion and time-series information fusion respectively. They use the mean of all input features as the query to estimate correlation across sensors / time intervals. They also adopt the multi-head design to learn the correlations from different latent spaces. In our implementation, three SA modules are deployed at spatial fusion, sensor fusion, and time fusion respectively based on their design in the paper. This design utilizes the local features as the query. Through comparing GlobalFusion and SADeepSense, we can see the advantages of using global features as the query in attention module.

attnLSTM [47]: They use two attentional modules to improve the classification performance of recurrent networks. We make minor enhancement to their framework: at the bottom of network, instead of using raw sensing input, we use a same three-layer convolutional module as GlobalFusion to extract the local features of each individual sensor within each time interval. The reason is that we want to compare the impact of different fusion mechanisms instead of lower-level feature extraction parts. Next lies the sensor attention part, where a shallow fully-connected module works as attention unit across different sensing modalities and sensor positions. Finally, after one LSTM layer, they use the output of last time interval to estimate importance of hidden state at each previous time interval, and take the weighted average of them for final prediction. In addition, they add a continuous constraint to both attention modules to regulate the attention weight to change smoothly over modalities and time intervals.

BANet [34]: This model exploits a different order of information fusion, where they first fuse information at different time intervals for same sensor, and then fuse information across sensing modalities. We still add an individual convolution module at the bottom of their architecture to extract low-level features. Temporal attention is computed by a 1 × 1 convolution followed by a softmax layer, which belongs to brute-force additive attention and no compatibility design is considered. Sensor attention is computed in a similar way: they use one fully-connected layer followed by a softmax layer.

4.4 Quantitative Classification Results

In this subsection, we present the evaluation results of GlobalFusion and aforementioned baseline frameworks on each selected dataset. We will use abbreviations for models in the tables and figures. GF represents the GlobalFusion with double attention modules, where both global position attention and global modality attention are deployed. GF-SGL or GlobalFusion-SGL both represent the GlobalFusion with global modality attention only. DS is short for DeepSense, while SA-DS is used for SADeepSense. In addition, DS-Acc, DS-Gyro, DS-Mag, and DS-Lig represent DeepSense-Accelerometer, DeepSense-Gyroscope, DeepSense-Magnetometer, DeepSense-Light respectively. Only one sensing modality is used in these individual models.

We mainly use accuracy and macro F1 score as the classification performance metrics, while micro F1 score is not used. In multi-class classification problems, micro precision = micro recall = micro F1 score = accuracy. In this case, true positives (TP) are defined as the samples that were predicted to have the correct label. Every time there is a false positive (FP), there will always also be a false negative (FN) and vice versa, because always one class is predicted. Therefore, the micro F1 score is redundant with accuracy. Instead, the macro F1 score can better reflect the model classification performance across all classes. It prefers more balanced classification results. This is a good complement to accuracy for capturing performance balance across classes. In our tables, "Acc." is short for accuracy, and "Mac. F1" represents macro F1 score.

PAMAP2 Results: We start with the PAMAP2 dataset. In Table 2, we give the accuracy and macro F1 score for all models compared, including both individual testing cases and the overall average values. Since the evaluation is performed under leave-one-user-out scenario, the model performance is supposed to be lower than random partition, considering the heterogeneity and transfer difficulty between training users and testing user. From Table 2, we can see that DeepSense obtains similar accuracy and macro F1 score as other state-of-the art attentional frameworks, while our global attention module further improves its performance by a clear margin, especially the global modality attention module (i.e., GF-SGL v.s. DS). Compared to modality-attention only GlobalFusion-Single

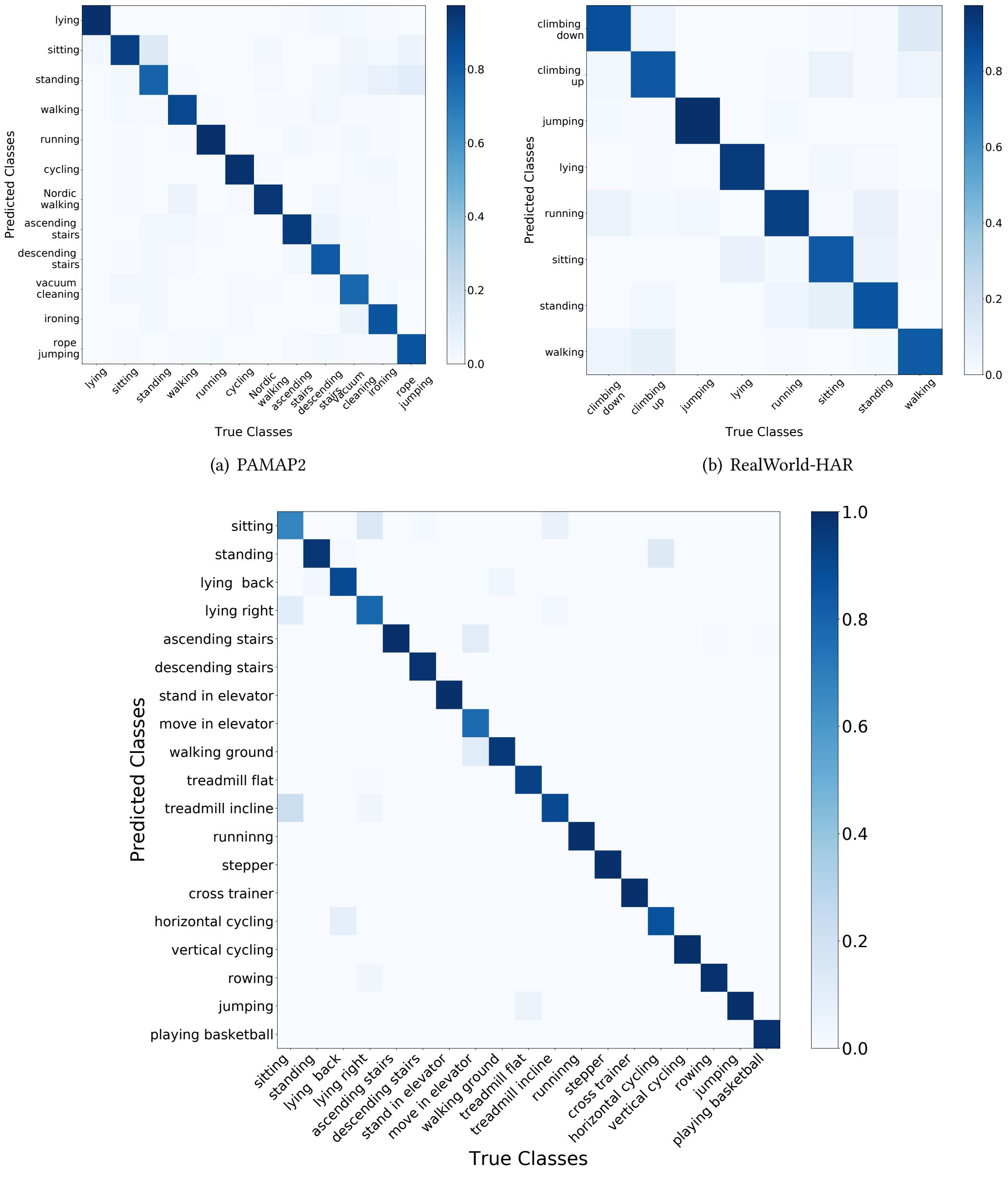
Table 2. PAMAP2 classification result.

