Privacy Issues Regarding the Application of DNNs to Activity-Recognition using Wearables and Its Countermeasures by Use of Adversarial Training

Yusuke Iwasawa
The University of Tokyo, Japan
iwasawa@weblab.t.u-tokyo.ac.jp

Kotaro Nakayama
The University of Tokyo, Japan
nakayama@weblab.t.u-tokyo.ac.jp

Ikuko Eguchi Yairi
Sophia University, Japan
i.e.yairi@sophia.ac.jp

Yutaka Matsuo
The University of Tokyo, Japan
matsuo@weblab.t.u-tokyo.ac.jp

Abstract

Deep neural networks have been successfully applied to activity recognition with wearables in terms of recognition performance. However, the black-box nature of neural networks could lead to privacy concerns. Namely, generally it is hard to expect what neural networks learn from data, and so they possibly learn features that highly discriminate user-information *unintentionally*, which increases the risk of information-disclosure. In this study, we analyzed the features learned by conventional deep neural networks when applied to data of wearables to confirm this phenomenon. Based on the results of our analysis, we propose the use of an adversarial training framework to suppress the risk of sensitive/unintended information disclosure. Our proposed model considers both an adversarial user classifier and a regular activity-classifier during training, which allows the model to learn representations that help the classifier to distinguish the activities but which, at the same time, prevents it from accessing user-discriminative information. This paper provides an empirical validation of the privacy issue and efficacy of the proposed method using three activity recognition tasks based on data of wearables. The empirical validation shows that our proposed method suppresses the concerns without any significant performance degradation, compared to conventional deep nets on all three tasks.

1 Introduction

The recognition of human activities through the use of wearable devices, such as smartphones, is a key technology for enabling next-generation applications; for example, it is possible for a smart phone to track exercise and sleep patterns [Consolvo *et al.*, 2008], to estimate emotional states [Rachuri *et al.*, 2010], and to be developed into a digital assistant [Lee *et al.*, 2012]. In the same way as with image and sound processing, deep neural networks are increasingly used in various activity recognition tasks, exhibiting excellent levels of performance [Plötz *et al.*, 2011; Yang *et al.*, 2015; Hammerla *et al.*, 2015; Ordóñez and Roggen, 2016].

Although recognition performance is obviously a very important aspect of wearable-based activity recognition, the privacy of the data produced by wearables, in that they could possibly contain rich personal information, is also important. From the viewpoint of privacy, the black-box properties of neural networks could lead to privacy issues of their application on data obtained from wearables; i.e., generally it is hard to pre-expect what neural networks learn from data, and as a result of the optimization, they possibly learn features that could accurately estimate user information, such as gender, age, or state of health, *without* any intentional design. This phenomenon potentially causes the privacy issues because such neural networks help to estimate the user-information without the informed consent with users. This issue is also problematic for application providers because it prevents users from using their applications, prevents data sharing with co-companies, and increases management costs.

In this study, we set out to analyze the features learned by conventional deep neural networks when applied to data of wearables, to confirm the relevance of privacy concerns. Specifically, we found that neural networks trained with cross-entropy regarding activity Y learn highly user-discriminative features *without* any intentional design. Based on the results of our analysis, we propose the use of *user-adversarial neural networks* that utilize an adversarial training to prevent the occurrence of this phenomenon. Our proposed model considers an adversarial user classifier in addition to a regular activity classifier during training, which allows the model to learn representations that help the activity classifier to distinguish activities but which, at the same time, prevent it from accessing user-discriminative information. In other words, the proposed model trains feature extractor to deceive virtual-adversarial classifier and explicitly penalize those features to become user-discriminative.

It should be noted that poor design of the privacy-protection significantly affects the utility of data. For example, the output of random mapping $f(X)$ apparently has no unintended/sensitive information, but it also gives no information about activities. Alternatively, as a more realistic example, a neural network with limited capacity might acquire less unintended information, but it may also lead to poorer

performance. This paper empirically validates that our proposed model establishes a better balance between the utility of the data and privacy protection compared to the simple neural networks and the networks with limited capacity.

The primary contributions of this paper are as follows:

- Our paper provides empirical validations of the privacy risk regarding unintended information disclosure for publicly available datasets (Opportunity and USC-HAD) regarding activity recognition with wearables. Specifically, empirical validations show that a simple logistic regressor achieves 0.847 of user-classification accuracy (Opportunity) and 0.657 of accuracy (USC-HAD) from features given by neural networks trained with cross-entropy regarding activity Y only, which is significantly higher than accuracy with raw-data (0.352 and 0.100 respectively).
- Our paper proposes the use of an adversarial training framework to alleviate the above phenomenon. The empirical validation shows the proposed model gives significantly less user-discriminative features (approximately 0.40 points lower than the simple neural networks on average) while maintaining a level of activity classification performance on three activity-recognition tasks.

