



Published in final edited form as:

Int Conf Distrib Comput Sens Syst Workshops. 2020 May ; 2020: 35–42. doi:10.1109/dcoss49796.2020.00019.

Wi-Fringe: Leveraging Text Semantics in WiFi CSI-Based Device-Free Named Gesture Recognition

Md Tamzeed Islam,

Department of Computer Science UNC at Chapel Hill

Shahriar Nirjon

Department of Computer Science UNC at Chapel Hill

Abstract

The lack of adequate training data is one of the major hurdles in WiFi-based activity recognition systems. In this paper, we propose Wi-Fringe, which is a WiFi CSI-based device-free human gesture recognition system that recognizes *named* gestures, i.e., activities and gestures that have a semantically meaningful name in English language, as opposed to arbitrary free-form gestures. Given a list of activities (only their names in English text), along with zero or more training examples (WiFi CSI values) per activity, Wi-Fringe is able to detect all activities at runtime. We show for the first time that by utilizing the state-of-the-art semantic representation of English words, which is learned from datasets like the Wikipedia (e.g., Google’s word-to-vector [1]) and verb attributes learned from how a word is defined (e.g, American Heritage Dictionary), we can enhance the capability of WiFi-based named gesture recognition systems that lack adequate training examples per class. We propose a novel cross-domain knowledge transfer algorithm between radio frequency (RF) and text to lessen the burden on developers and end-users from the tedious task of data collection for all possible activities. To evaluate Wi-Fringe, we collect data from four volunteers in a multi-person apartment and an office building for a total of 20 activities. We empirically quantify the trade-off between the accuracy and the number of unseen activities.

I. INTRODUCTION

The ubiquity of WiFi in indoor spaces and the availability of signal characteristics such as the channel state information (CSI) in commodity WiFi chipsets make WiFi an attractive technology for human activity monitoring. Alternate solutions that use wearables such as smartwatches and activity trackers are less effective due to their usage adherence issues, and systems that use cameras raise serious privacy concerns. WiFi sensing, on the other hand, is device-free, non-intrusive, and less privacy-invasive. Hence, we see an increase in WiFi-based sensing and inference systems whose feasibility has been demonstrated in applications such as home activity monitoring [2], sleep monitoring [3], controlling devices using gestures [4], [5], and tracking health vitals [6].

WiFi-based activity recognition systems employ either template matching algorithms or machine learning classifiers such as traditional support vector machines [7] as well as advanced deep convolutional neural networks [8]. These algorithms require a decent number of training examples for each class of activity in order for the system to accurately classify them. Furthermore, the capability of these systems are fundamentally limited by the number of activity classes for which the system has been trained for. When these systems are presented with a completely new type of activity, there is no built-in mechanism to make an educated guess about the possible class label for that unseen example.

Figure 1 illustrates this scenario. When a system is trained to recognize only {walk, drink}, but is presented with an example of an unseen activity, e.g., run, it is likely to detect the activity as either walk (based on the closest match) or it will determine that it is an unknown category (based on a distance threshold). At present, existing systems have no inherent mechanism to infer that the activity could be run, as these systems have no prior knowledge of how an activity called run might be. These systems require user labeled samples of run and retrain the model to recognize it.

In this paper, we propose the first system, called the *Wi-Fringe*, which can infer activities from WiFi data without requiring prior training examples for all of its activity classes. The principle behind Wi-Fringe is popularly known as the *zero-shot learning* [9], [10], which is an active research topic in computer vision and acoustics [11]. These techniques, however, are not directly applicable to RF-based gesture recognition problems, since gestures require tracking sequential properties of the signal and external knowledge about the attributes that defines an activity.

To the best of our knowledge, we are the first to apply zero-shot learning in RF-based device-free activity recognition problem, where the core idea is to exploit information or learned knowledge from other sources such as textual descriptions, rules, and logic. For example, to teach the concept of run to a system that has already learned to recognize walk, instead of training it with many examples of run, we can add a rule into the system, e.g., “run is just like walk but it’s 3 to 5 times faster.” At runtime, the system will use this additional information to classify a run activity correctly.

In Wi-Fringe, to embed such rules between *seen* classes (i.e., explicitly trained) and *unseen* classes (i.e., not explicitly trained) in an RF sensing system, we exploit attributes and context-aware representation of English words as the additional source of knowledge. Through a RF-domain to textdomain projection algorithm, we blur the difference between an activity’s RF signature and its corresponding English word/phrase by representing them in the same vector spaces, i.e., *word embedding* [12] and *word attribute* [13] spaces. The intuition behind Wi-Fringe is that the WiFi signature of an activity correlates with the corresponding verb’s semantic and attribute information. Like two similar activities perturb the WiFi signals similarly, when we describe these two activities in English sentences, we see a similar likeness between the sentences. We generalize this notion for an arbitrary number of activities represented in both RF and text domains, and find a projection between the two representations from the RF domain to the text domain. By learning this projection,

we gain the ability to find the corresponding English word from the RF representation of any arbitrary activity.

Projecting RF signals onto the space of textual representation is non-trivial and poses several challenges that are addressed in this paper. First, we propose context-aware RF features by explicitly learning the transition of *states* (i.e., micro-activities) in an activity. We show that such a representation is robust and yields better features for activity recognition in general. Second, we propose a neural network architecture to merge text- and RF-domain representations of activities so that WiFi CSI data are mapped to the attributes and distributional characteristics of English words. This results in the first *cross-modal RF embedding* work, and paves the way for device-free WiFi-based activity classification without requiring training data for all activities. Third, we propose a two-level classifier that is capable of classifying both *seen* and *unseen* activity types. This makes Wi-Fringe a generalized system for classifying a wide variety of human activities.

