Latent Independent Excitation for Generalizable Sensor-based Cross-Person Activity Recognition

Hangwei Qian, Sinno Jialin Pan, Chunyan Miao
Nanyang Technological University, Singapore
{hangwei.qian, sinnopan, ascymiao}@ntu.edu.sg

Abstract

In wearable-sensor-based activity recognition, it is often assumed that the training and the test samples follow the same data distribution. This assumption neglects practical scenarios where the activity patterns inevitably vary from person to person. To solve this problem, transfer learning and domain adaptation approaches are often leveraged to reduce the gaps between different participants. Nevertheless, these approaches require additional information (i.e., labeled or unlabeled data, meta-information) from the target domain during the training stage. In this paper, we introduce a novel method named Generalizable Independent Latent Excitation (GILE) for human activity recognition, which greatly enhances the cross-person generalization capability of the model. Our proposed method is superior to existing methods in the sense that it does not require any access to the target domain information. Besides, this novel model can be directly applied to various target domains without re-training or fine-tuning. Specifically, the proposed model learns to automatically disentangle domain-agnostic and domain-specific features, the former of which are expected to be invariant across various persons. To further remove correlations between the two types of features, a novel Independent Excitation mechanism is incorporated in the latent feature space. Comprehensive experimental evaluations are conducted on three benchmark datasets to demonstrate the superiority of the proposed method over state-of-the-art solutions.

Introduction

With the development of ubiquitous wearable sensors, such as mobile phones, smart watches and sports bracelets, human activity recognition has become one of the most crucial techniques in a wide range of real-world applications, such as healthcare, gait analysis, assisted living, security and smart homes (Chen et al. 2020; Wang et al. 2019; Janidarmian et al. 2017; Bulling, Blanke, and Schiele 2014; Lara and Labrador 2013). The goal of activity recognition is to classify the activities conducted by the participants via machine learning algorithms. Generally, there are two types of scenarios: wearable-sensor-based and video-based settings (Wang et al. 2019). In this work, we focus on wearable-sensor-based human activity recognition, in which the raw data is composed of streams of sensor readings received from wearable sensors. As the activity recognition applications become more and more pervasive, constructing a model that works well for different persons becomes increasingly important. Some state-of-the-art models improve the model capacity by improving feature learning capabilities with multiple neural network modules such as Convolutional Neural Networks, Recurrent Neural Networks and Autoencoders (Yang et al. 2015; Morales and Roggen 2016a; Qian et al. 2019). These models work well when the training and test data have overlapping data distributions, which are achieved by splitting data of each participant into both training and test sets (Micucci, Mobilio, and Napoletano 2016). Nevertheless, this kind of training/test split is rather impractical in real world applications, as it is almost impossible to annotate sufficient training data for every new user. What's worse, the accuracy of such models plummets when applied to data collected from new unseen participants, indicating lack of generality across different persons.

To tackle the distribution discrepancies among different persons, transfer learning and domain adaptation methods are proposed in the literature (Pan and Yang 2010). Transfer learning methods typically train a model from multiple source domains and then adapt to the target domain. In order to avoid negative transfer, additional information is more or less required from the target domain, such as annotated or unlabeled data, or other meta-information (e.g., the class label proportion) (Morales and Roggen 2016b; Wang et al. 2018; Khan, Roy, and Misra 2018; Wilson, Doppa, and Cook 2020). With the auxiliary information representing the target domain, it is possible to fine-tune the models to adapt to the target domain (Buffelli and Vandin 2020; Rokni, Nourollahi, and Ghasemzadeh 2018; Mazankiewicz, Böhm, and Bergés 2020). In the meanwhile, features can be mapped to a common subspace in order to reduce domain gap (Bai et al. 2020; Wang et al. 2018). From the literature, transfer-based approaches are demonstrated to be effective to reduce domain gaps. However, these existing approaches have a notable limitation, that is, in the case of multiple target domains, the models need to be re-trained for each target domain. Domain generalization approaches aim to learn generalizable features from several related source domains, to make the model work well on previously unseen domains during test time (Khosla et al. 2012; Li et al. 2017). While many methods have been designed for computer vision ap-

plications, these network architectures are incompatible with time series data. There is very limited research attention that focuses on wearable-sensor-based data (Wilson, Doppa, and Cook 2020). Taking this cue, as well as the uniqueness of people's activity characteristics, each person's data, in this paper, is treated as a single domain. Thus, we use the term `domain` and `person` interchangeably hereinafter.

In the applications of activity recognition, one desired model should be able to work well on unseen target data, which is arguably more challenging than domain adaptation, while with greater significance in practice. Take fall detection for the elderly as an example. Despite the goal being detecting falling activity of the elderly, it is inconvenient to collect training data from the elderly due to safety issues. However, it is less dangerous to collect training data by younger participants with safety-ensured equipments. The model is expected to train on the data collected from younger participants and be readily applicable to the elder users without collecting annotated training data from them. Empirically we demonstrate that the state-of-the-art approaches have a performance drop when encountered with unseen target data. Detailed observations are listed in Section . To tackle the problem, we propose to learn a deep generalizable model across different domains named Generalizable Independent Latent Excitation (GILE). Our proposed method is superior to existing methods, in the sense that it does not require any auxiliary information from an unseen target domain. After completing the training stage, our method can be directly applied to multiple target domains without re-training.

To incorporate the variations across domains, we develop our generative method on top of the variational autoencoder (VAE) framework (Kingma and Welling 2014). Specifically, two probabilistic encoders are utilized to induce two groups of latent representations, i.e., domain-agnostic and domain-specific representations. Ideally, the domain-agnostic representations capture the common information on conducting a certain class of activity, and the domain-specific representations can reflect the unpredictable factors that induce the variations among training domains, such as different environments, physical conditions and lifestyles of participants, etc. To effectively disentangle the two latent spaces, we develop a novel Independent Excitation mechanism. By removing domain-specific representations, the resulting domain-agnostic latent space is expected to be more invariant to different domains than the original data. As a result, the model is expected to generalize better to new unseen target domains. Our experimental evaluations on three activity datasets validate that our model can outperform state-of-the-art methods with enhanced generalization capability.