2 Related Works

2.1 Activity Recognition using Wearables

The primary requirement for an activity recognition system using wearables is to be able to accurately predict y_t from X_t, where X_t is a multichannel time series from the outputs of wearables from time t to time $t + T$, and y_t is an activity performed by the user with the wearable during that time slice. In this context, deep neural networks demonstrate significant performance improvement over classical feature-engineering approaches. Deep neural networks were first applied to human activity recognition with wearables by [Plötz *et al.*, 2011]. The authors demonstrate the superior performance of restricted Boltzmann machines, relative to traditional feature engineering, such as time-domain statistics and frequency coefficients. In the context of supervised representation learning, [Yang *et al.*, 2015] proposed a multi-channel convolutional neural network with a sensor unification layer to make CNN usable with multi-sensor inputs and, more recently, [Hammerla *et al.*, 2016] and [Ordóñez and Roggen, 2016] proposed the use of long- and short-term memory (LSTM) and its variants. As such, deep neural networks are becoming key to achieving high recognition performance.

Besides prediction accuracy, privacy protection is a core requirement for activity-recognition systems. One of the privacy risks peculiar to wearables is *inferring risks*, which is the risk associated with adversarial inference from the data provided by wearables [Raij *et al.*, 2011]. Since the wearables continuously sense the activity of users and the way activity taken differ among users' attributes, we need counter-plans for privacy-protection, or adversary could infer user-information such as age, gender, or possibly levels of health.

The simplest existing counter-plan involves stopping sensors based on user-intention by specifying the potential inference risks [Chakraborty *et al.*, 2014] or based on locations of users [Raghavan *et al.*, 2012]. Although such sensor-level corruption indeed reduces inference risks, it also directly reduces the utility of the data. Another approach involves transforming the representation of data to reduce privacy risks. For example, [Supriyo *et al.*, 2012] proposes an obfuscation strategy that selects features to help classify white-listed information but, at the same time, obstructs the classification of blacklisted information. This feature-level corruption potentially provides a better privacy-utility balance compared to naive sensor-level corruption, and it is similar to our proposal in principle. However, their proposals were limited to a feature-engineering paradigm and were not easily applicable to a representation-learning paradigm.

2.2 Adversarial Training

The central principle of the proposed method is the use of adversarial training for preventing representations from becoming user-discriminative. As far as the authors are aware, the adversarial training framework was first introduced by [Schmidhuber, 1992], and later re-invented by [Goodfellow *et al.*, 2014] in the context of an image-generation task. More recently [Ajakan *et al.*, 2014] introduced domain-adversarial neural networks (DANN), which use a framework for domain adaptation scenarios, while [Edwards and Storkey, 2016] proposed adversarial learned fair representations (ALFR), which are intended to remove sensitive information from representations. Although these studies are closer in principle to the proposed method, our work differs from the previous proposals; while both DANN and AFLR consider the case of the information sources being binary features $S = 1$ or $S = 0$, in our case there are typically many users and so U is categorical rather than binary. This paper first presents empirical validations of the categorical case, and investigates parameter sensitivity regarding the capacity of the adversarial classifier and that of the feature extractor, neither of which have been mentioned in previous studies. Also, this is the very first study that investigates potential privacy issues of application of deep nets on the data of wearables.

3 User-Adversarial Neural Networks

The proposed model *user-adversarial neural networks* designed to learn features that contain information about the activity label Y while, at the same time, *not* contain information about user U [1]. Formally, the networks solve a joint optimization problem of the loss $\mathrm{L} = \mathrm{L}_y(X, Y) + \lambda \mathrm{L}_u(R, U)$, where X indicate the input random variable, R is the random variable of representations, and, L_u and L_y are loss functions that represent how much information about Y and U are held, respectively. The λ is a weighting parameter between L_u and L_y, and the proposed method is equivalent to conventional method when $\lambda = 0$.

The proposed method consists of three feed-forward neural networks: feature extractor $f_f : X \to R$, label classifier

[1] Because of the dataset-limitation, this paper focuses on the removal of user-discriminative information, but a similar strategy could, in principal, be applied to the removal of other user-information. We will discuss this point later in this paper.