We develop Wi-Fringe using Intel Network Interface Card (NIC) 5300 [14] which captures the WiFi CSI data. To develop the machine learning models, we collect training and testing data from four volunteers in a multi-person apartment as well as in an office building for 20 activity classes — which is, to the best of our knowledge, the largest collection of activities used in any WiFi-based device-free gesture recognition system till date. We empirically quantify the trade-off between the accuracy and the number of unseen activities, and show that Wi-Fringe achieves 61%–90% accuracy, as we vary the number of unseen classes from 60% to 20%.

II. BACKGROUND

A. Channel State Information (CSI)

When wireless signals travel through the medium, they fade, they get reflected and scattered by obstacles on the way, and their power decays with the distance traveled. The *Channel State Information* (CSI) is a measure of all these phenomena of a wireless channel.

We express the relationship between the transmitted signal $X(f, t)$, the channel frequency response (CFR) $H(f, t)$, and the received signal $Y(f, t)$ as: $Y(f, t) = H(f, t) \cdot X(f, t) + N(f, t)$, where $N(f, t)$ denotes the noise. The CSI comprises of the CFR values, i.e., $\{H(f, t)\}$.

In WiFi, bits are transmitted simultaneously over 64 distinct frequencies or *sub-carriers* in parallel. The frequency response, $H(f, t)$ of each sub-carrier is a complex number. For N_{TX} transmitting antennas, N_{RX} receiving antennas, and N_s sub-carriers, we get a CSI matrix of complex numbers having the dimensions of $N_{TX} \times N_{RX} \times N_s$.

B. Word Embedding

The process of *Word Embedding* [12] maps words in a natural language to vectors of real numbers in a manner that words that are commonly used in the same textual context are positioned closely in the vector space. For example, consider the words: love and adore. Syntactically these two words are quite different, but they often appear in similar semantic contexts, i.e., with similar words. Hence, the word embedding process would map these two

words to two vectors whose distance is relatively closer than the embedding of two random words. We use *Word2Vec* [1] which is the most popular method to extract word embedding.

C. Attribute Embedding

While word embedding captures the co-occurrence information of words used in the same context, it does not describe the meaning of a word. Recently, natural language processing community has proposed an effective method to learn the attributes of English *verbs* from their dictionary definitions [13]. In this new method, verbs are expressed in terms of a set of attributes. Each verb is expressed as a vector of real numbers where each element of the vector corresponds to an attribute. Table I provides a simplified example. Three verbs: *Drink*, *Sip*, and *Drool* are expressed in terms of four attributes: *Motion*, *Social*, *Object*, *Head*, where the attributes correspond to the degree of motion, degree of social engagement, use of objects, and use of head, respectively. The process of attribute extraction is a supervised learning task where attributes are predicted from a word's dictionary definition. We refer to [13] for further details on the attribute learning process.

D. Zero-shot Learning

Recent branch of classification algorithms, known as *Zero Shot Learning* [15] do not require training data for all the classes to recognize them. Most Zero shot learning algorithms leverage knowledge from other domain such as text to classify in other domain such as images. The idea here is to project visual features and their corresponding labels in the same semantic space.

III. WI-FRIDGE SYSTEM DESIGN

Wi-Fridge takes a short-duration WiFi CSI stream (e.g., 5–8 seconds) and a list of possible activity types (i.e., a list of tags) encoded as one hot encoding [16] as the input, and processes the CSI stream through a signal processing pipeline to classify it as one of those given activity types. The duration is chosen short following [17], [18] as it is long enough to detect human activity. Although, Wi-Fridge takes possible activity types as input from the user, the labels do not have any influence on the training phase. This list is only used in classification stage for limiting the search space for better accuracy. If we do not have these additional labels, then the search space becomes too large (i.e., as big as having all the words in our database for a language) in the classification stage. Therefore, the user provided labels for unseen class is important in getting better accuracy for unseen activities.. The design of Wi-Fridge is modular. Computationally expensive modules such as the onetime offline training of the classifiers are run on a server, while the end-to-end activity classification pipeline—from sensing to classification—is runnable on embedded systems such as smartphones and tablets¹.

Figure 2 shows the signal processing pipeline of Wi-Fridge, which consists of three main steps: *State-Aware Representation*, *Cross-Modal Projections*, and *Two-Stage Classifier*. The State-Aware Representation step extracts local and contextual features from the CSI stream.

¹Recent developments [19] have shown how to extract CSI on smartphones. In this paper, we conduct experiments using an Intel NUC [20].

These features are projected onto the word and attribute spaces to incorporate external knowledge from the text domain in the Cross-Modal Projections step. The two-stage classifier determines if the input belongs to a seen or an unseen class and then classifies it accordingly. There are two classes of activities that Wi-Fringe may encounter at runtime: *seen* and *unseen* classes. The *seen* class refers to those activity types for which Wi-Fringe has labeled CSI streams for training. The *unseen* class, on the other hand, refers to activity types for which Wi-Fringe does not have any training CSI stream. The next three sections describe the algorithmic details of these three components of Wi-Fringe.

IV. STATE-AWARE REPRESENTATION

A. The Need for State-Aware Representation

The goal of this step is to obtain a state-aware representation of WiFi CSI values corresponding to an activity which encodes both the *local* features as well as the *contextual* features of an activity. The local features refer to the frequency response of an activity at a particular time-step. In contrast, the contextual features learn the temporal relationship among the local features. Existing works [21], [2] do not consider both local and temporal sequential nature of an activity when converting raw CSI values to feature vectors. To overcome the limitations of existing activity modeling techniques, we propose *state-aware* feature representation of CSI streams that captures complex, non-linear dependencies between micro-actions that constitute an action—without requiring a strict Markovian assumption or a predefined, fixed set of states. To achieve this, we employ a *Convolutional Neural Network* (CNN) [22] and a *Recurrent Neural Network* (RNN) [22] in tandem to capture the local and the contextual (temporal) features, respectively.