Test User	GlobalFusion		GlobalFusion-SGL		DeepSense		SADeepSense		attnLSTM		BANet	
	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1
Subject1	78.99%	75.61%	77.69%	75.76%	73.53%	69.57%	72.20%	67.43%	72.04%	67.21%	72.61%	69.32%
Subject2	92.22%	92.45%	91.75%	91.52%	74.77%	71.10%	83.28%	82.13%	74.84%	71.34%	69.77%	71.42%
Subject3	95.37%	75.67%	94.91%	68.70%	94.95%	68.52%	93.87%	62.32%	93.15%	67.41%	89.54%	65.00%
Subject4	95.26%	86.70%	95.74%	86.92%	94.02%	93.89%	88.24%	79.86%	88.79%	81.05%	91.91%	76.20%
Subject5	92.93%	92.36%	93.23%	92.75%	92.93%	92.67%	90.25%	88.94%	88.62%	87.45%	90.92%	89.79%
Subject6	92.88%	84.77%	92.27%	88.36%	90.79%	82.67%	89.72%	87.06%	91.78%	83.19%	85.44%	77.89%
Subject7	96.09%	94.65%	96.35%	94.72%	95.75%	94.28%	94.70%	86.26%	94.44%	85.30%	93.58%	84.71%
Subject8	83.09%	79.78%	80.05%	77.15%	59.69%	54.98%	63.13%	59.14%	68.86%	62.42%	55.23%	49.67%
Overall	**90.86%**	**85.25%**	**90.25%**	**84.48%**	**84.55%**	**78.46%**	**84.42%**	**76.64%**	**83.69%**	**75.67%**	**81.13%**	**73.00%**

Table 3. RealWorld-HAR classification result.

Test User	GlobalFusion		GlobalFusion-SGL		DeepSense		SADeepSense		attnLSTM		BANet	
	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1
Subject1	78.22%	79.36%	75.36%	75.81%	65.95%	65.66%	67.05%	66.31%	65.49%	64.58%	74.95%	74.54%
Subject2	94.39%	94.25%	95.29%	95.57%	80.42%	78.47%	79.20%	78.63%	80.91%	79.58%	93.07%	93.18%
Subject3	78.53%	71.37%	75.05%	62.06%	54.51%	51.91%	54.73%	51.89%	48.53%	42.10%	49.73%	50.41%
Subject4	92.61%	93.38%	91.17%	92.64%	86.62%	88.47%	82.23%	84.08%	89.26%	89.44%	84.18%	85.99%
Subject5	81.63%	81.11%	82.63%	82.22%	72.57%	72.29%	79.52%	78.83%	76.93%	74.89%	68.91%	69.30%
Subject6	84.14%	85.14%	84.47%	85.72%	67.52%	66.90%	73.77%	73.05%	69.51%	67.84%	73.39%	73.54%
Subject7	78.34%	70.88%	80.59%	78.05%	69.56%	70.10%	67.56%	62.77%	78.02%	79.30%	77.16%	78.79%
Subject8	63.54%	66.51%	62.79%	63.05%	39.39%	35.17%	50.08%	48.71%	24.30%	18.01%	47.94%	48.41%
Subject9	91.21%	92.07%	90.76%	91.44%	80.58%	74.77%	77.99%	74.24%	80.31%	77.89%	76.29%	75.64%
Subject10	96.18%	95.84%	96.33%	96.13%	95.07%	94.37%	85.35%	85.71%	80.27%	78.00%	82.76%	83.03%
Subject11	89.13%	89.82%	86.87%	87.46%	85.34%	85.92%	78.49%	76.16%	76.19%	74.54%	79.14%	77.95%
Subject12	95.12%	95.36%	95.88%	96.01%	83.85%	82.04%	85.70%	85.29%	86.46%	86.45%	81.72%	79.93%
Subject13	88.45%	89.48%	88.73%	89.68%	71.88%	73.51%	75.60%	75.19%	78.77%	80.13%	73.21%	73.83%
Subject14	90.51%	72.83%	83.62%	84.91%	71.82%	69.77%	75.06%	73.68%	50.45%	48.07%	79.85%	78.77%
Subject15	89.58%	89.61%	90.40%	90.42%	81.30%	79.92%	78.13%	77.34%	77.11%	76.99%	73.07%	73.82%
Overall	**86.11%**	**84.47%**	**85.33%**	**84.74%**	**73.76%**	**72.62%**	**74.03%**	**72.79%**	**70.83%**	**69.19%**	**74.36%**	**74.48%**

framework, GlobalFusion further improves the accuracy and macro F1 score by a small margin. We can regard it as a trade-off between efficacy and efficiency when considering model deployment on IoT devices. If we want to obtain better model efficiency, GlobalFusion-Single can be used; otherwise, GlobalFusion is a better choice for higher model efficacy. We also notice that SADeepSense does not show clear improvement compared to vanilla DeepSense here. The reason is that mean value of all sensors are used as the query to estimate the attention weight of each sensing modality in SADeepSense, which leads the model to attend to more homogeneous input, so that the model behaves similar as non-attentional DeepSense. Among the users, those with lower accuracies at back-bone DeepSense model (e.g., Subject2 and Subject8) typically have a higher chance to see a larger improvement by attention learning. According to the confusion matrix of GlobalFusion in Figure 4 (a), we can see an ambiguity between sitting and standing. This ambiguity among static gestures also exists in RealWorld-HAR and DSADS results. In later analysis, we will show that this results from the dominant effect of accelerometer features that share the same pattern under all static gestures.



(a) PAMAP2

(b) RealWorld-HAR

(c) DSADS

Fig. 4. Normalized confusion matrix of GlobalFusion on PAMAP2, RealWorld-HAR and DSADS. (Every figure is the average overall all subjects.)

Table 4. DSADS classification result.

Test User	GlobalFusion		GlobalFusion-SGL		DeepSense		SADeepSense		attnLSTM		BANet	
	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1
Subject1	91.67%	90.63%	91.67%	90.58%	89.06%	87.51%	91.82%	90.73%	90.07%	88.55%	83.46%	81.43%
Subject2	96.97%	96.87%	93.94%	93.36%	90.44%	88.23%	83.64%	78.72%	87.13%	83.96%	92.56%	92.10%
Subject3	93.09%	92.30%	91.38%	90.57%	90.81%	88.82%	89.52%	88.02%	88.79%	86.23%	88.33%	86.65%
Subject4	94.60%	93.99%	90.72%	89.26%	77.67%	74.36%	85.94%	83.70%	95.13%	94.75%	90.81%	89.99%
Subject5	96.78%	96.77%	96.97%	97.02%	88.14%	84.88%	88.14%	84.86%	82.81%	81.22%	84.93%	81.93%
Subject6	99.24%	99.23%	96.69%	96.57%	92.37%	91.77%	89.71%	87.38%	93.11%	91.59%	97.89%	97.90%
Subject7	94.51%	94.02%	96.40%	96.25%	88.60%	86.55%	91.45%	89.87%	94.49%	94.16%	86.76%	83.14%
Subject8	87.41%	85.54%	87.22%	85.26%	83.82%	80.74%	84.01%	80.72%	76.75%	72.48%	86.12%	84.84%
Overall	**94.28%**	**93.67%**	**93.12%**	**92.36%**	**87.61%**	**85.36%**	**88.03%**	**85.50%**	**88.53%**	**86.62%**	**88.86%**	**87.25%**

