Are you asleep when your phone is asleep? Semi-supervised methods to infer sleep from smart devices

Anonymous Author(s)
Affiliation
Address
email

Abstract

Sleep is a vital aspect of our life. Having a good quality sleep is necessary for our well-being and health. Therefore, sleep measurements can aid us in improving our sleep quality. While many users are reluctant to use intrusive sleep sensing techniques such as wearables, passive sensing such as network activity of smart phone devices can be utilized to measure the sleep duration of a user. However, to develop accurate sleep prediction models, we need large amounts of labeled data. In addition, due to heterogeneity in user behaviors, hardware and software of the devices used, a single model may not generalise to every user in a given population. Although ground truth data collection from a large population is costly and challenging, unlabelled network activity data is easy to gather using mobile applications or network logs. This motivates us to look for semi-supervised learning approaches to leverage unlabelled data from the users to develop accurate sleep prediction models. Our results show that semi-supervised learning techniques can be used to improve the accuracy of sleep duration estimation from smart devices.

1 Introduction

Despite the importance of sleep for human wellbeing, over a third of the human population sleeps less than seven hours per day and sleep disorders are commonplace among the adults (CDC, et al. 2009). Sleep monitoring via wearables has emerged as an approach for improving sleep health via daily tracking of sleep patterns. However, many users tend to be reluctant to wear trackers when sleeping, which has led to the emergence of contactless sensing methods to monitor sleep (Tauhidur Rahman, et al, 2015). One promising approach is to use smartphone activity as a proxy for users sleep and awake state, where long periods of phone inactivity, particularly in the night, are used to infer the user's nocturnal sleep durations. Recent research has shown the feasibility of using time series data of phone network WiFi activity (e.g. event rates) to infer bed time and wake-up times.

When good quality labelled ground truth data is available, such techniques can even approach the accuracy of wearable sleep trackers such as oura ring (Ghorbani S et al, 2022). However such approaches suffer from many challenges due to differences in data. Such differences arises due to user heterogeneity, device heterogeneity and study heterogeneity (Blunck et al, 2016). User heterogeneity refers to the differences in user's device usage practices and differences in sleeping patterns. Device heterogeneity refers to the differences in the hardware and software version of the device. Study heterogeneity arises due to the fact that users can change their devices over the time or new users can be added to the later stages of the study. Consequently, current approaches have several limitations. First of all, a supervised model trained on the labelled data from user studies will not be able to generalize to new users, which limits the scalability of the technique. Second, obtaining high quality

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

ground truth data from large number of users imposes high overheads and high cost for many user studies.

At the same time, unlabelled data is easy to gather since mobile app or standard network logs can be used to gather data of a phone's WiFi activity. This motivates the need for semi-supervised learning (SSL) approaches that can combine a small amount of labelled ground truth data with large amounts of unlabelled data to improve the efficacy of ML models to predict sleep using phone activity and thereby personalize the models using a user's phone's unlabelled data.

In this paper, we mitigate the limitation of current supervised methods by providing interesting ideas on developing semi-supervised methods for sleep duration estimation.

2 Motivation

2.1 Problem Statement

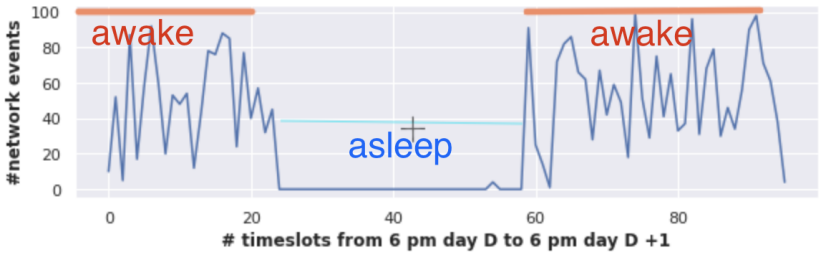

Figure 1: Time series data representing the network events from a smartphone, where the low event rate represents sleep.

We state the problem of sleep duration estimation as follows: For each user, we are given a time series of network activity (event rates) from a smartphone (refer Fig. 1). We split a 24 hour period starting from 6 p.m of day d to 6 p.m of day $d+1$ into 15 minute windows and classify each window into one of the two activity lables - "sleep" or "awake". We then find the longest continuous streak of sleep to calculate the bed time and wake-up time for that user. Although a user may have multiple sleep periods in a day, we are considering only the sleeping period with the longest duration to calculate the bed time and wake up time.

We are also considering the problem of generalizing our model for users with only unlabelled data. For a population of size N, we have labelled data D for n users and unlabelled data D' for $N-n$ users where $N-n \gg n$. Our goal is to train models in a generalized and personalized way that can leverage both D and D'.