B. Rationale Behind Deep Neural Networks

The CSI spectrogram exhibits rich, informative, and distinguishable patterns for different states within an activity. It contains local information, for which, a CNN is the most suitable choice [22]. CNNs contain a hierarchy of filters, where each filter's job is to detect the presence of a particular pattern in a small region on the spectrogram. In Figure 3, we use rectangular boxes on the spectrogram which are recognized by the different convolutional filters of the CNN shown on the left. To model the sequential variation of states within an activity, we choose *Long Short Term Memory* (LSTM) as the recurrent component since LSTMs are better at learning longterm dependencies between states [22], which makes them suitable for capturing relationship between the past states with the recent states. To preserve contextual information from both the future and the past states, we use a bi-directional LSTM model.

C. Detailed Algorithmic Steps

Figure 4 shows the integrated neural network architecture that takes CSI spectrogram as the input and produces the stateaware vector representation through a sequence of processing steps.

- **CSI Processing:** For a given CSI stream, X that corresponds to an instance of an activity, we divide the CSI values into n equal segments $\{X_1, X_2, \dots, X_n\}$, which are the

states. Here, n is empirically determined, we used $n = 5$, that corresponds to one seconds of CSI data, in all our experiments. Smaller value of n leads to little segment of states which may not carry rich local information. While larger value of n makes longer states, thus less information about the transition between them which eventually leads to poor temporal information. For each segment X_i , we take the spectrogram [23] to obtain S_i as follows: $S_i = STFT(X_i)$. Here, $STFT(.)$ denotes *Short-Time Fourier Transform* [24], which estimates the short-term, time-localized frequency content of X_i . Recent works [17], [2] show high accuracy with frequency domain feature for supervised WiFi based activity recognition.

- **CNN Processing:** We use a three-layer CNN, G_θ , where θ are the parameters of the network, which takes S_i as the input and produces a 1000 dimension vector, L_i that represents the local features of the input spectrogram: $L_i = G_\theta(S_i)$.

- **RNN Processing:** For a total of n segments, we obtain n local feature maps $G_\theta(S_i)$ from the CNN. Each state's local feature is fed into a bi-directional LSTM (bi-LSTM) to model the contextual property of the states of an activity. The bi-LSTM network has two unidirectional LSTMs, i.e., a *forward* and a *backward* LSTM. For the forward LSTM, each hidden state \vec{H}_i depends on the previous state H_{i-1} and the input S_i . On the other hand, for the backward LSTM, each hidden state \overleftarrow{H}_i depends on the future state H_{i+1} and the input S_i .

- **Representation:** The final hidden representation of the bidirectional LSTM is the concatenation of \vec{H}_i and \overleftarrow{H}_i .

V. CROSS-MODAL PROJECTIONS

In this section, we describe the cross-modal projection step of Wi-Fringe which brings external knowledge from the text domain to enable classification of unseen activities.

A. The Need for Cross-Modal Projections

The secret recipe behind Wi-Fringe's ability to classify unseen activities is the *cross modal projection*. Through this step, we blur the difference between an activity's CSI stream and its corresponding English word embedding and attributes, and make them (almost) equal in their feature representations. In other words, if an English word (say, run) has a known vector representation (e.g., obtained using Google's Word2Vec [1] on a large dataset like Wikipedia) and a vector of attributes of the word (e.g., *movement of legs* and *motion of hands*, which is obtained from [13]), the goal of the cross-modal projection is to generate the exact same vectors when Wi-Fringe is presented with a CSI stream of that word (i.e., CSI of running).

B. Rationale Behind Multiple Projections

The benefits of cross-modal projection from CSI to two latent spaces, i.e., word and activity-attribute spaces, are as follows:

- **Word Embedding Space:** There are over 150 thousand English words for which researchers in the natural language processing field have created semantically aware vector

representations, called the *Word Embedding*. Such an embedding preserves the contextual relationship among words and puts two words that are similar in meaning or are often used in the same context closer in the representation space. By projecting the activity representation to the word embedding space, Wi-Fringe is able to generate meaningful and context-aware representation of any CSI stream, irrespective of whether or not it has seen CSI of the same class before.

• **Activity-Attribute Space:** Human activities and bodily gestures comprise of movements by different body parts, i.e., arms, legs, head, and torso. Activities also involve external objects, e.g., an eating activity may involve the use of spoons and knives. These *attributes* create nuances in different RFbased activity feature representations. Projecting CSI onto the activity-attribute space embeds this information into the projection, as CSI implicitly gets affected by moving different sorts of objects due to their reflections. Using this embedded contextual and attribute information from CSI, Wi-Fringe recognizes activities without any training examples.

Wi-Fringe uses both word embedding and activity-attribute spaces for cross-modal projections to combine the predictive power of both. Combining these two help each other in the final prediction step.

C. Detailed Algorithmic Steps

The goal of cross-modal projection is to map state-aware representation of WiFi CSI streams (Section IV) to two latent spaces, i.e., the word embedding space and the activity-attribute space.

The projection operation is illustrated by Figure 5. The state-aware representation, i.e., the output of the LSTM from Figure 4, is fed to two neural networks having fully connected layers. The first network projects the state-aware representation onto the activity attribute space, and the second network projects the representation onto the word embedding space. We refer to the last layer of the neural networks that perform attribute space projection and word embedding projection as F_A and F_W , respectively.