$f_y : R \rightarrow Y$, and adversarial user classifier $f_u : R \rightarrow U$ (Figure 1). These networks are parameterized by θ_f, θ_y, and θ_u, respectively. The proposed methods measure L_y and L_u using the networks.

3.1 Quantifying $\mathrm{L}_y(X, Y)$

L_y is the same as in the case of a conventional neural network; generally it is the cross-entropy between the true probability distribution $P(Y)$ and $P(Y|X)$ estimated by neural networks with softmax activations. Formally, L_y is

$$\mathrm{L}_y = -\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M} y_{nm} \log f_y(f_f(x_n)), \tag{1}$$

where N is the number of samples, M is the number of activity classes, and y_{nm} is a binary variable that indicates whether the n-th sample belongs to the m-th class. By minimizing Eq. 1, the feature extractor and label classifier are trained to predict Y from X as correctly as possible.

3.2 Quantifying $\mathrm{L}_u(R, U)$

L_u is measured using adversarial training frameworks. First, assume that adversarial user classifier f_u tries to predict U from R. Specifically, this classifier is a neural network that is trained to maximize the log likelihood

$$\mathrm{L}_u = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} u_{nk} \log f_u(f_f(x_n)), \tag{2}$$

where K is the number of users, and u_{nk} is a binary variable that indicates whether the n-th sample is the k-th user or not. The user classifier is trained to predict U from R. Formally, this is done by updating θ_u to maximize Eq. 2 while fixing θ_f. On the other hand, feature extractor f_f is trained to obstruct user classifier f_u to predict user U from R. Formally, this is done by updating θ_f to minimize Eq. 2 while fixing θ_u. Intuitively, this optimization can be regarded as being a privacy-protection step, which assumes a virtual adversary and improves privacy by updating the feature extractor so that the adversary cannot retrieve any information from R.

3.3 Optimization

Overall, the proposed method consists of three neural networks f_f, f_y, and f_u, and the problem is defined as

$$\min_{\theta_f,\theta_y}\max_{\theta_u}[\mathrm{L}_y(X, Y; \theta_f, \theta_y) + \lambda\,\mathrm{L}_u(R, U; \theta_f, \theta_u)]$$

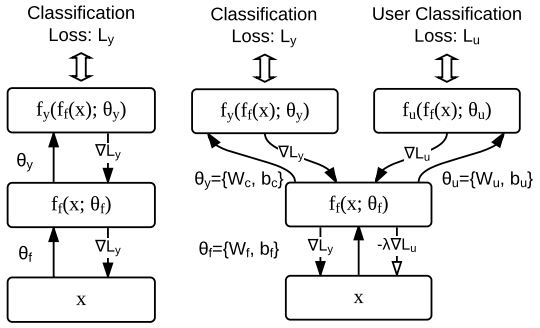
As the overall loss is not convex, single-step optimization is not possible. Rather, we alternatively optimize 1. θ_u to maximize L_u fixing other parameters and 2. θ_f and θ_y to minimize the entire loss fixing θ_u, in the same way as in related works that used adversarial training [Goodfellow *et al.*, 2014; Ajakan *et al.*, 2014; Edwards and Storkey, 2016].

Regarding this alternative optimization, a balance is sought in the min-max game between the adversary (user classifier) and feature extractor regarding L_u. If the adversary is too strong for the feature extractor, the feature extractor can never deceive the user classifier and therefore the features cannot be user-dependent; on the other hand, if the user classifier is too weak, deceiving the user classifier is meaningless. Especially, compared with other applications of adversarial training, the training adversary presents greater difficulties in our case, because adversary is a multi-class classification, not a binary classification as in other applications, thus it may be more difficult to distinguish between multiple users. To overcome this issue, we applied annealing heuristics to weighting parameter λ. the detailed scheduling was as follows:

$$\lambda(t) = \begin{cases} 0.0 & (t \leq \alpha) \\ (t-\alpha)\cdot\frac{\Lambda}{(\beta-\alpha)} & (\alpha < t \leq \beta) \\ \lambda & (\beta < t) \end{cases} \tag{3}$$

where t is epoch, $\alpha + 1$ is the start timing of annealing, β is the stop timing of annealing, and Λ is the target parameter. We found that this annealing sometimes improves the results, but further investigation is necessary.

The overall algorithm, is as shown in Algorithm 1.



(a) Feedforwad NN (b) User Adversarial NN

Figure 1: User-adversarial neural networks.