RealWorld-HAR Results: The evaluation results on RealWorld-HAR dataset is given in Table 3. Compared to other datasets, this dataset covers a wider range of subject diversity, reflected in their age, height, weight, gender, and dominant arm. Therefore, we can see a large difference across different models. attnLSTM is the worst model here. It only achieves an accuracy of 70.83%. The failure of attnLSTM indicates us that incorrect attention design can lead to a severe performance degradation. DeepSense also does not work well on this dataset. The reason is that the reliability of different sensing modalities differs in a large degree in this dataset (i.e., their single-modality performance are quite diverse as we will show in Figure 8 later). Thus, equally weighting and merging different sensing modalities leads to a severe performance degradation. Once again, the overall performance of SADeepSense and DeepSense are very close to each other, mainly due to the utilized mean query in sensor attention of SADeepSense. Considering the diverse reliability of sensing modalities, using a mean query for attention weight estimation is obviously not an optimal solution, which is even worse than the brute-force attention design in BANet. The best position still belongs to our GlobalFusion framework. The GlobalFusion-Single model improves the back-bone DeepSense by 11.57% in accuracy, and the GlobalFusion model further improves GlobalFusion-Single by 0.78%. From the confusion matrix of GlobalFusion in Figure 4 (b), we can see that most classes are classified accurately, but there is still an ambiguity between sitting and standing classes. This observation is similar as our finding in PAMAP2, which also results from the dominant effect of accelerometer features. One more point we want to mention is that GlobalFusion-SGL has a slightly better macro F1 score than GlobalFusion, which means it's more stable across classes in classification.

DSADS Results: According to the classification results on DSADS in Table 4, attentional models generally work better than non-attentional DeepSense framework. It means all attention designs have a positive influence in model performance, so we are expecting to see an unbalanced attention weight distribution here (as validated by Figure 10), in contrast to the homogeneous information fusion in DeepSense. SADeepSense outperforms DeepSense only by 0.42%, due to its defective mean query design, while our GlobalFusion-Single and GlobalFusion consistently improve the model performance, achieving an accuracy of 93.12% and 94.28% respectively. The normalized confusion matrix of GlobalFusion is given in Figure 4 (c). In this dataset, we can find the misclassification between sitting and lying right classes. We observe that most attention weights are still allocated to the accelerometer, which is not good at distinguishing between static gestures.

DG Results: At last, we analyze the classification results on DG. This is a binary classification problem where training data is distributed rather unbalanced between two classes. Most (over 95%) data samples belong to the 'Not freezing' class. We do not add any data augmentation techniques in data preprocessing, such as down-sampling or up-sampling of unbalanced classes. Since we only have accelerometer data from several data positions, only global position attention module is available here. As indicated in Table 5, the advantage of

Table 5. DG classification result.

Test User	GlobalFusion-SGL		DeepSense		SADeepSense		attnLSTM		BANet	
	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1
Subject1	92.71%	64.20%	89.40%	60.88%	86.16%	57.45%	90.63%	60.00%	92.41%	61.83%
Subject2	90.18%	71.10%	89.84%	65.07%	90.16%	66.74%	88.59%	60.04%	89.22%	56.52%
Subject3	86.15%	65.45%	86.46%	64.95%	86.56%	66.49%	84.48%	62.78%	85.94%	57.56%
Subject4	99.58%	49.90%	98.24%	49.56%	98.05%	49.51%	97.36%	49.33%	99.22%	49.80%
Subject5	83.65%	66.68%	80.42%	52.86%	81.98%	61.14%	81.56%	60.24%	81.56%	59.16%
Subject6	93.65%	56.57%	94.06%	51.75%	94.06%	55.92%	93.44%	48.30%	93.75%	48.39%
Subject7	96.22%	73.58%	95.18%	74.07%	93.36%	70.81%	92.32%	65.54%	94.40%	68.25%
Subject8	79.17%	72.96%	69.69%	66.33%	69.69%	64.23%	63.12%	61.16%	72.81%	61.58%
Subject9	88.15%	67.72%	85.94%	53.47%	86.42%	58.85%	86.78%	61.97%	85.94%	52.19%
Subject10	100%	100%	100%	100%	99.82%	49.95%	99.91%	49.98%	100%	100%
Overall	**90.94%**	**68.82%**	**88.92%**	**63.89%**	**88.62%**	**60.11%**	**87.82%**	**57.93%**	**89.52%**	**61.53%**

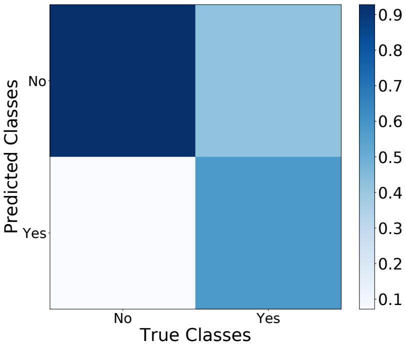


Fig. 5. Normalized confusion matrix of GlobalAttention-Single on DG. (The figure is the average result overall all subjects.)

GLobalFusion-SGL is not obvious as on other three datasets. Only 1.42% improvement in accuracy is observed compared to the best baseline model, BANet. The heterogeneity among different spatial positions is not as large as diversity among different sensing modalities. Furthermore, none of the models can overcome the unbalanced class problem, since all of them show very low F1 score. Same conclusion can also be drew from the normalized confusion matrix of GlobalFusion in Figure 5. A large portion of positive samples are still misclassified as negative, which is especially unacceptable in medical applications considering the safety of patients. In summary, the usage of global attention module is not as beneficial as in other datasets, and it is not powerful enough to overcome the unbalanced class distribution problem. Appropriate data augmentation techniques are still needed to address this issue, while the positive aspect is that attention design is independent of data augmentation so they can be applied together.