• **Projecting onto Activity-Attribute Space.**—From the attribute database provided by [13], we obtain a set of binary attributes associated with each activity. Each activity, a_i is represented as an m -dimension vector, $d_i \in \{1,0\}^m$, where m is the total number of attributes used to define an activity. Each element $d_i^{(k)}$ is a binary indicator of whether the k^{th} attribute is true or false for that activity.

$$d_i^{(k)} = \begin{cases} 1, & \text{if attribute } k \text{ is true for activity } a_i. \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

To get the likelihood of a CSI stream being predicted as an activity a_i , we project F_A to the attribute space. We take the dot product of the attribute vector d_i and F_A . The dot product demonstrates the similarity between the projection F_A with attribute vector d_i . For a CSI stream from activity a_i , our model's target is to increase the similarity between d_i and F_A .

The similarity which denotes the likelihood of F_A 's probability of belonging to activity a_i in attribute space is calculated as a dot product: $P_A^i = d_i \cdot F_A$.

• **Projecting onto Word-Embedding Space.**—For an activity, a_i whose word embedding is w_i , we want to project the CSI-based state-aware representation as close as possible to w_i . Therefore, for a CSI stream of an activity a_i , our target is to project F_W to be close to w_i in vector space. This results in a higher value of dot product between F_W and w_i . Therefore, the similarity in word embedding space is defined as: $P_W^i = w_i \cdot F_W$.

• **Projecting onto the Joint Space.**—To obtain a joint projection on both the attribute and the word embedding space, we employ an ensemble approach to combine the two projections from the previous steps as follows: $P_i = P_A^i + P_W^i$. where, P_i carries the confidence of a CSI segment's probability of belonging to class a_i . For —a— number of activities, given F_A and F_W extracted using state-aware representation for a CSI segment X as described in Section IV, the probability of $X \in a_i$ is calculated using the following softmax operation:

$$p(a_i | F_A, F_W) = \frac{e^{P_i}}{\sum_{j=1}^{|a|} e^{P_j}} \quad (2)$$

VI. TWO STAGE CLASSIFIER

Wi-Fringe employs a two-stage classifier to infer the most likely activity type for an input CSI stream segment. The first-stage classifier determines whether the input CSI stream segment belongs to a *seen* or an *unseen* class. The second-stage classifier makes the final determination of the most probable activity type for the input CSI stream segment.

A. The Need for Two-Stage Classifier

Neural networks tend to memorize patterns in data from training. Thus, even for an unseen category, the network tries to project an input close to one of the seen classes in the state-aware representation space. While this serves our purpose of classifying an unseen activity, it affects the classification performance of a generalized system where the system may encounter examples from both seen and unseen classes. Since the unseen classes are projected too close to the seen classes, they will be classified as one of those seen classes.

To overcome the problem posed by neural network based state-aware representation in distinguishing seen vs. unseen categories, we employ a classifier which is based on the signal characteristics such as the time-frequency representation. To be more precise, while some activities such as run and walk have similarity in their attributes, they still have distinguishable signal characteristics that are embedded in WiFi CSI. Past works [17], [2] have proven the distinguishable power of traditional signal-level features such as Short Time Fourier Transform (STFT) and Discrete Wavelet Transform (DWT). We use STFT to detect if a sample is from a seen or an unseen category. STFT gives us the changes in frequency components of the signal along the time axis.

B. Detailed Algorithmic Steps

The two steps of the algorithm are as follows:

• **Seen vs. Unseen Detection.**—We devise a simple thresholdbased decision algorithm to determine whether an input CSI stream segment belongs to an *unseen* class. We use Kmeans [25] clustering algorithm to cluster the STFT of the training samples of the seen category classes. This gives us K cluster centers, C_1, C_2, \dots, C_K for K seen classes. For an input CSI stream segment, u , it belongs to an *unseen* class if the following condition is true: $\min_{s \in S} \|C_s - \text{STFT}(u)\| > \Omega$. Here, S is the set of *seen* classes and Ω is an empirically determined threshold that maximizes the accuracy of *seen* vs. *unseen* class detection, and $\|\cdot\|$ is the Euclidean norm. When this condition is false, u belongs to a *seen* class.

• **Classification.**—If the CSI segment is recognized as from a *seen* category, only the labels from seen category are considered and the class label is obtained by applying the following equation:

$$\underset{a_i}{\operatorname{argmax}} p(a_i \mid F_A, F_W) \quad (3)$$

On the other hand, if the CSI segment is recognized as from an *unseen* category, we exclude all the labels from the *seen* category and only the labels from the *unseen* category are considered and the class label is obtained by applying Equation 3. Note that, the number of unseen category labels are dependent on the developers and the users who collect data for activity recognition model.

VII. EXPERIMENTAL RESULTS

A. Empirical Dataset

Our data collection setup is depicted in Figure 6. Based on our study of named activities from [26], we collect 20 most common named activities for our empirical evaluation. Our dataset contains activities collected from four volunteers in two different rooms with different orientations and furniture. Our dataset is diverse and it stresses out the algorithmic components of Wi-Fringe. In Table II, we provide the list of the 20 activities clustered with major attributes. On average, each class have 100 samples where the samples have on average 500 CSI values (5 seconds in duration).

B. Accuracy of Unseen Class Detection

In this experiment, we report the accuracy of Wi-Fringe for *unseen* classes. As state-of-the-art systems are not capable of recognizing activities without prior training examples, we are unable to compare them with our solution. Hence, we report Wi-Fringe's performance in this section by varying the number of unseen classes and compare it with two variants of our algorithm: (1) projecting State-Aware Representation (SAR) onto only word embedding (W2Vec) space, and (2) projecting State-Aware Representation (SAR) on only activity-attribute space. This comparison shows the performance boost due to joint space projection.