Algorithm 1 Optimization of the proposed model

Require: dataset $D = \{(x_n, y_n, u_n)\}_{i=1}^{N}$, parameter λ
Ensure: neural network $\{\theta_f, \theta y, \theta u\}$
  t $\leftarrow$ 1
  while training() **do**
    $L_u \leftarrow eq.2\ \forall(x_n, u_n) \in D$
    $\theta_u \leftarrow \theta_u + \frac{\delta L_u}{\delta \theta_u}$
    $L_y, L_u \leftarrow eq.1\ \forall(x_n, y_n) \in D, eq.2\ \forall(x_n, u_n) \in D$
    $\lambda_t \leftarrow eq.3$
    $\theta_f, \theta_y \leftarrow \theta_f - \frac{\delta(L_y + \lambda_t L'_u)}{\delta \theta_y}, \theta_y - \frac{\delta L_y}{\delta \theta_f}$
    t $\leftarrow$ t + 1
  end while

4 Experiments

4.1 Datasets

The Opportunity Recognition dataset [Sagha *et al.*, 2011] is a popular benchmark dataset in the field. The dataset contains data for four subjects (S1–4) and simulates a daily life activity, specifically, a breakfast scenario. A wide variety of body-worn, object-based, and ambient sensors, are used (see Figure 1 in the paper by Sagha et al. [2011] for more details). Each record consists of 113 real-value sensory readings, excluding time information. We used two recognition tasks: the

gesture recognition (Opp-G) and the locomotion recognition (Opp-L). Opp-G requires the recognition of 18 class activities [2], while Opp-L requires the recognition of 4 class locomotions: stand, walk, sit, and lie. Given a sampling frequency of 30 Hz, the sliding window procedure with 30 frames and a 50% overlap produced 57,790 samples.

The USC-HAD dataset [Zhang and Sawchuk, 2012] is a relatively new benchmark dataset, which contains a relatively large number of subjects (14 subjects, 7 males and 7 female). There are 12 activity classes, corresponding to the most basic and common activities in people's daily lives [3]. MotionNode, which is a 6-DOF inertial measurement unit specially designed for human motion sensing applications, is used to record the outputs from the accelerometers that record 6 real sensory values. The sliding window procedure, using 30 frames and a 50% overlap, produced 172,169 samples.

4.2 Experimental Setting

Figure 2 shows the network architectures used for the evaluations. With the exception of the adversarial user classifier connected to the full connected layer, the network architecture is same as that of the simple CNN. The architecture of the CNN was the same as that used in a previous study [Yang *et al.*, 2015] in which CNN was applied to the Opp-G task. The lower-right number of each layer represents the #filter of the convolution layers. Each convolution and fully connected layer was followed by ReLU activations and Dropout, and the output layer was followed by softmax. Every component of the network was trained with Adam algorithms [Kingma and Ba, 2015], using the default parameters for every optimization (150 epochs). We used a logistic regressor as the user classifier. The hyper parameter λ was set to 0.1, and the annealing parameters α and β were set to 15 and 135 so that λ will be fixed to the first and last 15 epochs.

We compared the models by using the leave-one-subject-out (LOSO) procedure. The procedure yields a pair of test users and $K-1$ training users K times. At each iteration, the classifier was trained using the data for $K-1$ users, and was tested using the one remaining user. All values reported in the evaluation section are the average of the K iterations.

As an evaluation metric, we used the activity classification accuracy for investigating the utility, and the user classification accuracy for investigating levels of the privacy-thread. Namely, we regarded a privacy thread as being high if we could build an accurate classifier over features. Otherwise, the privacy thread was low. The evaluation of user-classification is done by a new classifier γ that predicts U from the features, instead of the adversarial classifier f_u used in the training phase. We chose to build a new classifier to eliminate the possibility of underestimating the classification accuracy in the event of the failure of the optimization of f_u. Although the choice of classifier for γ apparently affects the accuracy, we used logistic regressor for γ, which is same as the adversarial classifier we used in the training phases. We later discuss how the choice of classifier for γ and f_u affects the results, by performing empirical evaluations.