Related Work

Feature learning methods on human activity recognition can roughly be grouped into two categories: conventional machine learning methods and deep learning methods (Chen et al. 2020). General machine learning methods include PCA, LDA, Fourier transformation and handcrafted features, such as mean, variance, median, maximum, minimum, etc (Janidarmian et al. 2017; Bulling, Blanke, and Schiele 2014). Other methods learn statistical and structural features that are specifically designed for activity recognition (Qian, Pan, and Miao 2018; Hammerla et al. 2013; Lin et al. 2007). Deep learning methods automatically extract features from raw signals, alleviating the feature design procedure. DeepConvLSTM model (Morales and Roggen 2016a) consists of four layers of convolutions and the succeeding two LSTM layers. DDNN model learns three types of features, i.e., statistical, temporal and spatial correlation features (Qian et al. 2019). Systematic comparisons on the performances of DNNs, CNNs and RNNs on activity data are provided in (Hammerla, Halloran, and Plötz 2016).

To enable transfer learning across domains, many approaches train a deep model on source domains and fine tune with fixed or sequential training data from the target domain (Morales and Roggen 2016b; Buffelli and Vandin 2020; Rokni, Nourollahi, and Ghasemzadeh 2018; Mazankiewicz, Böhm, and Bergés 2020). Wang et al. (2018) and Bai et al. (2020) reduced the distribution divergence between domains by minimizing the differences between domains. Khan, Roy, and Misra (2018) transferred the distributions of weights in source domains to target domains. The DSN model separates features into two subspaces and designs a loss function on feature space to encourage independence on the premise that unlabeled data from the target domain is available (Bousmalis et al. 2016). A Convolutional deep Domain Adaptation model for Time Series data (CoDATS) is a domain-adaptation method with weak supervision of target domain's label proportion (Wilson, Doppa, and Cook 2020). The model learns domain-invariant features by adversarial learning of the feature extractor and the classifier. Target domain's label proportion works as an extra constraint during the model training stage. Note that the above approaches more or less require access to the target domain in the training stage, limiting their potential capacities.

The Proposed Model

Problem Statement

In our setting of activity recognition, a domain is defined as a joint distribution $\mathbb{P}^d(x, y)$ on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}, \mathcal{Y}$ and $d \in \mathcal{D} = \{1, ..., D\}$ denote the activity instance space, activity class space and the index of a source domain. We are given labeled data from D source domains $\{(\mathbf{X}^d, \mathbf{y}^d) \sim \mathbb{P}^d(x, y)\}_{d=1}^{D}$ from D participants as training data. Note that $\mathbb{P}^i(x, y) \neq \mathbb{P}^j(x, y), \forall i \neq j$ where $i, j \in \mathcal{D}$. All the domains share the same label space, i.e., $y \in \{1, ..., n_c\}, \forall y \in \mathbf{y}^d, \forall d \in \mathcal{D}$ where n_c denotes the number of activity categories the participants have conducted. Each $\mathbf{X}^d$ contains N_d samples $\mathbf{x}_i^d$ where $i \in \{1, ..., N_d\}$, and each sample $\mathbf{x}_i^d \in \mathbb{R}^{M \times 1}$ represents a vector of signals received from wearable sensors at i-th timestamp. In the test phase, the test data $\mathbf{X}^{\tilde{d}}$ contains signals gathered from unseen participant(s), where $\tilde{d} \notin \mathcal{D}$. Our goal is to train a generalizable deep learning model parametrized as f from D source domains such that the trained model is able to generalize well to target domain data $\mathbf{X}^{\tilde{d}}$. Empirically, we want to minimize the target risk $\epsilon_{\tilde{d}}(f) = Pr_{(\mathbf{x},y)\sim\mathbb{P}^{\tilde{d}}}[f(\mathbf{x}) \neq y]$. Our setting is more challenging than conventional transfer learn-

ing setting, due to the fact that the model does not have access to any information from the target domain. Compared with standard machine learning settings, our setting requires D domain labels. Fortunately, the domain label is trivial to collect since the participants' identities are typically anonymously logged as the source of the training data.

Domain Agnostic and Specific Features

Generally, the data samples collected from different participants for activity recognition are determined by many factors, such as the class of activities being conducted, the environmental constraints for conducting activities, the moving patterns of participants, etc. In order to enable the generalization ability across different domains, it is crucial for a model to effectively identify different factors from raw data. Without loss of generality, we model the data generation process of observed activity data to be determined by two types of factors: domain-agnostic and domain-specific factors. The domain-agnostic factors contain commonality of conducting the same activity among different domains, i.e., the prototype or archetype of an activity. The domain-specific factors, on the contrary, are the latent factors that lead to the inter-domain differences of sensor readings, such as different participants' varying lifestyles, health conditions, moving patterns, etc.