In Figure 7(a) we report the accuracy of Wi-Fringe in recognizing the unseen classes of activities when 2–6 types of activities are from unseen classes. We evaluate with different combinations of *seen* and *unseen* activities and present the mean accuracy and variance in the plot. In Figure 7(a), we see that for two unseen classes, we achieve a classification accuracy of around 90%. With only word embedding and attribute space projection, the accuracy is 87% and 88%, respectively. For three unseen activities, we get an accuracy near 83% with Wi-Fringe. With only word embedding projection the accuracy is around 80%, but with attribute space projection the accuracy drops to 60%. This is due to the similarity of activities in attribute space, which results in very similar attribute vectors. Therefore, projecting only on attribute space makes the classification harder. As the number of unseen classes increase to 4, 5, and 6, the accuracy becomes to 73%, 67% and 62%, respectively. In all the cases, joint space projection boosts the performance in comparison with single space projection. As the number of unseen classes increase, the problem becomes harder since the model has to differentiate between more classes without training data. We report up to six unseen classes in the plot, however, for seven unseen classes, our accuracy is around 53%. With random selection, the accuracy for seven unseen class is 14.28%, so Wi-Fringe is still better by almost 40%.

C. Accuracy of Seen Class Detection

In this experiment, we evaluate Wi-Fringe's *seen* class detection performance by keeping all the classes in seen category. We compare Wi-Fringe with other baseline classification algorithms. Following [8], [27], we compare Wi-Fringe with a CNN classifier optimized for our dataset with five convolutional layers along with batch normalization and dropout layers. We also report the performance of state-aware representation (SAR) integrated with a softmax layer. In addition, we also show the performance of projecting state-aware representation only to word embedding space and attribute space. We use five-fold cross-validation by randomly selecting training and testing examples each time. We also report the classification performance of a shallow classifier with a traditional handcrafted feature (i.e., STFT).

In Figure 7(b), we find that Wi-Fringe achieves a mean accuracy of 82%. On the other hand, state-aware representation (SAR) along with softmax layer is able to achieve around 80% mean accuracy. The performance boost of Wi-Fringe is due to the fact that from joint space projection, our model is able to classify activities by integrating knowledge from both word embedding and attribute domain. Projecting state-aware representation onto only word embedding and attribute space yields accuracy of 78% and 76%, respectively. With CNN, we have an accuracy of 74%. The SAR with softmax layer has better performance than only using word embedding and attribute, as softmax layers are designed for learning decision boundary effectively. However, the softmax layer is not suitable for zero shot learning as it does not borrow knowledge from external domain and has no mechanism to classify samples without labeled examples. With an SVM, we see the accuracy is around 62%. Therefore, it is evident that Wi-Fringe is able to achieve better accuracy than other classifiers in seen class detection.

D. Accuracy of Seen vs Unseen Class Detection

In this section, we present the accuracy of our threshold based Seen vs. Unseen detection's performance. The accuracy is threshold dependent. In Figure 7(c), we plot the accuracy for $\Omega \in [4.0-5.25]$. We observe that setting a high threshold fails to detect many class as *unseen* and the accuracy drops for the *unseen* classes. With high threshold, the unseen class samples have to be very far apart from any cluster center of the seen class clusters. On the other hand, setting the threshold too small leads to poor results for *seen* classes as it determines majority CSI stream sample as *unseen*. Hence, there is a tradeoff between the *seen* and *unseen* class detection accuracy. The optimum threshold is 4.75, for which, the classification accuracy of the *seen* and *unseen* classes are around 80%.

E. End to End Evaluation

To quantify Wi-Fringe's end-to-end performance, we report its classification accuracy for an application scenario. We monitor a user's home activity for ten different classes: {push, pull, run, sit, rub, walk, stand, eat, scratch, drink}. We consider three different training scenarios. First, we consider that the user provides 8 out of 10 activity classes' examples to Wi-Fringe during training, Second, we consider the case where 5 out of 10 activity classes' examples are given to Wi-Fringe during training. The last and the hardest test case is a scenario where Wi-Fringe has only 2 activity classes' samples during training, i.e., 8 out 10 classes are unseen. In Figure 8, we report the performance of Wi-Fringe along with two baseline algorithms: a convolutional neural network (CNN) and a random forest classifier for all three aforementioned scenarios. For each scenario, we consider three cases where we vary the ratio between samples from seen and unseen classes in the test dataset in the following ways: a) $\frac{\# \text{ seen}}{\# \text{ unseen}} = 25\%$, b) $\frac{\# \text{ seen}}{\# \text{ unseen}} = 50\%$ and c) $\frac{\# \text{ seen}}{\# \text{ unseen}} = 75\%$ Here, # denotes number of samples.

In Scenario 1 (Figure 8(a)), where only 2 classes are in the unseen category, Wi-Fringe shows an accuracy around 80% for all the cases, whereas the baseline algorithms' accuracy drops below 20% when most of the samples are coming from the unseen category. Note that the unseen classes are chosen by keeping one of their closest neighbours in the word embedding and attribute space in the seen category.

In scenario 2 (Figure 8(b)), Wi-Fringe achieves an accuracy of 84% for case 3 with majority of the samples in test cases coming from the seen classes. However, when the ratio of seen classes in the test data gets decreased in case 1, the accuracy drops to 73%. Yet, Wi-Fringe's performance is better than both baselines by a margin of greater than 40%.

In scenario 3 (Figure 8(c)), where only two classes are in the seen category, the accuracy for the case where 75% of test samples are from seen classes reaches up to 72% for Wi-Fringe. However, for case 1, the accuracy drops to 36% where 75% of the test samples are from the unseen categories. This drop is due to the fact that most of the classes are now in unseen category and Wi-Fringe has very few classes to learn the mapping function from RF to text domain. For this case, baselines achieve a maximum accuracy of only 24%. Therefore, it is evident that Wi-Fringe's performance is better than traditional classification algorithms in all the cases.