[2] open door 1, open door 2, close door 1, close door 2, open fridge, close fridge, open dishwasher, close dishwasher, open drawer 1, close drawer 1, open drawer 2, close drawer 2, open drawer 3, close drawer 3, clean table, drink from cup, toggle switch, and Null

[3] walking forward, walking left, walking right, walking upstairs, walking downstairs, running forward, jumping, sitting, standing, sleeping, elevator up, elevator down

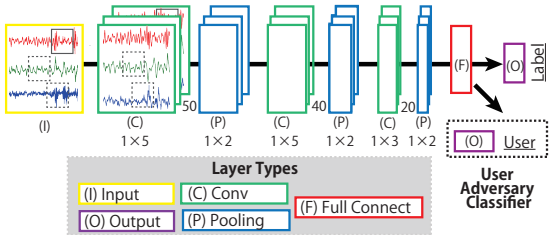


Figure 2: Network architecture used for the evaluations.

In addition to the quantitative evaluation, we also provide qualitative evaluations with clustering to show how the proposed method remove information about users. In particular, we cluster the representations R, and visualizing user-tendency matrix C, that is formed by the cross-tabulation of the cluster indexes and user indexes of each sample. Here, element c_{ij} of C indicates the ratio of the i-th user to the j-th cluster, and can takes a value from 0.0 to 1.0. As the value of c_{ij} increases, it indicates that the representation is highly user-discriminative, which implies a greater privacy risk.

4.3 Experimental Results

Table 1 shows the user-discriminability of raw-data, popular heuristic features (MV [Yang *et al.*, 2015] and ECDF [Hammerla *et al.*, 2013]), the features learned by CNN [Yang *et al.*, 2015], and the features learned by the proposed method (CNN+Adv). Each value indicates the accuracy of user classifier γ trained for predicting U from each feature, which means that a smaller value is a better value. As can be inferred from the results, the features learned by CNN become highly user-discriminative when compared with raw-data or heuristic features on all three recognition tasks. On average, the CNN-features achieve 0.798 accuracy, which is approximately 0.53 point higher than the result of raw-data (0.268). Note that the CNN is trained such that only the cross-entropy regarding Y is minimized, and it is not intended to be user-discriminative. The result also shows that the proposed method gives significantly lower classification accuracy relative to heuristic features and CNN, although it is still approximately 0.15 point higher than raw-data on average.

Figure 3 shows detailed comparison of the proposed method and CNN. The left column of Fig. 3 shows the relative performance based on $\lambda = 0.0$, which is correspond to CNN, for varied $\lambda = \{0.01, 0.05, 0.1, 0.2, 0.5, 1.0\}$. Each color denotes different metric (user accuracy, activity ac-

Table 1: The comparison of user-discriminability between different features.

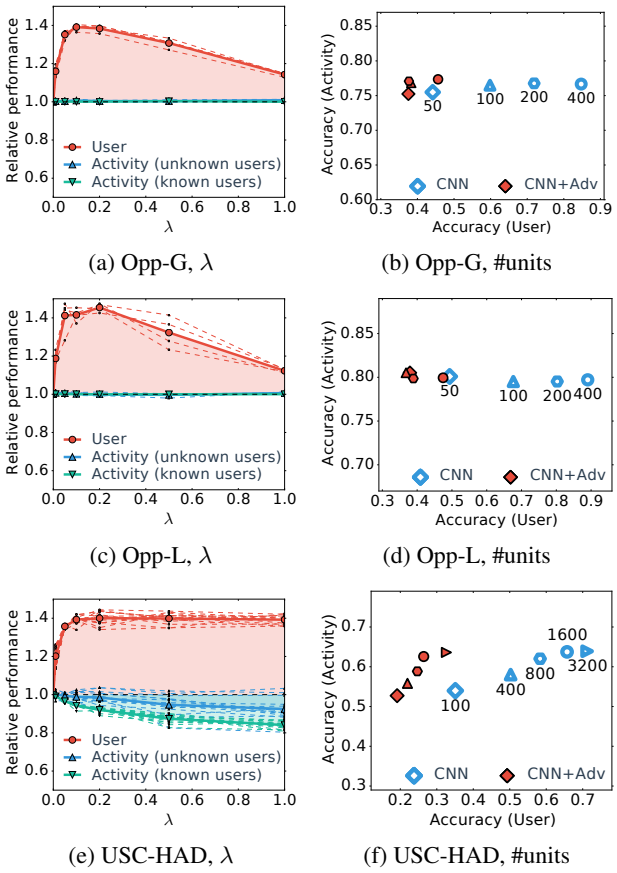
	Opp-G	Opp-L	USC	Ave.
Rawdata	0.352	0.352	0.100	0.268
MV	0.849	0.849	0.140	0.612
ECDF	0.914	0.914	0.204	0.678
CNN	0.847	0.890	0.657	0.798
CNN+Adv	0.456	0.475	0.264	0.398

curacy on unknown users, and activity accuracy on known users). The bold line indicates the average of LOSO iterations, and the dashed line indicates each iteration. We make the following observations. (1) $\lambda > 0$ tend to obstruct user classification, though too big λ relatively harms the performances. (2) With reasonable selection of λ, the performance drop on both activity recognition performance of unknown and known users is small or moreover, almost zero for Opp-L and Opp-G. (3) The parameter sensitivity curve is almost the same regardless of the iteration of the leave-one-subject-out iterations (the subset of users for training).