As inputs for neural networks, the training instances are partitioned by fixed-size sliding window with length L. Here the input of our model is denoted as $\mathbf{x}^d \in \mathbb{R}^{M \times L}$. We construct two probabilistic encoders $p(\mathbf{z}|\mathbf{x}^d)$ and $p(\mathbf{z}_d|\mathbf{x}^d)$ parameterized by ψ and ψ_d to extract latent features $\mathbf{z}$ and $\mathbf{z}_d$ that reflect the domain-agnostic and domain-specific factors respectively. Note that $\mathbf{z}_d$ is different for different domains. Different from deterministic autoencoders that encode an input instance as a single data point, we encode it as a distribution over the latent space in order to incorporate variations in the data. The marginal likelihood of the latent space becomes intractable due to the neural-network-based conditional probabilities. To solve this problem, variational approximations are adopted, and a prior is assumed on the latent representations. A common choice for the prior is a multivariate standard Gaussian distribution $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$. The imposed priors actually model the intra-domain uncertainties, since the generated data $\mathbf{x}^d$ is influenced by variations in the latent factors. As shown in the following equations, the encoders learn (μ_d, σ_d) and (μ, σ) such that $q_d(\mathbf{z}_d|\mathbf{x}^d)$ and $q(\mathbf{z}|\mathbf{x}^d)$ follow the data-driven Gaussian distributions, from which two latent vectors $\mathbf{z}_d$ and $\mathbf{z}$ are generated via the reparametrization trick in VAE and β-VAE (Kingma and Welling 2014; Higgins et al. 2017).

$$p(\mathbf{z}_d|\mathbf{x}^d) \approx q_d(\mathbf{z}_d|\mathbf{x}^d; \psi_d) = \mathcal{N}(\mathbf{z}^d|\mu_d(\mathbf{x}^d), \sigma_d^2(\mathbf{x}^d); \psi_d),$$
$$p(\mathbf{z}|\mathbf{x}^d) \approx q(\mathbf{z}|\mathbf{x}^d; \psi) = \mathcal{N}(\mathbf{z}|\mu(\mathbf{x}^d), \sigma^2(\mathbf{x}^d); \psi).$$

The network architecture of the proposed model is shown in Figure 1. The obtained latent representations are then concatenated to feed into a decoder $p(\mathbf{x}^d|\mathbf{z}, \mathbf{z}_d; \phi)$ to reconstruct $\mathbf{x}^d$, with ϕ being the parameter set for the decoder.

Therefore, the marginal distribution for domain d is

$$p(\mathbf{x}^d) = \iint p(\mathbf{x}^d|\mathbf{z}, \mathbf{z}_d)p(\mathbf{z})p(\mathbf{z}_d)\, d\mathbf{z}\, d\mathbf{z}_d. \quad (1)$$

The probabilistic encoders and decoder are learned with the following objective function:

$$\begin{aligned}
&\mathcal{L}_{elbo}(\mathbf{x}^d, \mathbf{z}_d, \mathbf{z}; \psi_d, \psi, \phi) \\
&= \mathbb{E}_{d, q_d(\mathbf{z}_d|\mathbf{x}^d;\psi_d), q(\mathbf{z}|\mathbf{x}^d;\psi)}[\log p(\mathbf{x}^d|\mathbf{z}_d, \mathbf{z}; \phi)] \\
&\quad - KL(q_d(\mathbf{z}_d|\mathbf{x}^d; \psi_d)||p(\mathbf{z}_d; \phi)) \\
&\quad - KL(q(\mathbf{z}|\mathbf{x}^d; \psi)||p(\mathbf{z}; \phi)),
\end{aligned}$$

where the first item is the reconstruction error, and the last two terms calculate KL divergence between the sampled latent features and corresponding priors, which are interpreted as regularizers on the latent feature spaces.

Note that the above feature learning procedure is in an unsupervised manner. Although the two latent factors $\mathbf{z}_d$ and $\mathbf{z}$ are designed to be independent to each other, simply using two separate probabilistic encoders is not sufficient to guarantee that non-overlapping and disentangled features are learned. In extreme cases, the two encoders may learn identical latent representations. As a result, we utilize the class label y_i^d and domain label d to guide the learning of features in the training stage. To incorporate the class and domain information, two disentangling classifiers $\{DC_{\mathbf{d}}, DC_{\mathbf{y}}\}$ with parameters $\mathbf{w_d}$ and $\mathbf{w_y}$ take the latent features as input and predict corresponding labels. The classifier $DC_{\mathbf{d}}$ is trained to correctly predict domain label d from the domain-specific features $\mathbf{z}_d$, and similarly, the classifier $DC_{\mathbf{y}}$ is trained to correctly predict activity labels from the domain-agnostic features $\mathbf{z}$. The loss function is thus defined as

$$\begin{aligned}
&\mathcal{L}_{\text{DC}}(\mathbf{z}, \mathbf{z}_d, y_i^d, d; \mathbf{w_d}, \mathbf{w_y}) \\
&= \frac{1}{N_S} \sum_{d=1}^{D} \sum_{i=1}^{N_d} [\ell(y_i^d, DC_{\mathbf{y}}(\mathbf{z}; \mathbf{w_y})) + \ell(d, DC_{\mathbf{d}}(\mathbf{z}_d; \mathbf{w_d}))],
\end{aligned}$$

where $N_S = \sum_{d=1}^{D} N_d$ and $\ell(\cdot)$ is a task-specific loss function, e.g., cross-entropy $\ell(y, \tilde{y}) = -\sum_{c=1}^{n_c} \mathbf{I}[y = c] \log \tilde{y}$. To minimize the loss, the domain-agnostic features are encouraged to contain domain-agnostic factors, and the domain-specific latent space are encouraged to capture domain-specific factors through the training process.