4.5 Information Fusion Capability

In order to further explore the impact of global attention module in general classification performance, in this part, we look into details about information fusion capabilities between different attention mechanisms. All

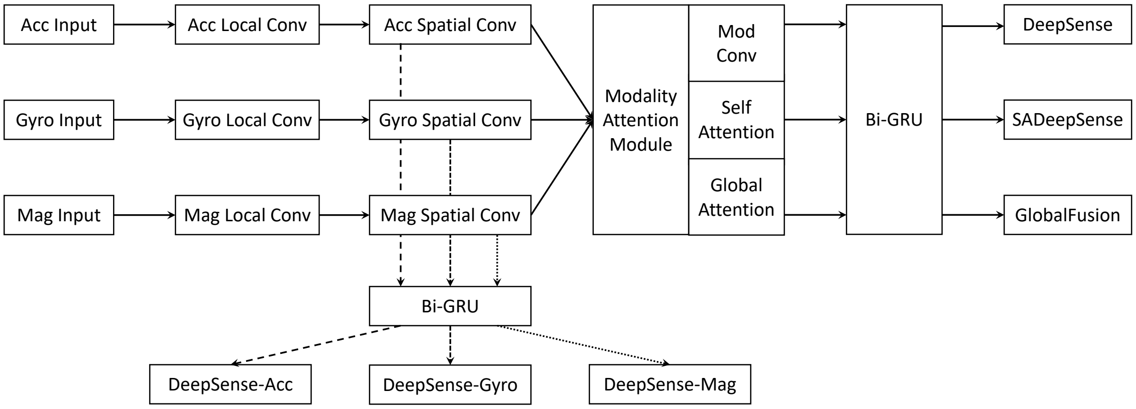
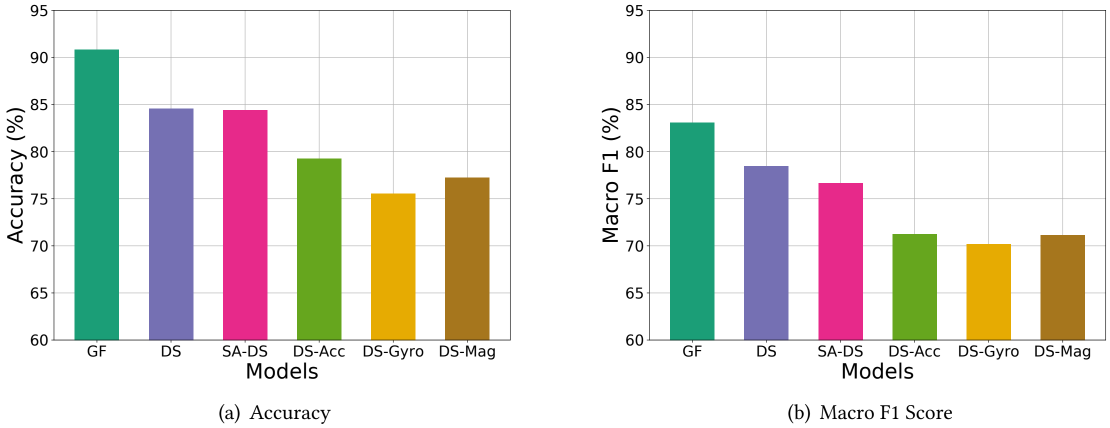


Fig. 6. Information fusion capability evaluation logic example. (We use different dash lines for each individual sensor model.)



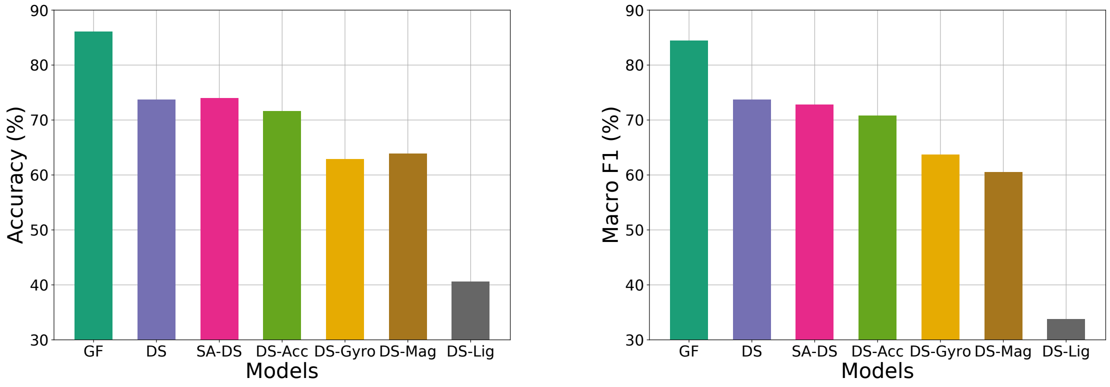
(a) Accuracy

(b) Macro F1 Score

Fig. 7. Information fusion comparison on PAMAP2.

models are based on the back-bone design of DeepSense. To make the comparison fair and straightforward, we use the same lower level structures (i.e., below modality convolution module) at each model. An example to illustrate the evaluation logic of this subsection is shown in Figure 6. We first show the prediction performance of each single-sensor model, and then compare the information fusion capability between different modality attention modules, including convolution operation (i.e., DeepSense), self-attention (i.e., SADeepSense), and our global attention (i.e., GlobalFusion-Single). No position attention module is applied here. Evaluations are performed on three datasets: PAMAP2, RealWorld-HAR, and DSADS. All figures are results based on the k-fold cross-validation on all subjects in each dataset.

PAMAP2: The comparison result of PAMAP2 is presented in Figure 7. Accuracy and macro F1 score of each single-sensor model and composite model are shown. We can see that when three sensing modalities are used



(a) Accuracy

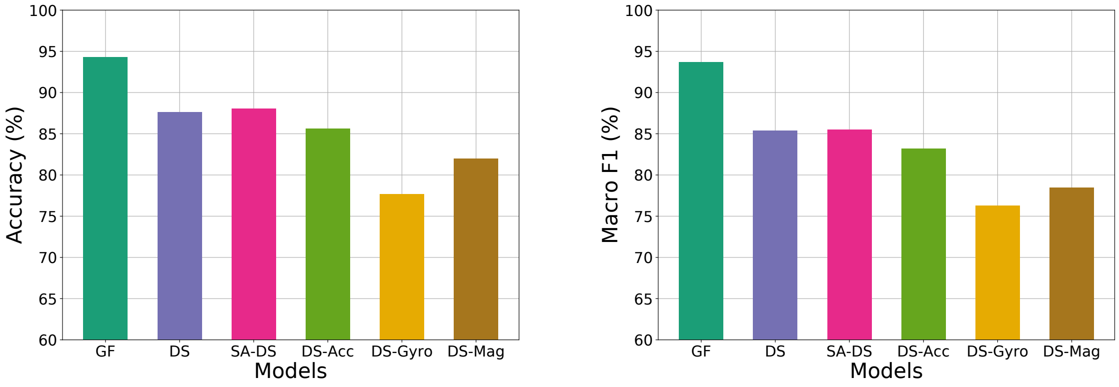
(b) Macro F1 Score

Fig. 8. Information fusion comparison on RealWorld-HAR.

individually, accelerometer and magnetometer achieve relatively better performance than gyroscope. All of three modality fusion modules can take use of the information from three sensing modalities, because all of them can beat the best single sensor model (i.e., DeepSense-Acc). Among the three referred attention modules, GlobalFusion has the best information fusion capability because it possesses the highest accuracy and macro F1, while the performance of SADeepSense and DeepSense are inferior to GlobalFusion but similar to each other. We also find that accelerometer data shows highest reliability in classification when used individually. Similar observations will also be given in other two datasets, and we will try to find the connection between single sensor performance and its corresponding attention weight returned by GlobalFusion in next subsection.