The balance between the user classification accuracy and label classification accuracy of the proposed method (CNN+Adv, red marker) and the baseline (CNN, blue marker) is illustrated in the right column of Fig. 3. The horizontal axis corresponds to the user classification accuracy while the vertical axis is the label classification accuracy, i.e., upper left side indicates that the model provides a better balance between user-discriminability and label classification performances. We show the results for a different-capacity feature extractor (specifically, a different unit size of the fully connected layer), which corresponds to each marker type in the figure. The unit size of each marker is indicated below (or above) each marker. The results show that, for any unit size and datasets, the proposed method indeed learns less user-discriminative features without significant performance drop on activity recognition performances compared to CNN with the same unit size. Moreover, the adversarial regularization provides better balance compared to CNN with smaller unit size. For example, on the result on USC dataset, the CNN+ADV with 1600 units provides better performance on both user classification accuracy and activity classification accuracy compared to CNN with 100 units, and similar conclusion could be made on the other datasets.

Figure 4 visualizes the user tendency matrix C. The depth of the blue increases with the user-tendency. The clusters are arranged in descending order of user-tendency. The visualization again indicates that the proposed method learns fewer user-discriminative features than CNN. For example, CNN produces a cluster that is almost occupied by a thrid user, but CNN+Adv did not give rise to such a cluster.

Table 2 lists the user classification accuracy when different adversarial user classifiers f_u and user classifiers γ are used for evaluations. We tested None (without adversarial training), logistic repressor (LR), multi-layer perceptron with 800 hidden units, and deep neural networks with 400-200 hidden units for f_f, and LR, multi-layer perceptron with 50 hidden units (MLP_{50}), and multi-layer perceptron with 800 hidden units (MLP_{800}) for γ. The table also includes the theoretical values for random guessing (Rand) as reference values. Rand is the logical limit as any classifier results in random guessing if the feature becomes perfectly user-independent. We can make the following observations. (1) With any value of γ, comparing the use of adversarial regularization with its non-use, the user identification performance was found to be relatively low when using adversarial regularization. (2) However, with any adversarial classifiers and γ, the proposed model results in a significantly better recognition performance, relative to random guessing. (3) The best adversarial user classifier depends on the dataset and γ, although MLP and DNN exhibit a relatively lower discrimination.

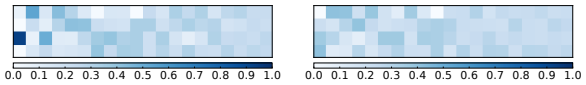


(a) Opp-G, λ (b) Opp-G, #units

(c) Opp-L, λ (d) Opp-L, #units

(e) USC-HAD, λ (f) USC-HAD, #units

Figure 3: Performance comparison of the proposed method with different parameters. Left: The results of various hyper parameter $\lambda = \{0.0, 0.01, 0.05, 0.1, 0.2, 0.5, 1.0\}$. Note that $\lambda = 0.0$ correspond to the normal CNN. Right: Balance between user-classification accuracy and label-classification accuracy for different unit sizes of fully connected layer.



(a) CNN (Opp-G) (b) CNN+Adv (Opp-G)

Figure 4: Visualization of user tendency of each cluster.

5 Discussion

This paper provides empirical validations of (1) the risks related to unintended information disclosure when applying deep nets to wearable-sensor data that contains rich user information, and (2) the efficacy of the proposed method for preventing the risk. Table 1 shows that the simple CNN learns highly user-discriminative features, despite not being trained to become user-discriminative. This phenomenon supports our assumption that the application of deep neural networks to data of wearables could lead to the appearance of new types of privacy issues, *unintended information disclosure*. The results also indicate that the installation of a trained model on

Table 2: Comparison of user classification accuracy between different user classifiers for training and evaluation.