The Independence Excitation Mechanism

Even though the above classifiers enable the the learning of two separate domain-agnostic and domain-specific features, overlapping features can still exist. For the domain-agnostic latent space, there may exist domain-specific features as long as these features do not alter the decision boundary of the classifier significantly, and vise versa, which will cause redundancy in neural networks. To further minimize the correlations between domain-agnostic and domain-specific features, we develop an extra Independence Excitation mechanism. This is inspired by the independent optical excitation of distinct neural populations (Klapoetke et al. 2014), which independently activate two distinct neural populations in the

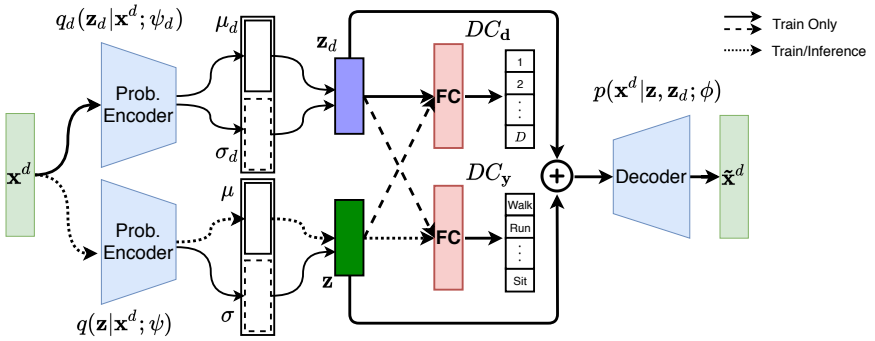


Figure 1: The network architecture of the proposed GILE model. The activity data $\mathbf{x}^d$ is the input of two probabilistic encoders $q_d(\mathbf{z}_d|\mathbf{x}^d;\psi_d)$ and $q(\mathbf{z}|\mathbf{x}^d;\psi)$, from which ($\mu_d$, σ_d) and (μ, σ) are learned. Then the domain-specific features $\mathbf{z}_d$ are sampled from μ_d and σ_d. Meanwhile, the domain-agnostic features $\mathbf{z}$ are sampled from μ and σ. The $\mathbf{z}_d$ is fed into a Fully-Connected layer $DC_\mathbf{d}$ to predict the domain label, and $\mathbf{z}$ is the input for activity classifier $DC_\mathbf{y}$. To minimize the correlations among two types of features, the Independent Excitation mechanism is designed, as shown in dashed arrow. Besides, the two features are concatenated to be input of the decoder $p(\mathbf{x}^d|\mathbf{z},\mathbf{z}_d;\phi)$ to reconstruct the input data. In the inference stage, only the modules corresponding to domain-agnostic features are used for activity class prediction, as shown in the dotted arrow.

brain tissue. Similarly, we encourage $\mathbf{z}$ to be directly responsible for class labels while making the other $\mathbf{z}_d$ totally uncorrelated. To this end, the Independence Excitation mechanism minimizes the accuracy of $DC_\mathbf{d}$ when $\mathbf{z}$ is fed into $DC_\mathbf{d}$. Likewise, the accuracy of $DC_\mathbf{y}$ is forced to be minimized when $\mathbf{z}_d$ is fed into $DC_\mathbf{y}$. In this way, the domain-agnostic features are encouraged to be irrelevant of the domain labels, and likewise, domain-specific features are expected to be non-informative of the activity labels. The Independence Excitation objective function is:

$$\mathcal{L}_{\text{IE}}(\mathbf{z},\mathbf{z}_d,y_i^d,d;\mathbf{w}_d,\mathbf{w}_y) = -\frac{1}{N_S}\sum_{d=1}^{D}\sum_{i=1}^{N_d}[\ell(y_i^d, DC_\mathbf{y}(\mathbf{z}_d;\mathbf{w}_\mathbf{y})) + \ell(d, DC_\mathbf{d}(\mathbf{z};\mathbf{w}_\mathbf{d}))].$$

Model Summary

In summary, our proposed model is trained in an end-to-end manner, and the total loss of the proposed model is formulated in weighted summation formula:

$$\mathcal{L} = \mathcal{L}_{elbo} + \alpha\mathcal{L}_{\text{DC}} + \gamma\mathcal{L}_{\text{IE}}, \tag{2}$$

where α and γ are the trade-off parameters between the three loss functions. Compared with VAE, our method considers two latent factors instead of a single factor, leading to more expressive capabilities of learned latent features. Compared with β-VAE which encounters trade-off between the complexity of latent features and the commonality towards the prior, our method alleviates this situation by utilizing available domain and activity labels as extra information to make the learned latent features meaningful and informative. Empirically, we find out that it is better to train the $\mathcal{L}_{elbo}$ separately from the other two loss functions. The underlying reason might be that the first loss function corresponds to the generative model part, while the remaining two loss functions correspond to classifiers. Therefore, we set one optimizer for the generative model and the other optimizer for the disentangling classifiers, and the two optimizers are trained iteratively. Also, α and γ are set to make three loss functions values in the similar order of magnitude.

When encountered with a new unseen target domain, the domain label space may be disjoint with source domains. Therefore, domain-specific features are not beneficial to the prediction task. Specifically, during the inference stage, a test data sample $\mathbf{x}^t$ drawn from an unknown target domain is fed into the probabilistic encoder $q(\mathbf{z}|\mathbf{x}^d;\psi)$ only, to generate the domain-agnostic features $\mathbf{z}^t$. Subsequently, $\mathbf{z}^t$ is fed into $DC_\mathbf{y}$ to predict the corresponding activity class labels.

Note that in this work, we simply apply both probabilistic encoders on the raw data $\mathbf{x}$. With this in mind, it is also possible to learn features $g(\mathbf{x})$ via existing feature extraction method g (any method from machine learning and deep learning models), and then feed the features into the encoders $q_d(\mathbf{z}_d|g(\mathbf{x}))$ and $q(\mathbf{z}|g(\mathbf{x}))$. Different feature extraction methods g and h can be applied to the raw data to construct features for the encoders separately, i.e., $q_d(\mathbf{z}_d|g(\mathbf{x}))$ and $q(\mathbf{z}|h(\mathbf{x}))$. It is also possible to incorporate more explicit domain expert knowledge to better define the domain-agnostic and domain-specific factors, serving as extra information for learning a generalizable model.