2.2 Dataset

We collected ground truth data from 20 undergraduate students who are on-campus residents under an IRB approved study. The data-collection for each participant lasted for a duration of 4 weeks. The participants were given Fitbit devices and manuals for logging the sleep data. We gathered logs of network activity from their smartphones. More details about the study is added in the Appendix

Ethical Considerations: This work's data collection and analysis were conducted under safeguards and restrictions approved by our Institutional Review Board (IRB) and Data Usage Agreement (DUA) with the campus network IT group. All device MAC addresses and authentication information are anonymized using a strong hashing algorithm. User identities were blinded by assigning numeric identifiers. Ground truth was collected within the IRB approved protocol. It is important to note that our population-scale analysis was performed on aggregate data of anonymous users. Individual analyses were performed on users who had consented to this study.

2.3 Supervised Learning

Below we describe the baseline architecture and the features that we used to perform supervised learning on labelled data.

Baseline Architecture: We use a simple Convolutional Neural Network as our baseline model. Our model has three temporal 1D convolutional layers with 32, 64 and 96 filters. Each of these layers has a kernel size of 24, 16, and 8 respectively with uniform a stride 1. Sigmoid activation function is used with L2 regularization with a factor of 0.0001 for the weights. A global 1D maximum pooling layer is connected to the last convolutional layer.

Features Used: We derived features from the WiFi syslog data of the devices. The features include Unique Access Points (AP) visited, Access Point Transitions, Location Type, Unique Buildings visited, Building Transitions and Time. The rationale for choosing these features is, when a user is active or awake, the value of these features is expected to be different compared to when a user is inactive or asleep.

Efficacy of the Model: We picked 10 users and did 80:20 splits of the labelled data for training and testing. The results are shown in Table 1 and we can see that the model is able to predict the bed time and wake up time with an average error of less than 30 mins.

Technique	T_{sleep} mins	T_{wake} mins
Supervised Learning	30 $\pm$7	19 $\pm$3

Table 1: Bed time and wake up time estimation errors

Inability to generalize: Next, we study the ability of the SL models to generalize. Data collection from various sources is an important step in mobile sensing. However, data collection and analysis is often challenged by the the differences in training data. In our experiment, we try to gradually increase the size of the test data by adding more users. We can see from Table. 2 that the models accuracy decreases as we bring in more variability in the data by adding new users. This shows that the model is not scalable.

no of users	T_{sleep} mins	T_{wake} mins
1	55 $\pm$10	52 $\pm$5
3	60$\pm$12	62 $\pm$11
5	62 $\pm$10	61 $\pm$11
7	67$\pm$5	61 $\pm$9
9	69$\pm$5	65 $\pm$7

Table 2: Bed time and wake up time estimation errors for test data. We can see that the accuracy of the model reduces as we increase the number of test users.

This motivates us to look for scalable and generalizable techniques that does not involve additional labelled data collection, leading us to the semi-supervised techniques.

3 Approach

In this section, we experiment with a state-of-the-art SSL technique to compare against the supervised learning model (baseline model) on sleep prediction.

3.1 Multi-head Single-view Co-training

In single-view co-training, we train multiple classifiers without splitting the features unlike multi-view co-training. In this paper we try to overcome the overhead of training multiple classifiers. We do this by using multiple heads in a shared module instead of training multiple classifiers separately. Thus in multi-head co-training, there will be a shared parametric module and multiple classification heads with identical structure. We use a similar architecture mentioned in (Chen, et al, 2022), however the training pipeline is different. Instead of augmenting the dataset using data perturbations, we use the unlabelled data from the new users or unseen users in a population. At first, we train the model with only one classification head with labelled data. We then use this model to generate the pseudo labels

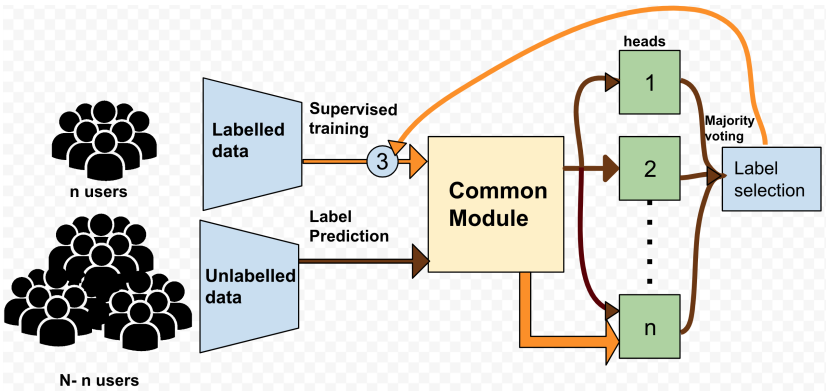

Figure 2: Schematic diagram of the proposed co-training pipeline. Here we first train the model using the labelled data with one classification head. Then we label the unlabelled data taking a majority vote from all the classification heads creating pseudo labels. As a third step, we again train the model with again using the mixed dataset

for the unlabelled data using a majority voting from all the classification heads. Each 15 min window is classified as "sleep" or "awake" if the majority is above a certain threshold. However, since we are interested in getting bed time and wake-up time of an entire day, knowing individual window's classification is not sufficient. We evaluate if the classification of the bins over a 24 hour period is good enough to mix with labelled data. We take the average majority scores of all the bins for 24 hour period and check if this average is above a certain threshold, we consider that days data good enough to be mixed with labelled data. After mixing the datasets, we again train the model using only one classification head using the new augmented dataset.