RealWorld-HAR: The comparison result of RealWorld-HAR is shown in Figure 8. Four sensing modalities achieve quite diverse performance, where the accuracy of DeepSense-Acc is clearly higher than DeepSense-Gyro and DeepSense-Mag, while the performance of DeepSense-Light is much worse than all other sensors. SADeepSense achieves similar performance as DeepSense, both of which are slightly better than DeepSense-Acc. It means that these two models is still able to extract complementary information from other sensing modality features excluding accelerometer feature. Our GlobalFusion beats these two models by a large margin, improving the performance of DeepSense by about 11%. Meanwhile, DeepSense-Acc shows an obviously better performance than other three single-sensor models. This observations is same as what we have saw in PAMAP2. We also dig into reasons behind the failure of light sensor model. After checking the raw data, we found that although the light sensor maintains a 50 Hz sampling rate, there is not much vibration in its readings. In most cases, there is no value change for light sensor within each time interval. Therefore, the information contained in light data is much lower than other sensor types.

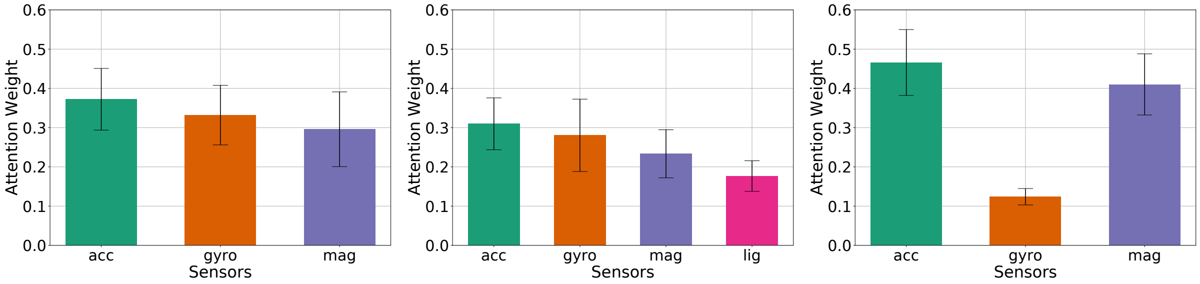
DSADS: At last, we compare the information fusion result on DSADS. The corresponding accuracy and F1 results are given in Figure 9. As we have observed in all previous datasets, regarding the single sensor model performance, DeepSense-Acc > DeepSense-Mag > DeepSense-Gyro. This result will be integrated into the attention weight analysis in next part. For the sensor fusion models, conclusion are similar: DeepSense and SADeepSense share similar accuracy and macro-F1 score, because they both prefer more homogeneous input from different sensing modalities. Since GlobalFusion does not follow this assumption, its performance is much better than the other two. In summary, we have shown the superior information fusion capability of GlobalFusion



(a) Accuracy

(b) Macro F1 Score

Fig. 9. Information fusion comparison on DSADS.



(a) PAMAP2

(b) RealWorld-HAR

(c) DSADS

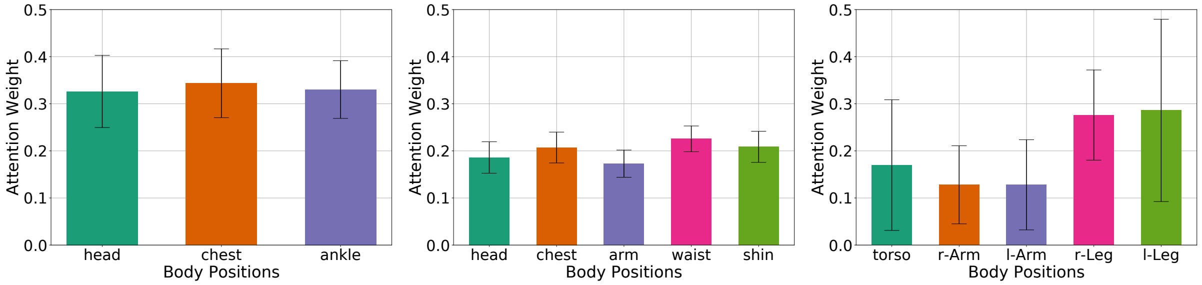
Fig. 10. Modality attention weight distribution on different datasets.

compared to DeepSense and SADeepSense from the pure data driven aspect. To further validate the reasonability of our design, we directly look at and analyze the returned attention weight distribution, and try to interpret it from the physical aspect.

4.6 Qualitative Analysis on Attention Weight

In this subsection, we give a qualitative analysis about how our global attention design deals with the heterogeneity in information fusion, as well as some observations we get through analyzing the attention weight distribution. We will also answer the question about how we decide the information fusion order from the perspective of heterogeneity level.

First, we only use modality attention module in GlobalFusion, and analyze the statistical distribution of attention weight among each sensing modality. The results are given in Figure 10. The accelerometer always has the best classification result, and it also receives most attention by GlobalFusion on all datasets, which is what we have



(a) PAMAP2

(b) RealWorld-HAR

(c) DSADS

Fig. 11. Position attention weight distribution on different datasets.

anticipated. However, on all three datasets, we observe a better performance on DeepSense-Mag than DeepSense-Gyro, but magnetometer has a higher attention weight than gyroscope in both PAMAP2 and RealWorld-HAR. We try to understand this observation from the physical principles: Accelerometer measures the absolute moving speed patterns, while magnetometer measures absolute orientation pattern in NESW plane. Gyroscope measures the changes in the orientation. If we take the derivative of consecutive magnetometer measures, we get the angular velocity in NESW plane; similarly, when we take the integral of consecutive gyroscope measures, we get the angular changes. Thus, gyroscope and magnetometer readings unavoidably contain overlapped information. At meantime, magnetometer lacks the vertical plane orientation information, while gyroscope lacks the absolute facing direction information of sensors to transferring the angular velocity from absolute earth coordinates to human body coordinates [30]. Therefore, the two sensors are also complementary to each other. The information contained by accelerometer readings is more independent of gyroscope and magnetometer, and is more critical in deciding human gesture patterns (i.e. activities), so it is supposed to receive highest attention weight. The information of gyroscope and magnetometer are both overlapped and complementary, so the relation between their attention weights is not fixed and can be affected by noise level of sensors. For the light sensor used in RealWorld-HAR, as we have mentioned before, although it has a completely different sensing principle compared to other inertial sensors, the lack of vibrations in its readings (at least in this dataset) restricts it to contain much information in frequency domain, so that it only gets a very low attention.