Experiments

Experimental Setup

Datasets. We evaluate the proposed method on three large-scale wearable-sensor-based benchmark datasets.

- The UCIHAR dataset (Anguita et al. 2012) contains six daily activities (walking, sitting, laying, standing, walking upstairs, walking downstairs) conducted by a group of 30 volunteers within an age range of 19 to 48. A smart phone is attached on the waist, with frequency of 50 Hz. There are 1,318,272 number of samples in total, and the instance dimension is 9.

- The Opportunity dataset (Chavarriaga et al. 2013) collects 4 participants' daily activities in an ambient-sensor home environment with different inertial sensor modalities. There are 18 fine-grained gesture classes, i.e., {Open / Close Dishwasher / Fridge / Drawer1 / Door1 / Drawer2 / Door2 / Drawer3, Move Cup, Clean Table and Null}. The Null class indicates the transition of every two adjacent activities. The total number of samples is 869,387, and the feature dimension is 113, with frequency of 30Hz.
- The UniMiB SHAR dataset (Micucci, Mobilio, and Napoletano 2016) records 9 types of activities of daily living (labeled as: StandingUpFL, LyingDownFS, StandingUpFS, Running, SittingDown, GoingDownS, GoingUpS, Walking, Null) and 8 types of falls (labeled as: FallingBackSC, FallingBack, FallingWithPS, FallingForw, FallingLeft, FallingRight, HittingObstacle, Syncope). 30 Subjects with ages ranging from 18 to 60 years conducted the 17 fine-grained activities partially or fully. The data is collected by an acceleration sensor embedded in an Android phone with sample frequency of 50 Hz. After pre-processing, each sample contains 3 vectors of 151 accelerometer values.

Baselines. We compare our proposed GILE model with the closely related baselines. The DeepConvLSTM (Morales and Roggen 2016a) and DDNN (Qian et al. 2019) are the state-of-the-art feature learning approaches on activity recognition. CoDATS (Wilson, Doppa, and Cook 2020) is the latest domain adaptation model for time series data. The class label proportions of target domain are provided to CoDATS as extra information for training. Since our method is based on VAE, we also compare with VAE (Kingma and Welling 2014), β-VAE (Higgins et al. 2017) and DIVA (Ilse et al. 2019), the latter of which is a state-of-the-art model for computer vision. We use the released code if it is available. Other methods without available codes are reproduced by us in Pytorch (Paszke et al. 2019).

Implementation Details. We conduct experiments with leave-one-domain-out strategy in each dataset: one of the domains is treated as the unseen target domain, and the rest domains are considered as available source domains. Due to the severe class imbalance problem in existing datasets, we set the probability of an activity being chosen in a training mini-batch to be the inverse of the number of the activity. F1 score is selected as performance measure. Data normalization is conducted on all datasets. Due to the large number of domains in UCIHAR, we simply utilize the first 5 domains. The UCIHAR data is pre-processed beforehand and is sampled in sliding window of 2.56 seconds and 50% overlap, resulting in 128 readings for each window. 77 out of 113 features are used for Opportunity, and sliding window of 1 second is applied. In UniMiB SHAR dataset, we select the first 4 subjects who have conducted all activities (ID: 1,2,3 and 5) in our experiments to keep the amount of subjects comparable to other datasets.

For our architecture, each probabilistic encoder consists of 4 layers of convolutions with a max-pooling layer after each convolution. A single fully-connected layer is used as each classifier. For methods which have l layers of LSTMs with h-dimensional hidden representations, we tune parameters $l \in \{1, 2, 3\}$ and $h \in \{32, 64, 128, 256, 512\}$. β is chosen from $\{0.002, 0.01, 0.1, 1, 5, 10, 100\}$. The batch size is set to 64, and the maximum training epoch is set to 100. For all methods except CoDATS, the Adam optimizer with learning rate 10^{-3} and weight decay 10^{-3} is used. For CoDATS method, the learning rate is reduced to 10^{-4} and the training epoch is set to 500. We report the best results among the different configurations for every method[1].

Domain Shift in Activity Recognition

We first conduct evaluations to investigate the domain shift problem in human activity recognition. To do so, we compare two settings. The first setting `random` is the traditional way, where each participant's data is randomly split into training and test data by a specific proportion. The second setting `cross-person` is the proposed way, where one participant's data is treated as unseen domain, and other participants' data are treated as labeled source domains data. To be fair, we set the proportion of training and test data in both settings to be identical. We apply two state-of-the-art methods, i.e., DeepConvLSTM and DDNN on both settings. For both methods, we tune the LSTM layers $l \in \{1, 2, 3\}$ and dimensions of hidden representations $h \in \{32, 64, 128, 256, 512\}$. The best performance are illustrated in Figure. 2 on the three datasets. As shown in the figure, the performance of both models drop significantly when the setting is changed from `random` setting to `cross-person` setting. These results favorably support our motivation that cross-person generalization is more challenging for human activity recognition. Among the three datasets, the UniMiB SHAR dataset has the most significant differences between the two settings, especially for the target domain 5, the accuracy drops from 83.89% to 34.9% for DeepConvLSTM approach and from 88.93% to 19.46% for DDNN method. The result actually fits the fact that the UniMiB SHAR dataset is designed to collect training data from diverse persons. Within each dataset, the performance gap between two settings are sometimes larger and sometimes smaller, indicating that the similarities among different persons may be different.

Experimental Results and Analysis

Source	DDNN	DeepConvLSTM	CoDATS	GILE
0	80.13	78.48	48.01	**82.49**
2	**93.05**	92.05	60.93	90.62
3	54.30	49.67	30.13	**56.56**
4	72.15	68.54	32.78	**76.56**
Ave.	74.91	72.18	42.96	**76.56**

Table 4: Performance comparisons of single source domain settings on UCIHAR. The target domain is domain 1.