3.2 Performance Evaluation

Experiments were run on Google colaboratory notebooks. We split the labelled dataset into $50:50$ splits, with training data set and test data set consisting of data from 10 users each. We also took one week of unlabelled data from users in the test dataset. The performance metrics we use here are sleep time and wake up time estimation errors which are defined as the difference between the estimated value and the true value. From Table 3, we can see that the Multihead single-view co-training is able to perform much better in comparison to the supervised learning without labelled data from test set users.

Technique	Labelled data from test set	T_{sleep} mins	T_{wake} mins
Supervised Learning	No	75 $\pm$7	70 $\pm$9
Multi-head Single-view Co-training	No	37$\pm$11	25 $\pm$10
Supervised Learning	Yes	30 $\pm$7	19 $\pm$3

Table 3: Bed time and wake up time estimation errors for test data set

4 Conclusions and Future Directions

Sleep sensing can potentially help us improve our sleep quality. As most of the users are reluctant to use intrusive sleep sensing techniques, network activity of the smartphones can be utilized to predict our sleep. However, due to limited labelled data and ground truth data collection overheads in a larger population, we have to rely on semi supervised learning techniques. Our experiments show that SSL techniques are viable options to improve the accuracy of sleep prediction models when there is limited labelled data. The results we presented are preliminary and as a future direction, we plan to extend the work by experimenting with different model configurations and other SSL techniques to achieve closer accuracy with supervised learning with labelled data.

References

[1] Blunck, Henrik, et al. "Activity recognition on smart devices: Dealing with diversity in the wild." GetMobile: Mobile Computing and Communications 20.1 (2016): 34-38.

[2] Chen, Mingcai, et al. "Semi-supervised learning with multi-head co-training." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 6. 2022.

[3] Centers for Disease Control, Prevention (CDC, et al. 2009. Perceived insufficient rest or sleep among adults-United States, 2008. MMWR. Morbidity and mortality weekly report 58, 42 (2009), 1175.

[4] Ghorbani S, Golkashani HA, Chee NIYN, Teo TB, Dicom AR, Yilmaz G, Leong RLF, Ong JL, Chee MWL. Multi-Night at-Home Evaluation of Improved Sleep Detection and Classification with a Memory-Enhanced Consumer Sleep Tracker. Nat Sci Sleep. 2022 Apr 14;14:645-660. doi: 10.2147/NSS.S359789. PMID: 35444483; PMCID: PMC9015046.

[5] Tauhidur Rahman, Alexander T Adams, Ruth Vinisha Ravichandran, Mi Zhang, Shwetak N Patel, Julie A Kientz, and Tanzeem Choudhury. 2015. Dopplesleep: A contactless unobtrusive sleep sensing system using short-range doppler radar. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 39–50.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We will make it available with the camera ready version.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We have included the mean and standard deviation
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] We do not use existing datasets
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]

A Appendix

A.1 User Study Procedure

We advertised about the study through flyers and using undergraduate mailing groups. We recruited 20 undergraduate students who are on-campus residents. The data-collection for each participant lasted for a duration of 4 weeks, while the overall recruiting process ran from October to mid-December. The participants were given Fitbit devices and manuals for logging the sleep data. We precisely identified the participants' hashed MAC addresses by monitoring their WiFi events from a dedicated AP on campus. We made sure the following guidelines were followed through out the study by sending periodic reminders to the participants :

- The students were asked to wear a fitness tracker (like a Fitbit or similar) for the duration of the study. They were advised to wear the device as much as possible, but it is critical that they wore it at night. We used the sleep-tracking function of the fitness tracker to help us understand the relationship between sleep and network activity. It is critical to keep the device charged and paired to the phone at night. Putting the device on and taking it off and charging as needed should take less than 10 minutes per day.
- They were asked to keep a sleep journal. This is used to verify the sleep data we get from the wearable, and vice versa. This should take 30 seconds or less per day.

In addition to the above data collection activities, the students made two visits to the lab:

- Participants attended a 30-minute orientation session at the beginning of the study. During the orientation session they are given a wearable fitness tracker along with instructions on how to use and charge it. The study's smartphone app was installed on your Android phone during the session, and they were instructed on how to use it.
- They came for a 30-minute debriefing session at the end of the study during which they turn over any study equipment they still own and deleted the study app from their phone. During this session they also received compensation for their participation in the study. The students received compensation in the form of Fitbits or Amazon coupons worth 100 USDs if they wore Fitbit every day for at least 4 weeks while on campus.