(a) Opp-G

	LR	MLP_{50}	MLP_{800}
None	0.857	0.971	0.998
LR	0.463	0.960	0.997
MLP	0.464	0.726	0.972
DNN	0.438	0.678	0.944
Rand	0.333	0.333	0.333

(b) Opp-L

	LR	MLP_{50}	MLP_{800}
None	0.895	0.968	0.998
LR	0.477	0.951	0.997
MLP	0.499	0.757	0.966
DNN	0.502	0.679	0.913
Rand	0.333	0.333	0.333

(c) USC-HAD

	LR	MLP_{50}	MLP_{800}
None	0.669	0.767	0.892
LR	0.266	0.516	0.731
MLP	0.249	0.334	0.612
DNN	0.313	0.424	0.695
Rand	0.077	0.077	0.077

the client side (such as a smartphone) could prove problematic, since the use of neural networks greatly increases the classification accuracy of user U. We need to determine why this phenomenon occurs, but we do not think that we can safely draw a clear conclusion based only on the results of this study. However, one possible explanation is that neural networks tend to learn the combination of *user-specific* features rather than *user-general* features, because the way in which activities are performed differs to a greater or lesser degree between the users. Therefore, the learning of a combination of user-specific features can be done at less cost than that incurred when learning user-general features, especially if the neural networks have a rich capacity. If the assumption is correct, it is also probable that neural networks learn features that clearly discriminate user-information such as ages, weights, or other physical characteristics, state of health, all of which can affect the way in which activities are performed. Further investigations with a dataset that provides information about user-attributes, and a deep analysis of how and why the phenomenon occurs, is required to clarify the privacy issues associated with the application of deep nets.

To prevent the occurrence of the above phenomenon, this paper has proposed the use of adversarial-training, and has provided empirical validations that the proposed method indeed reduces the possibility of said occurrence. Specifically, Table 1 shows that the proposed method significantly reduces the possibility of learning representations that incorporate user-discriminative information. Also, the left column of Fig. 3 shows that the proposed method prevents it with only small (USC) or even almost zero (Opp-G, Opp-L) performance drops in the label-classification accuracy (that is, the utility of the data), with reasonable hyper-parameter selection. Although the poor selection of λ possibly harm either the of user-discriminability or classification accuracy of activities, the parameter-sensitivity curve is almost the same for all users and therefore easy to optimize in practice. Moreover, the right column of Fig. 3 shows that adversarial training provides better balance between user-discriminability and the utility of features compared to the naive limitation of the capacity of the neural networks. All these results point to the suitability of the proposed method for balancing the prevention of unintended information disclosure with the level of performance of activity-recognition.

A limitation of this study is that it has only considered the user-discriminability of features to illustrate the efficacy of the proposed countermeasure, as well as the existence of privacy issues arising from the black-box property of neural networks. However, as mentioned in Section 3, the proposed method can, nevertheless, be used to remove more sensitive information about users. The only requirement needed to realize this method is the ability to construct classifiers of the target information, which means being able to prepare supervised data about the information. Also, the efficacy of the proposed method could be limited by the amount of information naturally possessed by a neural network. That is, the proposed method would not be necessary if the original CNN did not learn the features that discriminate the information, although our proposal would still help in preventing discrimination explicitly.

Another limitation of this paper is, as shown in Table 1 and Table 2, that the proposed method still produces those features that a user classifier could use to discriminate between, compared to completely random guessing or classification from raw-data. In other words, although the results show the superior performance of the proposed method relative to a conventional neural network, the learned features could still incorporate unintended information. Table 2 also shows that user-discriminability is highly affected, in terms of training and evaluation, by the choice of adversarial-user classifiers. In addition, there is the possibility that an adversary could retrieve information more accurately with the application of a better strategy, such as using the data from multiple windows to assemble user-classification results. In the future, we intend to improve the proposed method to attain better privacy protection, while considering the presence of a stronger adversary in the training phase instead of the naive adversary considered in this study.

6 Conclusion

Although deep nets currently offer a significant improvement in the performance of activity recognition based on data acquired from wearables, the black-box nature of such neural networks could lead to privacy concerns with the application of deep nets to data from wearables that could contain user-specific information. This study set out to analyze the features learned by conventional deep neural networks and confirm that the neural networks tend to learn user-discriminative information, despite this not being the intention of the network designer, further implying the relevance of the privacy concerns. Based on the results of our analysis, we propose the use of adversarial training to suppress these concerns, and verified the efficacy of the method by applying it to publicly available datasets related to activity-recognition using data gathered from wearables. The above two empirical validations imply the occurrence of privacy concerns, illustrate the efficacy of the proposed method, and give insights into better privacy protection.