[1]The source code and high-resolution figures are available at https://github.com/Hangwei12358/cross-person-HAR.

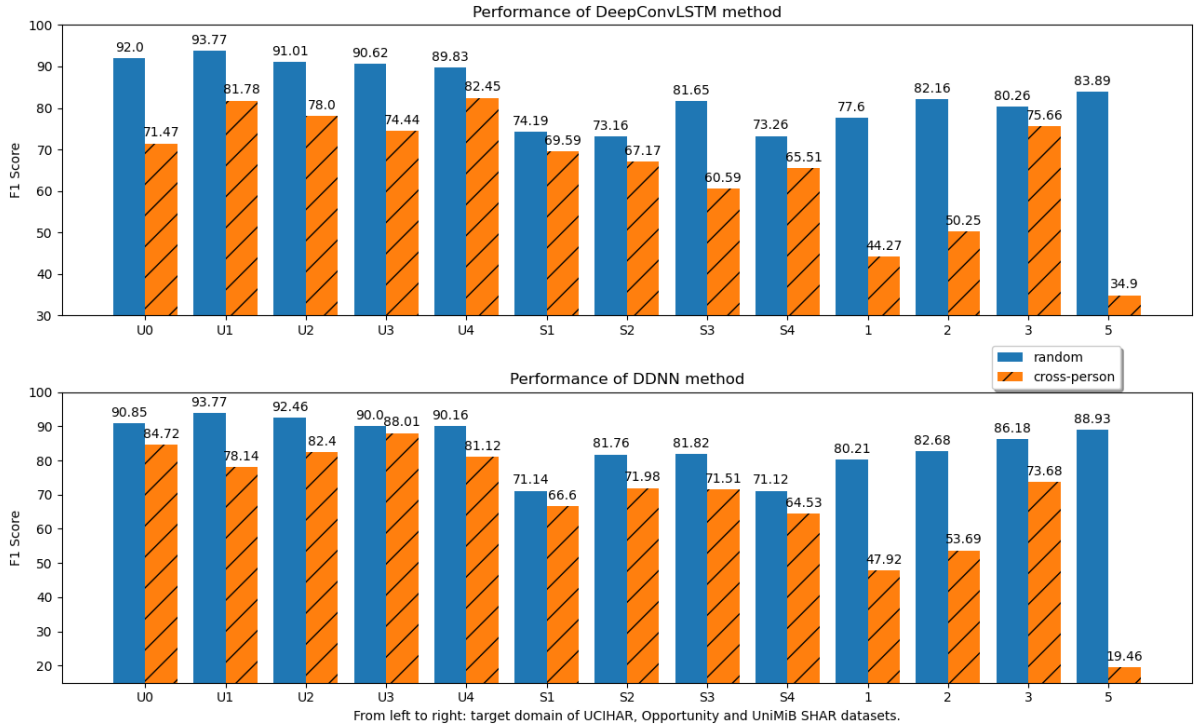


Figure 2: Comparisons of `random` and `cross-person` settings on 3 datasets. U0 to U4 denote target domains in UCIHAR. S1 to S4 are from Opportunity, and the rest are from UniMiB SHAR.

Source	Target	VAE	β-VAE	DIVA	DDNN	DeepConvLSTM	CoDATS	GILE
1,2,3,4	0	51.87	53.31	75.00	<u>84.72</u>	71.47	81.27	**85.15**
0,2,3,4	1	44.70	44.37	77.18	78.14	**81.78**	55.63	<u>81.56</u>
0,1,3,4	2	64.22	62.17	71.61	<u>82.40</u>	78.00	77.42	**86.97**
0,1,2,4	3	36.91	49.21	81.87	<u>88.01</u>	74.44	60.57	**94.37**
0,1,2,3	4	39.07	58.28	79.68	81.12	<u>82.45</u>	66.23	**92.81**
Ave.		47.35	53.47	77.07	<u>82.88</u>	77.63	68.22	**88.17**

Table 1: The overall performance on the UCIHAR dataset (unit: %). The best performance is highlighted in bold, and the second best performance is underlined.

Source	Target	VAE	β-VAE	DIVA	DDNN	DeepConvLSTM	CoDATS	GILE
S2,S3,S4	S1	77.21	11.48	75.86	66.6	69.59	<u>83.58</u>	**83.86**
S1,S3,S4	S2	73.94	61.02	73.54	71.98	67.17	<u>81.04</u>	**81.65**
S1,S2,S4	S3	15.65	31.72	65.81	71.51	60.59	<u>78.11</u>	**78.66**
S1,S2,S3	S4	75.86	13.65	73.43	64.53	65.51	<u>80.60</u>	**81.41**
Ave.		60.67	29.47	72.16	68.66	65.72	<u>80.83</u>	**81.40**

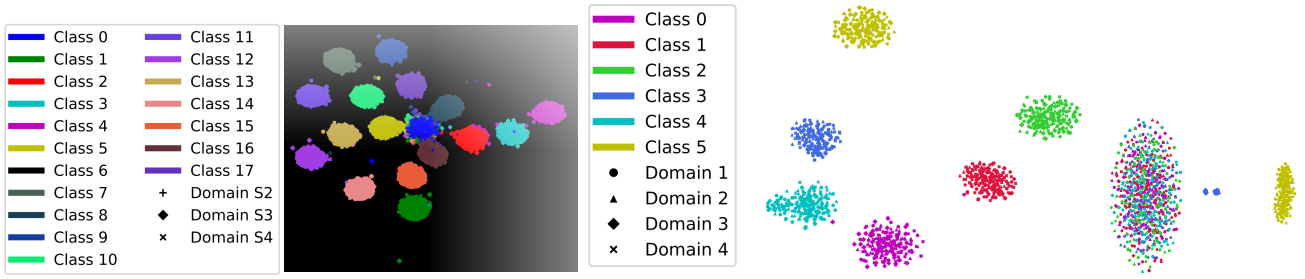
Table 2: The overall performance on the Opportunity dataset (unit: %). The best performance is highlighted in bold, and the second best performance is underlined.