Next, we only include position attention in GlobalFusion, and present the results in Figure 11. Our observations are in two folds. First, the divergence of attention distribution over different body positions is smaller than different sensing modalities, which means that the information contained across sensing modalities are more diverse compared to spatial positions. To further validate this hypothesis, we show the JS divergence between both attention distributions and uniform distribution on each dataset in Figure 12. JS divergence is defined based on KL divergence as follows:

$$
\begin{aligned}
JS(P\,||\,Q) &= \frac{1}{2}KL(P\,||\,M) + \frac{1}{2}KL(Q\,||\,M),\\
\text{where:}\qquad M &= \frac{1}{2}(P+Q),\\
KL(P\,||\,M) &= \sum_x P(x)\log\left(\frac{P(x)}{M(x)}\right).
\end{aligned}
$$

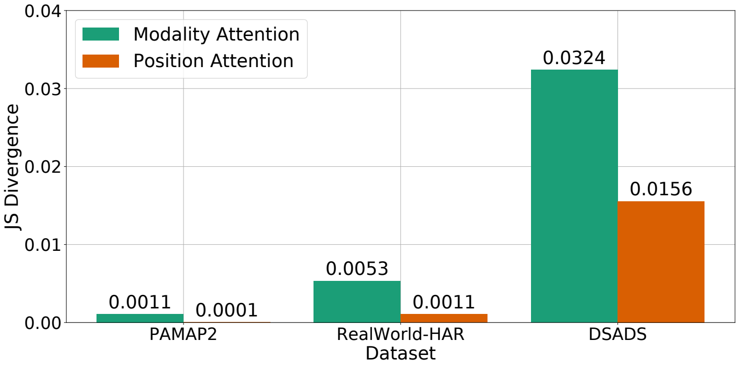


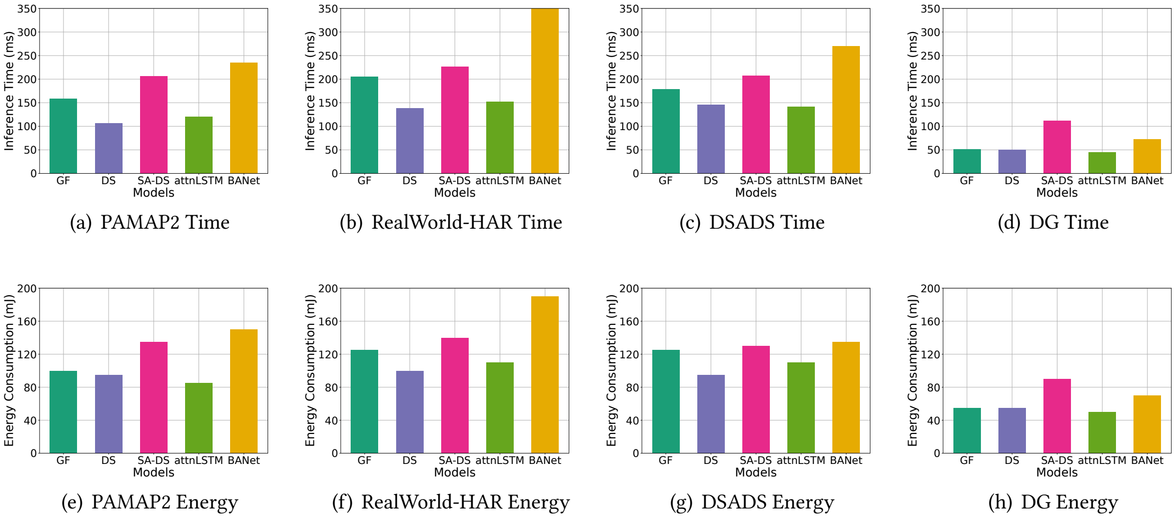
Fig. 12. Attention weight divergence on each dataset

P and Q are two normalized distributions. We can see from Figure 12 that the JS divergence in modality attention distribution is apparently larger than position attention distribution on each dataset. This is exactly the reason why we choose to merge position information first before sensing modalities. We want to push the fusion of more heterogeneous information to higher layers of network so that we do not break the information integrity at lower level feature extraction. The second observation is, in both RealWorld-HAR and DSADS, arm sensors both get lower attention weight than other positions. Our interpretation is that although the moving patterns of arm contains a lot of information because its movement range is larger than other body positions, this information is not necessarily closely related to the target classes (i.e., activities). Instead, the information contained in the arm movement can actually make confusion to the model. Therefore, arm features are not assigned large attention weight by the global attention module. By all the above observations and analysis, we prove that our global attention design is logically reasonable and agrees well with existing human knowledge.

4.7 Time and Energy Efficiency

In this part, we evaluate the time and energy efficiencies of GlobalFusion when deployed on IoT devices. The experiments are conducted on two types of IoT devices, LG Nexus 5 and Raspberry Pi Model B. Nexus 5 is powered by a 2.26 GHz quad-core Snapdragon 800 processor with 2 GB of RAM, 32 GB of internal storage, and a 2300 mAh battery. The installed operating system is Android 7.1.1. Raspberry Pi 3 Model B is powered by a quad core 1.2 GHz Broadcom BCM2837 64bit CPU with 1 GB RAM and 16 GB storage. The preinstalled Raspbian Jessie operating system is used for Raspberry Pi. For all the models, we only use on-chip CPU for inference. Every model is preloaded to IoT device before experiment, and any unnecessary application and service that may interfere model computation are closed in advance. During the runtime on Nexus 5, we use the Tensorflow Lite interpreter [2] as the inference engine, which is specially designed for running deep learning models on mobile, embedded, and IoT devices. Since TensorFlow Lite on Python is still under development during the paper writing, we use vanilla TensorFlow library for inference on Raspberry Pi. The energy consumption is measured by an external Monsoon High Voltage Power Monitor [1]. We independently run each model on each dataset for 20 times and take the average of their inference time and energy consumption.

The time and energy efficiency results of inference on Nexus 5 are shown in Figure 13, while the results on Raspberry Pi 3 Model B are shown in Figure 14. The results on both devices share the following common observations: First, although both GlobalFusion and SADeepSense are based on the backbone structure of



(a) PAMAP2 Time

(b) RealWorld-HAR Time

(c) DSADS Time

(d) DG Time

(e) PAMAP2 Energy

(f) RealWorld-HAR Energy

(g) DSADS Energy

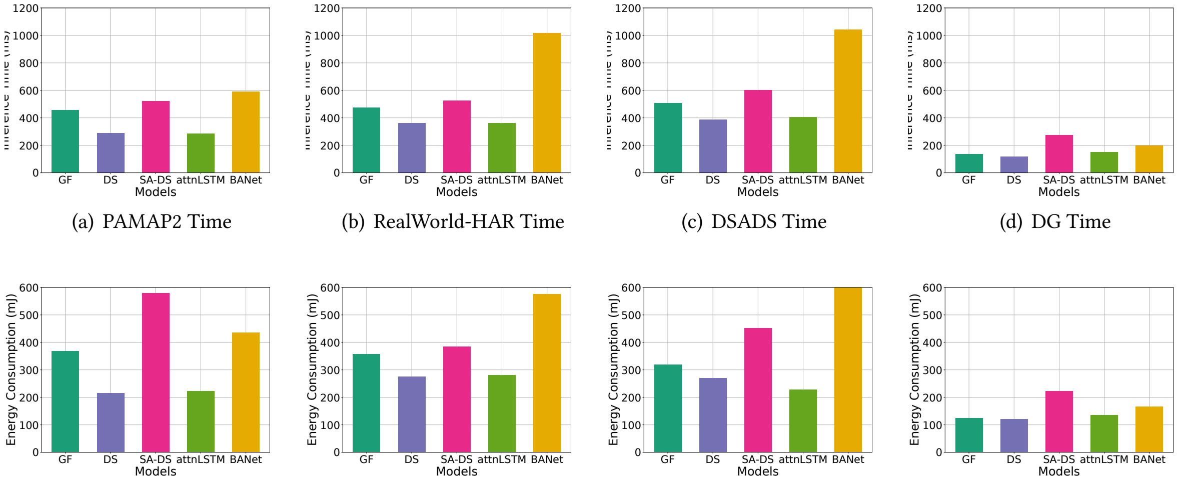
(h) DG Energy

Fig. 13. Time and energy efficiency on Nexus 5.