References

[Ajakan *et al.*, 2014] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. In *NIPS Workshop on Transfer and Multi-Task Learning: Theory meets Practice*, 2014.

[Chakraborty *et al.*, 2014] Supriyo Chakraborty, Chenguang Shen, Kasturi Rangan Raghavan, Yasser Shoukry, Matt Millar, and Mani Srivastava. ipShield: A Framework for Enforcing Context-Aware Privacy. In *Workshop on Secure Data Management*, pages 85–100, 2014.

[Consolvo *et al.*, 2008] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. Activity sensing in the wild: a field trial of ubifit garden. In *Proc. of CHI*, pages 1797–1806. ACM, 2008.

[Edwards and Storkey, 2016] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *Proc. of ICLR*, 2016.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. of NIPS*, pages 2672–2680, 2014.

[Hammerla *et al.*, 2013] Nils Y Hammerla, Reuben Kirkham, Peter Andras, and Thomas Ploetz. On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution. In *Proc. of International Symposium on Wearable Computers (ISWC)*, pages 65–68. ACM, 2013.

[Hammerla *et al.*, 2015] Nils Yannick Hammerla, James Fisher, Peter Andras, Lynn Rochester, Richard Walker, and Thomas Plötz. PD Disease State Assessment in Naturalistic Environments Using Deep Learning. In *Proc. of AAAI*, pages 1742–1748. AAAI Press, 2015.

[Hammerla *et al.*, 2016] Nils Y Hammerla, Shane Halloran, and Thomas Ploetz. Deep, convolutional, and recurrent models for human activity recognition using wearables. In *Proc. of IJCAI*, pages 1533–1540. Citeseer, 2016.

[Kingma and Ba, 2015] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of ICLR*, 2015.

[Lee *et al.*, 2012] Youngki Lee, SS Iyengar, Chulhong Min, Younghyun Ju, Seungwoo Kang, Taiwoo Park, Jinwon Lee, Yunseok Rhee, and Junehwa Song. Mobicon: a mobile context-monitoring platform. *Communications of the ACM*, 55(3):54–65, 2012.

[Ordóñez and Roggen, 2016] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[Plötz *et al.*, 2011] Thomas Plötz, Nils Y Hammerla, and Patrick Olivier. Feature learning for activity recognition in ubiquitous computing. In *Proc. of IJCAI*, number 1, pages 1729–1734, 2011.

[Rachuri *et al.*, 2010] Kiran K Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J Rentfrow, Chris Longworth, and Andrius Aucinas. Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In *Proc. of UbiComp*, pages 281–290. ACM, 2010.

[Raghavan *et al.*, 2012] Kasturi Rangan Raghavan, Supriyo Chakraborty, Mani Srivastava, and Harris Teague. OVERRIDE: A Mobile Privacy Framework for Context-Driven Perturbation and Synthesis of Sensor Data Streams. In *International Workshop on Sensing Applications on Mobile Phones - PhoneSense*, pages 1–5, New York, New York, USA, 2012. ACM Press.

[Raij *et al.*, 2011] Andrew Raij, Animikh Ghosh, Santosh Kumar, and Mani Srivastava. Privacy Risks Emerging From the Adoption of Innocuous Wearable Sensors in the Mobile Environment. In *Annual conference on Human factors in computing systems - CHI '11*, page 11, New York, New York, USA, 2011. ACM Press.

[Sagha *et al.*, 2011] Hesam Sagha, Sundara Tejaswi Digumarti, José del R Millán, Ricardo Chavarriaga, Alberto Calatroni, Daniel Roggen, and Gerhard Tröster. Benchmarking classification techniques using the opportunity human activity dataset. In *Proc. of INSS*, pages 36–40. IEEE, 2011.

[Schmidhuber, 1992] Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992.

[Supriyo *et al.*, 2012] Chakraborty Supriyo, Raghavan Kasturi Rangan, Srivastava Mani B., and Bisdikian Catschik. Balancing Value and Risk in Information Sharing Through Obfuscation. In *International Conference on Information Fusion (FUSION)*, page 2624. IEEE, 2012.

[Yang *et al.*, 2015] Jian Bo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Proc. of IJCAI*, pages 3995–4001, 2015.

[Zhang and Sawchuk, 2012] Mi Zhang and Alexander A Sawchuk. Usc-had: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proc. of UbiComp*, pages 1036–1043. ACM, 2012.