The overall experimental results of the proposed GILE method and baselines on UCIHAR, Opportunity and UniMiB SHAR datasets are listed in Table 1, Table 2 and Table 3, respectively. Overall, our proposed method has achieved the best average performance on all datasets, which greatly illustrates its robustness and generalization capability across different target domains.

On UCIHAR dataset, the feature-learning-based approaches, i.e., DDNN and DeepConvLSTM, have higher accuracy than VAE-based approaches and CoDATS. This may

Source	Target	VAE	β-VAE	DIVA	DDNN	DeepConvLSTM	CoDATS	GILE
2,3,5	1	11.72	15.63	<u>48.17</u>	47.92	44.27	42.71	**55.72**
1,3,5	2	32.76	32.76	39.06	<u>53.69</u>	50.26	46.66	**54.06**
1,2,5	3	22.37	26.97	61.87	<u>73.68</u>	**75.66**	61.51	70.31
1,2,3	5	29.19	30.20	<u>38.43</u>	19.46	34.90	31.88	**42.81**
Ave.		24.01	26.39	46.88	48.69	<u>51.27</u>	45.69	**55.61**

Table 3: The overall performance on the UniMiB SHAR dataset (unit: %). The best performance is highlighted in bold, and the second best performance is underlined.



(a) GILE model on Opportunity. (b) GILE model on UCIHAR. (c) VAE model on UCIHAR.

Figure 3: Visualization of the t-SNE embeddings of learned feature spaces. Best viewed in color. Different shapes denote different domains, and each class of activity is demonstrated by a distinct color.

be due to the extreme small number of features in UCIHAR, such that extracting more powerful features enable better learning of the classification. Our method achieves the best on 4 out of 5 scenarios, and the performance of the second scenario is only 0.22% inferior to the best performance.

On Opportunity dataset, the CoDATS and GILE is generally better than VAE-based and feature-learning-based methods. This also supports our observation on the importance of feature learning when raw data has only a few features. For Opportunity dataset, the raw data contains 77 dimensions of features, which enables our method and CoDATS to have superior performance.

On UniMiB SHAR dataset, our GILE method ranks the first when the target domain is 1, 2 and 5. These three tasks are more difficult than the rest one, according to Table. 3.

Therefore, the above results on the three datasets favourably demonstrate that our proposed GILE model is capable of generalizing well from several source domains to unseen target domain.

Source Domain Similarity Matters. We investigate why our proposed method achieves inferior performance on certain scenarios, such as the setting when person 1 is target domain in UCIHAR dataset. We conduct experiments on single source domain setting, and the results are listed in Table. 4. From the table, we find out that when the source data comes from person 2, both DDNN and DeepConvLSTM achieve relatively high performance compared with CoDATS. This indicates that person 2 and 1 have very little domain difference such that transfer learning is not necessary for the two persons. This also explains our inferior performance compared with feature-learning-based approaches when the target domain ID is 1.

Latent Feature Space Visualization The t-SNE embeddings of the learned latent domain-agnostic representations $\mathbf{z}$ on Opportunity and UCIHAR are plotted in Figure. 3(a) and 3(b), to validate whether our proposed GILE is able to successfully learn domain-agnostic features. For comparison, the t-SNE embedding of VAE on UCIHAR is shown in Fig. 3(c). We observe that for the proposed GILE model, features from different domains are mixed together, indicating that the learned latent space $\mathbf{z}$ is indeed not affected by domain-specific factors. In addition, the clusters of embeddings of GILE are more distinct and organized than those of VAE, and samples with the same activity class tend to group into the same cluster, resulting in that the number of clusters learned by GILE is exactly the number of classes.

Conclusion

In this paper, we propose a novel method named GILE for cross-person sensor-based human activity recognition. The proposed approach effectively learns generalizable feature representations across domains by means of disentangling domain-agnostic and domain-specific features. The two groups of features are split by the Independent Excitation mechanism. The proposed approach is shown to consistently achieve the best performance over the state-of-the-art methods on three datasets. In the future, we plan to combine source domain selection with the proposed method to investigate the similarities among domains.

Acknowledgments

This research is partially supported by the NTU Singapore Nanyang Assistant Professorship (NAP) grant M4081532.020 and Singapore MOE AcRF Tier-1 grant 2018-T1-002-143 (RG131/18 (S)). H.Qian thanks the support from the Wallenberg - NTU Presidential Postdoctoral Fellowship.

Ethics Statement

Our proposed method aims to develop a generalizable model in wearable-sensor-based activity recognition tasks such that it can be directly applied to different target domains without having access to any training data from target domains. Our starting point is that for sensors like mobile phones, sports bracelets, it is a common practice to fine-tune a pre-trained model with extra data collected from the target domain. Take the Garmin sports watch for example, it collects at least seven days' data from a new user before it is able to make accurate predictions on daily activities. From our perspective, it is practically important to 1) collect as less data as possible in real-world applications, 2) reduce the time required to customize the model for new users. To this end, we develop the GILE method to tackle the problem and we believe that it can be practically beneficial. In practical applications of activity recognition, the privacy of the collected training data should not be violated, and we will stay aware of this issue and further improve the model when necessary.

References

Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; and Reyes-Ortiz, J. L. 2012. Human Activity Recognition on Smartphones Using a Multiclass Hardware-Friendly Support Vector Machine. In *IWAAL*, 216–223. Springer.