DeepSense, GlobalFusion always leads to shorter inference time and less energy consumption, because we do not utilize the multi-head design in our global attention module. The additional time and energy overhead of GlobalFusion compared to DeepSense is within an acceptable range. Second, the attnLSTM is most time and energy efficient across all datasets, because it has less layers than all compared models. In addition, they first fuse information of different sensing modalities at each specific intervals (i.e., which is same as our design) before merging information across intervals, so that only one recurrent module is used. Third, SADeepSense and BANet show the worst time and energy efficiency. SADeepSense has the most layers in its architecture and utilize multi-head design in both sensor and time attention modules. BANet first aggregates information across intervals at each individual sensor before fusing information across the sensing modalities. One individual recurrent unit is required by each sensor. Since the RNN computations are typically time consuming, we can expect to see a larger time difference between BANet and all other models when utilizing GPU for computations, because GPU is optimized for parallel computation of convolutional operations. The inefficiency of BANet is amplified especially we have more sensor types and body positions to fuse information (i.e., BANet is slower than all other models on both RealWorld-HAR and DSADS with a large margin). Four, in most cases of our overhead evaluations, energy efficiency shares the similar results with time efficiency because we only use CPU for computations. A more complex relation between time and energy consumption will appear when GPU and DSP on the chip are involved in the computation, or cloud offloading is applied during the inference. This problem is beyond the scope of discussion in this paper.

5 DISCUSSION

In this section, we briefly discuss a few issues related to our design in global attention module and GlobalFusion framework. Some lessons and experiences we learned from this paper are also included.



(a) PAMAP2 Time

(b) RealWorld-HAR Time

(c) DSADS Time

(d) DG Time

(e) PAMAP2 Energy

(f) RealWorld-HAR Energy

(g) DSADS Energy

(h) DG Energy

Fig. 14. Time and energy efficiency on RaspBerry Pi 3 Model B.

5.1 Time Attention

We are aware that most attentional frameworks [34, 42, 47] for multisensor information fusion in IoT contain a time-attention module, which automatically learns weight among hidden state of Recurrent Neural Network (RNN) at each time interval. We did not choose to do so for the following reasons: First, we tried to add a time attention module after the time recurrent module in GlobalFusion, using the hidden state of last time interval as the query, but got no improvement at accuracy except slower training speed. Second, it is hard to distinguish the information corresponding to each time interval. Due to the recurrent property of RNN, the information contained in the hidden state of each time interval is accumulative. Each hidden state not only includes the information extracted from this time interval but also the information accumulated from all previous intervals. Third, since the number of time intervals is typically small in sensing data time-frequency representation, the information loss resulted by RNN when dealing with long sequence is not significant. In summary, we believe that temporal attention does not fit well with GlobalFusion architecture so we do not integrate it.

5.2 Integrating Human Intuitions

One aspect that we did not handle well is integrating human intuitions into attention design. In general, our global attention design still belongs to data-driven information fusion mechanisms. However, some human knowledge and insights are very helpful to attention design. For example, accelerometer makes dominating contribution to discriminating between different moving activities, but it fails to classify static postures because the acceleration is the same under all static postures. Instead, orientation information recorded by magnetometer is better at distinguish these static postures. Actually, as we observed in Section 4, much error of GlobalFusion comes from classifying static postures because acceleration features are paid too much attention. We are looking for effective ways to add more human insights into the global attention design as additional supervision or constraint to the model. We look forward to seeing these simple but critical human intuitions to help data-driven approach to avoid some unnecessary local optimal situations.

5.3 Complex Sensor Interactions

Our GlobalFusion framework independently computes the attention distributions among sensors within each time interval. We do not consider the correlations among sensor attention weight distributions across time intervals. Zeng et. al. [47] try to solve this problem from two aspects: first, when computing sensor attention at each time interval, they feed the sensor attention distribution from previous time interval as additional input; second, they add a continuous constraint to both time attention and sensor attention to make attention weight change smoothly. However, from our experience, these two tricks do not help improve the classification performance when used in GlobalFusion. We regard it as an interesting direction for future work. We hope to add a separate module to directly model the dynamic sensor interaction process and attention distribution transition. It would be better if we can integrate existing physical knowledge into the design of this interaction module.

5.4 Sharing Model among Datasets

Through comparing and analyzing attention distributions on different datasets in Section 4.6, we have observed some common results. For example, the heterogeneity among different sensor types is typically higher than that of different spatial positions. In addition, accelerometer information all occupies the highest attention weight, while arm features are less related to activity classes. Therefore, it is possible to extract shared knowledge from different datasets. From another perspective, unlike the large scale public dataset in computer vision, such as ImageNet [10], the publicly available IoT datasets are usually too small to conduct comprehensive test on deep and complex sensing data processing frameworks. It would be beneficial if we can combine these datasets together to train a general attentional framework for human activity recognition. Later, we only need a small number of data samples to fine tune the model before applying it to a specific problem. This learning pipeline has been proved successful in natural language processing (NLP) [11], so we also expect to see the success of big data training in IoT field.

5.5 Heterogeneous Sensing Modalities

The global attention module is originally designed for information fusion from heterogeneous sensing modalities and diverse spatial positions. However, due to the limitation of publicly available sensing dataset we can find, there are two main shortcomings in our current evaluation: first, we only evaluate the model performance on human activity recognition application; second, the covered sensors all belong to inertial sensors (i.e., accelerometer, gyroscope, and magnetometer). The heterogeneity among sensing modalities is not well addressed in our evaluation section. However, our design is not based on specific insight into activity recognition or inertial sensors, but it is a pure data-driven approach, so we can also extend it to other IoT applications and sensing modalities. In the future, we plan to continue to test the global attention module beyond human activity recognition application and inertial sensing system, either based on new public dataset or data collected by ourselves.

6 CONCLUSION

In this paper, we described attention mechanisms that improve accuracy of multisensor information fusion when applying deep learning techniques to IoT applications. We proposed a global attention mechanism for this problem. Different from previous attention mechanisms for sensor fusion, global attention correlates global context derived from higher level features of the neural network with local sensor features to find informative ones and filter out the unrelated noise. Through a residual connection between attention query and output, we successfully combined the emphasized local information and the extracted global information. The global attention module is a flexible and configurable unit that can be easily adapted to work with most sensing data processing frameworks. We integrated it into an end-to-end learning framework, named GlobalFusion, for IoT applications with multisensor inputs. It relies on two global attention modules to automatically evaluate the contribution from different spatial positions and sensors. Through extensive evaluations compared with the state-of-the-art information fusion mechanisms on four public human activity recognition (HAR) datasets, we demonstrated the effectiveness of our global attention mechanism. We have further shown that attention allocation in our design corresponds with intuition. Moreover, the inference time and energy consumption tests on commodity IoT devices revealed that the global attention module is a lightweight design that incurs a negligible overhead only.