VIII. RELATED WORK

WiFi based sensing have opened the doorway for device-free activity monitoring in the last couple of years. Researchers [4], [28] have used wifi signal characteristics such as signal strength (RSSI) for activity recognition. With the availability of CSI from network interface cards, multiple works [21], [29] have emerged which exploit CSI information for gesture and activity recognition. [30], [18], [31] use deep learning based models to recognize activities from CSI. All of these works rely on provided training examples to classify a particular class of activity. Wi-Fringe deals with classification of activity from WiFi CSI data without any training examples. This is significantly different from current state of the arts. The earlier practices of zero shot learning [9], [32] for image classification problem infer the labels of unseen classes using a two step algorithm. First, the attributes of the sample is inferred and then the class label is predicted from an attribute database. Recent works [33], [34] have explored the mapping between image features and semantic space. Although these papers propose zero shot learning method for images, none of them addresses the problem for RF domain and activity recognition. [13], [35] do activity recognition using zero shot learning for RGBD data. However, our work is the first paper to propose a zero shot learning method for WiFi based activity classification where we overcome the challenges for cross-modal learning between text and RF domain.

IX. CONCLUSION

In this paper, we present the first WiFi-based device-free activity recognition system that does not require training examples for all activities. We propose a novel way to embed contextual information from the text domain to the RF domain by projecting RF data onto the word embedding and attribute space. We use this cross-modal RF embedding and propose a general classifier to recognize both *seen* and *unseen* activities. We collect WiFi data for 20 different activities from four volunteers and show that Wi-Fringe is capable of inferring activities from WiFi without training examples with 62%–90% accuracy for 2–6 unseen classes. Wi-Fringe is able to detect an unseen activity with high accuracy if there is a seen class with similar class label property. We assume that in a large dataset, the chances that there is no semantically similar training example to an unseen class is relatively low.

ACKNOWLEDGEMENT









This paper was supported, in part, by NSF grants CNS-1816213 and CNS-1704469, NIH grant 1R01LM013329–01.

REFERENCES

- [1]. Mikolov T, Chen K, and Corrado DJ, Greg, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781.
- [2]. Wang Y, Liu J, Chen Y, Gruteser M, Yang J, and Liu H, “E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures,” in Proceedings of the 20th annual international conference on Mobile computing and networking. ACM, 2014.
- [3]. Liu X, Cao J, Tang S, and Wen J, “Wi-sleep: Contactless sleep monitoring via wifi signals,” in Real-Time Systems Symposium (RTSS), 2014 IEEE. IEEE, 2014, pp. 346–355.

- [4]. Abdelnasser H, Youssef M, and Harras KA, "Wigest: A ubiquitous wifi-based gesture recognition system," in Computer Communications (INFOCOM), 2015 IEEE Conference on. IEEE, 2015, pp. 1472–1480.
- [5]. Adib F and Katabi D, See through walls with WiFi! ACM, 2013, vol. 43, no. 4.
- [6]. Liu X, Cao J, Tang S, Wen J, and Guo P, "Contactless respiration monitoring via off-the-shelf wifi devices," IEEE Transactions on Mobile Computing, vol. 15, no. 10, pp. 2466–2479, 2016.
- [7]. Wang Y, Jiang X, Cao R, and Wang X, "Robust indoor human activity recognition using wireless signals," Sensors, vol. 15, no. 7, pp. 17195–17208, 2015. [PubMed: 26184231]
- [8]. Yue S, He H, Wang H, Rahul H, and Katabi D, "Extracting multi-person respiration from entangled rf signals," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2, no. 2, p. 86, 2018.
- [9]. Lampert CH, Nickisch H, and Harmeling S, "Attribute-based classification for zero-shot visual object categorization," Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 3, 2014.
- [10]. Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Mikolov T et al. , "Devise: A deep visual-semantic embedding model," in Advances in neural information processing systems, 2013, pp. 2121–2129.
- [11]. Islam MT and Nirjon S, "Soundsemantics: exploiting semantic knowledge in text for embedded acoustic event classification," in Proceedings of the 18th International Conference on Information Processing in Sensor Networks. ACM, 2019, pp. 217–228.
- [12]. Mikolov T, Sutskever I, Chen K, Corrado GS, and Dean J, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems.
- [13]. Zellers R and Choi Y, "Zero-shot activity recognition with verb attribute induction," arXiv preprint arXiv:1707.09468, 2017.
- [14]. "Intel ultimate n wifi link 5300," <https://www.intel.com/content/www/us/en/wireless-products/ultimate-n-wifi-link-5300-brief.html>.
- [15]. Xian Y, Schiele B, and Akata Z, "Zero-shot learning-the good, the bad and the ugly," arXiv preprint arXiv:1703.04394, 2017.
- [16]. Rodríguez P, Bautista MA, Gonzalez J, and Escalera S, "Beyond one-hot encoding: Lower dimensional target embedding," Image and Vision Computing, vol. 75, pp. 21–31, 2018.
- [17]. Wang W, Liu AX, Shahzad M, Ling K, and Lu S, "Understanding and modeling of wifi signal based human activity recognition," in Proceedings of the 21st Annual International Conference on Mobile Computing and Networking. ACM, 2015, pp. 65–76.
- [18]. Ma J, Wang H, Zhang D, Wang Y, and Wang Y, "A survey on wifi based contactless activity recognition," in Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016 Intl IEEE Conferences. IEEE, 2016, pp. 1086–1091.
- [19]. Schulz M, Wegemer D, and Hollick M, "Nexmon: Build your own wifi testbeds with low-level mac and phy-access using firmware patches on off-the-shelf mobile devices," in Proceedings of the 11th Workshop on Wireless Network Testbeds, Experimental Evaluation & CHaracterization, ser. WiNTECH '17, 10. 2017, pp. 59–66.
- [20]. "Intel nuc mini pc," <https://intel.ly/2Xa1pcp>.
- [21]. Wang Y, Wu K, and Ni LM, "Wifall: Device-free fall detection by wireless networks," IEEE Transactions on Mobile Computing, vol. 16, no. 2, pp. 581–594, 2017.
- [22]. Goodfellow I, Bengio Y, Courville A, and Bengio Y, Deep learning. MIT press Cambridge, 2016, vol. 1.
- [23]. W.-k. Lu and Q. Zhang, "Deconvolutive short-time fourier transform spectrogram," IEEE Signal Processing Letters, vol. 16, no. 7.
- [24]. Islam MT, Islam B, and Nirjon S, "Soundsifter: Mitigating overhearing of continuous listening devices," in Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services. ACM, 2017, pp. 29–41.
- [25]. Krishna K and Murty NM, "Genetic k-means algorithm," IEEE Transactions on Systems Man And Cybernetics-Part B: Cybernetics, vol. 29, no. 3, pp. 433–439, 1999.