Bai, L.; Yao, L.; Wang, X.; Kanhere, S. S.; Guo, B.; and Yu, Z. 2020. Adversarial Multi-view Networks for Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4(2): 42:1–42:22.

Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; and Erhan, D. 2016. Domain Separation Networks. In *NIPS*, 343–351.

Buffelli, D.; and Vandin, F. 2020. Attention-Based Deep Learning Framework for Human Activity Recognition with User Adaptation. *CoRR* abs/2006.03820.

Bulling, A.; Blanke, U.; and Schiele, B. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput. Surv.* 46(3): 33:1–33:33.

Chavarriaga, R.; Sagha, H.; Calatroni, A.; Digumarti, S. T.; Tröster, G.; del R. Millán, J.; and Roggen, D. 2013. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters* 34(15): 2033–2042.

Chen, K.; Zhang, D.; Yao, L.; Guo, B.; Yu, Z.; and Liu, Y. 2020. Deep Learning for Sensor-based Human Activity Recognition: Overview, Challenges and Opportunities. *CoRR* abs/2001.07416.

Hammerla, N. Y.; Halloran, S.; and Plötz, T. 2016. Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables. In *IJCAI*, 1533–1540. IJCAI/AAAI Press.

Hammerla, N. Y.; Kirkham, R.; Andras, P.; and Ploetz, T. 2013. On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution. In *ISWC*, 65–68.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR (Poster)*. OpenReview.net.

Ilse, M.; Tomczak, J. M.; Louizos, C.; and Welling, M. 2019. DIVA: Domain Invariant Variational Autoencoder. In *DGS@ICLR*. OpenReview.net.

Janidarmian, M.; Fekr, A. R.; Radecka, K.; and Zilic, Z. 2017. A Comprehensive Analysis on Wearable Acceleration Sensors in Human Activity Recognition. *Sensors* 17(3): 529.

Khan, M. A. A. H.; Roy, N.; and Misra, A. 2018. Scaling Human Activity Recognition via Deep Learning-based Domain Adaptation. In *PerCom*, 1–9. IEEE Computer Society.

Khosla, A.; Zhou, T.; Malisiewicz, T.; Efros, A. A.; and Torralba, A. 2012. Undoing the Damage of Dataset Bias. In *ECCV (1)*, volume 7572 of *Lecture Notes in Computer Science*, 158–171. Springer.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

Klapoetke, N. C.; Murata, Y.; Kim, S. S.; Pulver, S. R.; Birdsey-Benson, A.; Cho, Y. K.; Morimoto, T. K.; Chuong, A. S.; Carpenter, E. J.; Tian, Z.; et al. 2014. Independent optical excitation of distinct neural populations. *Nature methods* 11(3): 338–346.

Lara, O. D.; and Labrador, M. A. 2013. A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys and Tutorials* 15(3): 1192–1209.

Li, D.; Yang, Y.; Song, Y.; and Hospedales, T. M. 2017. Deeper, Broader and Artier Domain Generalization. In *ICCV*, 5543–5551. IEEE Computer Society.

Lin, J.; Keogh, E. J.; Wei, L.; and Lonardi, S. 2007. Experiencing SAX: a novel symbolic representation of time series. *Data Min. Knowl. Discov.* 15(2): 107–144.

Mazankiewicz, A.; Böhm, K.; and Bergés, M. 2020. Incremental Real-Time Personalization in Human Activity Recognition Using Domain Adaptive Batch Normalization. *CoRR* abs/2005.12178.

Micucci, D.; Mobilio, M.; and Napoletano, P. 2016. UniMiB SHAR: a new dataset for human activity recognition using acceleration data from smartphones. *CoRR* abs/1611.07688.

Morales, F. J. O.; and Roggen, D. 2016a. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* 16(1): 115. doi:10.3390/s16010115. URL https://doi.org/10.3390/s16010115.

Morales, F. J. O.; and Roggen, D. 2016b. Deep convolutional feature transfer across mobile activity recognition domains, sensor modalities and locations. In *ISWC*, 92–99. ACM.

Pan, S. J.; and Yang, Q. 2010. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 22(10): 1345–1359. doi:10.1109/TKDE.2009.191. URL https://doi.org/10.1109/TKDE.2009.191.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 8024–8035.

Qian, H.; Pan, S. J.; Da, B.; and Miao, C. 2019. A Novel Distribution-Embedded Neural Network for Sensor-Based Activity Recognition. In *IJCAI*, 5614–5620. ijcai.org.

Qian, H.; Pan, S. J.; and Miao, C. 2018. Sensor-Based Activity Recognition via Learning From Distributions. In *AAAI*.

Rokni, S. A.; Nourollahi, M.; and Ghasemzadeh, H. 2018. Personalized Human Activity Recognition Using Convolutional Neural Networks. In *AAAI*, 8143–8144. AAAI Press.

Wang, J.; Chen, Y.; Hao, S.; Peng, X.; and Hu, L. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters* 119: 3–11.

Wang, J.; Zheng, V. W.; Chen, Y.; and Huang, M. 2018. Deep Transfer Learning for Cross-domain Activity Recognition. In *ICCSE*, 16:1–16:8. ACM.

Wilson, G.; Doppa, J. R.; and Cook, D. J. 2020. Multi-Source Deep Domain Adaptation with Weak Supervision for Time-Series Sensor Data. In Gupta, R.; Liu, Y.; Tang, J.; and Prakash, B. A., eds., *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, 1768–1778. ACM. URL https://dl.acm.org/doi/10.1145/3394486.3403228.

Yang, J.; Nguyen, M. N.; San, P. P.; Li, X.; and Krishnaswamy, S. 2015. Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition. In *IJCAI*, 3995–4001. AAAI Press.