ACKNOWLEDGMENTS

Research reported in this paper was sponsored in part by the Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] [n.d.]. *Monsoon High Voltage Power Monitor.* https://www.msoon.com/online-store/High-Voltage-Power-Monitor-Part-Number-AAA10F-p90002590

[2] [n.d.]. *TensorFlow Lite Interpreter.* https://www.tensorflow.org/lite/guide/inference

[3] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 265–283.

[4] Kerem Altun, Billur Barshan, and Orkun Tunçel. 2010. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition* 43, 10 (2010), 3605–3620.

[5] Marc Bachlin, Daniel Roggen, Gerhard Troster, Meir Plotnik, Noit Inbar, Inbal Meidan, Talia Herman, Marina Brozgol, Eliya Shaviv, Nir Giladi, et al. 2009. Potentials of enhanced context awareness in wearable assistants for Parkinson's disease patients with the freezing of gait syndrome. In *2009 International Symposium on Wearable Computers*. IEEE, 123–130.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[7] Sourav Bhattacharya and Nicholas D Lane. 2016. Sparsification and separation of deep learning layers for constrained resource inference on wearables. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*. ACM, 176–189.

[8] Sneha Chaudhari, Gungor Polatkan, Rohan Ramanath, and Varun Mithal. 2019. An attentive survey of attention models. *arXiv preprint arXiv:1904.02874* (2019).

[9] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. 2018. Aˆ 2-Nets: Double Attention Networks. In *Advances in Neural Information Processing Systems*. 352–361.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[12] David Hall and James Llinas. 2001. *Multisensor data fusion*. CRC press.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[14] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[15] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. 2018. Learn to pay attention. *arXiv preprint arXiv:1804.02391* (2018).

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[18] Nicholas D Lane, Sourav Bhattacharya, Petko Georgiev, Claudio Forlivesi, Lei Jiao, Lorena Qendro, and Fahim Kawsar. 2016. Deepx: A software accelerator for low-power deep learning inference on mobile devices. In *Proceedings of the 15th International Conference on Information Processing in Sensor Networks*. IEEE Press, 23.

[19] Nicholas D Lane, Sourav Bhattacharya, Akhil Mathur, Petko Georgiev, Claudio Forlivesi, and Fahim Kawsar. 2017. Squeezing deep learning into mobile and embedded devices. *IEEE Pervasive Computing* 16, 3 (2017), 82–88.

[20] Xiaochen Liu, Pradipta Ghosh, Oytun Ulutan, BS Manjunath, Kevin Chan, and Ramesh Govindan. 2019. Caesar: cross-camera complex activity recognition. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. ACM, 232–244.

[21] Yang Liu, Zhenjiang Li, Zhidan Liu, and Kaishun Wu. 2019. Real-time Arm Skeleton Tracking and Gesture Inference Tolerant to Missing Wearable Sensors. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 287–299.

[22] Zuozhu Liu, Wenyu Zhang, Tony QS Quek, and Shaowei Lin. 2017. Deep fusion of heterogeneous sensor data. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5965–5969.

[23] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[24] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. 2015. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (2015), 1692–1706.

[25] Valentin Radu, Nicholas D Lane, Sourav Bhattacharya, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. 2016. Towards multimodal deep learning for activity recognition on mobile devices. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 185–188.

[26] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. 2018. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 157.

[27] Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers*. IEEE, 108–109.

[28] Oren Rippel, Jasper Snoek, and Ryan P Adams. 2015. Spectral representations for convolutional neural networks. In *Advances in neural information processing systems*. 2449–2457.

[29] Seyed Ali Rokni and Hassan Ghasemzadeh. 2017. Synchronous dynamic view learning: a framework for autonomous training of activity recognition models using wearable sensors. In *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*. ACM, 79–90.

[30] Sheng Shen, He Wang, and Romit Roy Choudhury. 2016. I am a smartwatch and i can track my user's arm. In *Proceedings of the 14th annual international conference on Mobile systems, applications, and services*. ACM, 85–96.

[31] Timo Sztyler and Heiner Stuckenschmidt. 2016. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 1–9.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[33] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. 2018. Multi-Level Sensor Fusion with Deep Learning. *CoRR* abs/1811.02447 (2018). arXiv:1811.02447 http://arxiv.org/abs/1811.02447

[34] Chongyang Wang, Min Peng, Temitayo A Olugbade, Nicholas D Lane, Amanda C De C Williams, and Nadia Bianchi-Berthouze. 2019. Learning Bodily and Temporal Attention in Protective Movement Behavior Detection. *arXiv preprint arXiv:1904.10824* (2019).

[35] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7794–7803.

[36] Xuyu Wang, Xiangyu Wang, and Shiwen Mao. 2018. RF sensing in the Internet of Things: A general deep learning framework. *IEEE Communications Magazine* 56, 9 (2018), 62–67.

[37] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[38] Weitao Xu, Yiran Shen, Neil Bergmann, and Wen Hu. 2016. Sensor-assisted face recognition system on smart glass via multi-view sparse representation classification. In *2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 1–12.

[39] Hongfei Xue, Wenjun Jiang, Chenglin Miao, Ye Yuan, Fenglong Ma, Xin Ma, Yijiang Wang, Shuochao Yao, Wenyao Xu, Aidong Zhang, et al. 2019. DeepFusion: A Deep Learning Framework for the Fusion of Heterogeneous Sensory Data. In *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 151–160.

[40] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. 2017. 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, 351–360.

[41] Shuochao Yao, Ailing Piao, Wenjun Jiang, Yiran Zhao, Huajie Shao, Shengzhong Liu, Dongxin Liu, Jinyang Li, Tianshi Wang, Shaohan Hu, et al. 2019. Stfnets: Learning sensing signals from the time-frequency perspective with short-time fourier neural networks. (2019), 2192–2202.

[42] Shuochao Yao, Yiran Zhao, Huajie Shao, Dongxin Liu, Shengzhong Liu, Yifan Hao, Ailing Piao, Shaohan Hu, Su Lu, and Tarek F Abdelzaher. 2019. SADeepSense: Self-Attention Deep Learning Framework for Heterogeneous On-Device Sensors in Internet of Things Applications. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 1243–1251.

[43] Shuochao Yao, Yiran Zhao, Aston Zhang, Shaohan Hu, Huajie Shao, Chao Zhang, Lu Su, and Tarek Abdelzaher. 2018. Deep learning for the internet of things. *Computer* 51, 5 (2018), 32–41.

[44] Shuochao Yao, Yiran Zhao, Aston Zhang, Lu Su, and Tarek Abdelzaher. 2017. Deepiot: Compressing deep neural network structures for sensing systems with a compressor-critic framework. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. ACM, 4.

[45] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in neural information processing systems*. 3320–3328.

[46] Ye Yuan, Guangxu Xun, Kebin Jia, and Aidong Zhang. 2017. 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, 213–222.

[47] Ming Zeng, Haoxiang Gao, Tong Yu, Ole J Mengshoel, Helge Langseth, Ian Lane, and Xiaobing Liu. 2018. Understanding and improving recurrent networks for human activity recognition by continuous attention. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. ACM, 56–63.

[48] Yiran Zhao, Shuochao Yao, Dongxin Liu, Huajie Shao, and Shengzhong Liu. 2019. GreenRoute: A Generalizable Fuel-Saving Vehicular Navigation Service. (2019), 1–10.