- [26]. Anguita D, Ghio A, Oneto L, Parra X, and Reyes-Ortiz JL, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in International workshop on ambient assisted living. Springer, 2012, pp. 216–223.
- [27]. Jiang W, Miao C, Ma F, Yao S, Wang Y, Yuan Y, Xue H, Song C, Ma X, Koutsonikolas D et al., "Towards environment independent device free human activity recognition," in Proceedings of the 24th Annual International Conference on Mobile Computing and Networking. ACM, 2018, pp. 289–304.
- [28]. Sigg S, Shi S, and Ji Y, "Rf-based device-free recognition of simultaneously conducted activities," in Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication. ACM, 2013, pp. 531–540.
- [29]. Wang G, Zou Y, Zhou Z, Wu K, and Ni LM, "We can hear you with wi-fi!" IEEE Transactions on Mobile Computing, vol. 15, no. 11, pp. 2907–2920, 2016.
- [30]. Chen Z, Zhang L, Jiang C, Cao Z, and Cui W, "Wifi csi based passive human activity recognition using attention based blstm," IEEE Transactions on Mobile Computing, 2018.
- [31]. Zou H, Zhou Y, Yang J, Jiang H, Xie L, and Spanos CJ, "Deepsense: Device-free human activity recognition via autoencoder long-term recurrent convolutional network," in 2018 IEEE International Conference on Communications (ICC). IEEE, 2018, pp. 1–6.
- [32]. Norouzi M, Mikolov T, Bengio S, Singer Y, Shlens J, Frome A, Corrado GS, and Dean J, "Zero-shot learning by convex combination of semantic embeddings," arXiv preprint arXiv:1312.5650.
- [33]. Socher R, Ganjoo M, Manning CD, and Ng A, "Zero-shot learning through cross-modal transfer," in Advances in neural information processing systems, 2013, pp. 935–943.
- [34]. Akata Z, Perronnin F, Harchaoui Z, and Schmid C, "Label-embedding for image classification," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 7, pp. 1425–1438, 2016. [PubMed: 26452251]
- [35]. Madapana N and Wachs JP, "A semantical & analytical approach for zero shot gesture learning," in Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on. IEEE, 2017, pp. 796–801.

	Training Phase		Testing(target) Phase	
Existing Systems	 walk	 drink	✓  drink	✗  run
Wi-Fringe	 walk	 drink	✓  drink	✓  run

✓ : able to detect ✗ : not able to detect

Fig. 1:

Unlike existing systems, Wi-Fringe recognizes *run* in the testing phase, even though it did not see any training example of *run* in the training phase.

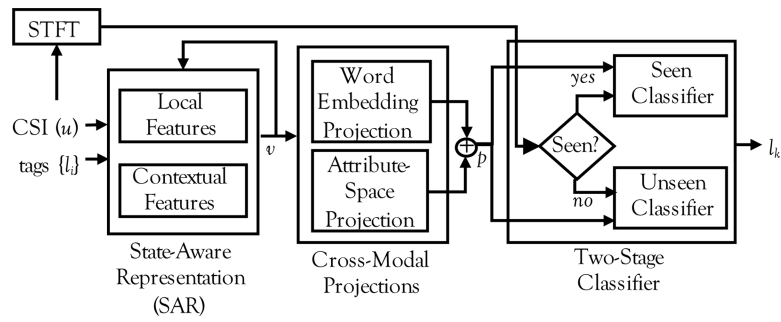


Fig. 2:
Wi-Fringe signal processing pipeline.

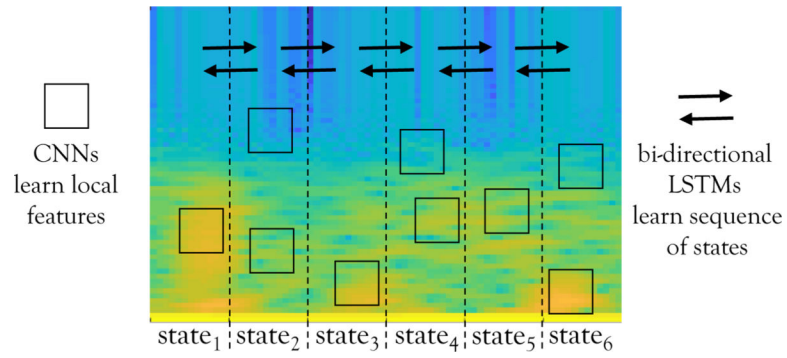


Fig. 3: CNNs learn local patterns that characterize each state, whereas RNNs (LSTMs) learn sequence of states.

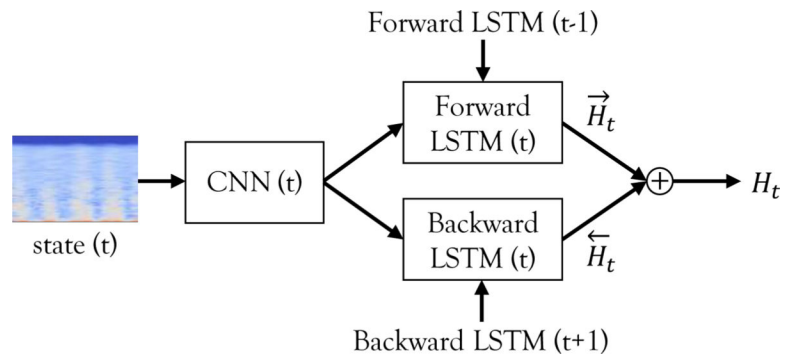
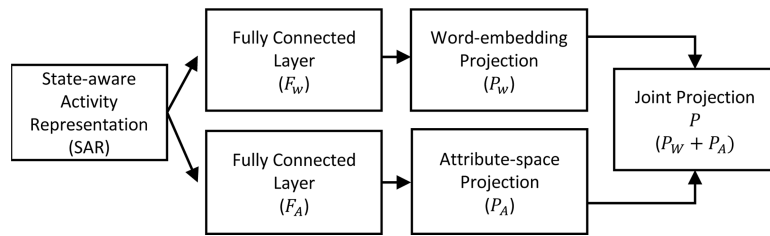


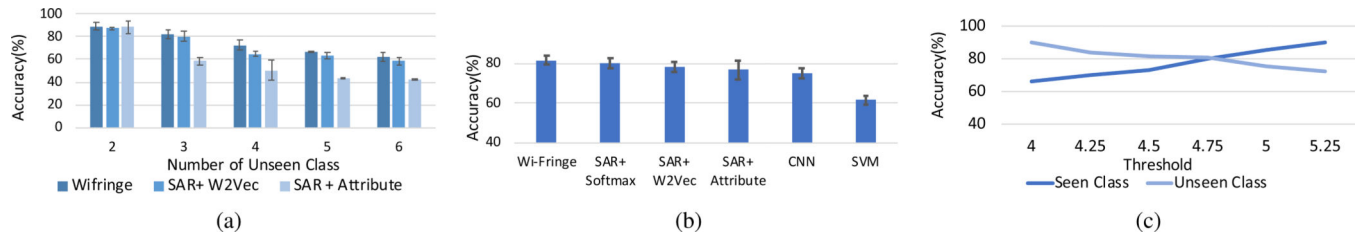
Fig. 4:
Network architecture for state-aware representation.

**Fig. 5:**

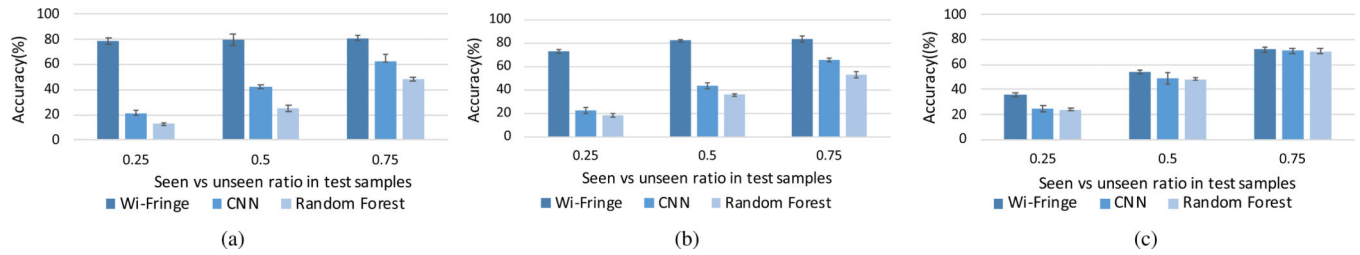
SAR is projected into both Word-embedding and Attribute space which are aggregated for a joint projection.



Fig. 6:
(a) Intel Nuc with Antennas. (b) Experimental Setup.

**Fig. 7:**

(a) Wi-Fringe's accuracy for unseen activity classification. (b) Wi-Fringe's accuracy is higher than baseline algorithms in seen class detection. (c) The accuracy of *seen* and *unseen* class detection depends on the threshold Ω 's value.

**Fig. 8:**

a) When 8 out of 10 classes are in seen category, Wi-Fringe has almost 80% accuracy for all cases. b) For 5 out of 10 classes in seen category Wi-Fringe outperforms all baseline algorithms for different cases. c) When 2 out of 10 classes are in seen class, for .25 fraction of test samples coming from seen category Wi-Fringe's performance drops below 40% which is still 1.5 times better than baselines.

TABLE I:

Word definitions and attributes.

Word	Representation
Drink	Dictionary: "To take into the mouth and swallow a liquid." Attributes: (<i>Motion, Social, Object, Head, ...</i>) = (<i>low, solitary, true, true, ...</i>)
Sip	Dictionary: "To drink in small quantities." Attributes: (<i>Motion, Social, Object, Head, ...</i>) = (<i>low, social, true, true, ...</i>)
Drool	Dictionary: "To let run from the mouth." Attributes: (<i>Motion, Social, Object, Head, ...</i>) = (<i>none, solitary, false, true, ...</i>)

TABLE II:

Twenty categories of activities.

Category	Activities
Freehand Gestures	Point, Raise, Rub, Scratch, Shake, Toss, Circle, Arc.
Object-Human Interactions	Drink, Eat, Push, Pull.
Upper/Lower-Body Gestures	Sit, Stand, Bow, Duck, Kick.
Mobility	Jump, Walk, Run.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript